Privacy in the Genomic Era

Muhammad Naveed[*], Erman Ayday[2], Ellen W. Clayton[3], Jacques Fellay[4], Carl A. Gunter[5], Jean-Pierre Hubaux[6], Bradley A. Malin[7], XiaoFeng Wang[8]

Abstract

Genome sequencing technology has advanced at a rapid pace and it is now possible to generate highly-detailed genotypes inexpensively. The collection and analysis of such data has the potential to support various applications, including personalized medical services. While the benefits of the genomics revolution are trumpeted by the biomedical community, the increased availability of such data has major implications for personal privacy; notably because the genome has certain essential features, which include (but are not limited to) *(i)* an association with traits and certain diseases, *(ii)* identification capability (e.g., forensics), and *(iii)* revelation of family relationships. Moreover, direct-to-consumer DNA testing increases the likelihood that genome data will be made available in less regulated environments, such as the Internet and for-profit companies. The problem of genome data privacy thus resides at the crossroads of computer science, medicine, and public policy. While the computer scientists have addressed data privacy for various data types, there has been less attention dedicated to genomic data. Thus, the goal of this paper is to provide a systematization of knowledge for the computer science community. In doing so, we address some of the (sometimes erroneous) beliefs of this field and we report on a survey we conducted about genome data privacy with biomedical specialists. Then, after characterizing the genome privacy problem, we review the state-of-the-art regarding privacy attacks on genomic data and strategies for mitigating such attacks, as well as contextualizing these attacks from the perspective of medicine and public policy. This paper concludes with an enumeration of the challenges for genome data privacy and presents a framework to systematize the analysis of threats and the design of countermeasures as the field moves forward.

I Introduction

The genomic era began with the announcement twelve years ago that the Human Genome Project (HGP) had completed its goals [Guttmacher and Collins, 2003]. The technology associated with genome sequencing has progressed at a rapid pace, and this has coincided with the rise of cheap computing and communication technologies. Consequentially, it is now possible to collect, store, process, and share genomic data in a manner that was unthinkable at the advent of the HGP. In parallel with this trend there has been significant progress on understanding and using genomic data that fuels a rising hunger to broaden the number of individuals who make use of their genomes and to support research to expand the ways in which genomes can be used. This rise in the availability and use of genomic data has led to many concerns about its security and privacy. These concerns have been addressed with efforts to provide technical protections and a corresponding series of demonstrations of vulnerabilities. Given that much more research is needed and expected in this

[*]University of Illinois at Urbana-Champaign. Work done in part at Ecole Polytechnique Federale de Lausanne. Email: naveed2@illinois.edu
[2]Bilkent University. Work done at Ecole Polytechnique Federale de Lausanne. Email: erman@cs.bilkent.edu.tr
[3]Vanderbilt University. Email: ellen.clayton@vanderbilt.edu
[4]Ecole Polytechnique Federale de Lausanne. Email: jacques.fellay@epfl.ch
[5]University of Illinois at Urbana-Champaign. Email: cgunter@illinois.edu
[6]Ecole Polytechnique Federale de Lausanne. Email: jean-pierre.hubaux@epfl.ch
[7]Vanderbilt University. Email: b.malin@vanderbilt.edu
[8]Indiana University at Bloomington. Email: xw7@indiana.edu

area, this seems like a good point to overview and systematize what has been done in the last decade and provide ideas on a framework to aid future efforts.

To provide context, consider that it was not until the early 1990s when sequencing the human genome was posited as a scientific endeavor. The first attempt for *whole genome sequencing*[1] (a laboratory process that maps the full DNA sequence of an individual's genome) was initiated at the U.S. National Institutes of Health (NIH) in 1990 and the first full sequence was released 13 years later at a total cost of $3 billion. Yet, sequencing technology has evolved and costs have plummeted, such that the price for a whole genome sequence is $5K[2] as of July 2014 and can be completed in two to three days. The "$1K genome in 1 day" will soon be a reality.

Decreases in sequencing costs have coincided with an escalation in genomics as a research discipline with explicit application possibilities. Genomic data is increasingly incorporated in a variety of domains, including healthcare (e.g., personalized medicine), biomedical research (e.g., discovery of novel genome-phenome associations), direct-to-consumer (DTC) services (e.g., disease risk tests), and forensics (e.g., criminal investigations). For example, it is now possible for physicians to prescribe the "right drug at the right time" (for certain drugs) according to the makeup of their patients' genome [Bielinski *et al.*, 2014; Overby *et al.*, 2010; Gottesman *et al.*, 2013a; Pulley *et al.*, 2012].

To some people, genomic data is considered (and treated) no differently than traditional health data (such as what might be recorded in one's medical record) or any other type of data more generally [Bains, 2010; Rothstein, 2005]. While genomic data may not be "exceptional" in its own right, it has many features that distinguish it (discussed in depth in the following section) and there is a common belief that it should be handled (e.g., stored, processed, and managed) with care. The privacy issues associated with genomic data are complex, particularly because such data has a wide range of uses and provides information on more than just the individual from which the data was derived. Yet, perhaps most importantly, there is a great fear of the unknown. Every day, we learn something new about the genome, whether it be knowledge of a new association with a particular disease or proof against a previously reported association. We have yet to discover everything there is from DNA, which makes it almost impossible to assign exact value, and thus manage DNA as a personal asset (or public good). So, as the field of genomics evolves, so too will the views on the privacy-sensitivity of genomic data. As this paper progresses, we review some of the common beliefs revolving around genome privacy. And, in doing so, we report on the results of a survey we conducted with biomedical specialists regarding their perspective on genome data privacy issues.

It should be recognized that there exist numerous publications on technical, ethical, and legal aspects of genomics and privacy. The research in the field covers privacy-preserving handling of genomic data in various environments (as will be reviewed in this paper). Yet, there are several challenges to ensuring that genomics and privacy walk hand-in-hand. One of the challenges that computer scientists face is that these views tend to be focused on one aspect of the problem in a certain setting with a certain discipline's perspective. From the perspective of computer science, there is a need for a complete framework which shows *(i)* what type of security and privacy requirements are needed in each step of the handling of genomic data, *(ii)* a characterization of the various threat models that are realized at each step, and *(iii)* open computational research problems. By providing such a framework in this paper, we are able to illustrate the important problems of genome privacy to computer science researchers working on security and privacy problems more generally.

Related Surveys and Articles. Privacy issues caused by forensic, medical, and other uses of genomic data have been studied in the past few years [Stajano *et al.*, 2008; Stajano, 2009; Malin, 2005a; Ayday *et al.*, 2013a; Naveed, 2014; Cristofaro, 2014]. A recent survey [Erlich and Narayanan, 2013] discusses privacy breaches using genomic data and proposes methods for protection. It addresses topics that we discuss in Sections VI and Section IX of this paper. In Section IX we present an end-to-end picture for the handling of genomic data in a variety of contexts as shown in Figure 9, while [Erlich and Narayanan, 2013] discusses how access control, data anonymization and cryptographic techniques can be used to prevent genetic privacy breaches. Moreover, [Erlich and

[1] In this study, we refer to the process of obtaining the Whole Genome Sequence (WGS) or the Whole Exome Sequence (WES) as *sequencing* and the process of obtaining the variants (usually only single nucleotide polymorphisms, or SNPs) as *genotyping*.

[2] http://www.genome.gov/sequencingcosts/

Narayanan, 2013] has been written for a general audience, whereas this paper is meant for computer scientists (and in particular security and privacy specialists).

Contributions. Following are the main contributions of this paper:

- We provide an *extensive and up-to-date (as of June 2015) literature survey*[3] of computer science as well as medical literature about genome privacy.
- We report concerns expressed by an opportunistically ascertained group of biomedical specialists about the security and privacy of genomic data.
- We develop an end-to-end *framework for the security and privacy of genomic data* in a variety of healthcare, biomedical research, legal and forensics, and direct-to-consumer contexts.
- We present what we believe to be the first document that reflects the *opinions of computer science, medical, and legal researchers* for this important topic.

We also provide an online tutorial[4] of biology and other related material to define technical terms used in this (and other) paper(s) on the security and privacy of genomic data. The remainder of this paper is organized as follows. Section II explains to what extent genomic data is distinct from data in general and health information in particular. Section III provides an overview of uses of genomic data for the non-specialist. Section IV emphasizes the relevance of genome privacy. Section V reports on the concerns of 61 opportunistically ascertained biomedical scientists regarding the importance of genomic data privacy and security. Sections VI and VII provide literature surveys, where the former summarizes the problem (i.e., the privacy risk) and the latter summarizes possible solutions. Section VIII summarizes the challenges for genomic medicine and privacy. Based on this analysis, Section IX offers a general framework for privacy-preserving handling of genomic data, including an extensive threat model that discusses what type of attacks are possible at each step of the data flow.

II Special features of genomic data

In this section, we discuss why genomic data is special. We have identified six features of genomic data, as shown in Figure 1, and while other data harbor some of these features, we are not aware of any data (including other molecular, such as proteomics, data) that have *all* of these features.

Consider the following scenario. Alice decides to have her genome sequenced by a service called MyGenome.com that keeps her data in a repository and gives Alice information about it over time. At first she uses information from MyGenome to explore parts of her family tree and contribute her genomic data, along with some facts about herself, to support medical research on diseases of her choosing. Many years after MyGenome performed the initial sequencing, Alice began experiencing health problems for which she visited a doctor who used her genomic data to help diagnose a likely cause and customize a treatment based on variation in her genome sequence. Alice was impressed by this experience and wondered what other conditions might be in her future. After some exploration she discovered that evidence (based on published research papers) suggested a high risk of dementia for people with her genomic profile. She worried that various parties, including MyGenome, the genealogy service, and research studies with whom she shared her data, might share this and other information in ways that she did not expect or intend and whether this might have undesired consequences for her.

Alice's story highlights several of the special features of genomic data. We depict six of them in Figure 1, which we review for orientation of the reader.

How does the result of a DNA-based lab test differ from that of other tests? One notable feature is how it is static and of long-lived value. Most tests, especially ones Alice could do for herself, like taking her temperature and blood pressure, are of relatively short term value, whereas genomic data changes little over a lifetime and may have value that lasts for decades. Of course, there are

[3] In this paper, the word "survey" is used to mean *literature survey* as well as *opinion poll*, however, the meaning will be clear from the context.

[4] Available at https://sites.google.com/site/genoterms

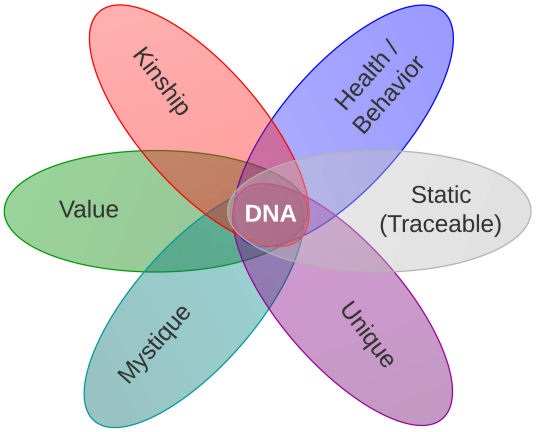

Figure 1: Properties of DNA that, in combination, may distinguish it from other data types. *Health/Behavior* means that DNA contains information about an individual's health and behavior. *Static(Traceable)* means that DNA does not change much over time in an individual. *Unique* means that the DNA of any two individuals can be easily distinguished from one another. *Mystique* refers to the public perception of mystery about DNA. *Value* refers to the importance of information content in DNA and that this importance does not decline with time (which is the case with other medical data e.g., blood pressure, glucose level, or a blood test). In fact, this importance will likely increase with time. *Kinship* means that DNA contains information about an individual blood relatives.

some exceptions to this longevity. For instance, sequencing techniques improve in accuracy over time, so tests may be repeated to improve reliability. Additionally, there are some modifications in DNA that accumulate over time (e.g., shortening of the ends of DNA strands due to aging [Harley *et al.*, 1990]). Most particularly, somatic mutations occur resulting in some degree of mosaicism in every individual: the most striking examples are the deleterious modifications of the DNA observed in cancer cells in comparison to DNA derived from normal cells. However, this long-lasting value means that holding and using genomic data over extended periods of time, as Alice did, is likely.

Alice's first use of her genomic data is expected to be a key driver for application development in the future. While DNA has been used for some time in parentage tests, it can be generalized from such studies to enable broader inference of kinship relations. Services such as Ancestry.com and 23andme.com already offer kinship services based on DNA testing. While a substantial portion of Alice's DNA is in common with that of her relatives, it is also unique to her (unless she has an identical twin). This has another set of implications about potential use of genomic data, like its ability to link to her personally, a property that makes DNA testing useful in criminal forensics.

Another of the special features of DNA relates to its ability for diagnosing problems in health and behavior. Tests are able to demonstrate increased likelihood for conditions such as macular degeneration in old age and Alzheimer's (the most common form of dementia) [Goldman *et al.*, 2011]. Although these are often probabilities, they can have diagnostic value as well as privacy ramifications [Seddon *et al.*, 2011]. For instance, if Alice's relatives learned about her increased risk of dementia, might they (consciously or unconsciously) trust her judgement a little less? Or might they instead help her to get timely treatment? This power for good and bad has led genomic data to have a certain "mystique", which has been promoted by scientists and the media [Tambor *et al.*, 2002]. The "mystique" surrounding the genomic data is evident from movies and books on the topic. Examples include the movie "GATTACA" and the book "The DNA mystique" [Nelkin and Lindee, 1995].

Although there are many other types of tests (e.g., protein sequence tests) that carry key common information with DNA tests, there is a special status that DNA data has come to occupy, a status that some have phrased as "exceptional" [Bains, 2010]. These special fears about the sharing of genomic data, whether founded or not, cannot be ignored when considering privacy implications.

Hence, while DNA data may or may not be exceptional [Evans *et al.*, 2010; Gostin and Hodge Jr, 1999], it is special in many ways, and hence warrants particular care.

III Uses of Genomic Data

An individual's genomic sequence contains over 3 billion base pairs, which are distributed across twenty-three chromosomes. Despite its size, it is estimated that the DNA of two individuals differ by no more than 0.5% [Venter *et al.*, 2001]; but it is these differences that influence an individual's health status and other aspects (as discussed in Section II). To provide further context for the importance of genomic data, this section reviews several of the major applications in practice and under development.

III.A Healthcare

First, it has been recognized that mutation in an individual's genomic sequence can influence his well being. In some cases, changes in a particular gene will have an adverse effect on a person's health immediately or at some point in the future [Botstein and Risch, 2003]. As of 2014, there were over 1,600 of these traits reported on in the literature[5], ranging from metabolic disorders (e.g., phenylketonuria, which is caused by a mutation in the PKU gene) to neurodegenerative diseases (e.g., Huntington's disease, which is caused by a mutation in the HD gene [MacDonald *et al.*, 1993]) to blood disorders (e.g., Sickle cell anemia, caused by a mutation in the HBB gene [Saiki *et al.*, 1985]). While some of these diseases are manageable through changes in diet or pharmacological treatments, others are not and have no known intervention to assist in the improvement of an individual's health status. Nonetheless, some individuals choose to learn their genetic status, so that they may order their affairs accordingly and contribute to medical research [Mastromauro *et al.*, 1987] (as elaborated upon below). Moreover, genetic tests can be applied in a prenatal setting to detect a variety of factors that can influence health outcomes (e.g., if a fetus is liable to have a congenital defect that could limit its lifespan, such as Tay-Sach's disease) [Lippman, 1991].

Yet, the majority of variations in an individual's genome do not follow the monogenic model. Rather, it has been shown that variation is associated with change in the susceptibility of an individual to a certain disease or behavior [Botstein and Risch, 2003]. Cancer-predisposing variants in genes such as BRCA1/2 or the Lynch Syndrome are well-known examples. Such variation may also modify an individual's ability to respond to a pharmaceutical agent. For instance, some individuals are slow (fast) metabolizers, such that they may require a different amount of a drug than is standard practice, or may gain the greatest benefit from a different drug entirely. This variation has been leveraged to provide dosing for several medications in practice, including blood thinners after heart surgery (to prevent clotting) and hypertension management (to lessen the severity of heart disease) [Pulley *et al.*, 2012]. Additionally, changes in an individual's genome detected in a tumor cell can inform which medications are most appropriate to treat cancer [Feero *et al.*, 2011].

III.B Research

While the genome has been linked with a significant number of disorders and variable responses to treatments, new associations are being discovered on a weekly basis. Technology for performing such basic research continues to undergo rapid advances [Brunham and Hayden, 2012]. The dramatic decrease in the cost of genome sequencing has made it increasingly possible to collect, store, and computationally analyze sequenced genomic data on a fine-grained level, as well as over populations on the order of millions of people (e.g., China's Kadoorie biobank [Chen *et al.*, 2011] and UK Biobank [Allen *et al.*, 2014] will each contain genomic data on 500,000 individuals by the end of 2014, while the U.S. National Cancer Institute is at the beginning of its Million Cancer Genome Project [Haussler *et al.*, 2012]). Yet, it should be recognized that computational analysis is separate

[5] http://www.ncbi.nlm.nih.gov/Omim/mimstats.html

from, and more costly than, sequencing technology itself (e.g., the $1K *analysis* of a genome is far from being developed).

Moreover, technological advances in genome sequencing are coalescing with a big data revolution in the healthcare domain. Large quantities of data derived from electronic health records (EHRs), for instance, are being made available to support research on clinical phenotypes that, until several years ago, were deemed to be too noisy and complex to model [Gottesman *et al.*, 2013b]. As a consequence, genome sequences have become critical components of the biomedical research process [Kohane, 2011].

III.C Direct-to-Consumer Services

Historically, genome sequencing was a complex and expensive process that, as a result, was left to large research laboratories or diagnostic services, but in the past several years, there has been a rise in DTC (Direct-to-consumer) genome sequencing from various companies [Prainsack and Vayena, 2013]. These services have made it affordable for individuals to become directly involved in the collection, processing, and even analysis of their genomic data. The DTC movement has enabled individuals to learn about their disease susceptibility risks (as alluded to earlier), and even perform genetic compatibility tests with potential partners. Moreover, and perhaps more importantly, DTC has made it possible for individuals to be provided with digital representations of their genome sequences, such that they can control how such information is disclosed, to whom, and when.

Of course, not all consumer products are oriented toward health applications. For example, genomic data is increasingly applied to determine and/or track kinship. This information has been applied for instance to track an individual's ancestral heritage and determine the extent to which individuals with the same surname are related with respect to their genomic variance [Jobling, 2001].

III.D Legal and Forensic

Given the static nature of genomic sequences, this information has often been used for investigative purposes. For instance, this information may be applied in contested parentage suits [Anderlik, 2003]. Moreover, DNA found at a crime scene (or on a victim) may be used as evidence by law enforcement to track down suspected criminals [Kaye and Smith, 2003]. It is not unheard of for residents of a certain geographic region to be compelled to provide tissue samples to law enforcement to help in such investigations [Greely *et al.*, 2006]. Given the kinship relationships that such information communicates, DNA from an unknown suspect has been compared to relatives to determine the corresponding individual's likely identity in order to better facilitate a manhunt.

One of the concerns of such uses, however, is that it is unclear how law enforcement may retain and/or use this information in the future. The U.S. Supreme Court recently ruled that it is permissible for law enforcement to collect and retain DNA on suspects, even if the suspects are not subsequently prosecuted [Maryland v. King, 2013]. Once DNA is shed by an individual (such as from saliva left on a coffee cup in a restaurant) it has been held as an "abandoned" resource [Joh, 2006], such that the corresponding individual relinquishes rights of ownership. While the notion of "abandoned DNA" remains a hotly contested issue, it is currently the case in the U.S. that DNA collected from discarded materials can be sequenced and used by anyone without the consent of the individual from which it was derived.

IV Relevance of Genome Privacy

As discussed in Section II, genomic data has numerous distinguishing features and applications. As a consequence, the leakage of this information may have serious implications if misused, as in genetic discrimination (e.g., for insurance, employment, or education) or blackmail [Gottlieb, 2001]. A true story exemplifying genetic discrimination was shared by Dr. Noralane Lindor at the Mayo Clinic's Individualizing Medicine Conference (2012) [Lindor, 2012]. During her study of a cancer patient, Dr. Lindor also sequenced the grandchildren of her patient, two of whom turned out to have

the mutation for the same type of cancer[6]. One of these grandchildren applied to the U.S. army to become a helicopter pilot. Even though genetic testing is not a required procedure for military recruitment, as soon as she revealed that she previously went through the aforementioned genetic test, she was rejected for the position (in this case legislation does not apply to military recruitment, as will be discussed below).

Ironically, the familial aspect of genomics complicates the problems revolving around privacy. A recent example is the debate between the family members of Henrietta Lacks and the medical researchers [Skloot, 2013]. Ms. Lacks (deceased in 1951) was diagnosed with cervical cancer and some of her cancer cells were removed for medical research. These cells later paved the way to important developments in medical treatment. Recently, researchers sequenced and published Ms. Lacks's genome without asking for the consent of her living family members. These relatives learned this information from the author of the bestselling book "The Immortal Life of Henrietta Lacks" [Skloot and Turpin, 2010], and they expressed the concern that the sequence contained information about her family members. After complaints, the researchers took her genomic data down from public databases. However, the privacy-sensitive genomic information of the members of the Lacks family was already compromised because some of the data had already been downloaded and many investigators had previously published parts of the cells' sequence. Although the NIH entered into an agreement with the Lacks family to give them a voice in the use of these cells [Ritter, 2013], there is no consensus about the scope of control that individuals and their families ought to have over the downstream of their cells. Thousands of people, including James Watson [Nyholt *et al.*, 2008], have placed their genomic data on the Web without seeking permission of their relatives.

One of the often voiced concerns regarding genomic data is its potential for discrimination. While, today, certain genome-disease and genome-trait associations are known, we do not know what will be inferred from one's genomic data in the future. In fact, a grandson of Henrietta Lacks expressed his concern about the public availability of his grandmother's genome by saying that "the main issue was the privacy concern and what information in the future might be revealed". Therefore, it is likely that the privacy-sensitivity of genomic data, and thus the potential threats will increase over time.

Threats emerging from genomic data are only possible via the leakage of such data, and, in today's healthcare system, there are several candidates for the source of this leakage. Genomic data can be leaked through a reckless clinician, the IT of a hospital (e.g., through a breach of the information security), or the sequencing facility. If the storage of such data is outsourced to a third party, data can also be leaked from such a database through a hacker's activity or a disgruntled employee. Similarly, if the genomic data is stored by the individual himself (e.g., on his smartphone), it can be leaked due to a malware. Furthermore, surprisingly, sometimes the leakage is performed by the genome owner. For example, on a genome-sharing website, openSNP[7] [Greshake *et al.*, 2014], people upload the variants in their genomes – sometimes with their identifying material, including their real names.

One way of protecting the privacy of individuals' genomic data is through the law or policy. In 2007, the U.S. adopted the Genetic Information Nondiscrimination Act (GINA), which prohibits certain types of discrimination in access to health insurance and employment. Similarly, the U.S. Presidential report on genome privacy [Presidential Commission for the Study of Bioethical Issues, 2012] discusses policies and techniques to protect the privacy of genomic data. In 2008, the Council of Europe adopted the convention concerning genetic testing for health purposes [Council of Europe, 2008]. There are, in fact, hundreds of legal systems in the world, ranging in scope from federal to state / province, and municipality level and each can adopt different definitions, rights, and responsibilities for an individual's privacy. Yet, while such legislation may be put into practice, it is challenging to enforce because the uses of data cannot always be detected. Additionally, legal regimes may be constructed such that they are subject to interpretation or leave loopholes in place. For example, GINA does not apply to life insurance or the military [Altman and Klein, 2002]. Therefore, legislation alone, while critical in shaping the norms of society, is insufficient to prevent privacy violations.

[6]Having a genetic mutation for a cancer only probabilistically increases the predisposition to the cancer.
[7]Hosted at http://www.openSNP.org

The idea of using technical solutions to guarantee the privacy of such sensitive and valuable data brings about interesting debates. On one hand, the potential importance of genomic data for mankind is tremendous. Yet, privacy-enhancing technologies may be considered as an obstacle to achieving these goals. Technological solutions for genome privacy can be achieved by various techniques, such as cryptography or obfuscation (proposed solutions are discussed in detail in Section VII). Yet, cryptographic techniques typically reduce the efficiency of the algorithms, introducing more computational overload, while preventing the users of such data from "viewing" the data. And, obfuscation-based methods reduce the accuracy (or utility) of genomic data. Therefore, especially when human life is at stake, the applicability of such privacy-enhancing techniques for genomic data is questionable.

On the other hand, to expedite advances in personalized medicine, genome-phenome association studies often require the participation of a large number of research participants. To encourage individuals to enroll in such studies, it is crucial to adhere to ethical principles, such as autonomy, reciprocity and trust more generally (e.g., guarantee that genomic data will not be misused). Considering today's legal systems, the most reliable way to provide such trust pledges may be to use privacy-enhancing technologies for the management of genomic data. It would severely discredit a medical institution's reputation if it failed to fulfill the trust requirements for the participants of a medical study. More importantly, a violation of trust could slow down genomic research (e.g., by causing individuals to think twice before they participate in a medical study) possibly more than the overload introduced due to privacy-enhancing technologies. Similarly, in law enforcement, genomic data (now being used in FBI's Combined DNA Index System – CODIS) should be managed in a privacy-preserving way to avoid potential future problems (e.g., mistrials, law suits).

In short, we need techniques that will guarantee the security and privacy of genomic data, without significantly degrading the efficiency of the use of genomic data in research and healthcare. Obviously, achieving all of the aforementioned properties would require some compromise. Our preliminary assessment of expert opinion (discussed in Section V) begins to investigate what tradeoffs users of such data would consider appropriate.

V Genomics/Genetics Expert Opinion

V.A Objective

We explored the views of an opportunistically ascertained group of biomedical researchers in order to probe levels of concern about privacy and security to be addressed in formulating guidelines and in future research.

V.B Survey Design

The field of genomics is relatively young, and its privacy implications are still being refined. Based on informal discussions (primarily with computer scientists) and our review of the literature, we designed a survey to learn more about biomedical researchers' level of concern about genomics and privacy. Specifically, the survey inquired about *(i)* widely held assertions about genome privacy, *(ii)* ongoing and existing research directions on genome privacy, and *(iii)* sharing of an individual's genomic data, using the following probes in Figure 2. The full survey instrument is available at http://goo.gl/forms/jwiyx2hqol. The Institutional Review Board (IRB) at the University of Illinois at Urbana-Champaign granted an exemption for the survey . Several prior surveys focused on genome privacy have been conducted and have focused on the perspectives of the general public [Kaufman *et al.*, 2009, 2012; Platt *et al.*, 2013; De Cristofaro, 2014] and geneticists [Pulley *et al.*, 2008]. Our survey is different because it investigates the opinion of biomedical researchers with respect to the intention of data protection by technical means.

The eight probes used to explore opinions about genome privacy follows:

1. Genome privacy is hopeless, because all of us leave biological cells (hair, skin, droplets of saliva,...) wherever we go.
2. Genomic data is not special and should be treated as any other sensitive health data e.g. health record or mental health notes.
3. Genome privacy is irrelevant, because genetics is non-deterministic
4. Genome privacy should be left to bioinformaticians, they can provide better privacy solutions than computer security, privacy and cryptography community can.
5. Genome privacy will be fully guaranteed by legislation
6. Privacy Enhancing Technologies are a nuisance in the case of genetics: genetic data should be made available online to everyone to facilitate research, as done e.g. in the case of the Personal Genome Project
7. Encrypting genomic data is superfluous because it is hard to identify a person from her variants
8. Advantages of genomic based healthcare justify the harm that genome privacy breach can cause.

Figure 2: Probes of attitudes

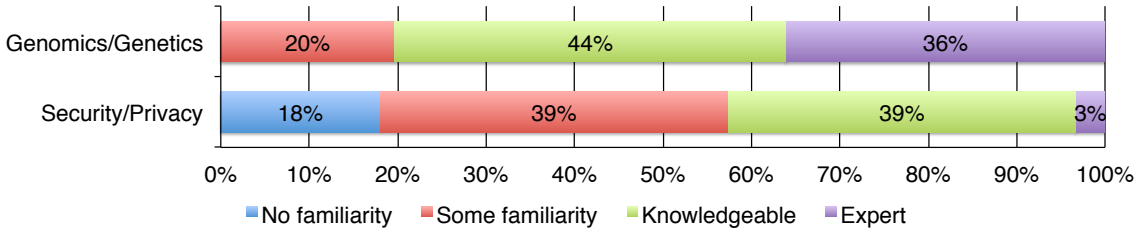

Figure 3: Self-identified expertise of the survey respondents

V.C Data Collection Methodology

We conducted our survey both online and by paper. Snowball sampling [Goodman, 1961] was used to recruit subjects for the online survey. This approach enables us to get more responses, but the frame is unknown and thus response rate cannot be reported. A URL for the online survey was sent to the people working in genomics/genetics area (i.e., molecular biology professors, bioinformaticians, physicians, genomics/genetics researchers) known to the authors of this paper. Recipients were asked to forward it to other biomedical experts they know. Email and Facebook private messages (as an easy alternative for email) were used to conduct the survey. Eight surveys were collected by handing out paper copies to participants of a genomics medicine conference. Overall the survey was administered to 61 individuals.

V.D Potential Biases.

We designed the survey to begin to explore the extent to which biomedical researchers share concerns expressed by some computer scientists. While not generalizable to all biomedical experts due to the method for recruiting participants, their responses do provide preliminary insights into areas of concern about privacy and security expressed by biomedical experts. More research is needed to assess the representativeness of these views.

V.E Findings

Approximately half of the participants were from the U.S. and slightly less than half from Europe (the rest selected "other"). The participants were also asked to report their expertise in genomics/genetics and security/privacy. We show these results in Figure 3.

We asked whether the subjects agree with the statements listed in Figure 2 above. Figure 4 shows the results. 20% of the respondents believe that protecting genome privacy is impossible as an individual genomic data can be obtained from his leftover cells (Probe 1). Almost half of the respondents consider genomic data to be no different than other health data (Probe 2). Even though genomic information is, in most instances, non-deterministic, all respondents believe that this fact does not reduce the importance of genome privacy (Probe 3). Only 7% of our respondents think that protecting genome privacy should be left to bioinformaticians (Probe 4). Furthermore, 20% of the respondents believe that genome privacy can be fully guaranteed by legislation (Probe 5). Notably, only 7% of the respondents think that privacy enhancing technologies are a nuisance in the case of genetics (Probe 6). According to only about 10% of the respondents, the confidentiality of genomic data is superfluous because it is hard to identify a person from her variants (Probe 7). And, finally, about 30% of the respondents think that advantages that will be brought by genomics in healthcare will justify the harm that might be caused by privacy issues (Probe 8).

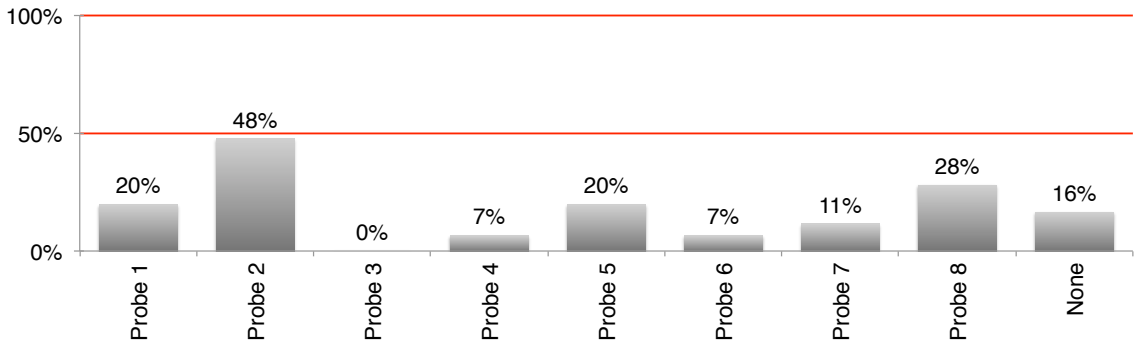

Figure 4: Response to the question: *Do you believe that: (Multiple options can be checked).* The probes are described in detail in Figure 2. "None" means the respondent does not agree with any of the probes.

We asked participants whether they would share their genomes on the Web (Figure 5). 48% of the respondents are *not* in favor of doing so, while 30% would reveal their genome anonymously, and 8% would reveal their identities alongside their genome. We also asked respondents how they think about the scope of the individual's right to share one's genomic data given that the data contain information about the individual's blood relatives. Figure 6 shows that only 18% of the respondents think that one should *not* be allowed to share, 43% of the respondents think that one should be allowed to share only for medical purpose, and 39% of the respondents think that one should have the right to share her genomic data publicly.

As discussed in Section IV, there is a tension between the desire for genome privacy and biomedical research. Thus, we asked the survey participants what they would trade for privacy. The results (shown in Figure 7) indicate that the respondents are willing to trade money and test time (duration) to protect privacy, but they usually do not accept trading accuracy or utility.

We also asked the respondents to evaluate the importance of existing and ongoing research directions on genome privacy (as discussed in detail in Section VII), considering the types of problems they are trying to solve (Figure 8). The majority of respondents think that genomics privacy is

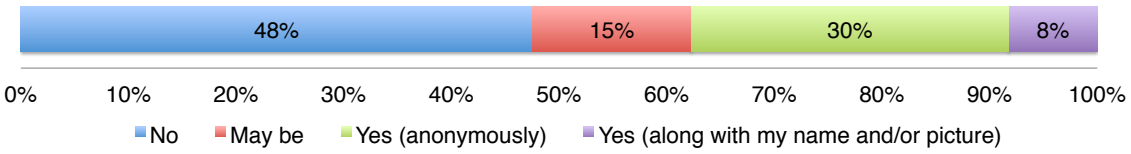

Figure 5: Response to the question: *Would you publicly share your genome on the Web?*

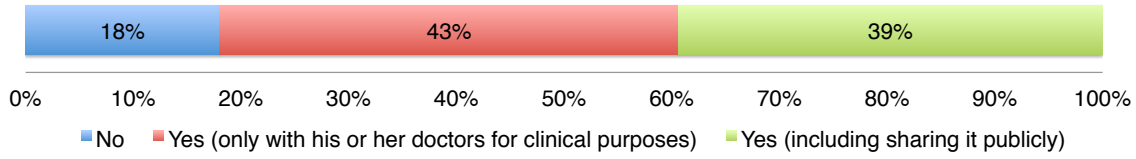

Figure 6: Response to the question: *Assuming that one's genomic data leaks a lot of private information about his or her relatives, do you think one should have the right to share his or her genomic data?*

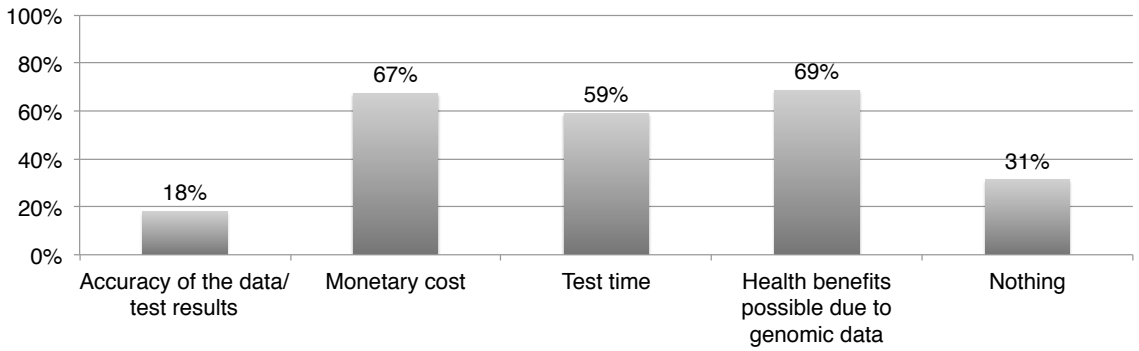

Figure 7: Response to the question: *What can we compromise to improve privacy of genomic data? (Multiple options can be checked)*

important.

V.F Discussion

Our results show that these biomedical researchers believe that genomic privacy is important and needs special attention. Figure 4 shows that, except for Probes 2 and 8, 80% of the biomedical experts do not endorse the statements listed in Figure 2. Approximately three-quarters of the biomedical experts believe that advantages of genome-based healthcare do not justify the harm that can be caused by the genome privacy breach. Probe 2 is an interesting result in that about half of the biomedical experts believe that genomic data should be treated as any other sensitive health data. This seems reasonable because, at the moment, health data can, in many instances, be more sensitive than genomic data. The biomedical community also agrees on the importance of current genome privacy research. Figure 8 shows that these biomedical researchers ranks the placement of genomic data to the cloud as their prime concern. Moreover, they agree with the importance of other genome privacy research topics shown in Figure 8.

We provide additional results stratified according to the expertise of the participants in the appendix.

VI Known Privacy Risks

In this section, we survey a wide spectrum of privacy threats to human genomic data, as reported by prior research.

VI.A Re-identification Threats

Re-identification is probably the most extensively studied privacy risk in dissemination and analysis of human genomic data. In such an attack, an unauthorized party looks at the published human genomes that are already under certain protection to hide the identity information of their donors

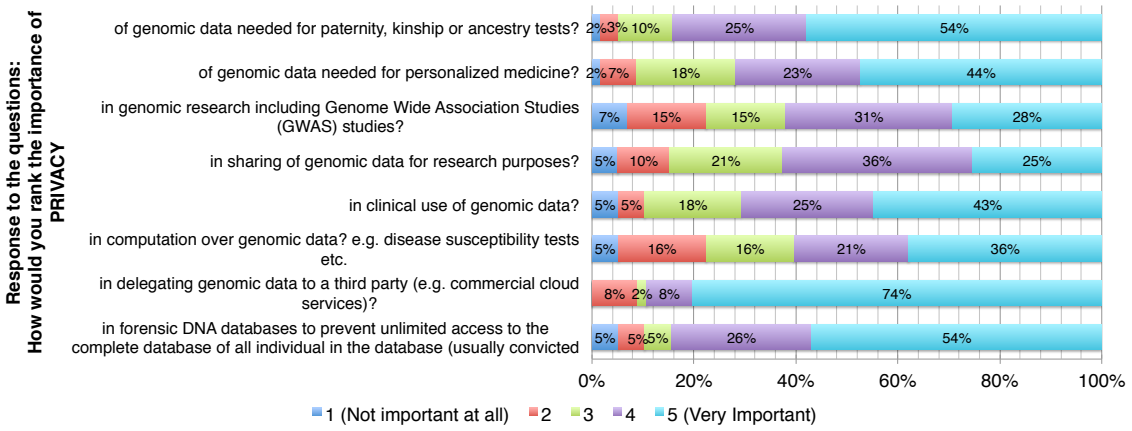

Figure 8: Relevance of genome privacy research done by the computer science community.

(e.g., patients), and tries to recover the identities of the individuals involved. Such an attack, once it succeeds, can cause serious damage to those donors, e.g., discrimination and financial loss. In this section, we review the weaknesses within existing privacy protection techniques that make this type of attack possible.

Pseudo-anonymized Data. A widely used method for protecting health information is the removal of explicit and quasi-identifying attributes (e.g., name and date of birth). Such redaction meets legal requirements to protect privacy (e.g., de-identification under the U.S. Health Insurance Portability and Accountability Act) for traditional health records. However, genomic data cannot be anonymized by just removing the identifying information. There is always a risk for the adversary to infer the *phenotype* of a DNA-material donor (that is, the person's observable characteristics like eye/hair/skin colors), which will lead to her identification, from her *genotypes* (her genetic makeup). Even though the techniques for this purpose are still rudimentary, the rapid progress in genomic research and technologies is quickly moving us toward that end. Moreover, re-identification can be achieved through inspecting the *background* information that comes with publicized DNA sequences [Gitschier, 2009; Gymrek *et al.*, 2013; Hayden, 2013]. As an example, genomic variants on the Y chromosome have been correlated with surnames (for males), which can be found out using public geneology databases. Other instances include identifying Personal Genome Project (PGP) participants through public demographic data [Sweeney *et al.*, 2013], recovering the identities of family members from the data released by the 1000 Genome Project using public information (e.g. death notices) [Malin, 2006], and other correlation attacks [Malin and Sweeney, 2004]. It has been shown that even cryptographically secure protocols leaks a lot of information when used for genomic data [Goodrich, 2009].

Attacks on Machine Learning Models. Most attacks on genomic data use the entire dataset for the attack. Recently, [Fredrikson *et al.*, 2014] showed that even the machine learning models trained on the genomic data can reveal information about the people whose data was used for training the model as well as any arbitrary person given some background information.

VI.B Phenotype Inference

Another critical privacy threat to human genome data is inference of sensitive phenotype information from the DNA sequence. Here we summarize related prior studies.

Aggregate Genomic Data. In addition to the aforementioned re-identification threats (discussed in Section VI.A, which comes from the possible correlation between an individual's genomic data and other public information, the identity of a participant of a genomic study can also be revealed by a "second sample", that is, part of the DNA information from the individual. This happens, for example, when one obtains a small amount of genomic data from another individual, such as a

small set of her single nucleotide polymorphisms (SNPs), and attempts to determine her presence in a clinical study on HIV (a phenotype), based on anonymized patient DNA data published online. This turns out to be rather straightforward, given the uniqueness of individual's genome. Particularly, in 2004, research shows that as few as 75 independent SNPs are enough to uniquely distinguish one individual from others [Lin *et al.*, 2004]. Based on this observation, the genomic researchers generally agree that such DNA raw data are too sensitive to release through online repositories (such as the NIH's PopSet resources), without proper agreements in place. An alternative is to publish "pooled" data, in which summary statistics are disclosed for the case and control groups of individuals in a study.

Yet, [Homer *et al.*, 2008] showed that when an adversary had access to a known participant's genome sequence, they could determine if the participant was in a certain group. Specifically, the researchers compared one individual's DNA sample to the rates at which her variants show up in various study populations (and a *reference* population that does not include the individual) and applied a statistical hypothesis test to determine the likelihood of which group she is in (i.e., case or reference). *The findings of the work led the NIH, as well as the Wellcome Trust in the UK, to remove all publicly available aggregate genomic data from their websites.* Ever since, researchers are required to sign a data use agreement (prohibiting re-identification) to access such data [Zerhouni and Nabel, 2008], the process of which could take several months. At the same time, such attacks were enhanced. First, Homer's test statistic was improved through exploitation of genotype frequencies [Jacobs *et al.*, 2009], while an alternative, based on linear regression was developed to facilitate more robust inference attacks [Masca *et al.*, 2011]. Wang et. al. [Wang *et al.*, 2009a] demonstrated, perhaps, an even more powerful attack by showing that an individual can be identified even from the aggregate statistical data (linkage disequilibrium measures) published in research papers. While the methodology introduced in [Homer *et al.*, 2008] requires on the order of 10,000 genetic variations (of the target individual), this new attack requires only on the order of 200. Their approach even shows the possibility of recovering part of the DNA raw sequences for the participants of biomedical studies, using the statistics including p-values and coefficient of determination (r^2) values.

Quantification of information content in aggregate statistics obtained as an output of genome-wide association studies (GWAS) shows that an individual's participation in the study and her phenotype can be inferred with high accuracy [Im *et al.*, 2012; Craig *et al.*, 2011]. Beyond these works, it has been shown that a Bayesian network could be leveraged to incorporate additional background information, and thus improve predictive power [Clayton, 2010]. It was recently shown that RNA expression data can be linked to the identity of an individual through the inference of SNPs [Schadt *et al.*, 2012].

Yet, there is debate over the practicality of such attacks. Some researchers believe that individual identification from pooled data is hard in practice [Braun *et al.*, 2009; Sankararaman *et al.*, 2009; Visscher and Hill, 2009; Gilbert, 2008]. In particular, it has been shown that the assumptions required to accurately identify individuals from aggregate genomic data rarely hold in practice [Braun *et al.*, 2009]. Such inference attacks depend upon the ancestry of the participants, the absolute and relative number of people in case and control groups, and the number of SNPs [Masca *et al.*, 2011] and the availability of the second sample. Thus, the false positive rates are much higher in practice. Still, others believe that publication of complete genome wide aggregate results are dangerous for privacy of the participants [Lumley and Rice, 2010; Church *et al.*, 2009]. Furthermore, the NIH continues to adhere to its policy of data use agreements.

Beyond the sharing of aggregate data, it should be recognized that millions of people are sequenced or genotyped for the state-of-the-art GWAS studies. This sequenced data is shared among different institutions with inconsistent security and privacy procedures [Brenner, 2013]. On the one hand, this could lead to serious backlash and fear to participate in such studies. On the other hand, not sharing this data could severely impede biomedical research. Thus, measures should be taken to mitigate the negative outcomes of genomic data sharing [Brenner, 2013].

Correlation of Genomic Data. Partially available genomic data can be used to infer the unpublished genomic data due to linkage disequilibrium (LD), a correlation between regions of the genome [Halperin and Stephan, 2009; Marchini and Howie, 2010]. For example, Jim Watson (the discoverer of DNA) donated his genome for research but concealed his ApoE gene, because it reveals

susceptibility to Alzheimer's disease. Yet, it was shown that the ApoE gene variant can be inferred from the published genome [Nyholt *et al.*, 2008]. Such completion attacks are quite relevant in DTC environments, where customers have the option to hide some of the variants related to a particular disease.

While all the prior genomic privacy attacks exploit low-order SNP correlations, [Samani *et al.*, 2015] show that high-order SNP correlations result in far more powerful attacks.

[Wagner, 2015] investigates 22 different privacy metrics to study which metrics are more meaningful to quantify the loss of genomic privacy due to correlation of genomic data.

Kin Privacy Breach. A significant part of the population does not want to publicly release their genomic data [McGuire *et al.*, 2011]. Disclosures of their relatives can thus threaten the privacy of such people, who never release their genomic data. The haplotypes of the individuals *not* sequenced or genotyped can be obtained using LD-based completion attacks [Kong *et al.*, 2008]. For instance, if both parents are genotyped, then most of the variants for their offspring can be inferred. The genomic data of family members can also be inferred using data that has been publicly shared by blood relatives and domain-specific knowledge about genomics [Humbert *et al.*, 2013]. Such reconstruction attacks can be carried out using *(i)* (partial) genomic data of a subset of family members, and *(ii)* publicly-known genomic background information (linkage disequilibrium and minor allele frequencies (MAFs). This attack affects individuals whose relatives publicly share genomic data (obtained using DTC services) on the Internet (e.g. on openSNP [Greshake *et al.*, 2014]). The family members of the individuals who publish their genomic data on openSNP can be found on social media sites, such as Facebook [Humbert *et al.*, 2013].

Note that "correlation of genomic data" and "kin privacy breach" attacks are based on different structural aspects of genomic data. While correlation attacks are based on the linkage disequilibrium (LD), which is a genetic variation within an individual's genome. A kin privacy breach is caused by genomic correlations among individuals. Moreover, a kin privacy breach can also be realized through phenotype information alone. For instance, a parent's skin color or height can be used to predict their child's skin color or height.

VI.C Other Threats

In addition to the above threats, there are a few other genome-related privacy issues. We describe them below:

Anonymous Paternity Breach. As mentioned above, the Y chromosome is inherited from father to son virtually intact and genealogy databases link this chromosome to the surname to model ancestry. Beyond the case discussed above, this information has been used to identify sperm donors in several cases. For example, a 15 year boy who was conceived using donor sperm, successfully found his biological father by sending his cheek swab to a genealogy service and doing Internet search [Motluk, 2005; Stein, 2005]. Similarly, an adopted child was able to find his real father with the help of a genealogy database (and substantial manual effort) [Naik, 2009]. In short, DNA testing has made tracing anonymous sperm donors easy and theoretically sperm donors can no more be anonymous [Lehmann-Haupt, 2010].

Legal and Forensic. DNA is collected for legal and forensic purposes from criminals[8] and victims[9]. On the one hand, forensic techniques are becoming more promising with the evolving technology [Kayser and de Knijff, 2011; Pakstis *et al.*, 2010]. On the other hand, abuse of DNA (e.g., to stage crime scenes) have already baffled people and law enforcement agencies [Bobellan, 2010]. Some people like Madonna (the singer) are paranoid enough about the misuse of their DNA that they hire DNA sterilization teams to clean up their leftover DNA (e.g., stray hairs or saliva) [Villalva, 2012]. We are not aware of any privacy risk assessment studies done primarily in legal and forensic context, in part because law enforcement agencies store a very limited amount of genetic markers. Yet, in the future, it could well happen that law enforcement agencies will have access to the database of

[8] http://www.justice.gov/ag/advancing-justice-through-dna-technology-using-dna-solve-crimes
[9] http://www.rainn.org/get-information/sexual-assault-recovery/rape-kit

whole genome sequences. We discussed sperm donor paternity breach above which is also relevant in legal context.

VII State-of-the-art Solutions

In this section, we provide an overview of technical approaches to address various privacy and security issues related to genomic data. Despite the risks associated with genomic data, we can find ways to mitigate them to move forward [Altman *et al.*, 2013]. Some solutions are efficient enough for practical use, while others need further improvement to become practical. In particular, practical solutions often exploit the special nature of the genomic data to find ways to be efficient under relevant domain assumptions.

VII.A Healthcare

Personalized medicine. Personalized medicine promises to revolutionize healthcare through treatments tailored to an individual's genomic makeup and genome-based disease risk tests that can enable early diagnosis of serious diseases. Various players have different concerns here. Patients, for instance, are concerned about the privacy of their genomes. Healthcare organizations are concerned about their reputation and the trust of their clients. And for-profit companies, such as pharmaceutical manufacturers, are concerned about the secrecy of their disease markers (proprietary information of business importance).

A disease risk test can be expressed as a regular expression query taking into account sequencing errors and other properties of sequenced genomic data. Oblivious automata enable regular expression queries to be computed over genome sequence data while preserving the privacy of both the queries and the genomic data [Troncoso-Pastoriza *et al.*, 2007; Frikken, 2009]. Cryptographic schemes have been developed to delegate the intensive computation in such a scheme to a public cloud in a privacy-preserving fashion [Blanton *et al.*, 2012].

Alternatively, it has been shown that a cryptographic primitive called Authorized Private Set Intersection (A-PSI) can be used in this setting [Baldi *et al.*, 2011; Cristofaro *et al.*, 2012]. In personalized medicine protocols based on A-PSI, the healthcare organization provides cryptographically-authorized disease markers, while the patient supplies her genome. In this setting, a regulatory authority, such as the U.S. Food and Drug Administration (FDA), can also certify the disease markers before they can be used in a clinical setting. Despite its potential, this protocol has certain limitations. First, it is not very efficient in terms of its communication and computation costs. Second, the model assumes that patients store their own genomes, which is not necessarily the case in practice.

To address the latter issue, it has been suggested that the storage of the homomorphically encrypted variants (e.g., SNPs) can be delegated to a semi-honest third party [Ayday *et al.*, 2013c]. A healthcare organization can then request the third party to compute a disease susceptibility test (weighted average of the risk associated with each variant) on the encrypted variants using an interactive protocol involving *(i)* the patient, *(ii)* the healthcare organization and *(iii)* the third party. Additive homomorphic encryption enables a party with the public key to add ciphertexts or multiply a plaintext constant to a ciphertext. Additive homomorphic encryption based methods can also be used to conduct privacy-preserving computation of disease risk based on both genomic and non-genomic data (e.g., environmental and/or clinical data) [Ayday *et al.*, 2013d]. One of the problems with such protocols, however, is that storage of homomorphically encrypted variants require orders of magnitude more memory than plaintext variants. However, a trade-off between the storage cost and level of privacy can be composed [Ayday *et al.*, 2013b]. A second problem is that when an adversary has knowledge of the LD between the genome regions and the nature of the test, the privacy of the patients will decrease when tests are conducted on their homomorphically encrypted variants. This loss of privacy can be quantified using an entropy-based metric [Ayday *et al.*, 2013e]. [Danezis and De Cristofaro, 2014] propose two cryptographic protocols using framework proposed by [Ayday *et al.*, 2013e]. The first protocol involves a patient and a medical center (MC).

MC encrypts the (secret) weights, sends them to the patient's smartcard, and operations are done inside the smartcard. This protocol also hides which and how many SNPs are tested. Second protocol is based on secret sharing in which the (secret) weights of a test is shared between the SPU and the MC. This protocol still relies on a smartcard (held by the patient) to finalize the computation. [Djatmiko *et al.*, 2014] propose a secure evaluation algorithm to compute genomic tests that are based on a linear combination of genome data values. In their setting, a medical center prescribes a test and the client (patient) accesses a server via his mobile device to perform the test. The main goals are to *(i)* keep the coefficients of the test (secret weights) secret from the client, *(ii)* keep selection of the SNPs confidential from the client, and *(iii)* keep SNPs of the client confidential from the server (server securely selects data from the client). They achieve these goals by using a combination of additive homomorphic encryption (Paillier's scheme) and private information retrieval. Test calculations are performed on the client's mobile device and the medical server can also perform some related computations. Eventually, client gets the result and shows it to his physician. As a case study, the authors implemented the Warfarin dosing algorithm as a representative example. They also implemented a prototype system in an Android App. In [Karvelas *et al.*, 2014], authors propose a technique to store genomic data in encrypted form, use an Oblivious RAM to access the desired data without leaking the access pattern, and finally run secure two-party computation protocol to privately compute the required function on the retrieved encrypted genomic data. The proposed construction includes two separate servers: cloud and proxy.

Functional encryption allows a user to compute on encrypted data and learn the result in plaintext in a non-interactive fashion. However, currently functional encryption is very inefficient. [Naveed *et al.*, 2014] propose a new cryptographic model called "Controlled Functional Encryption (C-FE)" that allows construction of realistic and efficient schemes. The authors propose two C-FE constructions: one for inner-product functionality and other for any polynomial-time computable functionality. The former is based on a careful combination of CCA2 secure public-key encryption with secret sharing, while the later is based on a careful combination of CCA2 secure public-key encryption with Yao's garbled circuit. C-FE constructions are based on efficient cryptographic primitives and perform very well in practical applications. The authors evaluated C-FE constructions on personalized medicine, genomic patient similarity, and paternity test applications and showed that C-FE provides much better security and efficiency than prior work.

Raw aligned genomic data. Raw aligned genomic data, that is, the aligned outputs of a DNA sequencer, are often used by geneticists in the research process. Due to the limitations of current sequencing technology, it is often the case that only a small number of nucleotides are read (from the sequencer) at a time. A very large number of these *short reads*[10], covering the entire genome are obtained, and are subsequently aligned, using a reference genome. The position of the read relative to the reference genome is determined by finding the approximate match on the reference genome. With today's sequencing techniques, the size of such data can be up to 300GB per individual (in the clear), which makes public key cryptography impractical for the management of such data. Symmetric stream cipher and order-preserving encryption [Agrawal *et al.*, 2004] provide more efficient solutions for storing, retrieving, and processing this large amount of data in a privacy-preserving way [Ayday *et al.*, 2014]. Order-preserving encryption keeps the ordering information in the ciphertexts to enable range queries on the encrypted data. We emphasize that order-preserving encryption may not be secure for most practical applications.

Genetic compatibility testing. Genetic compatibility testing is of interest in both healthcare and DTC settings. It enables a pair of individuals to evaluate the risk of conceiving an unhealthy baby. In this setting, PSI can be used to compute genetic compatibility, where one party submits the fingerprint for his or her genome-based diseases, while the other party submits her or his entire genome. In doing so, the couple learns their genetic compatibility without revealing their entire genomes [Baldi *et al.*, 2011]. This protocol leaks information about an individual's disease risk status to the other party and its requirements for computation and communication may make it impractical.

[10] A short read corresponds to a sequence of nucleotides within a DNA molecule. The raw genomic data of an individual consists of hundreds of millions of short reads. Each read typically consists of 100 nucleotides.

Pseudo-anyonymization. Pseudo-anonymization is often performed by the healthcare organization that collects the specimen (possibly by pathologists) to remove patient identifiers before sending the specimen to a sequencing laboratory. In lieu of such information, a pseudonym can be derived from the genome itself and public randomness, independently at the healthcare organization and sequence laboratory for symmetric encryption [Cassa *et al.*, 2013]. This process can mitigate sample mismatch at the sequencing lab. However, since the key is derived from the data that is encrypted using the same key, symmetric encryption should guarantee circular security (security notion required when cipher is used to encrypt its own key), an issue which is not addressed in the published protocol.

Confidentiality against Brute-force Attacks. History has shown that encryption schemes have limited lifetime before they are broken. Genomic data, however, has lifetime much longer than that of a state-of-the-art encryption schemes. A brute-force attack works by decrypting the ciphertext with all possible keys. Honey encryption [Juels and Ristenpart, 2014] guarantees that a ciphertext decrypted with an incorrect key (as guessed by an adversary) results in a plausible-looking yet incorrect plaintext. Therefore, HE gives encrypted data an additional layer of protection by serving up fake data in response to every incorrect guess of a cryptographic key or password. However, HE relies on a highly accurate distribution-transforming encoder (DTE) over the message space. Unfortunately, this requirement jeopardizes the practicality of HE. To use HE the message space needs to be understood quantitatively, that is, the precise probability of every possible message needs to be understood. When messages are not uniformly distributed, characterizing and quantifying the distribution is non-trivial. Building an efficient and precise DTE is the main challenge when extending HE to a real use case; [Huang *et al.*, 2015] have designed such a DTE for genomic data. We note that HE scheme for genomic data is not specific to healthcare and is relevant for any use of genomic data.

VII.B Research

Genome-Wide Association Studies (GWAS). Genome-Wide Association Studies (GWAS), [11] are conducted by analyzing the statistical correlation between the variants of a *case group* (i.e., phenotype positive) and a *control group* (i.e., phenotype negative). GWAS is one of the most common types of studies performed to learn genome-phenome associations. In GWAS the aggregate statistics (e.g., p-values) are published in scientific articles and are made available to other researchers. As mentioned earlier, such statistics can pose privacy threats as explained in Section VI.

Recently, it has been suggested that such information can be protected through the application of noise to the data. In particular, differential privacy, a well-known technique for answering statistical queries in a privacy preserving manner [Dwork, 2006], was recently adapted to compose privacy preserving query mechanisms for GWAS settings [Fienberg *et al.*, 2011; Johnson and Shmatikov, 2013]. A mechanism K gives ϵ-differential privacy if for all databases D and D' differing on at most one record, the probability of $K(D)$ is less than or equal to the probability of $\exp(\epsilon) \times K(D')$. In simple words, if we compute a function on a database with and without a single individual and the answer in both cases is approximately the same, then we say that the function is differentially private. Essentially, if the answer does not change when an individual is or is not in the database, the answer does not compromise the privacy of that individual. [Fienberg *et al.*, 2011] propose methods for releasing differentially private minor allele frequencies (MAFs), chi-square statistics, p-values, the top-k most relevant SNPs to a specific phenotype, and specific correlations between particular pairs of SNPs. These methods are notable because traditional differential privacy techniques are unsuitable for GWAS due to the fact that the number of correlations studied in GWAS is much larger than the number of people in the study. However, differential privacy is typically based on a mechanism that adds noise (e.g., by using Laplacian noise, geometric noise, or exponential mechanism), and thus requires a very large number of research participants to guarantee acceptable levels of privacy and utility. [Yu *et al.*, 2014] have extended the work of [Fienberg *et al.*, 2011] to compute differentially private chi-square statistics for arbitrary number of cases and controls. [Johnson and Shmatikov,

[11] http://www.genome.gov/20019523

2013] explain that computing the number of relevant SNPs and the pairs of correlated SNPs are the goals of a typical GWAS and are not known in advance. They provide a new exponential mechanism – called a distance-score mechanism – to add noise to the output. All relevant queries required by a typical GWAS are supported, including the number of SNPs associated with a disease and the locations of the most significant SNPs. Their empirical analysis suggests that the new approach produces acceptable privacy and utility for a typical GWAS.

A meta-analysis of summary statistics from multiple independent cohorts is required to find associations in a GWAS. Different teams of researchers often conduct studies on different cohorts and are limited in their ability to share individual-level data due to Institutional Review Board (IRB) restrictions. However, it is possible for the same participant to be in multiple studies, which can affect the results of a meta-analysis. It has been suggested that one-way cryptographic hashing can be used to identify overlapping participants without sharing individual-level data [Turchin and Hirschhorn, 2012].

[Xie *et al.*, 2014] proposed a cryptographic approach for privacy preserving genome-phenome studies. This approach enables privacy preserving computation of genome-phenome associations when the data is distributed among multiple sites.

Sequence comparison. Sequence comparison is widely used in bioinformatics (e.g., in gene finding, motif finding, and sequence alignment). Such comparison is computationally complex. Cryptographic tools such as fully homomorphic encryption (FHE) and secure multiparty computation (SMC) can be used for privacy-preserving sequence comparison. Fully homomorphic encryption enables any party with the public key to compute any arbitrary function on the ciphertext without ever decrypting it. Multiparty computation enables a group of parties to compute a function of their inputs without revealing anything other than the output of the function to each other. It has been shown that fully homomorphic encryption (FHE), secure multiparty computation (SMC), and other traditional cryptographic tools [Atallah *et al.*, 2003; Jha *et al.*, 2008] can be applied for comparison purposes, but they do not scale to a full human genome. Alternatively, more scalable provably secure protocols exploiting public clouds have been proposed [Blanton *et al.*, 2012; Atallah and Li, 2005]. Computation on the public data can be outsourced to a third party environment (e.g., cloud provider) while computation on sensitive private sections can be performed locally; thus, outsourcing most of the computationally intensive work to the third party. This computation partitioning can be achieved using *program specialization* which enables concrete execution on public data and symbolic execution on the sensitive data [Wang *et al.*, 2009b]. This protocol takes advantage of the fact that genomic computations can be partitioned into computation on public data and private data, exploiting the fact that 99.5% of the genomes of any two individuals are similar.

Moreover, genome sequences can be transformed into sets of offsets of different nucleotides in the sequence to efficiently compute similarity scores (e.g., Smith-Waterman computations) on outsourced distributed platforms (e.g., volunteer systems). Similar sequences have similar offsets, which provides sufficient accuracy, and many-to-one transformations provide privacy [Szajda *et al.*, 2006]. Although this approach does not provide provable security, it does not leak significant useful information about the original sequences.

Until this point, all sequence comparison methods we have discussed work on complete genomic sequences. Compressed DNA data (i.e., the variants) can be compared using novel data structure called Privacy-Enhanced Invertible Bloom Filter [Eppstein *et al.*, 2011]. This method provides communication-efficient comparison schemes.

Person-level genome sequence records. Person-level genome sequence records contrast with the previous methods which obscure sequences and report on aggregated data rather than that of a single person. Several techniques have been proposed for enabling privacy for person-level genome sequences. For instance, SNPs from several genomic regions can be generalized into more general concepts – e.g.; transition (change of A↔G or T↔C.), transversion (change of A↔C, A↔T, C↔G or G↔T.), and exact SNP positions into approximate positions) [Lin *et al.*, 2002]. This generalization makes re-identification of an individual sequence difficult according to a prescribed level of protection. In particular, k-anonymity can be used to generalize the genomic sequences such that a sequence is indistinguishable from at least other $k-1$ sequences. Also, the problem of SNP anonymization

can be expanded to more complex variations of a genome using multiple sequence alignment and clustering methods [Malin, 2005b; Li *et al.*, 2012]. However, such methods are limited in that they only work when there are a large number of sequences with relatively small number of variations.

Given the limitations of generalization-based strategies, it has been suggested that cryptographic techniques might be more appropriate for maintaining data utility. In particular, it has been shown that additive homomorphic encryption can be used to share encrypted data while still retaining the ability to compute a limited set of queries (e.g., secure frequency count queries which are useful to many analytical methods for genomic data) [Kantarcioglu *et al.*, 2008]. Yet, this method leaks information in that it reveals the positions of the SNPs, which in turn reveals the type of test being conducted on the data. Moreover, privacy in this protocol comes at a high cost of computation.

Cryptographic hardware at the remote site can be used as a trusted computation base (TCB) to design a framework in which all person-level biomedical data is stored at a central remote server in encrypted form [Canim *et al.*, 2012]. The server can compute over the genomic data from a large number of people in a privacy-preserving fashion. This enables researchers to compute on shared data without sharing person-level genomic data. This approach is efficient for typical biomedical computations; however, it is limited in that trusted hardware tends to have relatively small memory capacities, which dictate the need for load balancing mechanisms.

Sequence alignment. Sequence alignment is fundamental to genome sequencing. The increase in the quantity of sequencing data is growing at a faster rate than the decreasing cost of computational power, thus the delegation of read mapping to the cloud can be very beneficial. However, such delegation can have major privacy implications. Chen et. al. [Chen *et al.*, 2012] have shown that read mapping can be delegated to the public cloud in a privacy preserving manner using a hybrid cloud based approach. They exploit the fact that a sequence of a small number of nucleotides ($\approx$20) is unique and two sequences of equal length with edit distance of x, when divided into $x+1$ segments will have at least one matching segment. Based on this fact, computation is divided into two parts: *(i)* the public part is delegated to the public cloud, in which the public cloud finds exact matches on encrypted data and returns a small number of matches to the private cloud, whereas *(ii)* the private part takes place in a private cloud, which computes the edit distance using only the matches returned by the public cloud. This approach reduces the local computation by a factor of 30 by delegating 98% of the work to the public cloud.

VII.C Legal and Forensic

Paternity testing. Paternity testing determines whether a certain male individual is the father of another individual. It is based on the high similarity between the genomes of a father and child (99.9%) in comparison to two unrelated human beings (99.5%). It is not known exactly which 0.5% of the human genome is different between two humans, but a properly chosen 1% sample of the genome can determine paternity with high accuracy [Gibbs and Singleton, 2006]. Participants may want to compute the test without sharing any information about their genomes.

Once genomes of both individuals are sequenced, a privacy-preserving paternity test can be carried out using PSI-Cardinality (PSI-CA), where inputs to PSI-CA protocol are the sets of nucleotides comprising the genome. The size of the human genome, or even 1% of it, cannot be handled by current PSI and other SMC protocols. However, by exploiting domain knowledge, the computation time can be reduced to 6.8ms and network bandwidth usage to 6.5KB by emulating the Restriction Fragment Length Polymorphism (RFLP) chemical test in software, which reduces the problem to finding the intersection between two sets of size 25 [Baldi *et al.*, 2011]. Subsequent work demonstrates a framework for conducing such tests on a Android smartphone [Cristofaro *et al.*, 2012]. Since the ideal output of the privacy-preserving paternity test should be yes or no, it cannot be obtained using custom PSI protocols, whereas generic garbled circuit based protocols can be easily modified to add this capability [Huang *et al.*, 2011, 2012]. [He *et al.*, 2014; Hormozdiari *et al.*, 2014] propose cryptographic protocols for identifying blood relatives.

Criminal forensics. Criminal forensic rules enable law enforcement agencies to have unlimited access to the complete DNA record database of millions of individuals, usually of convicted criminals

(e.g., CODIS[12] in the US). The motivation behind creating such a database is to find a record that matches the DNA evidence from a crime scene. Yet, providing unlimited access to law enforcement agencies is unnecessary and may open the system to abuse. Cryptographic approaches have been developed to preserve the privacy of the records that fail to match the evidence from the crime scene [Bohannon *et al.*, 2000]. Specifically, DNA records can be encrypted using a key that depends upon certain tests, such that when DNA is collected from a crime scene, the scheme will only allow decryption of the records that match the evidence.

Finally, partial homomorphic encryption can be used for privacy-preserving matching of Short Tandem Repeat (STR) DNA profiles in an honest-but-curious model [Bruekers *et al.*, 2008]. Such protocols (described in Section VII.D) are useful for identity, paternity, ancestry, and forensic tests.

VII.D Direct-to-consumer (DTC)

Many DTC companies provide genealogy and ancestry testing. Cryptographic schemes can be used to conduct these tests in a privacy-preserving fashion. Partial homomorphic encryption can be cleverly used on STR profiles of individuals to conduct *(i)* common ancestor testing based on the Y chromosome, *(ii)* paternity test with one parent, *(iii)* paternity test with two parents, and *(iv)* identity testing [Bruekers *et al.*, 2008].

Despite the increasing use of DTC genome applications, less focus has been given to the security and privacy of this domain. In particular, genomic data aggregation issues require special attention because some companies allow people to publish high-density SNP profiles online in combination with demographic and phenotypic data.

Focusing on kin genome privacy, [Humbert *et al.*, 2014] build a protection mechanism against the kinship attack [Humbert *et al.*, 2013] that uses DTC genomic data from openSNP [Greshake *et al.*, 2014]. The main focus of the work is to find a balance between the utility and privacy of genomic data. Every family member has a privacy constraint that he wants to protect. At the same time, some family members want to publish (part of) their genomes mainly to facilitate genomic research. The paper proposes a multi-dimensional optimization mechanism in which the privacy constraints of the family members are protected and at the same time the utility (amount of genomic data published by the family members) is maximized.

VIII Challenges for Genome Privacy

While the value of genome sequencing in routine care has yet to be fully demonstrated, it is anticipated that the plummeting cost and commoditization of these analyses will change the practice of medicine. Data confidentiality and individual privacy will be central to the acceptance and widespread usage of genomic information by healthcare systems. However, a clear demonstration of the clinical usefulness of genomics is first needed for doctors and other healthcare providers to fully embrace genomics and privacy.

VIII.A Consumer-driven Genomics

An unprecedented aspect of contemporary genomics that comes with its own set of issues for data confidentiality is democratization, including facilitated access to large-scale personal health-related data. Whereas medical and genetic information used to be obtainable only through hospital or research laboratories, people can now access their own genotyping or sequencing results through direct-to-consumer companies such as 23andMe, as discussed before. On the research side, numerous participant-centric initiatives have recently been launched (notably by citizens' networks such as openSNP [Greshake *et al.*, 2014] and the Personal Genome Project). As a result, genomic data are increasingly found outside the controlled cocoon of healthcare systems or research. In particular, individual genetic results or aggregated datasets are available on the Internet, often with non-existent

[12] Combined DNA Index System (CODIS), The Federal Bureau of Investigation, http://www.fbi.gov/about-us/lab/biometric-analysis/codis

or minimal protection. On one hand, these crowd-based initiatives are very exciting, because they have the potential to stimulate biomedical research, accelerate discoveries and empower individuals. Yet, on the other hand, they raise a number of concerns about potential privacy risks (as highlighted in Section IX). For example, privacy risks must be assessed in the context of the extensive nature of information available on the Internet (including online social networks), and not only within the narrower confines of genomic research or clinical care delivery.

VIII.B Privacy and the Benefits of Genomic Data

It is important to note that both a lack and an excess of privacy have the potential to derail the expected benefits of genomics in healthcare and research. On one hand, the efficient and secure handling of individual genotype and sequence data will be central to the implementation of genomic medicine. The Hippocratic Oath[13] remains a pillar of medical deontology and one of the few stable concepts in the highly tumultuous history of medical practice. The Hippocratic Oath contains a clear statement about the patient's privacy. Trust is at the core of any successful healthcare system: any leakage of highly sensitive genomic information may raise concerns and opposition in the population and among political decision makers. Earning and conserving trust is essential for hospitals and private companies that deal with genomics. As a result, there is a potential for a service industry securing genomic data, either by providing ad hoc solutions or by fully supporting storage and delivery of raw/interpreted sequence information. Fortunately, as detailed in Section VII, there exist a variety of tools that can mitigate the problem.

On the other hand, an excess of privacy-related hurdles could slow down research and interfere with large-scale adoption of genomics in clinical practice. When designing privacy-preserving solutions for genomic data, security and privacy researchers should keep in mind that most end-users are not familiar with computer science and are almost exclusively interested in the clinical utility of test results. Education is again a fundamental requirement for privacy protection. However, in bioinformatics curricula, students are trained to maximize the information to be extracted from (biological) data. Usually, such curricula do not address security and privacy concerns, because adversarial scenarios are out of their scope. Conversely, computer scientists rarely have formal training in biology, let alone genomics. But, they are trained in security, notably because of the formidable challenge raised by the numerous vulnerabilities of the Internet. Consequently, to properly address the concerns about genomic data protection, there is a clear and strong potential of cross-fertilization between these two disciplines.

VIII.C Acceptable Utility vs. Privacy of Genomic Data

The balance between acceptable utility and privacy of genomic data needs to be considered in context.

Healthcare. Patient-level information must be as precise as possible. Because genomic data is used here to support clinical decision, including in life-threatening situations, any decrease in data accuracy must be avoided. Security of electronic medical records and other health-related information is therefore most often guaranteed through restricted access (e.g., intranet use, password and card identification) to unmodified data. It is important to note, however, that genetic testing is typically not urgent, and that privacy-preserving measures that would slightly delay a test result could be tolerated.

Research. Research on complex trait genomics relies on large datasets on which genotyping or sequencing association studies can be run. To gain statistical power and detect meaningful associations, it is often necessary to merge many such studies through meta-analyses that can include data from hundreds of thousands of individuals. Due to non-uniform use of technological platforms, variation in time and place of genotyping, and differences in analysis pipelines, some degree of noise

[13] See http://guides.library.jhu.edu/content.php?pid=23699&sid=190964 for a modern version of the Hippocratic Oath.

is unavoidable. An interesting avenue for research is here to empirically determine whether differential privacy strategies (e.g., [Johnson and Shmatikov, 2013]) can be applied without compromising discovery.

Legal and forensics. DNA collection and search for similarity pattern in genomic data are used in criminal investigations and for other legal purposes such as paternity testing. The accuracy of any test result is here again an absolute requirement to avoid legal prejudice. Extremely stringent data protection must also be ensured due to the highly sensitive nature of such cases.

Direct-to-consumer (DTC) genomics. DTC companies providing individual genomic data have a clear commercial incentive to protect customers' privacy, in order to maintain trust and attract new customers. For example, the 23andMe webpage states: *"Your personalized web account provides secure and easy access to your information, with multiple levels of encryption and security protocols protecting your personal information."*. Of course, these measures are ineffective when individuals choose to unveil their full identity online together with their genomic data, thereby putting their (and their blood relatives) genome privacy at risk, either knowingly (as in the case of Personal Genome Project participants) or out of naivety.

IX Framework for Privacy-Preserving handling of Genomic data

In this section, we provide a general framework for security and privacy in the handling of genomic data. The framework is illustrated in Figure 9. As has been done throughout this paper, we divide the framework into four categories: *(i)* healthcare, *(ii)* research, *(iii)* legal and forensics, and *(iv)* direct-to-consumer genomics. This classification is based on the most popular uses of genomic data; however, we recognize that the boundaries between these categories are blurred and there is significant overlap. For each of these we describe setting, threat model, and solutions and open problems. The setting provides the most general environment around the problem (e.g., we do not discuss the possibility of outsourcing the computation as one can easily extend our setting to another one involving a third party). In this section, we assume that the adversary is computationally bounded. We further assume that the adversary can leverage all publicly available information (e.g., data from the 1000 Genomes Project or public genealogy databases) to her advantage. Moreover, in some cases, the adversary might have access to private data. For instance, people can abuse their access to private data, an adversary can also steal the data, and data can be extracted from a lost laptop.

IX.A Biospecimen

DNA is obtained in chemical form and then digitized. This cyber-physical nature of DNA creates unique challenges for its protection.

IX.A.1 Threat Model

In our threat model, the adversary is capable of *(i)* obtaining DNA from an individual's biological cells either voluntarily (e.g., for research with informed consent) or involuntarily (e.g., leftover hairs or saliva on a coffee cup), *(ii)* sequencing or genotyping the DNA from biospecimen, *(iii)* interpreting the sequenced data to learn identity, disease, kinship, and any other sensitive information, and *(iv)* linking the genomic data (or biospecimen) to the identity, health record, or any arbitrary background information about the individual.

IX.A.2 Solutions and Open Problems

Legal protection is *necessary* to protect the biospecimen and the DNA (in its chemical form). However, a solution to this problem is a subject of public policy and is outside the scope of this paper.

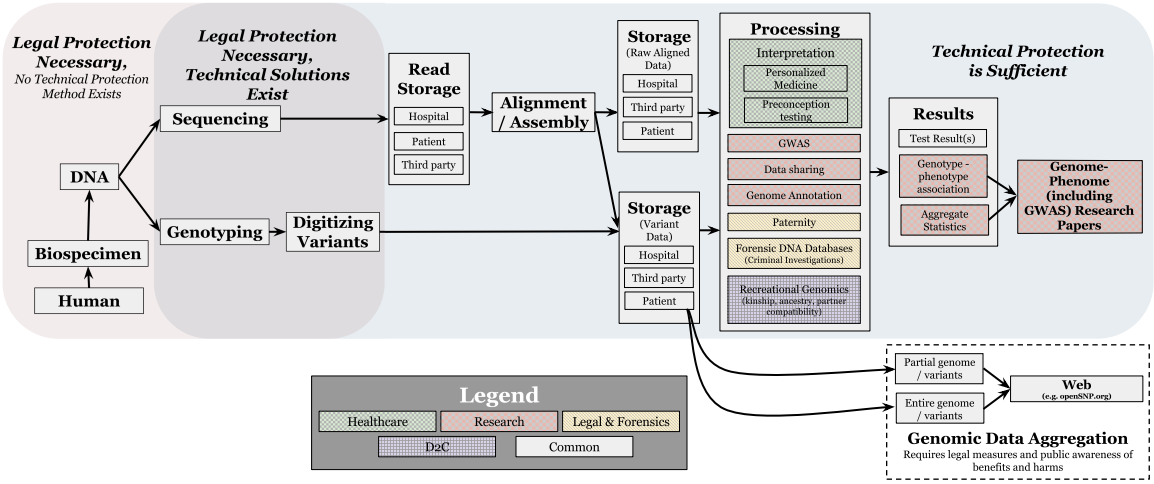

Figure 9: Genomic data handling framework: DNA is extracted from an individual's tissue or cells. DNA is digitized either using sequencing (to obtain Whole Genome Sequence (WGS) or Whole Exome Sequence (WES)) or genotyping (to obtain variants, usually only SNPs). Reads obtained from sequencing are aligned to form the complete genome, while genotyped variants are digitized from the microchip array directly. Read data may be stored for later analysis. The aligned genome can be either stored in raw form or compressed form (variations from a reference human genome). Medical tests and other types of computation shown in the figure can be performed either on raw aligned genome or just on variants. Possible outputs of computation are shown. Output depends on the type of computation and in some cases there is no output. The figure shows the genomic data aggregation problem caused by recreational genomics services. The figure is divided into three sections based on fundamental limitations of legal and technical measures for the protection of genomic data. Legal protection is required for the left section, legal as well as technical protection is required for the middle section, while, in theory, technical solutions would suffice for the protection of the right section. The legend shows which blocks are associated with different uses of genomic data. *We use the word "patient" in this paper to mean someone whose genome is sequenced or genotyped and not necessarily a person who is ill.*

IX.B Digitization: Sequencing/Genotyping

IX.B.1 Setting

A biospecimen is obtained by an agency (e.g., hospital) and is sequenced or genotyped either by the same agency or by an external agency (e.g., Illumina, 23andMe, etc.).

IX.B.2 Threat Model

Since a biospecimen is required for digitization, we assume the threat model discussed in Section IX.A.1 with the following extensions: *(i)* the adversary has full control over the entire sequencing or genotyping infrastructure, *(ii)* the adversary can be honest-but-curious and can attempt to learn partial or entire genomic data or any information derived from the genomic data, and *(iii)* the adversary can be malicious and can compromise the integrity of partial or entire genomic data.

IX.B.3 Solutions and Open Problems

Given the cyber-physical nature of DNA, it is not possible to address this issue with technical measures alone. Both legal and technical protections are required to protect against this threat. An external agency cannot derive genomic data without a biospecimen, and legal protection is required to prevent misuse. Sequencing machines are expensive and are manufactured by a limited number of companies. We envision a well-regulated process for the manufacturing, procurement, and use of sequencing machines and government regulations in place for it. The FDA already regulates the

manufacturing of medical devices[14] [Cheng, 2003]. Regular inspections would check for compliance. Under such legal protections, sequencing machines could have tamper resistant trusted computing base (TCB) that could output encrypted data such that the sequencing agency could not access the plaintext genomic data.

IX.C Storage

We assume that once the adversary has access to the read data, it is easy to get raw aligned data and variant data, hence we present storage of all three forms of data together.

IX.C.1 Setting

Genomic data can be stored by the *(i)* patient[15], *(ii)* healthcare organization (e.g., as part of patient's EHR), or *(iii)* a third party.

IX.C.2 Threat Model

For all settings, we assume that the lifetime of genomic data is much longer than the lifetime of a cryptographic algorithm. We consider the following threat models

Patient: Storage media or the device storing genomic data can be lost, stolen, or temporarily accessed. A patient's computer can be attacked by an adversary (curious or malicious) to compromise confidentiality and/or integrity of the genomic data. We further assume that an adversary can associate the identity and background information (including phenotype information) from any arbitrary source. And, we assume that the adversary can use the compromised genomic data for any arbitrary purpose.

Hospital: We consider all of the threats described for the patient and the following additional threats. An insider (or hacker) has full access to the infrastructure and can link genomic information to the phenotypic information in the patient's health record.[16] We also consider the threat of the healthcare organization communicating incidental findings that are harmful – violating *the right not to know*.[17] We assume that the adversary can intentionally try to figure out variants of regions of the genome from variants in other regions, e.g., to learn some sensitive SNPs that are removed to prevent incidental finding issues. We also assume that a healthcare organization could illegally collude with insurance agencies to facilitate discrimination based on genomic data.

Third party: We consider all of the threats discussed for the hospital and following additional threat. The adversary, as the third party itself, can be interested in de-anonymizing the anonymized genomic data or aggregate genomic data.

IX.C.3 Solutions and Open Problems

We report some solutions in Section VII. Users are generally not equipped with skills and equipment to protect the security and privacy of their genomic data. For the storage of genomic data, an option is to store it on a cloud in an encrypted fashion, which makes the adversary's job harder, as now it needs to circumvent cloud storage security measures and also require to hack into user's computer to steal the decryption key. Efficient encryption schemes allowing secure storage and computation are required.

[14] http://www.fda.gov/medicaldevices/deviceregulationandguidance/postmarketrequirements/qualitysystemsregulations/

[15] We use the word patient to mean an individual whose genome is sequenced or genotyped and not necessarily a person who is ill.

[16] We assume that data stored at the hospital is not anonymized.

[17] For instance, a doctor telling a patient his increased susceptibility to Alzheimer's disease, when he does not want to know. We emphasize that defining what is harmful is an ethical issue and is out of scope of this study.

IX.D Alignment/Assembly

As explained in Section VII, genomic data is obtained from the specimen in the form of short reads. These short reads are then assembled using alignment or assembly algorithms. Alignment is done by comparing the short reads to the reference genome, and is computationally very intensive. Hence, it can be economically beneficial to delegate the alignment to the cloud.

IX.D.1 Setting

Short reads are obtained locally from the sequencing machine and alignment is delegated to an untrusted third party.

IX.D.2 Threat Model

We assume that the third party can be honest-but-curious, as well as malicious, and can return incorrect results for economic or other malicious motives.

IX.D.3 Solutions and open problems

We presented some solutions to this problem in Section VII [Chen *et al.*, 2012]. However, there are several problems with the most efficient solution to date. First, it is not provably secure. Second, its security and efficiency requires that the read size be greater than 100 nucleotides. Third, this scheme only works in a hybrid cloud environment and requires local computation. Given that our opinion poll (described in Section V) shows that third party environments are of the greatest concern to biomedical researchers, a provably secure and efficient solution that is capable of aligning the human genome in a cloud computing environment is an important open research problem.

IX.E Interpretation

Interpretation depends upon two private inputs: the patient's genomic data and an interpretation algorithm (possibly from more than one party). Given the complexity of genomic data, it is unlikely that any single party will have a complete interpretation algorithm. This makes the computation a multiparty process between independent parties and the patient (or an agent of the patient, e.g., a hospital). Although each party with a piece of the interpretation algorithm can compute independently with the patient, collusion of any two parties may leak information about another party's inputs. Moreover, the interpretation of one party may depend upon the interpretation of another party. We assume that all of these parties can be malicious and can collude to learn information (e.g., the patient's genomic data or another parties' algorithm). In some cases, it is necessary to sequence one's parent to draw conclusions, in which case parents might also be concerned about their privacy.

Personalized medicine is a special case of interpretation and depends upon the patient's genomic data and disease markers (possibly distributed among multiple parties).

Preconception testing is another special case of interpretation. It is different from personalized medicine because it is a pre-pregnancy test and measures can be taken to conceive a healthy child (as evident from www.counsyl.com success stories). Additionally, the outcome of the preconception test almost *always* depends upon two people, each of whom might prefer to not disclose their genomic status to the other.

IX.E.1 Setting

The computation is typically performed on private data from multiple parties. The parties involved in the computation are those who input *(i)* their genomic data and *(ii)* interpretation algorithms. The output of the computation should only be released to the patient or authorized physician (possibly using the infrastructure of a healthcare organization).

IX.E.2 Threat Model

We assume that all parties can be honest-but-curious, malicious, or colluding (and possibly all at the same time). They can use arbitrary background knowledge to breach privacy. They may use falsified genomic data or a falsified interpretation algorithm an arbitrary number of times to ascertain another parties' private inputs. Furthermore, they may influence the results in an arbitrary manner.

IX.E.3 Solutions and Open Problems

We discussed some of the solutions for the personalized medicine scenario in Section VII. However, current solutions are limited to privacy-preserving disease susceptibility tests. It is clear that computational solutions that support a broad range of computation over genomic data are needed. At the same time, the design of such systems must be practical and provide reasonable usability, accuracy, security, and privacy.

IX.F Genome-Wide Association Studies (GWAS)

IX.F.1 Setting

The genomic data from two groups of people are collected, one being the case group (i.e., people with the disease or some other trait) and the other being the control group (i.e., people without the disease). Statistical analysis is then conducted to discover the correlation between the disease and genetic variants. The results are subsequently published in research papers and posted online possibly with restricted access (e.g., at dbGaP[18]).

IX.F.2 Threat Model

An adversary may wish to determine if the victim is a GWAS participant or blood relative of a GWAS participant. We assume that the adversary has access to the high density SNP profile of the victim and also to a reference population (which can be obtained from the same GWAS conducted on a different population). The attack succeeds if the adversary learns information from the data produced by GWAS, which she otherwise would not have learned.

IX.F.3 Solutions and Open Problems

There are various solutions that could be applied in this setting. We explained noise-based solutions, such as differential privacy, in Section VII.B. Yet, differential privacy-based solutions make data more noisy, which make adoption of these approaches difficult. This is particularly problematic because biomedical researchers and physicians want more (not less) accurate data than is available today. An ideal solution should preserve the utility of data while preserving the privacy of participants. We believe that more research is required in this area to determine if noise-based approaches can lead to more usable and pragmatic data publications. These approaches may, for instance, build upon well-established sets of practices from other communities. For example, the Federal Committee on Statistical Methodology (FCSM) has a long history of sharing information in a privacy-preserving manner. These practices obey multi-level access principles and, to the best of the authors' knowledge, no significant privacy breach from such domain has been reported.

While the data disclosed by federal agencies is quite different from high-dimensional genomic data, it might be possible to adapt these practices to balance the benefits and harms caused by public sharing of aggregate genomic data. These strategies may be composed of social and technical protections. From a social perspective, a popular method to mitigate risk is through contractual agreements which prohibit the misuse of such data. Such contracts could be complemented by cryptographic protocols that help preserve the privacy of the participants, particularly in settings in which the data is used in the secure computation and only the output of the computation is revealed to a specific party.

[18] http://www.ncbi.nlm.nih.gov/gap

IX.G Data sharing

The majority of genome-phenome discoveries come from very large populations, sometimes on the order of millions of participants. Given the costs and scarcity of such resources, sharing data would fuel biomedical research. However, sharing this data entails privacy implications as discussed earlier.

It should be noted that individual-level genomic data from a large number of people is needed to conduct genome-phenome studies (e.g., GWAS). Moreover, the results of a typical GWAS are published as aggregate genomic data, which is usually made available to researchers at large (e.g., NIH requires results of all NIH-sponsored GWASs to be uploaded to dbGaP and published in research papers — e.g., p-values of relevant SNPs). Therefore, we only focus on the sharing of individual level genomic data here.

IX.G.1 Setting

Genomic data needs to be shared among different research institutions, possibly under different jurisdictions. Privacy-preserving solutions can be built in the following settings: *(i)* all data delegated to and computation done on a trusted party (e.g., a governmental entity), *(ii)* all data delegated to and computation done on an untrusted party, *(iii)* all data stored at and computation done at the collection agency, *(iv)* sharing data using data use agreements, and *(v)* sharing anonymized data.

IX.G.2 Threat Model

We assume that data is being shared between untrusted parties. The parties with whom data is being shared may want to use it for any arbitrary purpose, including using it for participant re-identification, or for finding disease susceptibility of the patients or their blood relatives.

IX.G.3 Solutions and Open Problems

We described some solutions in Section VII.B. However, these solutions do not allow for arbitrary computations on encrypted data. Theoretically, many cryptographic solutions exist to solve this issue. For example, fully homomorphic encryption (FHE) can be used to encrypt the data and arbitrary computations can be done on it while preserving privacy, but data needs to be decrypted by the party that encrypted the data. Functional encryption (FE) could also be used, which allows computation on encrypted data and produces plaintext directly. However, FHE and FE are not sufficiently efficient to be practically useful. The performance of FHE and FE is progressing and these schemes might be usable in the future to support data sharing. Clearly though, specialized efficient solutions for genomic data exploiting nature of genomic data are needed to support specific analytics. Controlled Functional Encryption (C-FE) [Naveed *et al.*, 2014] described above is a promising approach to develop practical solutions.

IX.H Paternity

Genomic data is extensively used to determine parentage and test results are admissible in courts of law. Today, the biospecimen of the individuals involved in the tests are outsourced to a third party in the form of cheek swabs, where the DNA is extracted. Sending one's DNA to a third party could have serious implications for one's privacy.

IX.H.1 Setting

Two parties each have their sequenced genome or genotyped variants and one party wants to know whether the other party is the parent.

IX.H.2 Threat Model

The threat model in this case is the standard model defined for secure two-party computations. We assume that parties can be honest-but-curious or malicious.

IX.H.3 Solutions and Open Problems

In Section VII, we explain some of the solutions to the problem. A chemical test – RFLP – can be simulated for a neat and efficient privacy-preserving solution, given that genomes are stored by individuals themselves [Baldi *et al.*, 2011]. Yao's garbled circuits can be used instead of PSI to output a binary answer (YES or NO) instead of the number of matched segments in simulated RFLP test and therefore reveals the minimum amount of information.

IX.I Forensic DNA Databases

Many countries maintain a huge database of DNA profiles of convicted (and, in some cases, accused) criminals. Law enforcement agencies usually have unlimited access to such a resource, which makes it vulnerable to abuse. It is possible that in the near future, instead of concise DNA profiles, law enforcement agencies will be able to have access to full genome sequences of individuals, which further exacerbates the issues.

IX.I.1 Setting

Police officers collect DNA samples from a crime scene. Then, they want to check whether an individual with the same DNA profile/sequence is present in the DNA records database.

IX.I.2 Threat Model

We assume that the adversary can be honest-but-curious, interested in learning about other people in the database. In addition, the adversary can be malicious. If the adversary has write access to the database, he can also try to compromise the integrity of the record(s) in the database. We also assume that the adversary is able to affect the outcome of a query in an arbitrary manner.

IX.I.3 Solutions and Open Problems

We discussed some of the existing solutions to this problem in Section VII. Theoretically, this problem differs from interpretation and other types of computation, as the privacy for query is not required, only the privacy of the individuals other than the suspect is of concern here. This makes the problem more tractable, possibly making solutions scalable to large databases with millions of records.

IX.J Recreational Genomics

Several commercial companies offer direct-to-consumer genomics services. They include kinship, ancestry, and partner compatibility testing.

IX.J.1 Setting

The customer ships her specimen (usually saliva) to a company. This specimen is used to genotype typically one million SNPs and the data is then digitized and stored in digital form on the server. Some computation is done on this data for ancestry, disease susceptibility, kinship, or other tests. The data and the results are then available for the user to download or see through a browser. Some companies (e.g., 23andme) allow to search for people with common ancestors e.g., 3rd, 4th or 5th cousins.

IX.J.2 Threat Model

We assume the threat models for specimen collection, digitization and interpretation. There are also new threats. The owner of the data posts his data online along with his identity and some phenotypical data, as done for example on openSNP [Greshake *et al.*, 2014]. We assume that the

data owner makes an informed decision, so he willingly gives away his genome privacy. The major threat then is to the blood relatives of the data owner, whose private information is also leaked.

Genomic data aggregation is another issue caused by users posting their genomic data online. This data can be aggregated by companies and then used for commercial or other purposes. It is worth noting that some genome-sharing websites have achievement programs where users get rewards whenever they post their phenotype data (e.g., hair color, eye color, disease pre-disposition, etc.). Genomic data along with phenotypic data is much more valuable than genomic data alone.

IX.J.3 Solutions and Open Problems

All other solutions apply here, however recreational genomics presents new problems of public sharing and genomic data aggregation. We emphasize that public awareness is required to enable people to make an informed decision to share their genomic data publicly because such sharing compromises their own and their relatives' privacy. It should be made clear that in case of abuse of this publicly available data, people would be discouraged to share genomic data even for legitimate research purposes. However, this is a policy and ethics debate and is out of scope of this paper.

Conclusion

The confluence of cheap computing and high-throughput sequencing technologies is making genomic data increasingly easy to collect, store, and process. At the same time, genomic data is being integrated into a wide range of applications in diverse settings (e.g., healthcare, research, forensics, direct-to-consumer), such that privacy and security issues have yet to be sufficiently defined and addressed. For instance, massive computation capability is needed to analyze genomic data for research purposes, such that the cloud is likely to play a large role in the management of such data. At the same time, genomic data will be used in specific applications (e.g., forensics) where mobile computing environments (e.g., tablets and smartphones) will be routinely used to access those data. And, genomic data will be increasingly available on the Web, especially in citizen-contributed environments, (e.g., online social networks). While some individuals are sharing such information, there are significant privacy concerns because it is unknown what such information is capable of revealing, or how it will be used in the future.

As such, there is a clear need to support personalized medicine, genomic research, forensic investigation, and recreational genomics while respecting privacy. Computing is a crucial enabler, but can also be the first source of leakage if appropriate mechanisms are not put in place. Our survey (opinion poll from the biomedical community) provides some insight into what may be the most important aspects of the problem to study. And, along these lines, we have provided a review of the state-of-the-art regarding computational protection methods in this field, as well as the main challenges moving forward. To assist the data privacy and security community to develop meaningful solutions, we have provided a framework to facilitate the understanding of the privacy-preserving handling of genomic data.

References

Agrawal, R., Kiernan, J., Srikant, R., and Xu, Y. (2004). Order preserving encryption for numeric data. In *ACM International Conference on Management of Data*, pages 563–574.

Allen, N. E., Sudlow, C., Peakman, T., Collins, R., Dal-Ré, R., Ioannidis, J. P., Bracken, M. B., Buffler, P. A., Chan, A.-W., Franco, E. L., *et al.* (2014). UK biobank data: come and get it. *Science translational medicine*, 6(224).

Altman, R. B. and Klein, T. E. (2002). Challenges for biomedical informatics and pharmacogenomics. *Annual Review of Pharmacology and Toxicology*, 42(1), 113–133.

Altman, R. B., Clayton, E. W., Kohane, I. S., Malin, B. A., and Roden, D. M. (2013). Data re-identification: societal safeguards. *Science*, 339(6123), 1032.

Anderlik, M. R. (2003). Assessing the quality of DNA-based parentage testing: findings from a survey of laboratories. *Jurimetrics*, pages 291–314.

Atallah, M. J. and Li, J. (2005). Secure outsourcing of sequence comparisons. *International Journal of Information Security*, 4(4), 277–287.

Atallah, M. J., Kerschbaum, F., and Du, W. (2003). Secure and private sequence comparisons. In *ACM Workshop on Privacy in the Electronic Society*, pages 39–44.

Ayday, E., Cristofaro, E. D., Hubaux, J.-P., and Tsudik, G. (2013a). The chills and thrills of whole genome sequencing. *Computer*.

Ayday, E., Raisaro, J. L., and Hubaux, J.-P. (2013b). Personal use of the genomic data: privacy vs. storage cost. In *IEEE Global Communications Conference, Exhibition and Industry Forum*.

Ayday, E., Raisaro, J. L., and Hubaux, J.-P. (2013c). Privacy-enhancing technologies for medical tests using genomic data. In *(short paper) Network and Distributed System Security Symposium*.

Ayday, E., Raisaro, J. L., McLaren, P. J., Fellay, J., and Hubaux, J.-P. (2013d). Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. In *Workshop on Health Information Technologies*.

Ayday, E., Raisaro, J. L., Hubaux, J.-P., and Rougemont, J. (2013e). Protecting and evaluating genomic privacy in medical tests and personalized medicine. In *Workshop on Workshop on Privacy in the Electronic Society*, pages 95–106.

Ayday, E., Raisaro, J. L., Hengartner, U., Molyneaux, A., and Hubaux, J.-P. (2014). Privacy-preserving processing of raw genomic data. In *Data Privacy Management and Autonomous Spontaneous Security*, pages 133–147.

Bains, W. (2010). Genetic exceptionalism. *Nature Biotechnology*, 28(3), 212–213.

Baldi, P., Baronio, R., Cristofaro, E. D., Gasti, P., and Tsudik, G. (2011). Countering gattaca: efficient and secure testing of fully-sequenced human genomes. In *ACM Conference on Computer and Communications Security*, pages 691–702.

Bielinski, S. J., Olson, J. E., Pathak, J., Weinshilboum, R. M., Wang, L., Lyke, K. J., Ryu, E., Targonski, P. V., Van Norstrand, M. D., Hathcock, M. A., *et al.* (2014). Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time: using genomic data to individualize treatment protocol. In *Mayo Clinic Proceedings*, volume 89, pages 25–33.

Blanton, M., Atallah, M. J., Frikken, K. B., and Malluhi, Q. (2012). Secure and efficient outsourcing of sequence comparisons. In *European Symposium on Research in Computer Security*, pages 505–522.

Bobellan, M. (2010). DNA's dirty little secret. In *Washington Monthly*.

Bohannon, P., Jakobsson, M., and Srikwan, S. (2000). Cryptographic approaches to privacy in forensic DNA databases. In *Public Key Cryptography*, pages 373–390.

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 33, 228–237.

Braun, R., Rowe, W., Schaefer, C., Zhang, J., and Buetow, K. (2009). Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS Genetics*, 5(10), e1000668.

Brenner, S. E. (2013). Be prepared for the big genome leak. *Nature*, 498(7453), 139.

Bruekers, F., Katzenbeisser, S., Kursawe, K., and Tuyls, P. (2008). privacy-preserving matching of DNA profiles. *IACR Cryptology ePrint Archive*, 2008, 203.

Brunham, L. R. and Hayden, M. R. (2012). Whole-genome sequencing: the new standard of care? *Science*, 336, 1112–1113.

Canim, M., Kantarcioglu, M., and Malin, B. (2012). Secure management of biomedical data with cryptographic hardware. *IEEE Transactions on Information Technology in Biomedicine*, 16(1), 166–175.

Cassa, C. A., Miller, R. A., and Mandl, K. D. (2013). A novel, privacy-preserving cryptographic approach for sharing sequencing data. *Journal of the American Medical Informatics Association*, 20(1), 69–76.

Chen, Y., Peng, B., Wang, X., and Tang, H. (2012). Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. In *Network and Distributed System Security Symposium*.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., and Li, L. (2011). China kadoorie biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology*, 40(6), 1652–1666.

Cheng, M. (2003). *Medical device regulations: global overview and guiding principles*. World Health Organization.

Church, G., Heeney, C., Hawkins, N., de Vries, J., Boddington, P., Kaye, J., Bobrow, M., Weir, B., *et al.* (2009). Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genetics*, 5(10), e1000665.

Clayton, D. (2010). On inferring presence of an individual in a mixture: a bayesian approach. *Biostatistics*, 11(4), 661–673.

Council of Europe (2008). Additional protocol to the convention on human rights and biomedicine, concerning genetic testing for health purposes.

Craig, D. W., Goor, R. M., Wang, Z., Paschall, J., Ostell, J., Feolo, M., Sherry, S. T., and Manolio, T. A. (2011). Assessing and managing risk when sharing aggregate genetic variant data. *Nature Reviews Genetics*, 12(10), 730–736.

Cristofaro, E. D. (2014). Genomic privacy and the rise of a new research community. *IEEE Security and Privacy*, 12(2), 80–83.

Cristofaro, E. D., Faber, S., Gasti, P., and Tsudik, G. (2012). Genodroid: are privacy-preserving genomic tests ready for prime time? In *ACM Workshop on Privacy in the Electronic Society*, pages 97–108.

Danezis, G. and De Cristofaro, E. (2014). Fast and private genomic testing for disease susceptibility. In *Workshop on Privacy in the Electronic Society*, pages 31–34.

De Cristofaro, E. (2014). An exploratory ethnographic study of issues and concerns with whole genome sequencing. In *Workshop on Usable Security*.

Djatmiko, M., Friedman, A., Boreli, R., Lawrence, F., Thorne, B., and Hardy, S. (2014). Secure evaluation protocol for personalized medicine. In *Workshop on Privacy in the Electronic Society*, pages 159–162.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming*, pages 1–12.

Eppstein, D., Goodrich, M. T., and Baldi, P. (2011). Privacy-enhanced methods for comparing compressed DNA sequences. *arXiv preprint arXiv:1107.3593*.

Erlich, Y. and Narayanan, A. (2013). Routes for breaching and protecting genetic privacy. *arXiv*, abs/1310.3197v1.

Evans, J. P., Burke, W., and Khoury, M. (2010). The rules remain the same for genomic medicine: the case against "reverse genetic exceptionalism". *Genetics in Medicine*, 12(6), 342–343.

Feero, W. G., Guttmacher, A. E., McDermott, U., Downing, J. R., and Stratton, M. R. (2011). Genomics and the continuum of cancer care. *New England Journal of Medicine*, 364, 340–350.

Fienberg, S. E., Slavkovic, A., and Uhler, C. (2011). Privacy preserving gwas data sharing. In *IEEE Data Mining Workshops*, pages 628–635.

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., and Ristenpart, T. (2014). Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*, pages 17–32.

Frikken, K. B. (2009). Practical private DNA string searching and matching through efficient oblivious automata evaluation. In *Data and Applications Security*, pages 81–94.

Gibbs, J. R. and Singleton, A. (2006). Application of genome-wide single nucleotide polymorphism typing: Simple association and beyond. *PLoS Genetics*, 2(10).

Gilbert, N. (2008). Researchers criticize genetic data restrictions. *Nature News*.

Gitschier, J. (2009). Inferential genotyping of y chromosomes in latter-day saints founders and comparison to utah samples in the hapmap project. *American Journal of Human Genetics*, 84(2), 251–258.

Goldman, J. S., Hahn, S. E., Catania, J. W., Larusse-Eckert, S., Butson, M. B., Rumbaugh, M., Strecker, M. N., Roberts, J. S., Burke, W., Mayeux, R., *et al.* (2011). Genetic counseling and testing for alzheimer disease: joint practice guidelines of the american college of medical genetics and the national society of genetic counselors. *Genetics in Medicine*, 13(6), 597–605.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, 32(1).

Goodrich, M. T. (2009). The mastermind attack on genomic data. In *IEEE Symposium on Security and Privacy*, pages 204–218.

Gostin, L. O. and Hodge Jr, J. G. (1999). Genetic privacy and the law: an end to genetics exceptionalism. *Jurimetrics*, 40, 21–58.

Gottesman, O., Scott, S. A., Ellis, S. B., Overby, C. L., Ludtke, A., Hulot, J.-S., Hall, J., Chatani, K., Myers, K., Kannry, J. L., *et al.* (2013a). The clipmerge pgx program: clinical implementation of personalized medicine through electronic health records and genomics-pharmacogenomics. *Clinical Pharmacology and Therapeutics*, 94(2), 214.

Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., Sanderson, S. C., Kannry, J., Zinberg, R., Basford, M. A., *et al.* (2013b). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in Medicine*, 15(10), 761–771.

Gottlieb, S. (2001). Us employer agrees to stop genetic testing. *BMJ: British Medical Journal*, 322(7284), 449.

Greely, H. T., Riordan, D. P., Garrison, N., and Mountain, J. L. (2006). Family ties: the use of dna offender databases to catch offenders' kin. *The Journal Of Law, Medicine & Ethics*, 34(2), 248–262.

Greshake, B., Bayer, P. E., Rausch, H., and Reda, J. (2014). opensnp–a crowdsourced web resource for personal genomics. *PloS One*, 9(3), e89204.

Guttmacher, A. E. and Collins, F. S. (2003). Welcome to the genomic era. *New England Journal of Medicine*, (349), 996–998.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., and Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339(6117), 321–324.

Halperin, E. and Stephan, D. A. (2009). SNP imputation in association studies. *Nature Biotechnology*, 27(4), 349–351.

Harley, C. B., Futcher, A. B., and Greider, C. W. (1990). Telomeres shorten during ageing of human fibroblasts.

Haussler, D., Patterson, D. A., Diekhans, M., Fox, A., Jordan, M., Joseph, A. D., Ma, S., Paten, B., Shenker, S., Sittler, T., *et al.* (2012). A million cancer genome warehouse. Technical report, DTIC Document.

Hayden, E. C. (2013). Privacy protections: The genome hacker. *Nature*, 497, 172–174.

He, D., Furlotte, N. A., Hormozdiari, F., Joo, J. W. J., Wadia, A., Ostrovsky, R., Sahai, A., and Eskin, E. (2014). Identifying genetic relatives without compromising privacy. *Genome Research*, 24(4), 664–672.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8).

Hormozdiari, F., Joo, J. W. J., Wadia, A., Guan, F., Ostrosky, R., Sahai, A., and Eskin, E. (2014). Privacy preserving protocol for detecting genetic relatives using rare variants. *Bioinformatics*, 30(12), i204–i211.

Huang, Y., Evans, D., Katz, J., and Malka, L. (2011). Faster secure two-party computation using garbled circuits. In *USENIX Security Symposium*.

Huang, Y., Evans, D., and Katz, J. (2012). Private set intersection: Are garbled circuits better than custom protocols. In *Network and Distributed System Security Symposium*.

Huang, Z., Ayday, E., Fellay, J., Hubaux, J.-P., and Juels, A. (2015). Genoguard: Protecting genomic data against brute-force attacks. In *IEEE Symposium on Security and Privacy*, pages 447–462.

Humbert, M., Ayday, E., Hubaux, J.-P., and Telenti, A. (2013). Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *ACM Conference on Computer and Communications Security*, pages 1141–1152.

Humbert, M., Ayday, E., Hubaux, J.-P., and Telenti, A. (2014). Reconciling utility with privacy in genomics. In *Workshop on Privacy in the Electronic Society*, pages 11–20.

Im, H. K., Gamazon, E. R., Nicolae, D. L., and Cox, N. J. (2012). On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *American Journal of Human Genetics*, 90(4), 591–598.

Jacobs, K. B., Yeager, M., Wacholder, S., Craig, D., Kraft, P., Hunter, D. J., Paschal, J., Manolio, T. A., Tucker, M., Hoover, R. N., *et al.* (2009). A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genetics*, 41(11), 1253–1257.

Jha, S., Kruger, L., and Shmatikov, V. (2008). Towards practical privacy for genomic computation. In *IEEE Symposium on Security and Privacy*, pages 216–230.

Jobling, M. A. (2001). In the name of the father: surnames and genetics. *Trends in Genetics*, 17(6), 353–357.

Joh, E. E. (2006). Reclaiming “abandoned" DNA: the Fourth Amendment and genetic privacy. *Northwestern University Law Review*, 100(2), 857–884.

Johnson, A. and Shmatikov, V. (2013). Privacy-preserving data exploration in genome-wide association studies. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 1079–1087.

Juels, A. and Ristenpart, T. (2014). Honey encryption: Security beyond the brute-force bound. In *Advances in Cryptology–EUROCRYPT*, pages 293–310.

Kantarcioglu, M., Jiang, W., Liu, Y., and Malin, B. (2008). A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on Information Technology in Biomedicine*, 12(5), 606–617.

Karvelas, N., Peter, A., Katzenbeisser, S., Tews, E., and Hamacher, K. (2014). Privacy-preserving whole genome sequence processing through proxy-aided oram. In *Workshop on Privacy in the Electronic Society*, pages 1–10.

Kaufman, D., Bollinger, J., Dvoskin, R., and Scott, J. (2012). Preferences for opt-in and opt-out enrollment and consent models in biobank research: a national survey of veterans administration patients. *Genetics in Medicine*, 14(9), 787–794.

Kaufman, D. J., MurphyBollinger, J., Scott, J., and Hudson, K. L. (2009). Public opinion about the importance of privacy in biobank research. *American Journal of Human Genetics*, 85(5), 643–654.

Kaye, D. H. and Smith, M. E. (2003). DNA identification databases: legality, legitimacy, and the case for population-wide coverage. *Wisconsin Law Review*, page 413.

Kayser, M. and de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics*, 12(3), 179–192.

Kohane, I. S. (2011). Using electronic health records to drive discovery in disease genomics. *Nature Review Genetics*, 12(6), 417–428.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., *et al.* (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, 40(9), 1068–1075.

Lehmann-Haupt, R. (2010). Are sperm donors really anonymous anymore? *Slate*.

Li, G., Wang, Y., and Su, X. (2012). Improvements on a privacy-protection algorithm for DNA sequences with generalization lattices. *Computer Programs in Biomedicine*, 108(1), 1–9.

Lin, Z., Hewett, M., and Altman, R. B. (2002). Using binning to maintain confidentiality of medical data. In *American Medical Informatics Association Annual Symposium*, page 454.

Lin, Z., Owen, A. B., and Altman, R. B. (2004). Genomic research and human subject privacy. *Science*, 305(5681), 183.

Lindor, N. M. (2012). Personal autonomy in the genomic era. In *Video Proceedings of Mayo Clinic Individualizing Medicine Conference*.

Lippman, A. (1991). Prenatal genetic testing and screening: constructing needs and reinforcing inequities. *American Journal of Law in Medicine*, 17(1-2), 15–50.

Lumley, T. and Rice, K. (2010). Potential for revealing individual-level information in genome-wide association studies. *Journal of the American Medical Association*, 303(7), 659–660.

MacDonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., *et al.* (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*, 72(6), 971–983.

Malin, B. (2006). Re-identification of familial database records. In *American Medical Informatics Association Annual Symposium*, volume 2006, pages 524–528.

Malin, B. A. (2005a). 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), 28–34.

Malin, B. A. (2005b). Protecting DNA sequence anonymity with generalization lattices. *Methods of Information in Medicine*, 44, 687–692.

Malin, B. A. and Sweeney, L. (2004). How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*, 37(3), 179–192.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7), 499–511.

Maryland v. King (2013).

Masca, N., Burton, P. R., and Sheehan, N. A. (2011). Participant identification in genetic association studies: improved methods and practical implications. *International Journal of Epidemiology*, 40(6), 1629–1642.

Mastromauro, C., Myers, R. H., Berkman, B., Opitz, J. M., and Reynolds, J. F. (1987). Attitudes toward presymptomatic testing in Huntington disease. *American Journal of Medical Genetics*, 26(2), 271–282.

McGuire, A. L., Oliver, J. M., Slashinski, M. J., Graves, J. L., Wang, T., Kelly, P. A., Fisher, W., Lau, C. C., Goss, J., Okcu, M., *et al.* (2011). To share or not to share: a randomized trial of consent for data sharing in genome research. *Genetics in Medicine*, 13(11), 948–955.

Motluk, A. (2005). Anonymous sperm donor traced on internet. *New Scientist*.

Naik, G. (2009). Family secrets: an adopted man's 26-year quest for his father. *The Wall Street Journal*.

Naveed, M. (2014). Hurdles for genomic data usage management. In *International Workshop on Data Usage Management*.

Naveed, M., Agrawal, S., Prabhakaran, M., Wang, X., Ayday, E., Hubaux, J.-P., and Gunter, C. A. (2014). Controlled functional encryption. In *ACM Conference on Computer and Communications Security*, pages 1280–1291.

Nelkin, D. and Lindee, S. (1995). *The DNA mystique*. WH Freeman & Company.

Nyholt, D. R., Yu, C.-E., and Visscher, P. M. (2008). On Jim Watson's APOE status: genetic information is hard to hide. *European Journal of Human Genetics*, 17(2), 147–149.

Overby, C. L., Tarczy-Hornoch, P., Hoath, J. I., Kalet, I. J., and Veenstra, D. L. (2010). Feasibility of incorporating genomic knowledge into electronic medical records for pharmacogenomic clinical decision support. *BMC Bioinformatics*, 11(Suppl 9), S10.

Pakstis, A. J., Speed, W. C., Fang, R., Hyland, F. C., Furtado, M. R., Kidd, J. R., and Kidd, K. K. (2010). SNPs for a universal individual identification panel. *Human Genetics*, 127(3), 315–324.

Platt, J., Bollinger, J., Dvoskin, R., Kardia, S. L., and Kaufman, D. (2013). Public preferences regarding informed consent models for participation in population-based genomic research. *Genetics in Medicine*.

Prainsack, B. and Vayena, E. (2013). Beyond the clinic: 'direct-to-consumer' genomic profiling services and pharmacogenomics. *Pharmacogenomics*, 14(4), 403–412.

Presidential Commission for the Study of Bioethical Issues (2012). Privacy and progress in whole genome sequencing.

Pulley, J. M., Brace, M. M., Bernard, G. R., and Masys, D. R. (2008). Attitudes and perceptions of patients towards methods of establishing a DNA biobank. *Cell and Tissue Banking*, 9(1), 55–65.

Pulley, J. M., Denny, J. C., Peterson, J. F., Bernard, G. R., Vnencak-Jones, C. L., Ramirez, A. H., Delaney, J. T., Bowton, E., Brothers, K., Johnson, K., *et al.* (2012). Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. *Clinical Pharmacology and Therapeutics*, 92(1), 87–95.

Ritter, M. (2013). Henrietta lacks' family, feds reach settlement on use of DNA info.

Rothstein, M. A. (2005). Genetic exceptionalism & legislative pragmatism. *Hastings Center Report*, 35(4), 27–33.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732), 1350–1354.

Samani, S. S., Huang, Z., Ayday, E., Elliot, M., Fellay, J., Hubaux, J.-P., and Kutalik, Z. (2015). Quantifying genomic privacy via inference attack with high-order snv correlations. In *Workshop on Genome Privacy*, pages 32–40.

Sankararaman, S., Obozinski, G., Jordan, M. I., and Halperin, E. (2009). Genomic privacy and limits of individual detection in a pool. *Nature Genetics*, 41(9), 965–967.

Schadt, E. E., Woo, S., and Hao, K. (2012). Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genetics*, 44(5), 603–608.

Seddon, J. M., Reynolds, R., Yu, Y., Daly, M. J., and Rosner, B. (2011). Risk models for progression to advanced age-related macular degeneration using demographic, environmental, genetic, and ocular factors. *Ophthalmology*, 118(11), 2203–2211.

Skloot, R. (2013). The immortal life of henrietta lacks, the sequel.

Skloot, R. and Turpin, B. (2010). *The immortal life of Henrietta Lacks*. Crown Publishers New York.

Stajano, F. (2009). Privacy in the era of genomics. *netWorker*, 13(4), 40–ff.

Stajano, F., Bianchi, L., Liò, P., and Korff, D. (2008). Forensic genomics: kin privacy, driftnets and other open questions. In *ACM Workshop on Privacy in the Electronic Society*, pages 15–22.

Stein, R. (2005). Found on the web, with DNA: a boy's father. *The Washington Post*.

Sweeney, L., Abu, A., and Winn, J. (2013). Identifying participants in the personal genome project by name (a re-identification experiment). *CoRR*, abs/1304.7605.

Szajda, D., Pohl, M., Owen, J., Lawson, B. G., and Richmond, V. (2006). Toward a practical data privacy scheme for a distributed implementation of the smith-waterman genome sequence comparison algorithm. In *Network and Distributed System Security Symposium*.

Tambor, E. S., Bernhardt, B. A., Rodgers, J., Holtzman, N. A., and Geller, G. (2002). Mapping the human genome: an assessment of media coverage and public reaction. *Genetics in Medicine*, 4(1), 31–36.

Troncoso-Pastoriza, J. R., Katzenbeisser, S., and Celik, M. (2007). Privacy preserving error resilient DNA searching through oblivious automata. In *ACM Conference on Computer and Communications Security*, pages 519–528.

Turchin, M. C. and Hirschhorn, J. N. (2012). Gencrypt: one-way cryptographic hashes to detect overlapping individuals across samples. *Bioinformatics*, 28(6), 886–888.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., *et al.* (2001). The sequence of the human genome. *Science*, 291(5507), 1304–1351.

Villalva, B. R. (2012). Madonna sterilization, star hires DNA team on tour. In *The Christian Post*.

Visscher, P. M. and Hill, W. G. (2009). The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genetics*, 5(10), e1000628.

Wagner, I. (2015). Genomic privacy metrics: a systematic comparison. In *Workshop on Genome Privacy*, pages 50–59.

Wang, R., Li, Y. F., Wang, X., Tang, H., and Zhou, X. (2009a). Learning your identity and disease from research papers: information leaks in genome wide association study. In *ACM Conference on Computer and Communications Security*, pages 534–544.

Wang, R., Wang, X., Li, Z., Tang, H., Reiter, M. K., and Dong, Z. (2009b). Privacy-preserving genomic computation through program specialization. In *ACM Conference on Computer and Communications Security*, pages 338–347.

Xie, W., Kantarcioglu, M., Bush, W. S., Crawford, D., Denny, J. C., Heatherly, R., and Malin, B. A. (2014). Securema: protecting participant privacy in genetic association meta-analysis. *Bioinformatics*, page btu561.

Yu, F., Fienberg, S. E., Slavkovic, A. B., and Uhler, C. (2014). Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50(0), 133 – 141.

Zerhouni, E. A. and Nabel, E. G. (2008). Protecting aggregate genomic data. *Science*, 322(5898), 44a.

APPENDIX

In this appendix, we provide additional results from our expert opinion poll discussed in Section V. Results shown here are stratified according to the security/privacy and genetics/genomics expertise of the participants. Results for the cases where the number of participants was small are omitted.

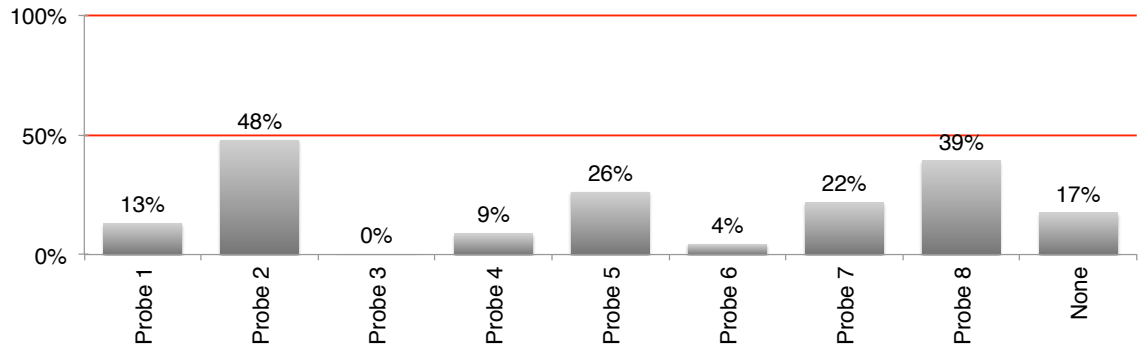

Figure 10: Response to the question: *Do you believe that: (Multiple options can be checked).* The probes are described in detail in Figure 2. "None" means the respondent does not agree with any of the probes: **Only "Expert" biomedical participants (sample size 23).**

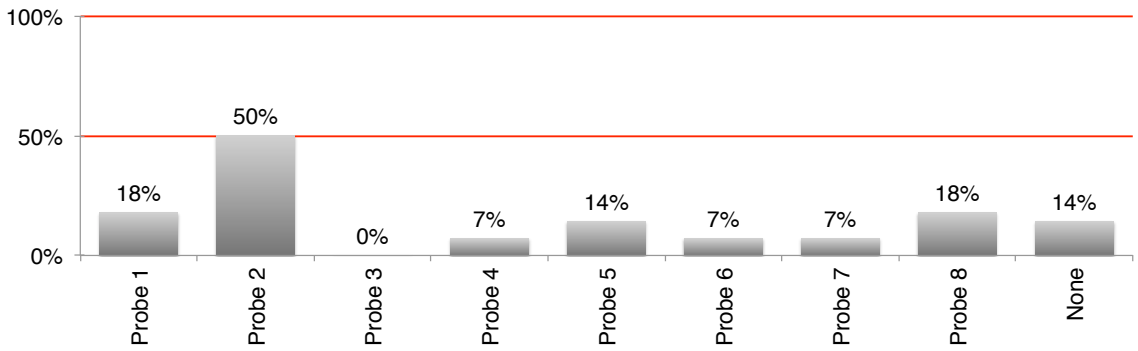

Figure 11: Response to the question: *Do you believe that: (Multiple options can be checked).* The probes are described in detail in Figure 2. "None" means the respondent does not agree with any of the probes: **Only "Knowledgeable" biomedical participants (sample size 28).**

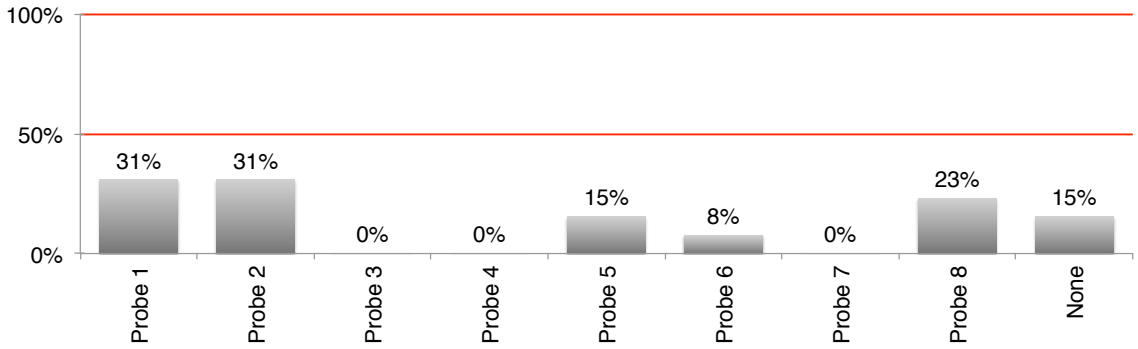

Figure 12: Response to the question: *Do you believe that: (Multiple options can be checked).* The probes are described in detail in Figure 2. "None" means the respondent does not agree with any of the probes: **Only "Some familiarity" biomedical participants (sample size 13).**

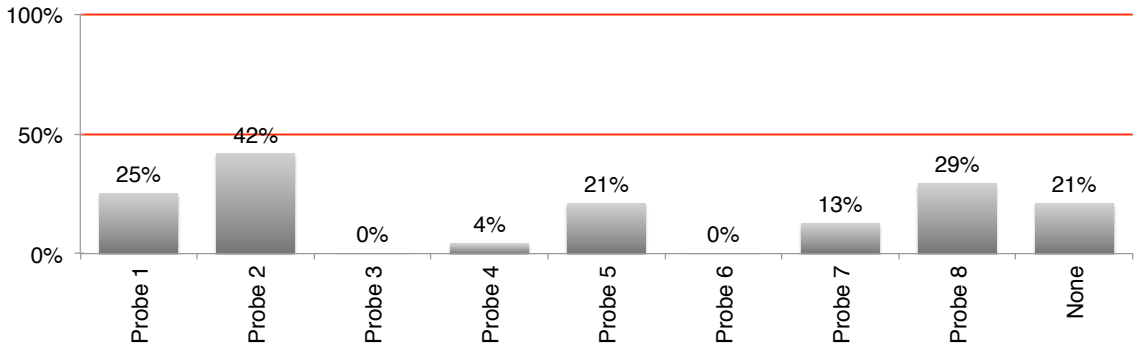

Figure 13: Response to the question: *Do you believe that: (Multiple options can be checked).* The probes are described in detail in Figure 2. "None" means the respondent does not agree with any of the probes: **Only "Knowledgeable" security and privacy participants (sample size 24).**

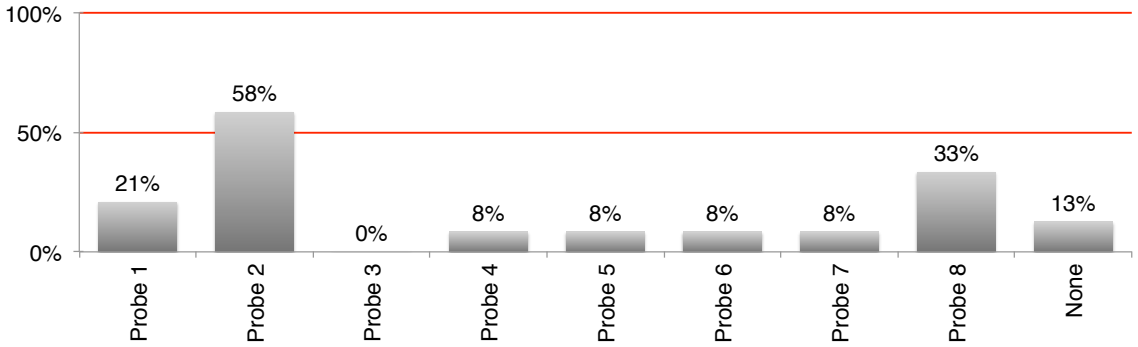

Figure 14: Response to the question: *Do you believe that: (Multiple options can be checked).* The probes are described in detail in Figure 2. "None" means the respondent does not agree with any of the probes: **Only "Some familiarity" security and privacy participants (sample size 24).**

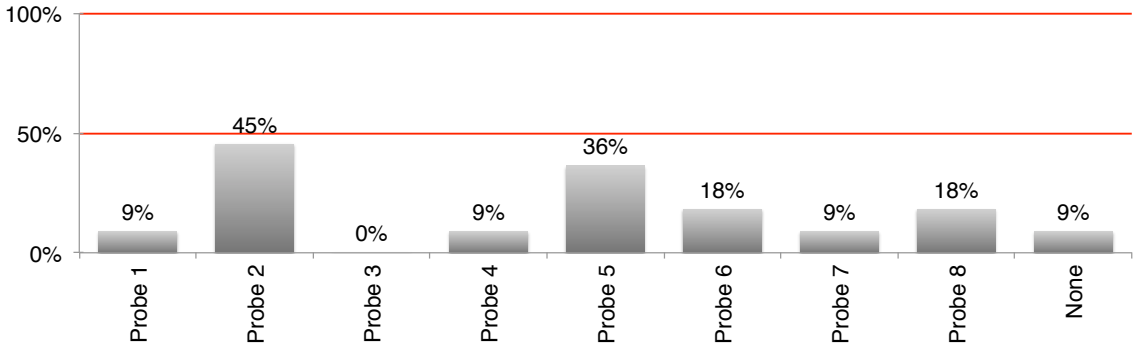

Figure 15: Response to the question: *Do you believe that: (Multiple options can be checked).* The probes are described in detail in Figure 2. "None" means the respondent does not agree with any of the probes: **Only "No familiarity" security and privacy participants (sample size 11).**

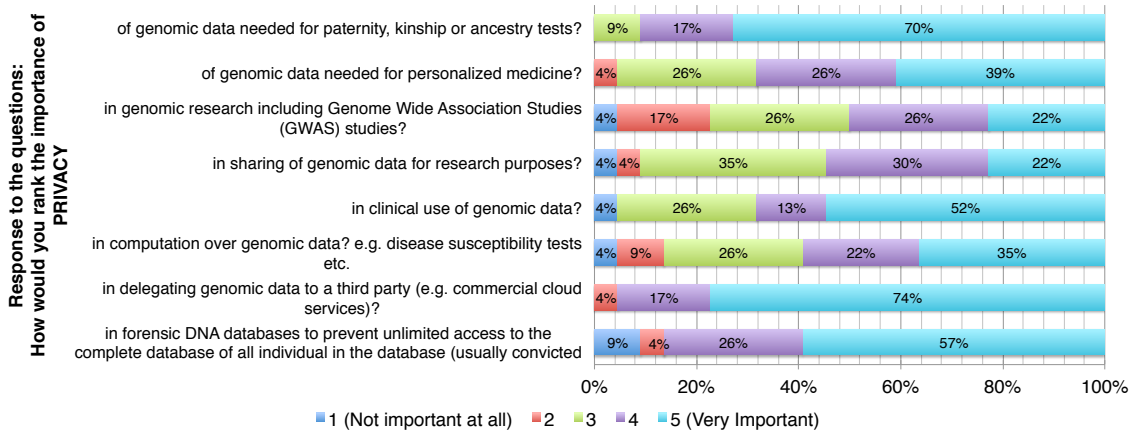

Figure 16: Relevance of genome privacy research done by the computer science community: **Only "Expert" biomedical participants (sample size 23).**

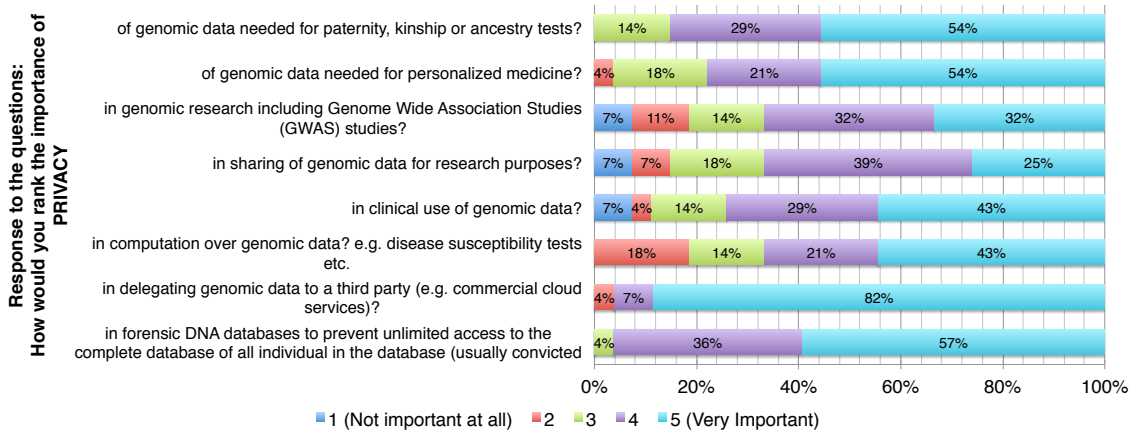

Figure 17: Relevance of genome privacy research done by the computer science community: **Only "Knowledgeable" biomedical participants (sample size 28).**

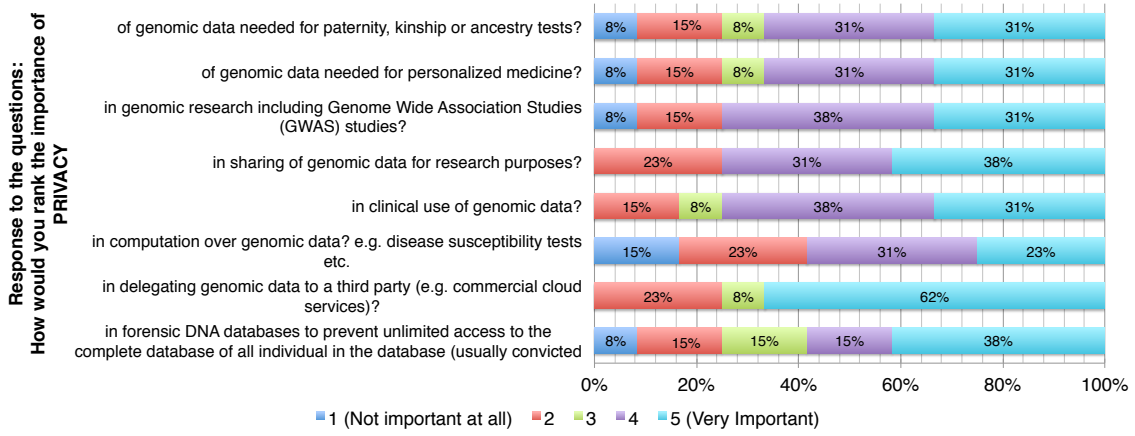

Figure 18: Relevance of genome privacy research done by the computer science community: **Only "Some familiarity" biomedical participants (sample size 13).**

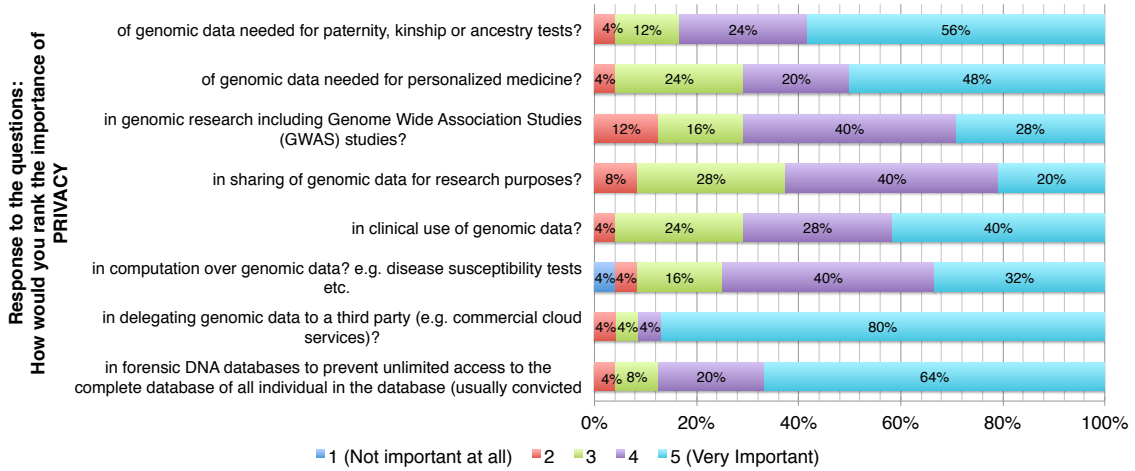

Figure 19: Relevance of genome privacy research done by the computer science community: **Only "Knowledgeable" security and privacy participants (sample size 24).**

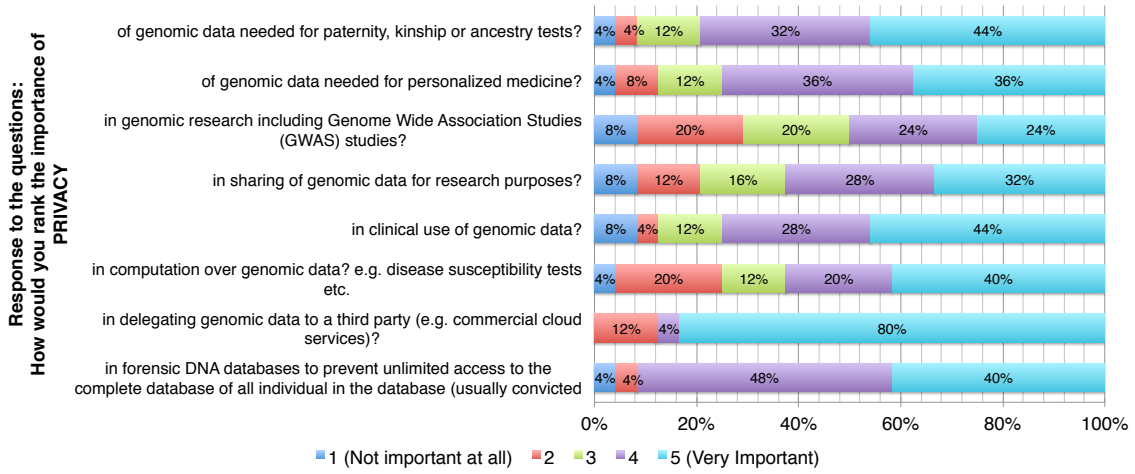

Figure 20: Relevance of genome privacy research done by the computer science community: **Only "Some familiarity" security and privacy participants (sample size 24)**.

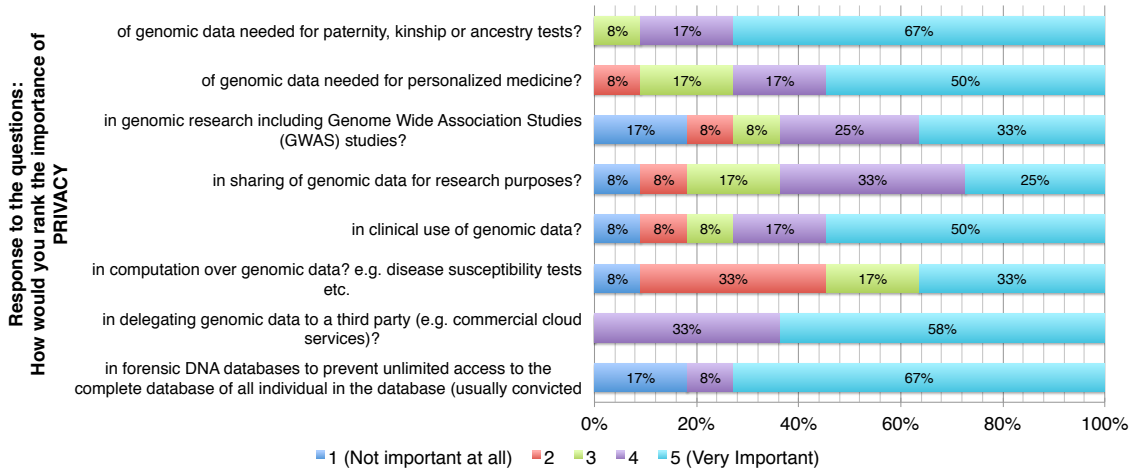

Figure 21: Relevance of genome privacy research done by the computer science community: **Only "No familiarity" security and privacy participants (sample size 11).**

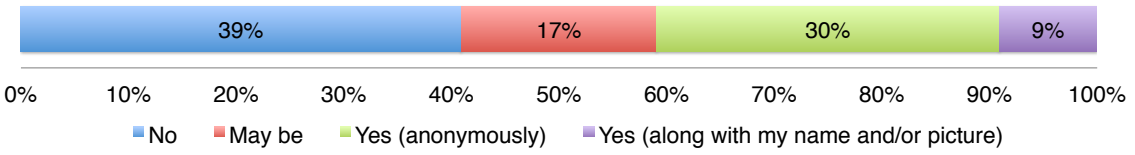

Figure 22: Response to the question: *Would you publicly share your genome on the Web?*: **Only "Expert" biomedical participants (sample size 23).**

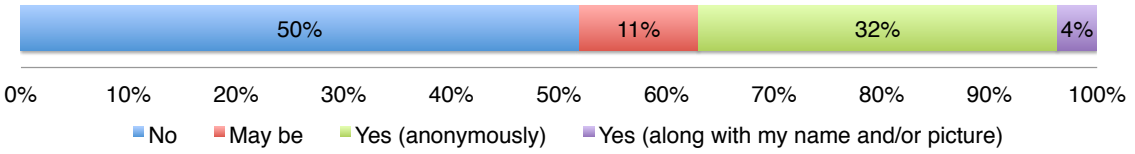

Figure 23: Response to the question: *Would you publicly share your genome on the Web?*: **Only "Knowledgeable" biomedical participants (sample size 28).**

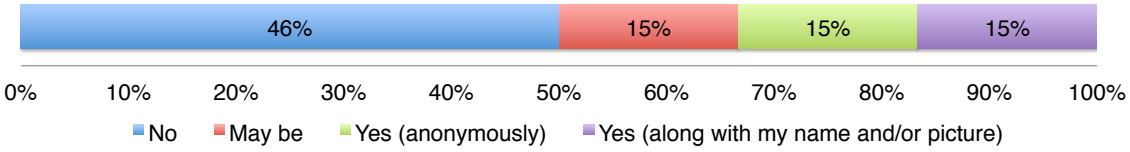

Figure 24: Response to the question: *Would you publicly share your genome on the Web?*: **Only "Some familiarity" biomedical participants (sample size 13).**

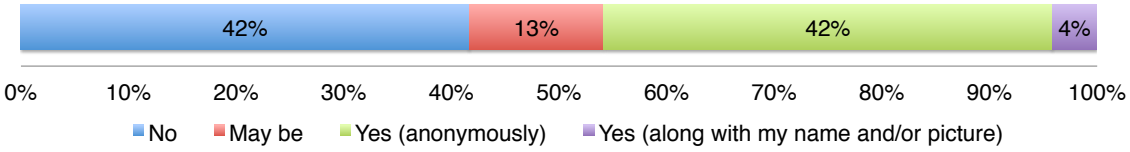

Figure 25: Response to the question: *Would you publicly share your genome on the Web?*: **Only "Knowledgeable" security and privacy participants (sample size 24).**

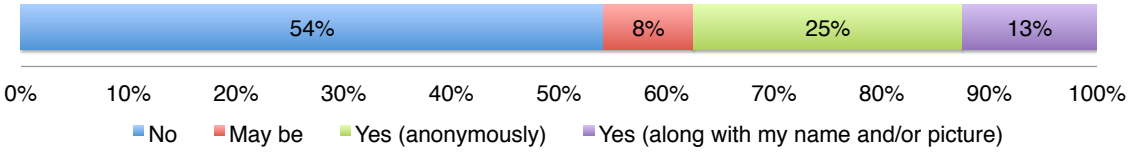

Figure 26: Response to the question: *Would you publicly share your genome on the Web?*: **Only "Some familiarity" security and privacy participants (sample size 24).**

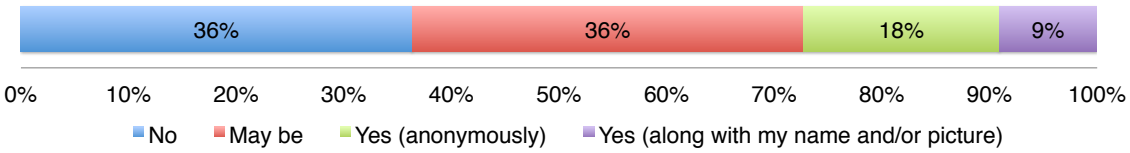

Figure 27: Response to the question: *Would you publicly share your genome on the Web?*: **Only "No familiarity" security and privacy participants (sample size 11).**

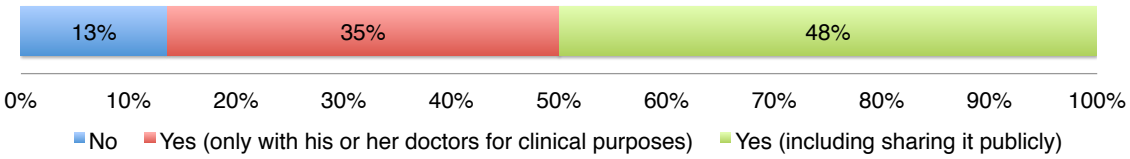

Figure 28: Response to the question: *Assuming that one's genomic data leaks a lot of private information about his or her relatives, do you think one should have the right to share his or her genomic data?*: **Only "Expert" biomedical participants (sample size 23).**

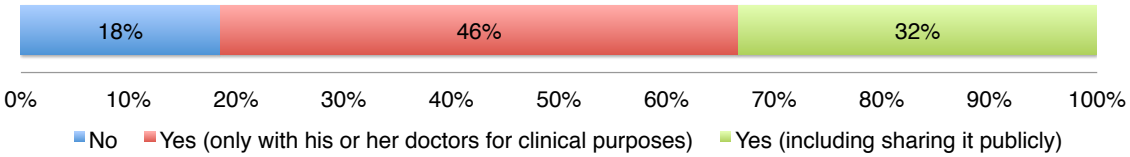

Figure 29: Response to the question: *Assuming that one's genomic data leaks a lot of private information about his or her relatives, do you think one should have the right to share his or her genomic data?*: **Only "Knowledgeable" biomedical participants (sample size 28).**

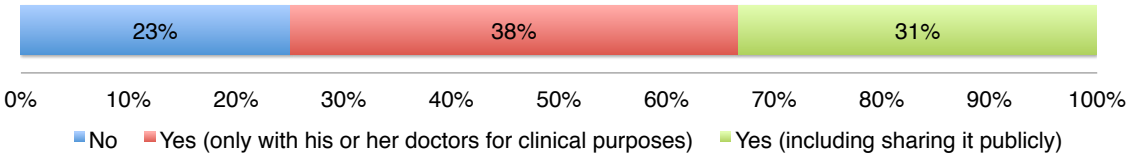

Figure 30: Response to the question: *Assuming that one's genomic data leaks a lot of private information about his or her relatives, do you think one should have the right to share his or her genomic data?*: **Only "Some familiarity" biomedical participants (sample size 13).**

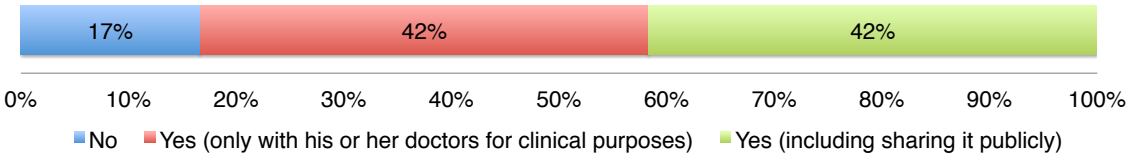

Figure 31: Response to the question: *Assuming that one's genomic data leaks a lot of private information about his or her relatives, do you think one should have the right to share his or her genomic data?*: **Only "Knowledgeable" security and privacy participants (sample size 24).**

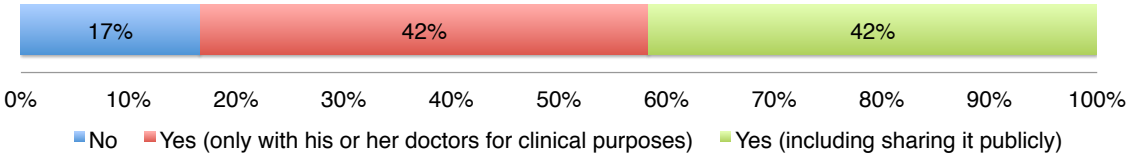

Figure 32: Response to the question: *Assuming that one's genomic data leaks a lot of private information about his or her relatives, do you think one should have the right to share his or her genomic data?*: **Only "Some familiarity" security and privacy participants (sample size 24).**

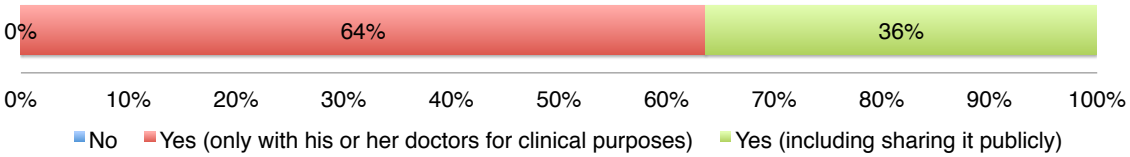

Figure 33: Response to the question: *Assuming that one's genomic data leaks a lot of private information about his or her relatives, do you think one should have the right to share his or her genomic data?*: **Only "No familiarity" security and privacy participants (sample size 11).**