BIOPRESERVATION AND BIOBANKING
Volume 15, Number 6, 2017
Mary Ann Liebert, Inc.
DOI: 10.1089/bio.2017.0033

How Sensitive Is Genetic Data?

Murat Sariyar,[1] Stephanie Suhr,[2] and Irene Schlünder[3,4]

The rising demand to use genetic data for research goes hand in hand with an increased awareness of privacy issues related to its use. Using human genetic data in a legally compliant way requires an examination of the legal basis as well as an assessment of potential disclosure risks. Focusing on the relevant legal framework in the European Union, we discuss open questions and uncertainties around the handling of genetic data in research, which can result in the introduction of unnecessary hurdles for data sharing. First, we discuss defining features and relative disclosure risks of some DNA-related biomarkers, distinguishing between the risk for disclosure of (1) the identity of an individual, (2) information about an individual's health and behavior, including previously unknown phenotypes, and (3) information about an individual's blood relatives. Second, we discuss the European legal framework applicable to the use of DNA-related biomarkers in research, the implications of including both inherited and acquired traits in the legal definition, as well as the issue of ''genetic exceptionalism''—the notion that genetic information has inherent characteristics that require different considerations than other health and medical information. Finally, by mapping the legal to specific technical definitions, we draw some initial conclusions concerning how sensitive different types of ''genetic data'' may actually be. We argue that whole genome sequences may justifiably be considered ''exceptional'' and require special protection, whereas other genetic data that do not fulfill the same criteria should be treated in a similar manner to other clinical data. This kind of differentiation should be reflected by the law and/or other governance frameworks as well as agreed Codes of Conduct when using the term ''genetic data.''

Keywords: data protection, genetic data, genomic data, sensitive data

Introduction

The rising demand for using genetic data in different contexts, such as healthcare and biomedical research, where it is used for precision medicine or the discovery of novel genome–phenome associations, goes hand in hand with an increased awareness—of researchers as well as research participants and the public—concerning privacy issues related to its use.[1–9] To ensure human genetic data are provided or used by individual researchers or research consortia in a legally compliant way requires an examination of the legal basis for using genetic data, the meaning accorded to the term genetic data in the legal sphere (including rights and obligations related to it), and an assessment of disclosure risks and data security mechanisms.

To understand the use and meaning of the term genetic data, it is essential to note that the term is frequently used synonymous to genomic data, which often results in a primary focus on whole genome sequences (WGSs) when discussing and assessing risks. In practice, however, there are many other data types that would fall under genetic data, such as for example single nucleotide polymorphism (SNP) data, that are relevant for clinical decisions and research.[10] In the European Union (EU), the term genetic data is associated with uncertainty contributed by the recently adopted EU General Data Protection Regulation (GDPR). The GDPR frequently refers to the term genetic data throughout its recitals, and—possibly more importantly—genetic data are included in the list of sensitive data in Article 9, without any kind of differentiation. Listing genetic data with other sensitive data—that is, data that may be used for example to discriminate against specific individuals based on certain characteristics—indicates that the legislator tries to reflect the increasing use of genetic data (not only for research purposes) and the growing concerns around such use.

The essential challenge with genetic data is the type of information it may contain, such as ethnicity and health-relevant information, together with their reidentification potential. Generally, genetic data can be globally unique—and thus used as an identifier—as soon as data are linkable to a

[1]Institute of Medical Informatics, Bern University of Applied Sciences, Bienne, Switzerland.
[2]Independent Researcher, Cambridge, United Kingdom.
[3]TMF–Technologie- und Methodenplattform e.V., Berlin, Germany.
[4]BBMRI-ERIC, Graz, Austria.

specific person. The fact that the disclosure of genetic data might not only affect a single individual but also the individual's relatives renders the problem of privacy even more complex. Maximum risks to the privacy of individuals or data subjects can arise in the context of open science or open data sharing, where different kinds of genetic data are made available to a large number of users for a variety of purposes. It is a major challenge for databases and service providers dedicated to open science to develop regulatory frameworks that reconcile the needs of easy access to important research resources to the widest possible number of researchers with the protection of research participants' privacy and the right to self-determination.

An additional emerging challenge is faced by sample collections or data resources that are actively used in research today and which hold supposedly anonymized samples and/or data that have been acquired in the past, before the introduction of specific legislation and widely accepted best practices such as, for example, around participant consent. These so called legacy data have been published in the past and are often available without restrictions. Given their huge value for scientific research, it would be inappropriate and even damaging to research not to continue offering and using this valuable resource. However, in light of technical developments and changing legislation, there are questions around how to deal with these data and/or samples as some of the research participants have never given consent for the many different and previously unforeseen uses possible today.

It is a challenge, for example in concrete terms of technical implementation, to ensure research participants' control over their genetic data. While attitudes toward participants' rights have fundamentally changed over time (see e.g., the HeLa case),[11–13] disclosure techniques as well as the availability of context information have also increased immensely over even just the past few years. As a consequence, the concept of anonymization of biomaterial and genetic data to meet privacy constraints—which is required under certain jurisdictions—has become more and more challenging. Even when genetic data are disseminated in an aggregated form, large amounts of information can still be deduced, in response to which the NIH and Wellcome trust, for example, removed genomic summaries from public databases.[14–17] Therefore, most institutions now rely on informed consent and pseudonymization instead of anonymization when capturing and storing material and data, as deletion of the link to specific research participants removes the possibility of contacting them again in case this should become necessary.[18,19]

The open questions around the handling of genetic data in research can lead to unnecessary hurdles for data sharing. To resolve them it is necessary to take a deeper look into specific privacy implications of different types of data that fall into this category, and whether and how the types of genetic data in question differ from other types of clinical data. In this study, we first discuss defining features and disclosure risks of some genetic data types, distinguishing between disclosure of (1) the identity of an individual, (2) information about an individual's health and behavior, including previously unknown phenotypes (attribute disclosure), and (3) information about an individual's blood relatives (group disclosure). Second, we discuss the European legal framework applicable to the use of genetic data in research. Finally, we map the legal to the technical definition and draw some preliminary conclusions concerning how sensitive and/or exceptional genetic data are.

Genetic Data and Associated Privacy Risks

Genetic data are especially relevant for healthcare and biomedical research when they indicate an increased or decreased likelihood of developing certain diseases. While there are many forms of DNA-related markers that are relevant for certain phenotypes and identification of individuals, generally only a tiny fraction of DNA is relevant for healthcare and research. We will focus here on SNPs, short tandem repeats (STRs), copy number variation (CNV), CpG methylation, and WGSs. We have chosen these important markers because they are directly related to the DNA without any translation step in between (in contrast to, for example, RNA variants), which allows us to focus on the more stable characteristics at the molecular level. As a primer on basic genetics terminology and concepts for those readers unfamiliar with that subject, we recommend the following introductions: https://en.wikiversity.org/wiki/Genetics/Introduction and https://ghr.nlm.nih.gov/primer

Many variations have no or unknown effects (i.e., they are silent variations). SNPs, STRs, and CNVs are important representatives of relevant DNA sequences that are not genes. While SNPs are the smallest possible variation in the sequence (e.g., changing A to C in position X of the DNA sequence), the other two DNA-related marker types cover variations in the number of repetitions of certain sequences. Although CNVs include STRs, we discuss them separately: STRs are highly relevant in their own right because they are frequently used for genetic fingerprinting of individuals, which is an important privacy issue that may be overlooked if CNVs are examined in a more general manner. CpG methylation is an epigenetic phenomenon that affects phenotypes without replacing specific bases in the DNA sequence. There is no hierarchy with respect to the amount of information contained in these markers, other than that a WGS covers all of this information.

We will discuss the following characteristics of these genetic markers:

- Their defining features,
- How they are used in healthcare and biomedical research, and
- Associated disclosure risks.

Single nucleotide polymorphisms

The human genome contains about 6,190,000,000 nucleotides (A, C, G, and T) and ~3,000,000 nucleotide variations. Defined as a variation of a single nucleotide that occurs to a significant degree within a population, SNPs frequently act as markers for specific genes. These variations are alleles for a specific base position and can, for example, have effects on the protein or small interfering RNA made by that coding region. Since SNPs are mostly markers for genes, which themselves are markers for phenotypes, SNPs are frequently used in combination with other gene-related information in healthcare. One example for using SNPs in healthcare as part of a diagnostic device is the Food and Drug Administration (FDA)-approved *in vitro* diagnostic device BRACAnalysis CDx™ used for the companion drug

Lynparza™ (olaparib).[20,21] The intention is to improve treatment of ovarian, fallopian tube, and primary peritoneal cancer in women who carry mutations in *BRCA1* or *BRCA2*, and who have already received three or more chemotherapy treatments. SNPs and small insertions and deletions (indels) are measured together with larger deletions and duplications in *BRCA1* and *BRCA2*, detected using multiplex PCR. Hence, the number of SNPs being tested is neither fixed nor sufficient to reach any conclusions. The test is intended to be performed on specific serial number-controlled instruments at one single laboratory (Myriad Genetic Laboratories, Inc.).

In contrast to the healthcare setting, there are many cases where SNPs are used as sole genetic explanation factors in biomedical research. Especially genome-wide association studies (GWAS) are conducted to find relevant associations between SNPs and certain diseases. We will not refer to such studies here, but emphasize that the main aim of GWAS is to produce validated SNP panels that discriminate between cases and controls. Thus, the results of GWAS are usually groups of 10 and 500 relevant SNPs that can be used in clinical practice to specify diagnoses. Due to the limitations of GWAS and the decreasing price of whole genome sequencing, studies based on SNPs seem to be less attractive than previously.[22]

Main disclosure risks of SNPs are related to reidentification. For example, Pakstis et al. showed that a carefully chosen set of 45 SNPs is often sufficient to single out a specific individual with a type 1 error of 10^{-5}.[23] Attribute disclosure—that is, inferring sensitive information from SNP data in addition to information related to specific questions—is less a problem, as 10–500 SNPs cover only a very tiny fraction of human DNA. In contrast, membership disclosure risks can be relevant in the case of biomedical research when the investigated SNPs allow, for example, the inference that someone belongs to a specific group even without allowing reidentification of an individual. This is especially a problem when diseases such as HIV, which have significant stigmatization potential, are the focus of the research conducted.

Short tandem repeats

STRs or microsatellites are adjacent repetitions (5–50 times) of DNA motifs (short, recurring patterns in DNA that are presumed to have a biological function www.nature.com/nbt/journal/v24/n4/full/nbt0406-423.html), consisting of 2–5 base pairs. Similar to SNPs, they can indicate the location of a gene or a mutation that is responsible for a disease, and they are distributed throughout the genome in both noncoding and coding areas.[24] We are not aware of any approved diagnostic tool based on STRs or any other usage of STRs in healthcare. However, STRs (especially Y-chromosome STRs) are used for genetic fingerprinting of individuals, especially in forensics, and there is research into STR mutations that can lead to phenotypic changes and diseases.[25] Huntington's disease, spinobulbar muscular atrophy, and fragile X syndrome are diseases in which STRs are deemed to play an important role.[26] Disclosure risks are similar to those of SNPs.

Copy number variations

CNV extends STR by including long repeats (e.g., entire genes) and repeats that are not adjacent, that is, interspersed throughout the genome. One example of a whole gene repeat is the alpha-amylase 1 gene (*AMY1*), which has diet-related functions. The main difference to STRs with respect to applications is the fact that CNVs can contain many genes as well as affecting genes and gene products more directly. Even if there is no direct application in healthcare, there is an increasing body of research on long repeats in the context of disorders such as Danon disease, ventricular tachycardia, hypertension susceptibility, or abnormal arteriogenesis.[27]

With respect to disclosure risks, longer repeats lead to an increased risk of inference of additional sensitive attributes compared with STRs. Reidentification, group and membership disclosure risks depend on the size and number of the repeats, without there being a linear relationship between size and risk (greater size does not necessarily mean greater risk).

DNA methylation

In addition to the nucleotide information in DNA, methylation of DNA is another essential element shaping our phenotype. DNA methylation is different from mutations (i.e., alteration of the nucleotide sequence), since it does not change the nucleotide sequence. It represents an epigenetic phenomenon that is subject to possibly frequent changes within an individual lifespan and influenced by certain external factors. CpG methylation is the most common methylation type: 60%–90% of all Cytosines in CpGs are methylated. It is usually measured by (whole) genome bisulfite sequencing.[28] High CpG methylation in promoter regions is frequently associated with reduced gene expression, whereas methylation of the gene body generally is not.[29] In cancer cells for example, there can be a loss of methylation in repetitive sequences within gene bodies. We are not aware of any diagnostic usage of CpG methylation patterns in healthcare. However, there are FDA and European Medicines Agency (EMA)-approved hypomethylating agents (inhibiting the maintenance of DNA methylation) for the treatment of myelodysplastic syndrome and acute myeloid leukemia.[30,31] In biomedical research, methylation patterns for different diseases, especially different types of cancer, are searched for and investigated. For example, the ''CpG island methylator phenotype'' (CIMP) denotes an aberrant methylation pattern in the promoter region compared with normal cells.[29,32] These are found for example in colon cancer as well as breast cancer; however, there is still a lack in validated CIMPs.

Some risk characteristics of CpG methylation differ from those of mutation-based markers. First, reidentification risks based on methylation patterns cannot be estimated easily, as methylation changes within an individual lifespan are much more likely than mutations. Second, because of this dynamic aspect of CpG methylation, it is also difficult to associate certain patterns with groups of individuals (e.g., relatives). Third, as methylation reflects exposure to different environmental factors, the attribute disclosure risk can be increased compared with other DNA-related markers as the methylation patterns can reveal many behavior-and exposition-related characteristics.

Whole genome sequences

In the strict sense, a WGS entails the entirety of an organism's chromosomal DNA as well as DNA contained in the mitochondria. Thus, WGSs cover a lot of information that can guide precision medicine on the level of diagnosis

and therapy as well as population screening for different kinds of diseases (e.g., by preconception carrier screening for autosomal recessive disorders). In healthcare settings, gene panels (i.e., lists of specific genes investigated at the same time), rather than WGSs, are used to guide decisions. Examples for approved diagnostic tools based on genes are the cobas® *KRAS* Mutation Test for the use of Cetuximab for treatment of colorectal cancer patients and the cobas 4800 *BRAF* V600 Mutation Test for the use of Vemurafenib for treatment of melanoma patients.[33,34] In research, WGSs (for our purposes here identical to whole exome sequences, i.e., sequences of all genes) are generated to allow the discovery of new gene–phenotype associations.[35] Because of this, the term incidental findings is mostly inadequate in the context of WGSs in research. WGSs are rather the basis for nonhypothesis-driven research in genetics, which is associated with an openness for the associations that might be found.[36,37]

Compared with the other types of DNA-related markers discussed so far, WGSs have a maximum inherent risk of disclosing information. Consequently, DNA profiling aims at identifying individuals based on characteristics of their DNA, gains a lot by WGSs, as they increase the validity that the genetic material came from specific individuals or groups. Compared with CpG methylation, WGSs can infer many new gene–phenotype associations, and the attributes that can be inferred are less associated with concrete behavior or exposure to environmental factors of an individual. In addition, WGSs are relatively stable, although their variable parts (i.e., mutations) are of primary interest in research and healthcare.

Legal Requirements for Genetic Data

In this study, we focus on one specific aspect of the sensitivity of genetic data: legal concerns related to the disclosure of sensitive information contained in DNA that may require different handling than other types of data associated with an individual (other issues, for example related to the problems of informed consent when using genetic data, are out of our scope here). In the context of EU legislation, sensitive data have always been subject to a higher level of protection (see Article 8 of the Data Protection Directive): they are special categories of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, and data concerning health or sex life. In addition, Art. 9 of the GDPR now expressly refers to genetic data among other sensitive data as special category of data. The assumption that genetic information has specific characteristics that require different handling and risk assessment compared with other types of medical information is known as genetic exceptionalism. Our main contribution in this debate is the differentiation between specific types of genetic data when making decisions on exceptionality (and, consequently, any specific treatment of such data in terms of data protection). One central reason for assuming such a position is the uncertainty about the information that can be inferred from genetic data. Besides the risk of loss of anonymity for specific individuals, certain data may also be used for different forms of discrimination, which may even affect relatives. Thus, if genetic data are deemed exceptional, establishing fit for purpose policies, governance structures, terms of use and, where necessary, new norms, are essential for protecting privacy. One concrete example for a national approach on this topic is the Genetic Information Nondiscrimination Act of 2008 for protecting genetic privacy, which prohibits the use of genetic information in health insurance and employment in the United States.

Legal definition of genetic data

With the GDPR, EU data protection law provides for a definition of genetic data. However, when examined in more detail, the definition raises some questions instead of offering clarity since the law unambiguously states that the special provisions for the handling of genetic data under the GDPR are not limited to inherited genetic characteristics (traits that are passed from parents to their children, such as eye color, ethnic origin, or the predisposition for certain diseases), but are also applicable to acquired genetic characteristics (genetic characteristics that are caused by mutations that occur during an individual's lifespan—such mutations can, for example, lead to different types of cancer).

The EU GDPR states:

> *'genetic data' means personal data relating to the inherited or acquired genetic characteristics of a natural person which give unique information about the physiology or the health of that natural person and which result, in particular, from an analysis of a biological sample from the natural person in question* (Art. 4 para 13 GDPR)

Before assessing whether any or all of the DNA-related markers discussed here fall under this definition, it is helpful to analyze the main characteristics it contains. In many earlier definitions of genetic data, heritability is a core element:

- *All data of whatever type concerning the hereditary characteristics of an individual or concerning the pattern of inheritance of such characteristics within a related group of individuals* (Council of Europe Recommendation No. R(97)5)
- *Any data concerning the hereditary characteristics of an individual or group of related individuals* (Art 2 (g) of the 2 August 2002 law of Luxembourg on the protection of persons with regard to the processing of personal data)
- *Information about heritable characteristics of individuals obtained by analysis of nucleic acids or by other scientific analysis* (International Declaration on Human Genetic data, UNESCO).

While these definitions emphasize hereditary characteristics, the GDPR—referring explicitly to inherited or acquired genetic characteristics—seems to take a broader approach. Even epigenetic markers such as CpG methylation could conceivably fall under this definition, as long as they are considered to be a genetic characteristic of a natural person. The question of which criteria the ''acquired genetic characteristics'' must fulfill to fall under the definition of genetic data remains unanswered.

The criterion of ''unique information'' could be interpreted to mean that the information about the physiology or health of an individual must be unambiguous, but a more convincing interpretation is to assume that data must be rich enough to single out an individual—that is, the data in question must be so rare in a population that it is possible to assign it to a single person, as is the case for WGSs. This is a feature that is related to many other kinds of data, for example social security numbers or the globally unique

combination of attributes. The dependence of singling out possibilities on the context and governance structures is part of the discussion that is happening at an institutional level (e.g., the Wellcome Trust policy changes with respect to genetic data mentioned above), the national level (e.g., Estonia using Blockchain technology for healthcare data security), and the supranational level (e.g., the results of the Working group article 29 at the EU level).

An analysis of the other elements of the definition of genetic data in the GDPR (''information about the physiology or health'' and ''in particular result from the analysis of a biological sample'') does not provide much help to understand the element genetic characteristics, since they are intended to establish additional general requirements. For example, ''providing information about the physiology'' just makes clear that the origin of the genetic data must be the body of a human being.

A systematic assessment of whether the DNA-related markers discussed in this study fall under the EU GDPR definition for genetic data based on using the four main characteristics of the definition reveals that WGSs possess all characteristics of the definition, and are therefore definitively covered (Table 1). For all other markers, it *is unclear under which circumstances* they might fall under the definition as, for example, unique information (related to the disclosure risk) and information about health characteristics (related to attribute disclosure) depend on the number, the position, and the stability of the markers under consideration. Hence, all of these non-WGS genetic data should be treated in the same way, that is, they should be part of case-by-case evaluations that compute reidentification and discrimination risks.

One disclosure risk that is only implicitly covered by the elements of the definition is group disclosure, which is related to an inherited genetic characteristic exhibiting information about the physiology or health of an individual. For this disclosure risk, the same can be stated as for attribute and disclosure risks. Decisions for these markers require statistical knowledge and representative samples. It might be useful to categorize genetic datasets according to their inherent risks (attribute, group, and identity disclosure), which would require an assessment of the amount of information contained in them, for example, based on an entropy-related index.

Discussion

The issues discussed here raise questions regarding the utility of attempts to define genetic data in a legal context. The list of sensitive data in Article 9 GDPR does not only include genetic data but also health data, some of which were already classified as a special data category under the Data Protection Directive of 1995. In some countries, such as Germany for example, genetic data were already considered to be health data as long as they have an impact on the health status of an individual. From this perspective, the term genetic data adds nothing to the list of sensitive data in Article 9 GDPR. Even Article 9 para 4 only refers to both genetic data and health data. Hence, the aim of expressly referring to genetic data seems to be to emphasize its increased importance in healthcare and biomedical research while avoiding the addition of new requirements concerning legal compliance compared with other sensitive data. We understand the necessity to raise the level of protection in particular for WGS data (and to some extent whole DNA methylation data), as this is more stable and implies many more potential phenotype inferences than nongenetic data, and suspect that this may be the main aim and motivation behind the specific addition of the term genetic data in the law. However, without further clarification on how the definition may be interpreted in practice—for example through agreed codes of conduct as endorsed by the GDPR and informed discussions on the technical implementation of such protections that take the needs of scientific research into account—the definition will likely become problematic to work with.

What could such a higher level of protection look like in terms of technical implementation? As discussed above,

Table 1. Mapping of DNA-Related Markers to the Relevant Elements of the European Union General Data Protection Regulation Definition of Genetic Data

	Elements of EU GDPR definition			
Marker	Inherited or acquired genetic characteristics (in the sense of human DNA-related marker)	Unique information (disclosure of the identity)	Information about the physiology or health (attribute disclosure)	In particular result from the analysis of a biological sample
SNPs	Yes	Depends on the number and position in the DNA	Depends on the number and position in the DNA	Yes, but can also be derived from WGS
Short tandem repeats	Yes	Depends on the number and position in the DNA	Depends on the number and position in the DNA	Yes, but can also be derived from WGS
Copy number variations	Yes	Depends on the number and position in the DNA	Depends on the number and position in the DNA	Yes, but can also be derived from WGS
CpG methylation	Yes	Depends on the number, position in the DNA, and stability	Depends on the number, position in the DNA, and stability	Yes (typically derived from bisulfite sequencing)
WGSs	Yes	Yes	Yes	Yes

EU, European Union; GDPR, General Data Protection Regulation; SNP, single nucleotide polymorphism; WGS, whole genome sequence.

anonymization should not be the preferred solution because it involves the removal of information from data that might be needed for certain research purposes, especially if the research purpose is not clear before the data is anonymized. Therefore, technical alternatives are necessary: for example, fully homomorphic encryption, secure multipart computation, but also tight access policies and elaborated security mechanisms (regarding authentication, authorization, etc.). Besides this, the involvement of advisory boards, new types of codes of conducts—as mentioned above, new data governance models (such as data cooperatives, enabling continuous involvement of the individuals whose data are at stake), and a hierarchy of informed consent principles (the higher the amount of genetic data to be used, the higher the involvement of the affected individuals, including family members) are important measures for achieving higher protection levels.[38,39] Codes of Conduct should cover such measures by way of listing principles and examples how these can be fulfilled.

It is interesting to note that the revised Common Rule in the United States will require that subjects be informed specifically if specimens collected during the research will undergo WGS. In addition, it also introduces a process whereby the U.S. Department of Health and Human Services will evaluate on a regular basis whether the definition of identifiability should be modified or whether certain technologies should be placed on a list of technologies that render data inherently identifiable. Such processes offer increased transparency and may increase the motivation to donate samples and data for genetic research. However, as they are highly abstract, they can only serve as guiding principles. In the European context, such principles could be evaluated to support the implementation of the GDPR requirements, but they also should be extended by protection measures mentioned above.

We argue that defining the sensitivity of genetic data based on their related risks would be useful. For a risk assessment, stability and distinguishability of markers as inherent characteristics must be considered, together with the risk of these markers of being available outside of the specific context for which they were collected. WGS data and those data that are based on statistical assessments—are deemed to be of a similarly sensitive nature (e.g., a set of CNVs covering all genes) might need a higher level of protection than other sensitive data. Regarding further risks related to genetic data, the following considerations of the Art. 29 WP are interesting: taking into account the developments in research, genetic data may in future reveal more information and be used by an ever increasing number of agencies for various purposes; often unknown to the bearer him/herself. This points to an increasing public sensitivity that future uses of ''genetic data'' and related risks for the affected individuals should not be overlooked.[40] Once the full genome of an individual has become public, it can be used in a variety of settings for unforeseeable and discriminatory purposes. In the worst case, such illicit use could remain undetected or at least unknown to the affected individual, since the means to interpret the data are normally not accessible to the average citizen.

In the Legal Requirements for Genetic Data section, we stated that non-WGS data should be treated in the same way when calculating privacy risks associated with them. This, however, does not mean that there are no relevant differences between them. SNPs are mostly useful when combined with other gene-related information, which make them in their own less critical than other forms of genetic data. This is one reason for having public-domain databases such as *dbSNP*. STRs are relevant for genetic fingerprinting, especially in forensics. As they are less relevant for inferring further information, it is just important to prevent reidentification by STRs without losing information of other genetic data. CNVs can contain many genes and are therefore useful for many purposes. In this case, the assessment of disclosure risks will probably reveal more critical aspects than SNPs and STRs. Finally, DNA methylation as an epigenetic phenomenon is less stable than the other kinds of genetic features. On the hand one, this makes such data less vulnerable to reidentification attacks; on the other hand, they can reveal many behavior- and exposition-related characteristics. Hence, the primary goal with respect to DNA methylation data should be the prevention of attribute disclosure.

The main argument against genetic exceptionalism is that the characteristics of genetic data are similar to those of other medical or health-related information (e.g., it can be used to disclose an individual's identity, disease risks, or drug response) or are not important in terms of privacy risks (e.g., disclosure of information on features and risks of blood relatives based on low penetrance genes).[41,42] As our considerations above imply, in the case of individual genes it is indeed plausible to argue against genetic exceptionalism: other medical information, for example, history of previous hospitalizations due to immunodeficiency or psychotic episodes, can be used to extract the same information. However, the situation is different when considering whole genomes as these allow the deduction of many unknown future phenotype associations, which might further the risk of stigmatizations. Hence, the sheer amount of potential information contained in genomes makes a qualitative difference, although, as for many quantity-to-quality transformations, it is not possible to give a concrete number of genome subsets (e.g., genes) that result in this difference. This may in fact be one reason why the distinction of different genetic data with respect to a potential privacy breach is often omitted.

Conclusion

The definition of genetic data in the GDPR does not favor any form of genetic exceptionalism, meaning they are equally sensitive as biometric or clinical data. For SNPs, CNVs, DNA methylation, and STR, this perspective seems plausible. However, WGS data cover much latent information, which—in most cases—cannot be captured by a manageable amount of other medical information. In addition to this, the potential information richness of WGS data is maintained or even increases over longer time periods as new technologies and scientific methodologies to analyze and interpret them develop. Hence, we see a necessity to incorporate the need for a higher level of protection for such exceptional data. The requirements resulting from such protection should be based on the amount of information, the stability of this information, potential risks, and the probability of risk realizations.

One approach that seems to gain momentum (see, e.g., the dbGaP database) is to strictly control the availability of certain types of genetic data, especially genomic data. Importantly, this need for control should not be translated into making the data unavailable or difficult to obtain for

research purposes in general, but rather into mechanisms that ensure they are only used for research in line with the purposes for which the data were collected and, if applicable, consented. Inevitably, this leads to some constraints in how data can be shared; however, such constraints should not be used to justify the avoidance of data sharing based on other motives, such as monopolizing them for exclusive scientific use by a small group of researchers.

To enable such fair (rather than open) sharing requires transparent governance structures that support the widest possible sharing of data for research while providing the best possible protection of research participants. Core elements of such governance structures must be data access committees or their equivalents—acting based on transparent sharing policies—and ethics committees and institutional review boards, which are well established internationally and control research involving humans and human data. To date, there is no widely accepted single governance model, nor are transparent sharing policies available everywhere. It seems to be a major challenge as well as a pressing need that global governance standards are developed to foster data sharing in research—in particular also across international borders—while treating research participants' interests in protecting their data against abuse with the greatest possible respect.

It is important to reiterate that—just as with any other legal requirements—there can never be absolute protection against data leakage or misuse. Data protection law may always be incomplete or even insufficient to ensure that research participants are protected against unfair treatment on the basis of their sensitive genetic information. Consequently, the legislator will need to address the potential for abuse of genetic data that may affect life opportunities for the individuals in question. Maybe it is time to discuss new methods such as regulating the use of certain data once they have been leaked. As John Wilbanks said: ''harm is not the act … of distributing data. Harm comes from actions that are taken once the data has been distributed.''[43]

Author Disclosure Statement

No conflicting financial interests exist.

References

1. Naveed M, Ayday E, Clayton EW, et al. Privacy in the genomic era. ACM Comput Surv 2015;48:6.
2. Ayday E, Raisaro JL, McLaren, et al. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. In: *Proceedings of the USENIX Conference on Safety, Security, Privacy and Interoperability of Health Information Technologies.* Washington, DC. USENIX Association; 2013.
3. Ayday E, Cristofaro ED, Hubaux JP, et al. The chills and thrills of whole genome sequencing. EPFL-REPORT 2013: 186866. Berlin. ACM; 2013.
4. Knoppers BM, Harris JR, Tasse AM, et al. Towards a data sharing Code of Conduct for international genomic research. Genome Med 2011;3:46.
5. Kaye J. The tension between data sharing and the protection of privacy in genomics research. Annu Rev Genomics Hum Genet 2012;13:415–431.
6. Gymrek M, McGuire AL, Golan D, et al. Identifying personal genomes by surname inference. Science 2013;339: 321–324.
7. Lin Z, Owen AB, Altman RB. Genomic research and human subject privacy. Science 2004;305:183–183.
8. Heatherly RD, Loukides G, Denny JC, et al. Enabling genomic-phenomic association discovery without sacrificing anonymity. PLoS One 2013;8:e53875.
9. Budin-Ljøsne I, Isaeva J, Knoppers BM, et al. Data sharing in large research consortia: Experiences and recommendations from ENGAGE. Eur J Hum Genet 2014;22:317–321.
10. Goh CL, Saunders EJ, Leongamornlert DA, et al. Clinical implications of family history of prostate cancer and genetic risk single nucleotide polymorphism (SNP) profiles in an active surveillance cohort. BJU Int 2013;112:666–673.
11. Callaway E. Deal done over HeLa cell line. Nature 2013; 500:132–133.
12. Culliton BJ. HeLa Cells: Contaminating cultures around the world. Science 1974;84:1058–1059.
13. Humbert M, Ayday E, Hubaux JP, et al. Addressing the concerns of the lacks family: quantification of kin genomic privacy. In: *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security.* 2013: 1141–1152.
14. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 2008;4:1–9.
15. Clayton EW, Smith M, Fullerton SM, et al. Confronting real time ethical, legal, and social issues in the electronic medical records and genomics (eMERGE) Consortium. Genet Med 2010;12:616–620.
16. Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science 2008;322:44.
17. Malin B, Loukides G, Benitez K, et al. Identifiability in biobanks: Models, measures, and mitigation strategies. Hum Genet 2011;130:383–392.
18. Caenazzo L, Tozzo P, Pegoraro R. Biobanking research on oncological residual material: A framework between the rights of the individual and the interest of society. BMC Med Ethics 2013;14:17.
19. Elger BS, Iavindrasana J, Lo IL, et al. Strategies for health data exchange for secondary, cross-institutional clinical research. Comput Methods Programs Biomed 2010;99:230–251.
20. Tutt A, Robson M, Garber JE, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: A proof-of-concept trial. Lancet 2010;376:235–244.
21. Audeh MW, Carmichael J, Penson RT, et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: A proof-of-concept trial. Lancet 2010;376:245–251.
22. Visscher PM, Brown MA, McCarthy MI, et al. Five years of GWAS discovery. Am J Hum Genet 2012;90:7–24.
23. Pakstis AJ, Speed WC, Kidd JR, et al. Candidate SNPs for a universal individual identification panel. Hum Genet 2007; 121:305–317.
24. Wooster R, Cleton-Jansen AM, Collins N, et al. Instability of short tandem repeats (microsatellites) in human cancers. Nat Genet 1994;6:152–156.
25. Moretti TR, Baumstark AL, Defenbaugh DA, et al. Validation of short tandem repeats (STRs) for forensic usage: Performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. J Forensic Sci 2001;46:647–660.
26. Usdin K. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. Genome Res 2008;18:1011–1019.

27. Pollex RL, Hegele RA. Copy number variation in the human genome and its implications for cardiovascular disease. Circulation 2007;115:3130–3138.
28. Deng J, Shoemaker R, Xie B, et al. Targeted bisulfite sequencing reveals changes in DNA methylation associated with nuclear reprogramming. Nat Biotechnol 2009;27:353–360.
29. Ziller MJ, Hongcang G, Müller F, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature 2013;500:477–481.
30. Bejar R, Lord A, Stevenson K, et al. TET2 mutations predict response to hypomethylating agents in myelodysplastic syndrome patients. Blood 2014;24:2705–2712.
31. Issa JPJ, Kantarjian HM, Kirkpatrick P. Azacitidine. Nat Rev Drug Discov 2005;4:275–276.
32. Roessler J, Ammerpohl O, Gutwein J, et al. The CpG island methylator phenotype in breast cancer is associated with the lobular subtype. Epigenomics 2015;7:187–199.
33. Angulo B, Lopez-Rios F, Gonzalez D. A new generation of companion diagnostics: Cobas BRAF, KRAS and EGFR mutation detection tests. Expert Rev Mol Diagn 2014;14:517–524.
34. Halait H, Demartin K, Shah S, et al. Analytical performance of a real-time PCR-based assay for V600 mutations in the BRAF gene, used as the companion diagnostic test for the novel BRAF inhibitor vemurafenib in metastatic melanoma. Diagn Mol Pathol 2012;21:1–8.
35. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010;42:30–35.
36. Green RC, Berg JS, Grody WW, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 2013;15:565–574.
37. Schuol S, Schickhardt C, Wiemann S, et al. So rare we need to hunt for them: Reframing the ethical debate on incidental findings. Genome Med 2015;7:83.
38. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39:1181–1186.
39. Vayena E, Gasser U. Between openness and privacy in genomics. PLoS Med 2016;13:e1001937.
40. Wang R, Li YF, Wang, X. Learning your identity and disease from research papers: Information leaks in genome wide association study. In: *Proceedings of the 16th ACM Conference on Computer and Communications Security.* Chicago. ACM; 2009:534–544.
41. Evans JP, Burke W. Genetic exceptionalism. Too much of a good thing? Genet Med 2008;10:500–501.
42. Witt MM, Witt MP. Privacy and confidentiality measures in genetic testing and counselling: Arguing on genetic exceptionalism again? J Appl Genet 2016;57:483–485.
43. Gutmann A, Wagner JW. Found your DNA on the web: Reconciling privacy and progress. Hastings Cent Rep 2013;43:15–18.

Address correspondence to:
Murat Sariyar, PhD
Institute of Medical Informatics
Bern University of Applied Sciences
Höheweg 80
Bienne 2502
Switzerland

E-mail: murat.sariyar@bfh.ch