INNOVATIONS IN MEDICAL GENOMICS: WHAT ARE THE PRIVACY AND SECURITY RISKS?

Forrest Briscoe (fbriscoe@psu.edu) and Barbara Gray (b9g@psu.edu)*
with assistance from Celeste Diaz Ferraro
Penn State University

MARCH 2017

"How do we support the very noble cancer moonshot and a healthier society and environment, without eroding what are truly our fundamental values of individualism and privacy and unreasonable search and seizure?" (Genomics company CEO)

AS GENOMICS IS INTEGRATED INTO HEALTH CARE DELIVERY, NEW PRIVACY AND DATA-SECURITY RISKS NEED TO BE CONSIDERED

Human genome sequencing is transforming health care. Sequencing is the innovation at the heart of the shift to precision medicine, reflected in the 21st Century Cures Act (2016) and other recent initiatives. As sequencing becomes incorporated into routine medical care, millions of Americans will have their genomic data generated in order to diagnose, treat, or prevent disease. At the same time, that data will be added to genome databases that are proliferating across the country. These databases are necessary to realize progress in medical care – but they also carry particular privacy and security risks that are not widely appreciated. In particular, genomic data reveal detailed physical and mental characteristics for each person sequenced, as well as for his or her family members and offspring for generations. These databases are inevitably vulnerable to breaches, and their contents cannot be anonymized effectively.

In this whitepaper, we summarize privacy and security risks associated with medical genomics.[1] Some options for risk mitigation would have non-trivial consequences, and could potentially slow further innovation and progress in science and medicine. Currently, leaders in the genomics field are debating these risks and benefits; we think the broader public needs to be engaged as well, since their genomes are entering the genome databases. Citizens and policy makers need to grapple with important trade-offs inherent in how we treat genomic data. Hence we conclude with three questions about genomics that need to be debated by informed American citizens and policy makers: Who should have the right to make decisions about your genome? How closely should you hold onto your genome? What standards are desirable for securing genomic databases? Addressing these questions will help insure that we strike the right balance of benefits and risks in this rapidly developing field.

ADVANCES IN GENOME SEQUENCING TECHNOLOGY ARE TRANSFORMING HEALTH CARE AND THE LIFE SCIENCES

The ***genome*** is that unique sequence of deoxyribonucleic acid (DNA) molecules which represents each individual's biochemical code.[2] The first human genome was sequenced over 15 years ago, and now the promise of genomics is being realized. In that time, advances in the technologies used to generate and digitally store that sequence have made it feasible to incorporate genomic data into a wide range of medical, scientific and commercial activities. In medicine, genomics is moving to center stage in the search for effective therapies and drugs; it is also being incorporated into the clinical diagnosis, treatment, and prevention of disease. Genomics is also transforming many scientific fields, from biology and anthropology to public health and the social sciences, shedding new light on research topics from human evolution to societal inequality. And a range of commercial firms, such as Google, Apple, IBM, Amazon and Alibaba, aim to use genomics in order to tailor consumer products and services according to users' genetic profiles, allowing these companies to more precisely manage relations with users and anticipate those users' needs and activities.

The field of genomics is large, complex, and rapidly expanding. A diverse ecosystem of organizations is involved, including hospitals and health care organizations, academic departments and research teams, pharmaceutical and biotechnology firms, consumer technology companies, genomics device manufacturers and sequencing labs, grant funding organizations, venture capitalists, regulatory agencies, citizen science and patient advocacy organizations, and more. Major genomics entities are organized as for-profit companies, non-profit organizations, and government agencies, reflecting a range of differing interests and governance structures. A web of inter-organizational partnerships and alliances crisscrosses the field. A 2013 study estimated the genomic sector's contribution to the economy at $25 billion in direct output and $40 billion in indirect output; observers anticipate the footprint of the sector will grow rapidly in the coming decade.[3]

In this white paper, we focus on medical genomics, which is the largest component of the overall genomics field. Changes now occurring in the health care system suggest that in the next few years a significant portion of Americans will be asked to have their genome sequenced as part of routine medical care.[4] This development has far reaching consequences that are only dimly appreciated at present. Our aim in this paper is to provide an initial risk assessment focused on the privacy and security consequences of this development. In particular, we focus on privacy risks. Medical genomics has great potential to improve the human condition, and our aim is not to question that potential. Rather, we hope that in sensitizing the public to more of the implications of medical genomics, we can stimulate public discussion and deliberation on the best ways to balance privacy concerns with other interests driving this field.

CLINICAL HEALTH CARE IS THE GATEWAY THROUGH WHICH A MAJORITY OF AMERICANS ARE LIKELY TO BE INTRODUCED TO SEQUENCING

With rapid advances in the feasibility of genome sequencing, leaders in health care are now forging a key role for ***medical genomics*** in the delivery of clinical patient care. Indeed, in major hospitals and health care organizations, medical genomics is already being integrated with patient data and other personal health information, so they can be used together in diagnosing and treating patients.[5] At a recent conference, a panel of leading experts concluded that within a decade, over 1 billion humans worldwide will be sequenced, and that sequencing will have become part of routine patient care in the U.S.[6] For example, a patient presenting symptoms of diabetes will have her genome analyzed for mutations known to make common diabetes drugs ineffective; if found, the patient's physician would be informed and a different drug prescribed. A patient arriving in the E.R. with chest pains will be analyzed for genetic heart-rhythm mutations which, if found, would alter critical care delivery in life-saving ways.[7]

This incorporation of genomic data into clinical care takes us toward an individualized approach to health care that is being called ***precision medicine***. Proponents argue that as the field advances, the practice of medicine will be revolutionized through targeted therapies and custom drugs based on knowledge of a specific patient's genome. Already, prior to being prescribed with a drug such as warfarin and Plavix, a patient can have his or her genome checked for mutations that are known to reduce the efficacy of those drugs.[8] In addition, genomic testing for prostate cancer can now identify patients at risk of the more aggressive form of the disease, leading to more targeted screening and treatment of those patients.[9] Routine "cell-free DNA" sequencing on healthy patients is aimed at early detection of a variety of cancers. Powerful new ***gene editing*** capabilities are also now making it possible to curtail, and even cure, certain diseases by identifying the offending mutations in an individual and then using gene therapy to fix them. In 2016, such approaches were found successful against diseases including hemophilia and non-Hodgkins lymphoma.[10] As these examples illustrate, precision medicine involves the intensive use of genomics in biomedical research, as well as the routine use of genomics in clinical diagnosis and treatment.

GENOMICS DATABASES PLAY A CRUCIAL – AND CONTROVERSIAL – ROLE

The linchpin of this revolution is the genomics database. The reuse, combination, and sharing of human genomic data in order to create these databases is an essential requirement for realizing the medical advances chronicled above.

To see why this is the case, consider how a scientist discovers the genetic mutations (called *variants*) that cause a specific disease. At a minimum, she needs to compare genomes from many people *with* the disease, against genomes from many people *without* the disease. Both groups need to be present in sufficient numbers in order to be statistically confident about the association between the variants and the disease. If the disease is rare, the database needs to be larger in order to include enough disease-carrying individuals. If the disease has *many*

variants associated with it, and they occur in complex combinations – a situation that is common – then an even larger database is needed to ensure all those combinations are included. This statistical comparison process is made more challenging because of the sheer size of the human genome. Although most of the 6 billion base pairs in each genome are identical across individuals, there are still a few *million* **base pairs** that vary between people that may need to be analyzed for potential disease-causing variants.

The net result of this situation is that researchers require access to large databases comprised of DNA data from thousands - and in some cases millions - of individuals. What's more, those databases are most useful to researchers when they are linked to additional information about the individuals within them. For example, to identify the genomes of disease-carrying individuals and disease-free individuals, genomic data may need to be linked to medical records that report the disease status of each person in the database. Hence efforts are underway to compile and connect multiple databases together, to create ever-larger pools of genomic data coupled with other medical and personal information on the individuals contained in them.

Many people have - wittingly or unwittingly - contributed their DNA sequences to these databases simply by participating in medical testing or by using direct-to-consumer genomics testing services, often without realizing their data has now become available for academic and medical research, and in some cases also for additional commercial research. Although contributors sign consent forms mentioning their data will be reused, few seem aware of the implications, as we describe below. On the other hand, actors in the genomics field are well aware of the value of genomics databases. As a result, health care organizations, pharmaceutical companies, scientific teams, government agencies, and entrepreneurs are all engaged in a genomics database gold-rush, trying to assemble, secure, monetize and mine these databases for the secrets they hold. Such databases are quietly regarded as strategic assets by forward-thinking medical organizations such as Kaiser Permanente, Geisinger Health System, and the Veterans Health Administration, by pharmaceutical firms such as Roche and AstraZeneca, and by consumer technology firms such as Google and 23andMe.

DISCLOSED GENOMIC DATA REVEAL FUTURE HEALTH, BEHAVIORAL TENDENCIES, AND OTHER TRAITS, PLUS DETAILS OF FAMILY MEMBERS

Despite the increase in and popularization of genomics, the risks associated with this field in general, and genomic databases in particular, are largely underappreciated by patients. The latent value (and latent harm) of disclosing one's genomic information is already significant. Consider, for example, that genomic information now reveals an individual's susceptibility to substance abuse, odds of developing depression and early Alzheimer's disease, a baseline for innate mathematical ability, tendency toward aggression, and probabilities for developing a wide range of both common and rare genetic diseases.[11] Further, advances in statistical analysis mean that even when one part of a person's genome is disclosed, data from other parts of that person's genome can be inferred with relative certainty. In one famous example, when James Watson (the co-discoverer of the structure of DNA) made his genome publicly available, he

withheld data for the variant that predicts early Alzheimer's disease. But scientists showed how to infer that variant using other parts of his sequence.[12]

Further, disclosure that occurs today will also increase one's vulnerability to future risks that are as-yet unknown. As the genomic revolution advances, both the value – and the potential harm – of an individual's genomic data will continue to increase, long after the moment when he or she was initially sequenced. Within 5 years, expanded sequencing of a person's metagenome, which includes his or her personal ***microbiome***, will reveal further details of a "genetic fingerprint," encompassing fine-grained information about ethnicity/national origin, places an individual has visited, and even recent contact with other people. Clinical collection of expanded metagenomic data will increase in the near future, as scientists uncover the role of the microbiome in triggering and regulating disease, and in serving as indicators of the environmental exposures affecting disease.[13] Metagenomic data will also improve on existing abilities to predict an individual's socioeconomic circumstances from his or her genome.[14]

Importantly, genomic disclosure risks also extend to family members. Because so much of one person's DNA is identical to that of his or her relatives, including current and future descendants, these disclosures affect not only an individual but also their family members. Knowledge of one individual's genome, or even just some of their variant data, can reveal useful information on the disease probabilities and personal characteristics of that person's parents, siblings, children, and even extended kin up to 5 degrees of separation.[15]

Given the inherent value in personal genomic data, many entities may have an interest in acquiring it - including employers, educational institutions, insurance and financial firms, and even romantic partners. Although the 2008 Genetic Information Nondiscrimination Act (GINA) currently protects against abuse by some of these parties (specifically, employers and health insurers are prohibited from discriminating against individuals based on genomic data), it does not cover life, disability or long-term care insurance. Educational or financial discrimination is also possible, if educators or lenders - which are not subject to GINA's non-discrimination provisions - begin to screen applicants according to desirable genomic features. Genomics data is already being used to determine the eligibility of some students to participate in collegiate athletics.[16] State and non-state actors also have a strong interest in obtaining genomic data (see text box).[17]

Foreign State and Non-State Actors Also Seek Genomic Data

Beyond these market actors, who else is interested in genomic data? Foreign state and non-state actors also have a strong interest in gaining access to the U.S. citizens' genomic data. Foreign entities have conducted large-scale hacking of health care databases. Such data provide valuable medical and behavioral insights on a target country's future business leaders and governmental officials. As such, these data will have increasing value on the black market, to both criminals and states. If genomic signatures become incorporated into cryptographic and information security tools, the value of obtaining such genomic data will increase further.

A potential parallel to the current moment in genomic data disclosure might exist in the early years of online social media. Early social media users failed to anticipate how disclosure of their embarrassing or dubious pictures on their Facebook page might jeopardize their future employment options. As people became more aware of how online data could be used against them, social media users appear to have become more careful, and social media companies have responded with more options for individuals to control their data and limit disclosure risks. For DNA data, although GINA currently offers some protections against the use of disclosed genetic information, those protections are incomplete, and unlikely to keep pace with future attributes yet to be revealed from previously-released DNA.

In short, genomic data appears to be uniquely informative of an individual's future life expectations, challenges and opportunities - and ominously - those of their progeny as well. This suggests that an individual's data should be treated with unique care. But as we describe in more detail below, there are many scenarios in clinical use and beyond in which identifiable genomic data can potentially be released - intentionally or not - and cause harm to an individual and their family members.

THE LEGAL AND REGULATORY FRAMEWORK TO PROTECT GENOMIC DATA IS LIMITED

As the value and risk associated with genomic data disclosure comes into focus, the U.S. legal-regulatory framework protecting it has not kept up. Even when genomic data is generated in the health care setting, it is not protected in the same manner as a patient's personal health information. Personal health information is protected under ***Health Insurance Portability and Accountability Act*** (HIPAA), which circumscribes how that information can be stored, used, and shared. No clear regulatory framework for protecting genomic data exists. In its absence, organizations appear to be taking an approach with three basic pillars – consent, de-identification, and cybersecurity – each of which suffers from significant limitations.[18]

How does the U.S. Government's Involvement in Genomics Compare with Other Countries?
Governmental involvement in genomics varies greatly across countries. Relative to the U.S., there is greater involvement in the United Kingdom, China, Iceland, France, and other countries where the government is collecting genomic data on a many or all of its citizens. Why the differences across countries? In part, differences may reflect societal valuation of personal privacy. However, differences also reflect the state's pre-existing role in the financing and provision of health care; where that role is large (such as the UK), there are is a mandate for the government to invest in the collection and use of genomics databases to control the nation's health care costs.

Consent is the legal process through which individuals provide approval for their genomic data to be generated and used. In part to facilitate genomics research, a new model of "broad consent" has been developed, under which patients and others agreeing to DNA sequencing are told something like, "Your genomic data and health information will be studied

along with information from other participants in this research, and it will be stored for future studies by this and other research teams."[19] This approach is legally adequate, because under current U.S. law, genomic data itself is regarded as a form of property, subject to contract law, rather than governed by any universal privacy right of the individual who the data describes. Under this approach, patients are rarely informed of specific ways their data will be subsequently used or the consequences of such use.[20]

De-identification is intended to ensure that people who access a genomics database cannot tell who the individuals are within it. HIPAA regulations detail a set of steps needed to de-identify data, involving removing 18 specified individually-identifying data items, such as name, telephone number, and fingerprint details, which make it possible to uniquely identify a person. If a genomics database is free-standing, or linked to items *other* than those 18 specified ones, it falls outside of the jurisdiction of HIPAA, and in that state it can be reused and shared relatively easily. In fact, strictly speaking, in that non-HIPAA format, genomic data can be reused and shared for research purposes without patient consent.[21]

Unfortunately, genomic data itself appears to allow for re-identification, whether or not it is linked to any other items. Indeed, genomic data appears to be one of the best ways to uniquely identify an individual that has ever been discovered. Researchers have demonstrated a number of ways that a person can be re-identified by anyone with access to a (de-identified) genomic database. For example, if you know just a fraction of a person's DNA sequence from another source, you can use that information to re-identify them in the genomic database. Or, if you know fragments of a person's phenotype information, such as eye and skin color, height, and a few other items, you can also use that to re-identify that person in a genomic database.[22]

While a primary concern with re-identification is disclosure of private information manifest in the person's genome, as described above, there is also a privacy risk in just learning that a person is included within a given database. Because many databases focus on specific diseases, such as autism spectrum disorder or mitochondrial diseases, knowing that a person is in such a database reveals that this individual is a carrier for that disease, and their family members are likely carriers as well.

How Often Do Data Breaches Occur in the U.S. Health Care System?
According to the Institute for Critical Infrastructure Technology, nearly half of the U.S. population had personal healthcare data compromised in just one year. Anthem's data breach, disclosed in 2015, involved 78 million records with personally identifiable information. This suggests that personal data protections mandated by HIPAA are not providing adequate protection, even of non-genomic data. These sensitive personal health data are being successfully targeted by nation state actors, cyber criminals, and hacktivists.

The third pillar is ***cybersecurity***. However, the framework for ensuring cybersecurity in genomics is flexible, and depends in large part on the varying policies of each organization involved. In health care, HIPAA shapes cybersecurity, but genomics data can be kept outside of HIPAA jurisdiction. Further, even HIPAA-protected data is often breached (see text box).[23] For research projects and organizations receiving federal funding

(including many medical organizations, academic institutions, and pharmaceutical companies), an internal ***Institutional Review Board*** (IRB) approves genomics cybersecurity practices, along with other aspects of clinical or academic genomics research. IRBs, in turn, try to ensure that best practice guidelines for data security are followed, and they can enforce internal penalties for non-compliance as they deem appropriate.[24]

UNCOVERING THE FULL LIFECYCLE OF DNA DATA

To further explain how genomic data are being used, we developed a portrait of the "lifecycle of DNA data," tracing an arc from its creation during clinical diagnosis and treatment to its subsequent reuse for clinical and research purposes in other settings. While the process is by no means standardized, this general lifecycle model nonetheless captures the broad set of steps that are common across many settings. The lifecycle model (see ***Table 1***) offers a useful means for identifying where privacy and security vulnerabilities associated with the use of clinical DNA data can originate, and what steps need to be in place to ensure the privacy and security of the data as it is transferred from an individual to a large database and beyond.

The top row of Table 1 shows four basic steps in how DNA data is handled: (1) the generation of the sequenced data, (2) the analysis of variants in the data, (3) clinical interpretation of data, and (4) subsequent data sharing and reuse. The next two rows in the table summarize the key activities involved in each step, and specify who handles the data. The bottom two rows summarize the implications of each step for the privacy and security of DNA data. In the sections below, we first move left-to-right across the top half of the table to explain each step in the lifecycle. Then we go left-to-right through the bottom half of the table, discussing the privacy and security issues, and measures to address them, at each step.

Step 1. Generation of DNA data

The first step is the decision to create DNA data for the patient. This decision originates in a clinical setting in consultation with an individual's health care provider because the patient has symptoms the clinician believes may be better understood through genomic analysis. This step involves obtaining a ***biosample*** and getting it sequenced. For genomics research, the biosample is usually obtained via a blood sample or a "spit kit" – a small container in which the individual deposits saliva– which is then sent to the lab for sequencing.

Prior to submitting their biosample, patients are asked to sign a consent form that indicates they understand how their data will be used and the risks involved. In addition, in many cases, patients are asked to meet with a genetic counselor, and to read additional materials, in order to increase their understanding of the implications of having their genome sequenced, the information it may reveal, and the decisions they may face as a result. The people "touching the data" in this step include nurses, lab technicians and ***bioinformaticians*** who are either employees of the medical organization that is requesting the DNA data creation, or employees of a sequencing lab contracted to collect and sequence the data.

Steps in the lifecycle of DNA Data, and attendant privacy and security risks

	1. Generating DNA data	***2. Analysis of Variants***	***3. Clinical interpretation***	***4. Further data sharing and reuse***
Summary	Individual makes decision to collect DNA data and raw sequence data is generated	Individual's DNA data processed & analyzed to identify variants	DNA database linked to individual's other health data, and analyzed by clinicians for insight into medical condition	Findings from analysis may be shared, and data are shared for subsequent reanalysis
Key activities	Patient seeks medical care, or joins health care system Patient signs consent form Biosample collected at clinic and sent to sequencing lab -Biosample is prepared and run through sequencing machine -Resulting raw DNA data is stored temporarily	Data moves to bioinformaticians Patient's raw DNA data processed and annotated using variant database	DNA data analysis moves to clinicians DNA data analysis linked to other data in patient's health record, and to larger genomics & medical databases. May include: -visualization of specific patient's DNA -scrutiny of specific patient's genotype-phenotype associations -Sharing and discussion of data, within local clinical team and possibly other external clinical or research consultants Diagnosis and treatment options given to patient	Patient's DNA data merged with others' DNA data in larger database Findings from analysis shared with other researchers and clinicians working in similar areas Data may be shared for reuse with: - clinicians treating other patients; -NIH, which requires funded projects to deposit data for academic reuse -pharma firms for drug discovery -commercial firms and other types of organizations
Who handles data?	Sequencing lab employees	Employees of sequencing lab, medical organization, and/or third-party bioinformatics company	Employees of sequencing lab, medical organization, and/or third-party bioinformatics company -Also other clinicians, researchers and genetic counselors consulted for diagnosis/treatment	Other clinicians, academic researchers, pharmaceutical employees, and governmental entities that acquire access to databases
Data privacy & security risks	Consent form does not allow individual to easily appreciate implications for self and family Sample or data crossing state/legal jurisdictions	*Behavior of lab employees and clinicians* -carelessness or theft of data -clinician's urge to diagnose patient increases incentive for rapid data sharing, careless use of communication technology *Technology* -Storage and transit of data via internet, local computers or cloud		Individuals handling data are not subject to privacy & security policies of original organization; may not be as careful with data; goals may also differ Re-identification of data may be possible despite efforts to anonymize it
Ways to address data privacy & security	Require expanded genetic counseling Increase individual control over data Update consent process with "opt-in" to data reuse	*Behavior of lab employees and clinicians* -screening, compliance training and sanctions -monitoring and auditing *Technology* -database access: encryption, authentication, authorization -secure platform for communication of genomics & health information		Make sharing contingent on recipient organization replicating privacy and security standards of sharing organization Maintain individual's opt-in/opt-out rights, enabled by technology and organizational processes

Examples of Genomic Sequencing Devices

Illumina MiSeq

https://www.illumina.com/systems/sequencing-platforms.html

Oxford Nanopore Technologies Minion

https://nanoporetech.com/products/minion

Most often, the sequencing lab is separate from the clinical organization, although some hospitals or clinics have an in-house sequencing lab. The regulatory oversight of sequencing labs is not clearly defined currently. Some sequencing occurs in labs that have obtained ***CLIA certification*** by the Centers for Medicare and Medicaid Services (CMS), but this is not a requirement to conduct sequencing. The initial result of sequencing an individual is the creation of raw data files, which then move to step 2 for analysis.

Step 2. Individual's DNA data is processed and analyzed to identify variants

In this step, the individual's data is largely in the hands of bioinformaticians who work for the in-house or contracted sequencing lab, or for an external consulting organization specializing in DNA analysis. In its entirely raw form, DNA data is not very useful. And although sequencing technology has advanced greatly in the last decade, the processing of raw data into usable form requires a considerable amount of skill, experience, and time on the part of bioinformaticians.

Depending on the type of sequencing done (e.g., ***genotyping***, ***whole genome*** or ***whole exome sequencing***), the raw data files will have different formats, and their processing and analysis will require different numbers of steps. For example, for whole genome and exome sequencing, the raw data produced by the sequencing of a single individual consists of a large number of small datafiles, each of which contains a tiny section of the genome, and which together form a ***sequencing library***. That library is then subjected to a 3-step process of "assembly" and "alignment" (in which the original DNA sequence is reconstructed in its proper order) and finally "annotation" (comparison with a well-known ***reference genome*** or ***variant database***) to identify and catalogue variants found along the reconstructed sequence. Although this process is automated, bioinformaticians often visually inspect the results using a ***gene browser*** to help increase accuracy.

Step 3. DNA data is used for clinical interpretation

In the third step, the analytic results are provided back to clinicians who interpret it for the patient. The lab results include key findings about variations from clinical cases of similar patients in other genomic databases. It is worth noting that even though they are based on digital data, these findings are not free from potential errors, and in fact they always involve a degree of subjective interpretation. The variant information helps clinicians interpret the data of the current patient being treated. In this step, variant information is also linked to health records, so that clinicians can gain insight into the patient's condition and recommend treatment options.

At this point in the process, if health care providers are dealing with a difficult diagnosis or treatment situation, they are likely to consult with in-house and/or external clinicians in order to gather additional information to assist them in making a diagnosis. For this purpose, they may seek genomic data on patients with similar symptoms, or medical record data on patients with similar gene variants. This step requires sharing and discussion of the patient genomic data and ***phenotypic data*** (i.e. medical records) from multiple individuals. Who is involved at this step depends on the specific diagnostic situation, but it would usually include the patient's primary clinical team, clinicians at other medical facilities, and even internal or external bioinformaticians if subsequent scrutiny of the genomic data is deemed necessary to reach a diagnosis.

Step 4. Further data sharing and reuse

In the final step, the patient's DNA data is nearly always reused and shared beyond the original purpose for which it was created. Initially the focus of attention in the patient's data was likely on variants tied to his or her specific symptoms or disease. However, the patient's larger genome is valuable for myriad other purposes and this data is highly sought after by other clinicians and researchers. Hence it is common practice for an individual's DNA data (and its associated phenotypic data) to be de-identified and then added to a larger DNA database for further research and clinical uses. Although in many cases the patient has signed a consent form that authorizes additional uses in general, they rarely are informed about the specific ways their data will be reused. In addition, genomic data that is considered de-identified can be shared without consent, as noted above.

Databases that contain merged DNA and clinical data from many individuals are critical for researchers aiming to discover new variants and tease out their clinical implications. Variant discoveries from new data can then be compared with prior knowledge from pre-existing variant databases, gradually growing and refining the overall pool of genomic knowledge.

As the initial patient's data is merged into a larger de-identified database, that database may in turn be shared further afield. Clinical organizations such as health systems creating these databases may enter into partnership or licensing agreements with other organizations (such as other clinics, academic research institutes or drug companies searching for new **pharmacogenomic drugs**) to enable re-use of their data. Such partnerships facilitate the transfer DNA data from clinical to commercial settings. In doing so, these partnerships also blur the line between clinical, research, and commercial genomics (see text box).[25] If the patient's sequencing was done as part of a federally-funded research study, their data will also become part of a database deposited with the National Institutes of Health (NIH) and made available for other researchers.[26] At this point in the life cycle, individual patients no longer have control of their individual DNA data and most likely will never learn the results of subsequent analyses that utilize it.

Data Sharing Blurs the Lines Between Clinical Care, Research, and Consumer Firms

Because genomics databases are seen as valuable for a range of purposes, including clinical care, biomedical research, and commercial purposes, partnerships are forged that allow data sharing and reuse across organizations in all three of these sectors. For example, Foundation Medicine – a publicly traded corporation – offers a service to patients and their doctors involving sequencing coupled with tailored treatment recommendations based on that sequencing. Using client data, Foundation is assembling a growing database of cancer patient and tumor DNA, from which it hopes to derive new insights for tailored treatments. In November 2015, Roche, one of the world's largest pharmaceutical firms, acquired a majority share in Foundation, giving it access to their data for drug research. At the same time, another partnership links Foundation's genomics data on 20,000 cancer patients to Flatiron Health's clinical data on the drugs, treatments, and health outcomes of those patients from medical records. They promise eventual public disclosure of these de-identified data. And through ongoing ownership and governance control in Foundation, the consumer technology giant Google is also accessing Foundation's data for their commercial research and development purposes.

In sum, the life cycle of DNA data shows how genomes sequenced for individual patient care become critical to new research discoveries, as well as profit-making endeavors by entrepreneurial organizations, through database creation, sharing, and reuse.

PRIVACY AND SECURITY RISKS AT EACH STEP IN THE DNA DATA LIFECYCLE

Given the nature of genetic data, and specifically how much it can reveal about an individual and his or her family, many people may want to maintain their privacy regarding this kind of information. However, across this DNA data lifecycle, many privacy and security risks can be identified. In this section, we consider the risks at each step, along with potential remedies that may help to ameliorate them. Before taking stock of the specific risks, it is worth pausing to consider the meaning of ***privacy,*** both in general and in the specific context of DNA data. In broad strokes, privacy involves being able to choose what parts of oneself to disclose to different audiences.[27] Although the United States lacks a broad legal right to privacy, many Americans, as well as the United Nations, regard a degree of privacy to be a fundamental human right.[28]

In the context of personal information, such as DNA data, privacy means having control over that information, such that a person can confidently choose to disclose it to one trusted party for one specific purpose without concern that it will be passed to other parties or used for other purposes. Maintaining privacy therefore also logically requires mechanisms to ensure the ***security*** of that information while it is being used by the other party, preventing unauthorized disclosure and further uses. Of course, when an individual's information becomes part of a larger database that encompasses data on many other people, the maintenance of privacy and security no longer depend just on how their specific information is treated; instead, much depends on the policies and practices related to the larger database.[29] Hence, much of the risk – and potential remedy – concerning DNA data involves the databases and how they are managed.

In ***step 1*** of the DNA life cycle, when genomic data are collected for sequencing, patients must give their consent for the process to begin. However, they are unlikely to be fully aware of the pathways their data are going to travel. Patients often consent to sequencing because their doctor hopes it can help in their diagnosis and treatment; that usage is naturally their priority. But in most situations, their data will also be added to a DNA database, and reused for other research and clinical purposes. Patients typically sign consent forms that say their data will be reused in unspecified ways (see text box).[30] Given an inability to truly anonymize DNA data

Update on Federal Consent Rules for Research
A January 2017 update to the U.S. Common Rule - which applies to research projects and organizations receiving federal funding - has implications for the treatment of genomics data. The update allows researchers to be in compliance if they obtain '**broad consent**' from individuals whose biosamples are collected (as described above). An alternative approach, requiring researchers to inform individuals about each specific way their data are being reused, was rejected due to the burden it would have placed on researchers.[1] Many clinical and commercial genomics organizations are adopting this broad consent approach as well.

through de-identification, that onward journey carries disclosure risks that rise as more individuals handle the data. All those individuals could accidentally disclose the data, or themselves engage in its theft, making the data available to parties who would use it against them. Although those risks may be reduced, as we discuss below, they can never be eliminated. The consent forms used by commercial genomics companies and other DNA data services tend to be similarly non-specific in detailing the reuses of an individual's data (see text box).[31]

Excerpts from a commercial genomics consent form

The company 23andMe, which shares genomic data in some form with more than a dozen other entities, provides its customers with a consent form, privacy statement, and terms-of-service. Their consent "key points" 500-word summary describes some aspects of their genomic data sharing and associated risks:

"23andMe researchers who conduct analyses will have access to your genetic and other personal information, but not to your name, contact, or credit card information."

"23andMe may share some data with external research partners and in scientific publications. These data will be summarized across enough customers to minimize the chance that your personal information will be exposed."

The 23andMe privacy form elaborates on data sharing:

"We may share anonymized and aggregate information with third-parties; anonymized and aggregate information is any information that has been stripped of your name and contact information and aggregated with information of others or anonymized so that you cannot reasonably be identified as an individual."

The longer terms-of-service form also states:

"Genetic Information that you share with family, friends or employers may be used against your interests. Even if you share Genetic Information that has no or limited meaning today, that information could have greater meaning in the future as new discoveries are made."

What can be done to enhance privacy and security of DNA data at this initial step? At a minimum, during the consent process, patients need to be educated about the benefits and risks. Before providing consent and being sequenced, the patient can be required to meet with a ***genetic counselor***, to learn about the potential benefits and risks associated with DNA data generation and analysis – as well as about the potential benefits and risks of further sharing and reuse of data.

Patients can also be given the choice to ***opt-out*** of data reuse later in the life cycle. The consent process itself can be modified so that patients are able to choose the degree of database inclusion and further uses they want. For example, they could opt for inclusion in the health care organization's internal database, but not for further sharing with external partners. They could also choose to opt-in to reuse in instances that they personally support (for example, certain medical or scientific studies they care about), but stay out of reuse in instances they do not support (for example, certain commercial uses).

Without such genomic counseling and opt-out provisions, consent in the clinical setting represents an individual loss of control and creates potential privacy risks. At the same time, it is important to recognize that these provisions would come at a non-trivial cost. In particular,

researchers would lose the freedom to easily re-analyze the data that came from individuals who opted-out. A further issue is the treatment of infants and children in genomics databases. These minors cannot provide informed consent, but after sequencing in the clinical setting, their data is being reused in the same ways that adult genomic data is being reused.

In ***steps 2 and 3***, data security risks come to the fore, as activities related to the storage and handling of DNA data present many points of vulnerability. Although health care organizations are relatively sophisticated when it comes to protecting personal health data, because of HIPAA laws, nonetheless they are frequent targets of attack, and suffer frequent data spills. Indeed, 90% of health care organizations and associated firms responding to a 2016 Ponemon survey reported they had experienced a data breach, and 64% reported a breach that involved leaking patient medical records, in just the last two years.[32] It is too early to know about the occurrence of genomic spills, but they are likely to be occurring now and in the near future as well. In one recent incident, Quest Diagnostics, which handles genomics data for a large IBM Watson initiative, reported a major spill of patient lab data (see inset).[33]

> *"Quest Diagnostics, a New Jersey-based medical laboratory company, disclosed a data breach affecting about 34,000 people on Monday. Digital intruders stole personal and medical information of customers—including names, dates of birth, lab results…. Attackers gained access to the data on November 26 through an improperly secured mobile app that lets patients share and store electronic health records."*

For genomic data management, there are many hardware and software options for the storage and transit of DNA data. Data can be stored on an organization's local servers or in a cloud service such as Amazon Web Services or Google Cloud. Data can be moved via dedicated internal cables within an organization, or over the Internet, or physically via portable hard drives. For DNA data that is stored only (or largely) in the cloud, data transit can be limited through cloud computing practices, essentially requiring any analysis software to be "brought to the data" in the cloud rather than the data being analyzed by software on a local machine.

Each of these alternative data management options can be made relatively more secure using a number of organizational best-practices:

- *Encryption*: Ensuring that data is encrypted in transit and at rest
- *Authentication*: Verifying the identity of individuals accessing the data. Two-factor authentication, involving a token sent to a mobile device or key fob, provides additional validation. For sensitive data, in-person authentication could be required.
- *Authorization*: Narrowing the number of people with access to data based on the project or task, and limiting the duration of that access.
- *Monitoring and auditing*: Assessing and improving system security, and tracking details of use, unauthorized use and compliance. Routine vulnerability assessments and penetration tests. Using a **blockchain leger system** may be one way to verify history of data access throughout lifecycle.
- *De-identification*: Stripping an individual's identity from the DNA data is useful, but it does not achieve true anonymization in many circumstances (as we discussed above).

Beyond these current best-practices, there are also efforts to allow users to query databases as to their contents, and to aggregate and analyze data across multiple databases, all without exposing the user to the actual underlying data. However, these efforts have proven to be challenging. For example, in 2015 a "beacon" system was developed by Global Alliance for Genomic Health (GA4GH) to allow people to quickly query many databases (via their beacons) about whether they contain individuals with certain specific variants. However, this system was quickly shown to be vulnerable to re-identification attacks.[34] In another example, "Datashield" software was developed to allow pooled statistical analyses across many genomic databases, returning pooled results without revealing information about any specific dataset that went into the pooling.[35]

At the same time, despite these technological protections, all these options are vulnerable to risks that arise through human ***compliance behavior***. In general, many data spills result from (non-) compliance behavior among individuals who are authorized to use sensitive data.[36] Carelessness and theft by employees pose series risks to data management. Minimizing such behavioral threats to data privacy and security are best handled through robust staff screening and training, coupled with sanctions for employees and subcontractors who violate policies and procedures, and physical workplace security. In designing training and sanctions, it is critical to recognize that as the intrinsic value of a genomics database increases, so does the financial temptation for insiders to become involved in a data spill.

In step 3, data security risks also arise as physicians and other clinicians share *individual* patient genomic data and interpretations with colleagues in the course of their work. Out of a desire to help their patients, physicians and other providers often search for other patients whose DNA variants and clinical symptoms appear similar to the patient they are treating. In this process, they will share data about a patient with other clinicians, and request that these clinicians share data on other patients with them. In the era of genomic sequencing, this sharing can involve genomic data – variant lists, interpretations, genome snippets or sections of interest, or even whole exome or genome data. In this process, clinicians are often motivated to act quickly and efficiently, given both their professional desire to help a suffering patient, and their own busy schedules.

These forces increase the incentive to share a patient's genomic data, along with other personal health information, in a manner that is not secure. In particular, the electronic ***communication platforms*** used for this purpose are an important consideration. In recent years, it has become increasingly common for physicians to share information from their mobile devices, using email, text, or apps that do not comply with HIPAA requirements for data security, and this has been a major cause of health data breaches, as in the Quest data breach reported above.[37] Mobile communications transmitted over wireless networks can be particularly vulnerability to interception. When genomic data is linked with other patient information, it is clearly sensitive; even in isolation, as we have shown, genomic data must be treated with care.

Hence all communication of patient data needs to take place using a secure platform – one that complies with HIPAA requirements, maintaining encryption in transit and at rest. Of course, this poses a complication for sharing among clinicians who do not use the same secure platform. And the security of these communications is still limited by the extent to which clinicians are compliant in maintaining the security of their own mobile devices.

Step 4 involves the DNA data's onward journey beyond the original setting and purpose for which it was generated. In many ways, this step represents the greatest risk. The data is changing hands and moving across organizational boundaries, being entrusted to more individuals, exposed to multiple organizational policies and routines, all of which increase the chances of loss, accidental disclosure or theft. Organizations receiving DNA data may have different goals, and their employees may be subject to different privacy and security practices, and these differences may grow with the passage of time, as illustrated in a recent court case involving DNA collected for medical research (see text box).[38]

A court case involving disputed DNA data reuse
In 1993, DNA samples from members of a small American Indian tribe, the Havasupai, were obtained for medical research purposes, using broad consent. The initial study approved by the tribe involved diabetes, but subsequently the DNA data were used by other researchers in other studies, on topics related to mental health, migration, and inbreeding. In 2003, a tribal member learned about other research while attending a university lecture, leading to a lawsuit that was ultimately settled out of court in 2010. Issues in the lawsuit, *Arizona Board of Regents v. Havasupai Tribe*, included lack of informed consent, violation of civil rights, unapproved use of data, and violation of medical confidentiality/re-identification.

When genomics databases are shared across organizational boundaries – i.e. when data is being transferred to reside in another organization, or data access is being provided to members of another organization – there needs to be an accountability structure that ensures data privacy and security in the receiving organization. An obvious template for this accountability structure is provided by HIPAA. Under HIPAA, after it was updated with the 2010 HITECH Act, when a hospital or health company shares data that contains personal health information, the receiving organization has to sign a ***HIPAA Business Associate*** (BA) contract, in which they agree to a set of data privacy and security practices that largely mirror those required within the hospital or health company itself. The BA entity becomes liable for responding to data breaches as well. Such BA contracts are signed by third-party organizations for insurance claims processing, hospital consultants, and even transcriptionists handling personal health information. Such an approach does not apply to genomics data, but it could.

As in the primary organization, data security practices in the organizations that receive access to shared data should also be set to a high standard. If data are transferred to a partner organization, the management of genomics data at that partner organization naturally pose risks that are similar to those we inventoried above in steps 2 and 3. In particular, there are

data breach risks arising in the storage and use of the data, regardless of whether it is physically located on local servers or in the cloud, and whether or not it is moved between locations over the internet. The same organizational best-practices can help mitigate those risks. If data is not transferred, but access is provided to members of a partner organization, then the same human compliance behavior concerns arise for those partner organization employees who are given access.

The severity of concerns with genomic database sharing was underscored in the conclusions of a recent assessment by the American Association of Arts and Sciences (AAAS) and the Federal Bureau of Investigation on this topic (see inset).[39]

> *"Beyond access controls, encryption, and other common data and cyber security technologies, no solutions exist that prevent or mitigate attacks on databases or the cyber infrastructure that support Big Data in the life sciences, which could result in consequences to the life science, commercial, and health sectors."*

What factors are likely to be associated with an increased risk of privacy and security compromise in partner organizations? Research on the incidence of wrongdoing and accident events in organizational and scientific fields provides a useful guide.[40] In particular, the research suggests heightened risk when data sharing includes:

- *Emerging innovators*. Small and recently-founded organizations tend to lack the resources, experience and scale to have developed and funded internal compliance systems. And the culture of innovation in emerging ventures often encourages breaking with industry rules. Examples of emerging innovators include new specialized sequencing labs and bioinformatics firms, consumer-facing startups, and citizen-science organizations that lack funding and experience. (Of course, larger organizations may be more visible, to hackers as well as everyone else, but smaller organizations tend to be more vulnerable.)

- *Organizations based in weaker regulatory jurisdictions*. Examples include community hospitals and medical offices that have not previously engaged in clinical research, which could fall outside federal Common Rule jurisdiction and lack IRB experience. Other examples include commercial genomics companies and patient advocacy organizations that do not handle patient medical records, implying they fall outside HIPAA jurisdiction and lack experience handling sensitive patient data. Partner organizations may also be based in state jurisdictions with looser regulations.[41] An extreme (but common) example is partner organizations based outside the legal jurisdiction of the United States altogether. For example, a major U.S. genomics company, Human Longevity Inc., recently formed partnerships with the British-Swiss pharmaceutical firm AstraZeneca and the South Africa-based health and life insurance company Discovery Ltd.[42]

- *Use of partners and outside contractors*. As the number of partner organizations increases, accountability structures become decentralized and more diffuse, and there is a greater chance that compliance standards will conflict between organizations. More partners also add complexity, which increases the chance for errors or gaps in procedure that create vulnerabilities. For example, a clinical organization may have crisscrossing partnerships with pharmaceutical companies, genomics startups, academic researchers, and patient advocacy groups – all of which are common in the genomics field. Data brokers are also used to share genomic data, adding another layer of organizational complexity (see text box).[43]

Genomic Data Brokers
In response to growing demand for human genomic data, a new crop of start-up companies is serving as data brokers, offering to pay individuals for access to their genomic data, which the broker then sells to research studies. One example of this broker function, a startup venture called DNASimple, offers to give consumers control over which research studies their data are given to, and to later destroy a customer's data upon their request.

- *Links between for-profit and non-profit entities*. These partnership arrangements will mix conflicting legal jurisdictions, and they are also likely to mix conflicting financial and societal objectives. Those factors increase the chances that compliance standards will differ, and complicate compliance oversight.

- *Organizations experiencing rapid change*. Organizations that are restructuring, merging or being acquired, rapidly expanding or contracting, experiencing financial hardship or going through bankruptcy are all more likely to experience breakdowns and gaps in compliance, as standard operating procedures are suspended (see text box).[44] They are also likely to have increased staff turnover, bringing an elevated risk for newly hired employees who lack training, and outgoing employees who may be more willing to retaliate against their former employer.

What happens to a genomics database in bankruptcy?
As genomics databases proliferate in the private sector, the question arises concerning how they would be treated in a bankruptcy proceedings. Although personal data is partly protected under Federal Trade Commission rules, and enforced by state consumer protection authorities, this does not stop the sale of genomic data to another entity during bankruptcy. The rule of thumb is that whatever privacy policies the bankrupt company has in place will have to be replicated in the entity that that buys the data. HIPAA can also constrain the sale of personal health data, but de-identified genomic data would probably be exempt from this constraint.

Professional training. Across the entire lifecycle, there is a broad need for clinicians, bioinformaticians, and others who handle genomic data to be trained in the privacy and security risks associated with these data. Currently, the primary vehicle for this is employee training – a highly decentralized and therefore uneven vehicle. Educational institutions can also

play an important role, by incorporating privacy and security topics into the curriculum of degree programs in medicine, bioinformatics, and related fields. There is also likely to be a need for commercial training programs, and advisory services to help disseminate and establish best-practices in organizational training programs across health care organizations. Many practicing physicians, for example, will need continuing medical education (CME) related to genomics in coming years, representing an opportunity to ensure they are exposed to privacy and security concerns.

Self-regulation in the private sector. Market forces and self-regulation can play an important role in reducing these risks. Yet in the absence of an external accountability mechanism, self-regulating market actors only bear part of the cost of a genome data breach that occurs. Instead, further costs are shouldered by individuals whose data are disclosed (and their family members, whose data are indirectly disclosed), who incur harm if those data are used against them subsequently. Through this lens, genomic data-spill risks are a negative externality that may merit regulation.[45] While the threat of competition or theft from rivals should incentivize genomic database security, at the same time competitive pressure increases the incentive to enter partnerships to accelerate discovery ahead of competitors - increasing the data sharing risks posed by those partnerships.

Public data sharing initiatives. In the public sector, commitments to open science are aimed at accelerating scientific advances. These initiatives, while laudable, also pose risks to the extent that public sharing of genomic datasets increase their exposure. For example, a genomic database of cancer patients, linked to their medical record data, was recently released by the American Association for Cancer Research, in collaboration with Sage Bionetworks, for public research access. While data privacy provisions were included in the design of the public data release, it remains to be seen whether these data are vulnerable to disclosure and reidentification attacks.[46] On a larger scale, the National Institutes of Health collects genomics data into several centralized archives designed for sharing and reuse by other researchers, increasing the benefits as well as the risks of data reuse (see text box).[47]

Centralized Collection and Sharing of U.S. Human Genomics Data

The central collection and integration of many databases occurs within a single public sector entity, the U.S. National Center for Biotechnology Information's (NCBI) database of Genotypes and Phenotypes (dbGaP). Such central collections may bring additional exposure risks associated with the scale of the database and the number of sharing events to be managed (20,178 approved data requests as of July 1 ,2015). The dbGaP database has already experienced several known data security incidents, and is likely to be a target for future hacking. There are other risks associated with unintended future uses of such data collections, including future governmental reuse, including reuse for forensic investigation, to identify victims in the wake of mass casualty events, for citizenship verification.

THE BIG PICTURE: WE NEED GREATER SOCIETAL ATTENTION TO PRIVACY AND SECURITY OF GENOMIC DATA

Genomics offers bold promises for revolutionizing medicine as we know—some of which have already been realized. There are tremendous potential benefits from genomics, revealing new life-saving and life-enhancing discoveries for precision medical care. Genomics databases will play an important role in achieving those breakthroughs. At the same time, we also need to be appreciate the serious risks that the disclosure of sequenced DNA results pose for individuals.

If indeed the genomics field is at a critical inflection point, as many believe, then this is a crucial point for us to wrestle with the tensions inherent in promoting future research while at the same time safeguarding individual privacy. The lifecycle view we offered in this paper – showing how DNA data is generated and processed for use in precision medicine – identifies potential risks to privacy and security at each step. Our intention in setting out this lifecycle perspective is to provide a cautionary tale, indicating where data breaches could occur in clinical practice, despite breach prevention efforts currently employed.

Our intention is to animate broad public deliberation about how DNA privacy and security issues can best be addressed to achieve the twin goals of both facilitating on-going research while ensuring that state-of-the art privacy and security measures are adopted. This public discussion should include all relevant stakeholders – not just those who are already involved and invested in the current genomics field, but also representatives of those ordinary citizens who are soon to be affected by the genomics revolution (whether they like it or not).

Our analysis revealed three fundamental questions that we believe warrant broad societal reflection and deliberation if we are to reach for the promise of medical genomics while simultaneously mitigating risks of disclosure. Exploration of these questions should be at the core of public deliberations about the privacy and security of genomic data. There are of course many other pressing questions – involving legal-regulatory frameworks, economic impacts, and national interests, among other things. But we believe these three questions should take priority because they start from a recognition of the fundamental nature of the genome.

Three Fundamental Questions

Question #1: "Who should have the right to make decisions about <u>your</u> genome?"

The tension underlining this question pits individual vs. societal rights. Basically, the question is to what extent can I control what happens to my sequenced DNA results? Although no one "owns" their DNA data, an individual-centric approach to DNA data means that individuals retain control over when their DNA data is extracted and to what uses it is put. Presumably this would entail individuals making informed decisions about when they choose to be sequenced, what their sequenced data can be used for, and under what circumstances their data can be used by others. While this is the purpose of informed consent, at the time that consent is given, individuals are asked to agree to - but cannot possibly fully appreciate - the potential, unnamed uses to which their data may be put in the future. Indeed, the current common practice of obtaining "broad consent" limits individual control further by requiring people being sequenced to agree to a broad range of unspecified future reuse and sharing scenarios.

The most obvious answer to this question from a privacy and security perspective is that individuals ought to retain control over their DNA and its use. However, as we noted above, there are many research and commercial organizations that depend on large databases of DNA data to do their work. "Nearly three-quarters of all genomics companies provide tools (both physical and in the cloud) to pharmaceutical companies and academic research institutions."[48] In addition, medical organizations are scrambling to build bigger (and better) DNA databases to enable new research and improved clinical practice, and the National Institutes of Health actively promote the sharing of DNA databases when funding innovative medical research.

Just as the *HeLa* cell line that originated from Henrietta Lacks was instrumental in advancing biomedical research,[49] these sequence databases – in this case originating from many people rather than just one – are likely to be instrumental for progress in medicine and health care. Indeed, creating and sharing these databases is believed to be critical for success in the White House Precision Medicine Initiative, the Cancer Moonshot, and the 21st Century Cures Act.[50] So a compelling counter argument to individual control is that implementing full individual control could grind genomic research to a halt, and greatly slow progress in medicine and science. In this view, the scientific community should be empowered and trusted to decide how individual genomic data should be used, because they are best positioned with the relevant expertise to weigh the potential benefits and costs of its use overall.

The scientific community should decide how DNA databases are handled

vs

Everyday citizens should have control of the uses of their own genomic data

Question #2: "How closely should you hold onto your genome?"

How proprietary individuals feel about their sequenced DNA data runs the gamut from those who want to zealously guard it to those who are willing, and even eager, to circulate it freely, either unaware of or indifferent to potential future risks to themselves and their families. Still others cite a sense of inevitability of widely-shared DNA data and even the potential likelihood of peer pressure to share it that could emanate from suspicion that those who do not must be hiding damaging evidence about themselves. Indeed, within a culture of widespread disclosure of personal data through social media, arguing in favor of privacy may be a losing battle. There are even persuasive arguments that under some conditions there may be a obligation to reveal your genomic information (see text box).[51]

A Duty to *Reveal* Your Genome?
Genomic data often provide information of relevance to relatives, including information about increased disease risk, and biological parent/child/sibling status. This is giving rise to a range of unsettled ethical and legal questions. Under what conditions does this information create a duty for an individual to actually *reveal* information from his or her genomic data to relatives who are impacted? And beyond that individual, could a physician or other clinician treating both the person sequenced and his or her relatives have a duty to inform the relatives? What responsibility is born by the sequencing labs, research organizations, or medical organizations that acquire and store this information? Beyond ethical considerations, could these actors be exposed to claims of negligence, malpractice, or other legal liabilities?

Still, if the societal response to question #1 is that individuals should have the right to protect their own DNA from use by others and, instead, retain it as a treasured private possession that should be kept secure in a digital safety deposit box, then attention to how we are going to enable that is needed. At a minimum, this deserves an informed public discussion, which brings us to the third question.

DNA data is just another bit of personal data to share with relative openness during a wide range of social transactions

vs

DNA is a treasured possession that needs to be kept secure in a digital safety deposit box

Question #3: "What standards are desirable for securing genomic databases?"

The privacy considerations not withstanding, as we have shown above, the risks of data breaches of sequenced DNA data remain as real possibilities either because of human ignorance or carelessness in the handling of DNA results or nefarious activities such as theft through hacking. With sequenced data only falling under HIPAA protections once it is linked to a patient's medical record, the potential for unsecure transmission of these data accidentally or intentionally is greater than zero, and the potential for re-identification of anonymized data certainly exists and becomes even higher when the data is linked to health records.

As our life cycle model shows, DNA passes through a lot of hands in its trajectory from collection to reuse. The responsibility for securing DNA currently falls to those individual organizations doing the sequencing, interpreting the results, presenting them to patients or reanalyzing big DNA data bases. While technical protocols for ensuring data security secure are available, they only work as deterrents if organizations (and their employees) are assiduous in their use and enforcement. An unanswered question that remains is whether specific standards (similar to HIPAA) for the handling of genomic data are needed at a societal level to ensure that individual organizations take this responsibility seriously. Those who oppose such regulations (such as research agencies) make the reasonable claim that strict regulations will slow the progress in realizing the promise of genomics.

Could a model for best practices in the privacy and security of sensitive data exist in another industry? Some observers believe the financial services industry may provide an instructive example (see text box).[52]

Learning from the Financial Services Industry?
The experiences of the investment banking sector could provide a degree of road mapping for data policies and practices in the field of medical genomics. Financial service firms handle large volumes of sensitive customer data, and share data across institutions on a regular basis. A combination of government regulation and voluntary initiatives have led to relatively uniform and rigorous data security practices which, thus far, appear to have limited the scope and success of data breaches. As investment banks adapt practices for new technological platforms, including cloud computing and mobile access, there may be valuable lessons for medical genomics.

Unrestricted databases will help our nation be at the forefront of genomic science and innovation

vs

Tight monitoring of the use of genomic databases will help protect individual and family privacy

What Action is Needed Now?

It is our contention that answering these questions necessitates careful, well-informed societal deliberation at the broadest level possible. The only way that the myriad stakeholders interested in genomics can determine the most efficacious way forward is through a public, collaborative exploration of these questions and how best to resolve them. As a society, we need to consider how best to structure such a dialogue to ensure that all interested stakeholders (patients, family members, clinicians, researchers, insurers, advocacy groups, businesses, regulators etc.) can thoroughly explore and debate alternative scenarios for how to safeguard the privacy and security of DNA data.

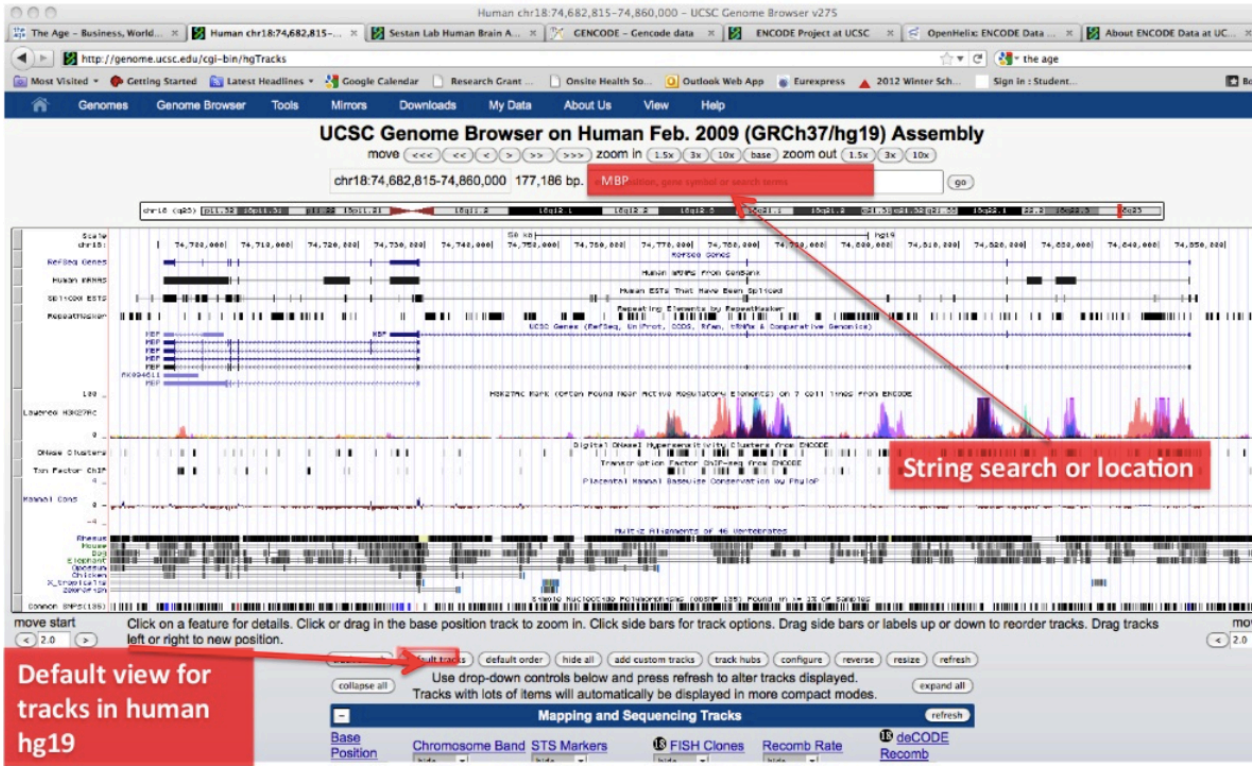
***About the authors**: We are accomplished social scientists who study the development and transformation of industries and scientific fields. We teach and conduct research on the emergence of new standards and practices in these fields. Our research has addressed questions such as: How do a diverse set of organizations in a field come to agree on collective standards and governance procedures? How do the risks and benefits of new technologies and practices come to be perceived and communicated among these organizations? How do voluntary industry associations, formal state regulations, and social movements influence this process? What facilitates collaboration within scientific communities? And within organizations, what are the structural and leadership characteristics that affect adoption of new standards and practices? We both have undergraduate degrees in science, and doctorates in social science – specifically the field of Management & Organization.

We wrote this white paper because, from our vantage as social scientists who study field transformation, we believe the field of medical genomics is in a watershed moment. Many of the issues the field faces involve ethical decisions with uncertain outcomes for many. Given these stakes, we also believe it is important to increase public understanding and involvement in decisions about the standards for this rapidly developing field. Even though the issues are complex and technical, we need the involvement of both insiders and outsiders. We hope to stimulate widespread discussion of the issues raised here.

We can be reached at fbriscoe@psu.edu and b9g@psu.edu.

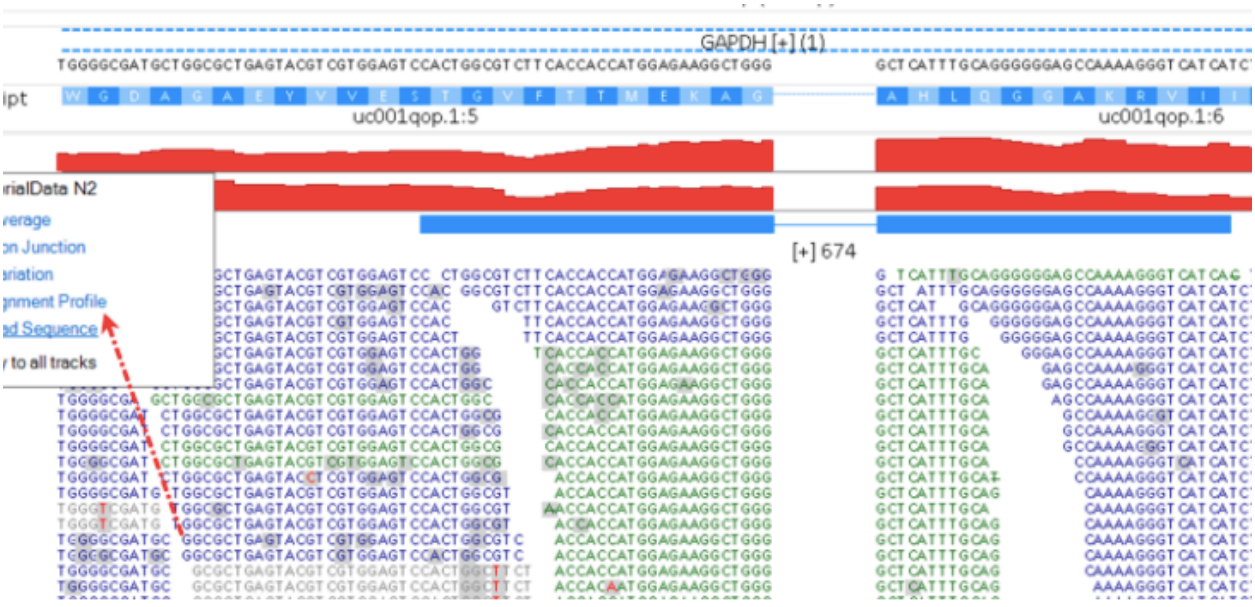
APPENDICES: Examples of genome browsers

USCS Genome Browser



http://genome.ucsc.edu

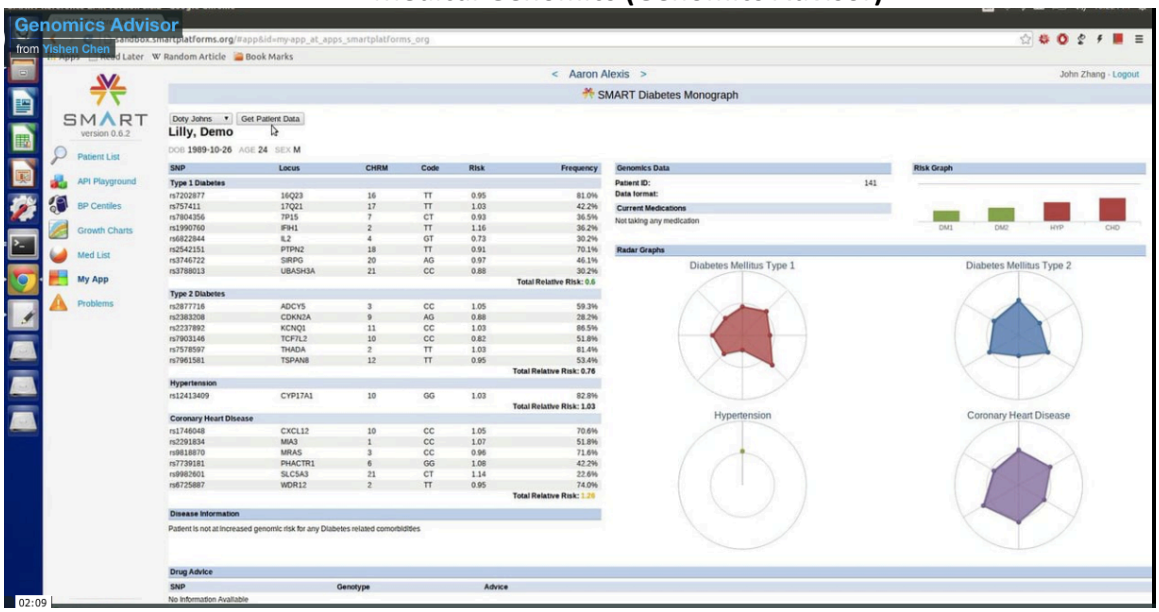
Omicsoft Genome Browser



http://www.omicsoft.com/genome-browser/

APPENDIX: Examples of Individual Genomics Reports

Medical Genomics (Genomics Advisor)



http://projects.iq.harvard.edu/smartgenomics

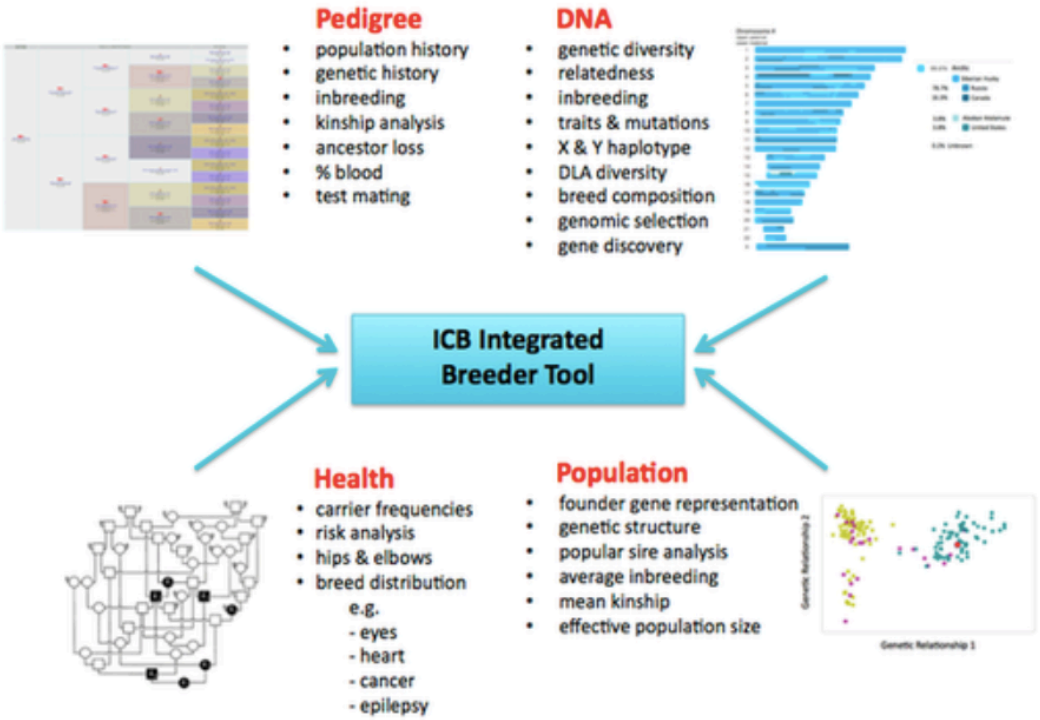
Metagenomic Profile (Biome Organisms)

Sample Composition

Name	Readcount (% of classified reads)
Mycobacterium abscessus	959 (11.65%)
Vibrio cholerae	937 (11.38%)
Salmonella enterica subsp. enterica	738 (8.96%)
Enterobacter cloacae subsp. cloacae NCTC 9394	704 (8.55%)
Rhodobacter sphaeroides 2.4.1	613 (7.44%)
Staphylococcus aureus	567 (6.89%)
Klebsiella	563 (6.84%)
Bacillus cereus ATCC 10987	434 (5.27%)
Bacillus cereus	347 (4.21%)
Rhodobacter sphaeroides	334 (4.06%)
(Remaining organisms)	2038 (24.75%)

https://www.onecodex.com

ICB Breeder Tool (Canine Genomics)



http://www.instituteofcaninebiology.org/blog/the-icb-breeder-tool-available-now

ENDNOTES

[1]This paper is based on in-depth interviews with more than 25 leaders in the genomics field, and review of over 200 archival documents, conducted during the fall of 2016. Our informants included leaders in genomics biomedical research, clinical health care delivery, health care regulation, commercial genomics, venture capital, data security, research oversight.

[2]The shift in medical terminology from "genetics" to "genomics" reflects a shift from the use of tailored genetic tests for specific mutations toward the use of sequencing technology to generate data on a person's whole genome (or significant sections of it).

[3] Battelle 2013 report: "The Impact of Genomics on the US Economy." http://web.ornl.gov/sci/techresources/Human_Genome/publicat/2013BattelleReportImpact-of-Genomics-on-the-US-Economy.pdf

[4] Topol, E. 2015. *The patient will see you now: The future of medicine. Philadelphia:* Perseus Books.

[5] Examples of specific initiatives integrating genomics into clinical care include Kaiser Permanente's lung cancer initiative (https://share.kaiserpermanente.org/article/new-clinical-trials-use-genetic-testing-to-personalize-lung-cancer-treatment/), Geisinger Health System's autism initiative (http://www.geisinger.org/for-researchers/initiatives-and-projects/pages/simons-vip.html), InterMountain Healthcare's Precision Genomics Cancer Initiative (https://intermountainhealthcare.org/services/cancer-care/precision-genomics/research/), Partners HealthCare's service for patients with suspected but undiagnosed rare genetic diseases (http://personalizedmedicine.partners.org/laboratory-for-molecular-medicine/tests/genome.aspx), and the Texas Medical Center's Clinical Cancer Genetics Program which coordinates genetic testing and high-risk cancer surveillance for families with hereditary cancer syndromes (https://www.mdanderson.org/prevention-screening/family-history/hereditary-cancer-syndromes.html). Several hospitals and medical centers are participating in the NIH-funded Electronic Medical Records and Genomics (eMERGE) Network initiative to integrate genomics and patient medical records into clinical care (https://www.genome.gov/27540473/electronic-medical-records-and-genomics-emerge-network/).

[6] Precision Medicine World Conference, Closing Panel, January 24, 2017, Mountain View, CA.

[7] Schwartz, P. J., Crotti, L., & Insolia, R. (2012). Long-QT syndrome from genetics to management. *Circulation: Arrhythmia and Electrophysiology*, *5*(4), 868-877.

[8] FDA Drug Safety Communication: Reduced effectiveness of Plavix (clopidogrel) in patients who are poor metabolizers of the drug. http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm203888.htm ; Dean L. Warfarin Therapy and the Genotypes CYP2C9 and VKORC1. 2012 Mar 8 [Updated 2016 Jun 8]. In: Medical Genetics Summaries [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2012-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK84174/

[9] Cooperberg, M. R., Davicioni, E., Crisan, A., Jenkins, R. B., Ghadessi, M., & Karnes, R. J. (2015). Combined value of validated clinical and genomic risk stratification tools for predicting prostate cancer mortality in a high-risk prostatectomy cohort. *European urology*, *67*(2), 326-333.

[10] Grupp, S. A., et al. (2014). T cells engineered with a chimeric antigen receptor (CAR) targeting CD19 (CTL019) have long term persistence and induce durable remissions in children with relapsed, refractory ALL. *Blood*, *124*(21), 380-380.
[11] Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, *187*(2), 367-383. Robinson, M. R., Wray, N. R., & Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics*, *30*(4), 124-132.
[12] D. Nyholt, C. Yu, and P. Visscher. On Jim Watson's APOE status: Genetic information is hard to hide. European Journal of Human Genetics, 17:147–149, 2009.
[13] Hooper, L. V., Littman, D. R., & Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science*, *336*(6086), 1268-1273. Honda, K., & Littman, D. R. (2016). The microbiota in adaptive immune homeostasis and disease. *Nature*, *535*(7610), 75-84.
[14] Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., ... & Albrecht, E. (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *science*, *340*(6139), 1467-1471. Krapohl, E., & Plomin, R. (2016). Genetic link between family socioeconomic status and children's educational achievement estimated from genome-wide SNPs. *Molecular psychiatry*, *21*(3), 437-443.
[15] Humbert, M., Ayday, E., Hubaux, J. P., & Telenti, A. (2013, November). Addressing the concerns of the lacks family: quantification of kin genomic privacy. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (pp. 1141-1152). ACM.
[16] Wagner, J. K. (2013). Playing with heart and soul… and genomes: sports implications and applications of personal genomics. *PeerJ*, *1*, e120.
[17] https://www.bloomberg.com/news/articles/2015-06-05/u-s-government-data-breach-tied-to-theft-of-health-care-records
[18] Other regulatory actors could also become more involved in genomics in the future. The FDA has been working to develop a "flexible, adaptive regulatory approach" to ensuring the accuracy and safety of genomic sequencing. However, after significant stakeholder input on proposed regulations over several years, as of January 2017, they have not issued new regulations, nor is there a timeline for issuing them. The Federal Communications Commission (FCC) also has a particular interest in privacy in the digital age, which could apply to genomics. In October 2016, the FCC passed new regulations that ensure consumer privacy by limiting how internet provider companies use and sell customer data. The FCC has also expressed an interest in consumer health data, but since the FCC does not have jurisdiction over non-profit organizations, and many genomics entities in health care and academia are non-profit, their jurisdictional reach is limited.
[19]NHGRI IRB Guide to Writing Consent Forms - Version 2.0 (November 25, 2015). Available at https://www.genome.gov/27528182/irb-forms-templates-and-guides/
[20]In contrast, European privacy laws provide individuals with greater rights to their personal data, but those laws are being updated, and their application to genomics is not yet clear. See Townend, D. (2016). EU Laws on Privacy in Genomic Databases and Biobanking. *The Journal of Law, Medicine & Ethics*, *44*(1), 128-142.
[21] U.S. Department of Health and Human Services. 2016. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance

Portability and Accountability Act (HIPAA) Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#coveredentities ; Presidential Commission for the Study of Bioethical Issues. 2012. *Privacy and Progress in Whole Genome Sequencing*. http://bioethics.gov/sites/default/files/PrivacyProgress508_1.pdf

[22] Nyholt, D. R. 2012. Using Genomic Data to Make Indirect (and Unauthorized) Estimates of Disease Risk. *Public Health Genomics* 15: 303–311; Zaaijer, S., Gordon, A., Piccone, R., Speyer, D., & Erlich, Y. (2016). Democratizing DNA Fingerprinting. *bioRxiv*, 061556.

[23] Institute for Critical Infrastructure Technology (ICIT) (2016). Hacking Healthcare IT in 2016: Lessons the Healthcare Industry Can Learn from the OPM Breach. http://icitech.org/wp-content/uploads/2016/01/ICIT-Brief-Hacking-Healthcare-IT-in-2016.pdf

[24] Global Alliance for Genomics and Health. Security Technology Infrastructure: Standards and Implementation Practices for Protecting the Privacy and Security of Shared Genomic and Clinical Data. Version 2.0, August 9, 2016. https://genomicsandhealth.org/category/search-topics/security

[25] Flatiron Health and Foundation Medicine Unveil Powerful Oncology Information Resource to Advance Precision Medicine. November 3, 2016 Press Release. http://investors.foundationmedicine.com/releasedetail.cfm?releaseid=997385

[26] NIH's Genomic Data Sharing Policy, document available at https://gds.nih.gov/03policy2.html.

[27] The Right to Privacy in the Digital Age, United Nations Office of the High Commissioner, http://www.ohchr.org/EN/Issues/DigitalAge/Pages/DigitalAgeIndex.aspx.

[28] Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509-514. For attitudes about privacy of health information, see Dimitropoulos, L., Patel, V., Scheffler, S. A., & Posnack, S. (2011). Public attitudes toward health information exchange: perceived benefits and concerns. *American Journal of Managed Care*, *17*(12 Spec No.), SP111-6.

[29] Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

[30] Jaschik, S. U.S. Issues Final Version of 'Common Rule' on Research Involving Humans. Inside Higher Ed, January 19, 2017. Available at http://insidehighered.com. National Academy of Sciences. Optimizing the Nation's Investment in Academic Research: A New Regulatory Framework for the 21st Century. Available at http://www.nap.edu/download/21824

[31] https://www.23andme.com/about/consent/

[32] Ponemon Institute/IBM, Sixth Annual Benchmark Study on Privacy & Security of Healthcare Data. May 2016. http://www2.idexpertscorp.com/ponemon2016.

[33] http://fortune.com/2016/12/13/quest-diagnostics-data-breach-health/

[34] Shringarpure, S. S., & Bustamante, C. D. (2015). Privacy risks from genomic data-sharing beacons. *The American Journal of Human Genetics*, *97*(5), 631-646.

[35] Heatherly, R. (2016). Privacy and Security within Biobanking: The Role of Information Technology. *The Journal of Law, Medicine & Ethics*, *44*(1), 156-160.

[36] Ponemon Institute, Sixth Annual Benchmark Study on Privacy & Security of Healthcare Data. May 2016. http://www2.idexpertscorp.com/ponemon2016.

[37] Thomson, L. 2015. Health Care Data Breaches and Information Security: Addressing Threats and Risks to Patient Data. Chapter 15 in Peabody, A. (Ed.). (2013). *Health Care IT: The essential lawyer's guide to health care information technology and the law*. American Bar Association. Available at http://www.americanbar.org/content/dam/aba/publications/books/healthcare_data_breaches.authcheckdam.pdf

[38] Eveleth, R. 2015. Genetic Testing and Tribal Identity. *The Atlantic*, January 26. Also see Havasupai Tribe and the Lawsuit Settlement Aftermath. http://genetics.ncai.org/case-study/havasupai-Tribe.cfm

[39] Berger, K. M., & Roderick, J. (2014). National and transnational security implications of Big Data in the life sciences. *Washington, DC: American Association for the Advancement of Science*. http://www.aaas.org/sites/default/files/AAAS-FBI-UNICRI_Big_Data_Report_111014.pdf

[40] Trevino, L. K., & Nelson, K. A. (2010). *Managing business ethics*. John Wiley & Sons. D. Palmer, K. Smith-Crowe and R. Greenwood (2016), *Organizational Wrongdoing: Key Perspectives and New Directions* (Cambridge, UK: Cambridge University Press), including chapters 5 ("Bad Apples, Bad Barrels and Bad Cellars: A 'Boundaries' Perspective on Professional Misconduct.") and 7 ("She blinded me with Science: The Sociology of Scientific Misconduct." Leveson, N., Dulac, N., Marais, N., Carroll, J. 2009. Moving Beyond Normal Accidents and High Reliability Organizations: A Systems Approach to Safety in Complex Systems. *Organization Studies* Vol 30, Issue 2-3, pp. 227 – 249. Roberts, K. H., Bea, R., & Bartles, D. L. (2001). Must accidents happen? Lessons from high-reliability organizations. *The Academy of Management Executive*, *15*(3), 70-78. Kochan, T. A., Smith, M., Wells, J. C., & Rebitzer, J. B. (1994). Human resource strategies and contingent workers: The case of safety and health in the petrochemical industry. *Human Resource Management*, *33*(1), 55-77.

[41] There is increasing variation across states in the extent to which genomics is regulated in research, clinical, and commercial settings. See https://www.genome.gov/policyethics/legdatabase

[42] http://www.humanlongevity.com/human-longevity-inc-and-discovery-ltd-to-offer-whole-exome-whole-genome-and-cancer-genome-sequencing-to-discovery-insurance-clients-in-south-africa-and-the-united-kingdom/

[43] https://www.dnasimple.org

[44] Thomson, L. L. (2015). Personal Data for Sale in Bankruptcy. *American Bankruptcy Institute Journal*, *34*(6), 32.

[45]Acquisti, A., C. Taylor and L. Wagman. 2016. The Economics of Privacy. Journal of Economic Literature, Vol. 52, No. 2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2580411##

[46] American Association for Cancer Research. http://www.aacr.org/RESEARCH/RESEARCH/PAGES/AACR-PROJECT-GENIE-DATA.ASPX

[47] Compliance Statistics for Policies that Govern Data Submission, Access and Use of Genomic Data. NIH Genomic Data Sharing (GDS). Available at https://gds.nih.gov/20ComplianceStatistics_dbGap.html

[48] Devos, L., T. Wang and S. Iyer. 2016. The Genomic In ection Point: Implication for Healthcare. Rock Health Special Topics Report. https://rockhealth.com/reports/the-genomics-inflection-point-implications-for-healthcare/

[49]Skloot, Rebecca (2010). The Immortal Life of Henrietta Lacks, New York City: Random House.

[50]See https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative and https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative

[51] Conley, J. 2017. Williams and Beyond: Legal and Policy Issues in the Regulation of Genetic Testing. UNC Center for Genomics and Society. Presentation, February 9, 2017. http://www.genomicslawreport.com.

[52] Cyber Security: Confronting the Threat. 2015. Accenture Consulting. https://www.accenture.com/_acnmedia/Accenture/next-gen/top-ten-challenges/challenge9/pdfs/Accenture-2016-Top-10-Challenges-09-Cyber-Security.pdf