Review

Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review

Mara Thomas[1], PhD; Nuria Mackes[2], PhD; Asad Preuss-Dodhy[3], PhD; Thomas Wieland[4], PhD; Markus Bundschus[3], PhD

[1]F. Hoffmann-La Roche AG, Basel, Switzerland
[2]xValue GmbH, Ratingen, Germany
[3]Roche Diagnostics GmbH, Penzberg, Germany
[4]Foundation Medicine GmbH, Penzberg, Germany

Corresponding Author:
Mara Thomas, PhD
F. Hoffmann-La Roche AG
Grenzacherstrasse 124
Basel, 4070
Switzerland
Phone: 41 616881111
Email: mara.thomas@roche.com

Abstract

Background: Genetic data are widely considered inherently identifiable. However, genetic data sets come in many shapes and sizes, and the feasibility of privacy attacks depends on their specific content. Assessing the reidentification risk of genetic data is complex, yet there is a lack of guidelines or recommendations that support data processors in performing such an evaluation.

Objective: This study aims to gain a comprehensive understanding of the privacy vulnerabilities of genetic data and create a summary that can guide data processors in assessing the privacy risk of genetic data sets.

Methods: We conducted a 2-step search, in which we first identified 21 reviews published between 2017 and 2023 on the topic of genomic privacy and then analyzed all references cited in the reviews (n=1645) to identify 42 unique original research studies that demonstrate a privacy attack on genetic data. We then evaluated the type and components of genetic data exploited for these attacks as well as the effort and resources needed for their implementation and their probability of success.

Results: From our literature review, we derived 9 nonmutually exclusive features of genetic data that are both inherent to any genetic data set and informative about privacy risk: biological modality, experimental assay, data format or level of processing, germline versus somatic variation content, content of single nucleotide polymorphisms, short tandem repeats, aggregated sample measures, structural variants, and rare single nucleotide variants.

Conclusions: On the basis of our literature review, the evaluation of these 9 features covers the great majority of privacy-critical aspects of genetic data and thus provides a foundation and guidance for assessing genetic data risk.

(JMIR Bioinform Biotech 2024;5:e54332) doi: 10.2196/54332

KEYWORDS

genetic privacy; privacy; data anonymization; reidentification

Introduction

Privacy Risks of Genetic Data

Genomics is a rapidly developing field with exabytes of genetic data being generated, stored, and analyzed by public and private institutions per year. These data drive scientific progress, especially when they are shared with the scientific community or among institutions. However, genetic data can provide information about an individual's identity together with sensitive details, such as their ethnic background [1]; physical traits such as eye color [2], hair and skin color [3], height [4]; and diseases or susceptibility to diseases [5]. Therefore, even if personal identifiers (eg, name, date of birth, or others) are removed, sharing genetic data may violate the individual's right to privacy. In 2018, a seminal study demonstrated that it is possible to reidentify individuals by name from genetic data alone [6]. The authors matched genetic data of an anonymous female study

participant to the genetic genealogy database GEDmatch and identified her surname from matches with relatives who had uploaded their data on GEDmatch. Such reidentification of genetic data records using publicly available databases is highly problematic and a growing threat to privacy as publicly available genetic genealogy databases continue to grow. It is estimated that a genetic database needs to cover "only 2% of the target population to provide a third-cousin match to nearly any person" in a matching attack, similar to the one demonstrated by Erlich et al [6]. As of 2018, the probability for such a match was estimated to be 60% for the platform GEDmatch. Through similar methods of familial DNA searches, multiple individuals have been identified in criminal cases, despite never having shared their genetic data themselves [7,8]. Other attacks aim to reveal sensitive information from genetic data. In 2009, researchers discovered a genetic predisposition for Alzheimer disease in the public genome of the famous molecular biologist and Nobel laureate James Watson, although he had attempted to prevent such an attack by withholding certain parts of the data [9]. The high identifiability potential of genetic data together with its sensitive content with regard to health (eg, susceptibility to diseases such as Alzheimer disease or cancer) and physical traits (refer to the studies by Erlich and Narayanan [10], El Emam et al [11], and Mohammed Yakubu and Chen [12] for a review) has raised public concern that genetic data that are shared or published in the context of research or health care could be misused [13]. For example, attackers could exploit genetic data to obtain personal and sensitive information about individuals, and this information could be misused by insurance companies, mortgage providers, or employers to discriminate on the basis of genetic information (eg, about disease susceptibility) [14]. As an additional complication, DNA sequence is heritable; therefore, leakage of an individual's genetic data can violate the privacy of whole families [15,16].

The Challenge of Anonymizing Genetic Data

Genetic data can be used to identify individuals because each person's DNA sequence differs uniquely from the standard human reference genome. Although more than 99% of the DNA sequence is identical across all humans, the remaining <1% consists of distinct combinations of insertions, deletions, duplications, translocations, and inversions of short or long DNA fragments (refer to the study by Trost et al [17] for a review). These genetic variations are not randomly distributed across the genome but occur more frequently in specific variable regions. Some variations are rare, while others (ie, polymorphisms) are shared by a significant proportion of the population. While some variations have no observable effect, others influence gene transcription, expression, or the amino acid sequence of a protein and have an effect on the phenotype, for example, physical traits, metabolism, and disease susceptibility. These variable regions with an effect on the phenotype are of great interest to research; however, these can also be effectively used for individual identification and the inference of sensitive attributes. Even a small genetic data set of only 30 highly variable genetic loci is likely to contain unique records, and these could not only be linked to genetic records in other data sets but also provide insights into health and physical traits (refer to the studies by Erlich and Narayanan [10], El Emam et al [11], and Mohammed Yakubu and Chen [12] for a review). Furthermore, genetic variation is highly intercorrelated (variation in one genomic region correlates with variation in another) and correlated to other modalities (genetic variation is associated with transcription, expression, epigenetic regulation, etc), making it possible to link data records of the same individual even across databases that do not contain the same type of data (eg, match a genetic data sequence to a gene expression record). Anonymizing genetic data while maintaining its full utility remains an unsolved challenge, and there is no consensus on whether it is even possible [18]. Many privacy-enhancing technologies aim to reduce the information content of genetic data or restrict access to it, such that only a minimal amount of information is shared. An example is genomic beacons, which allow only simple yes or no queries to determine whether a specific variant is present in a study cohort [19]. However, it has become evident that even this limited amount of information can be exploited for privacy attacks, and few queries to genomic beacons can suffice to determine whether individuals (whose genome is known) are present in a study cohort [20-23]. Similarly, proposals for encryption and differential privacy approaches [24,25] have often been countered by demonstrations of attacks [26-28], and even synthetic genetic data may not fully protect the study participants from privacy attacks [29] (refer to the study by Mittos et al [30] for a review of privacy-enhancing technologies). Thus, even a substantial reduction in information content can often not completely eliminate all privacy risks of genetic data [31].

The Risk Minimization Approach for Genetic Data Privacy

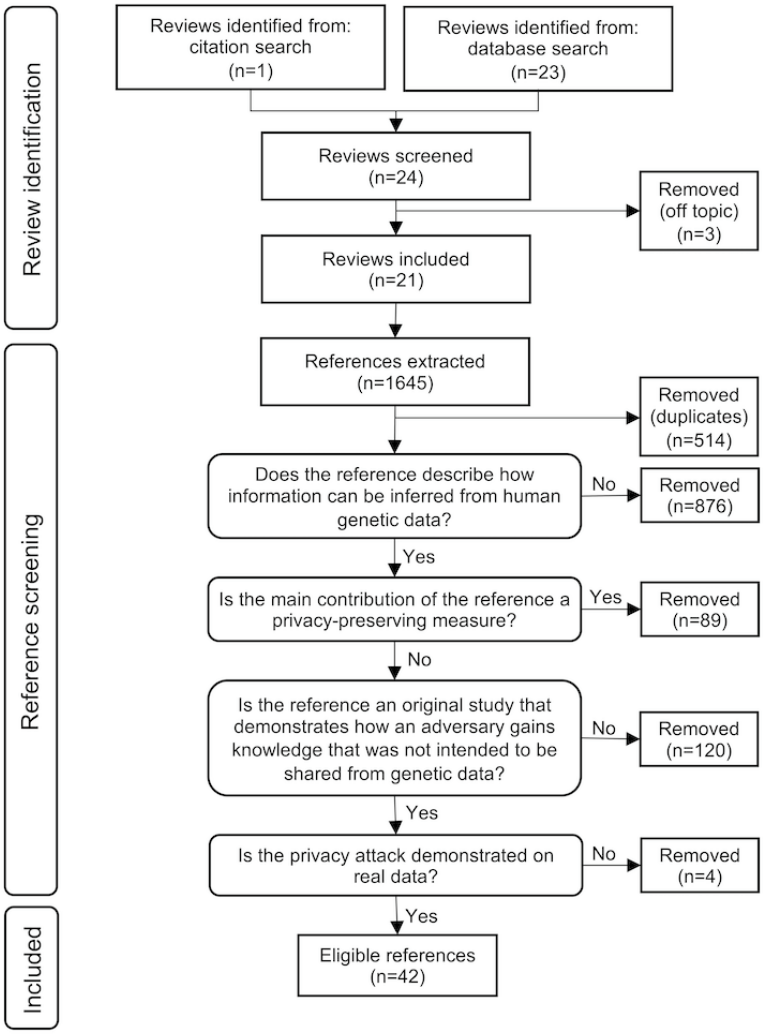
Most legislations do not require to reduce the risk of individual identification to zero, and several jurisdictions have decided to take a risk-based approach and consider genetic data anonymous if the risk of successful reidentification is below a predefined acceptable threshold [32]. Therefore, genetic data processors must find the balance between reducing information such that reidentification is no longer reasonably likely, while maintaining as much utility of the data as possible [33]. The challenge in adopting this approach lies in the correct assessment of the reidentification probability. Genetic data are complex and come in various shapes or forms, making it difficult to standardize reidentification assessments. Established methods such as assessing k-anonymity are difficult to apply to genetic data because of their high uniqueness, and many other methods fall short because of the high intercorrelation of genetic data. Simple measures such as assessing the number of single nucleotide polymorphisms (SNPs) in genetic data ignore the importance of the location of the SNPs in the genome, their frequencies in the population, and the actual feasibility of cross-linking the specific SNPs to identifiable information. For example, the reidentification risk is much higher for SNPs that are commonly included in the SNP assays used by direct-to-consumer genetic testing (DTC-GT) providers than for less frequently studied SNPs, as these are more difficult to link to publicly available identifying information. In addition, genetic data may contain SNP information even if this is not immediately evident, for example, in the raw data of sequencing-based gene expression

studies. Data processors who are not familiar with the intricacies of genetic data find little guidance on performing an assessment on genetic data that considers these factors. While several genomic privacy metrics have been proposed, the great majority focus on evaluating SNPs only [34] and neglect other known privacy-critical aspects of genetic data as well as aspects of feasibility (eg, the expertise, time, effort, availability of external resources, and other requirements required for an attack). However, the risk of severe privacy attacks on genetic data (ie, where the identity of the data subject is revealed) greatly depends on the specific content of the data as well as "soft factors," such as the availability of publicly accessible resources to cross-link and infer quasi-identifying information and the time, cost, and knowledge required to perform such an attack. Given the foundational potential of genetic data to advance research and health care, a risk-based approach that carefully evaluates the true risk of reidentification on a case-by-case basis for each data set in question is warranted, or else any type of genetic data must be considered identifiable.

Methods

To get a comprehensive overview of the types and aspects of genetic data sets that are vulnerable to reidentification attacks, as well as the methods, databases, and know-how used for these attacks, we searched for studies that demonstrate a privacy attack on genetic data. We did not aim to establish an exhaustive overview of all published privacy attacks but aimed to get a comprehensive understanding of the most vulnerable features of genetic data. Therefore, we first searched for recent reviews published on the topic of genomic privacy using ProQuest. Using the search terms (ti(*genom* OR *genetic*) AND ti(privacy OR re-identification OR reidentification OR "data security")) and (pd(>20170101)) and (at.exact("Review")), we identified 23 reviews, of which 3 (13%) were discarded because they were off topic. One additional review was identified during the literature research and added to the selection (refer to Multimedia Appendix 1 [35-55] for an overview of the included and excluded reviews), resulting in a final sample of 21 reviews. In a second step, we extracted all references cited in the reviews (n=1645) and identified all original research studies that demonstrate a privacy attack on genetic data. After the removal of 514 duplicates and 876 reference studies that did not contain any description of information inference from human genetic data, we first excluded 89 studies whose main contribution was the presentation of privacy-preserving measures to exclude privacy attacks that were performed only for the purpose of proving the efficiency of the proposed counter methods. Next, we excluded 120 studies that did not present original research and were purely associative (ie, did not demonstrate how an adversary gains knowledge that was not intended to be shared from genetic data) as well as 4 studies that did not demonstrate the attack on real data. This process resulted in the selection of 42 unique studies (refer to Figure 1 for an overview of the process and Table S1 in Multimedia Appendix 1 for an overview of the eligible attack studies). Extending on the framework by Mohammed Yakubu and Chen [12] and Lu et al [56], we categorized attacks into (1) identity tracing (attacker triangulates the identity of an individual), (2) inference (attacker uses an individual's genetic data to infer sensitive attributes such as disease or drug abuse or to infer additional data or cross-link records across databases), and (3) membership attacks (attacker uncovers membership of an individual in a data set). We evaluated the type and components of genetic data exploited for this attack as well as the effort and resources used for it (time, expertise, databases, and computation power) and its success rate if sufficient information was reported in the study. The initial evaluation was conducted by one reviewer and independently verified by another. Table S1 in Multimedia Appendix 1 presents a detailed overview of the attack studies.

Figure 1. Flowchart overview of the 2-step literature review process: identification of relevant reviews, followed by extraction and screening of references.



Results

A Comprehensive Overview of Privacy Risks in Genetic Data Sets

On the basis of our literature review, we created an overview of the parts and aspects of genetic data that are commonly exploited in privacy attacks and that should therefore be taken into consideration when performing a risk assessment on genetic data. The goal of this overview is to provide data processors, who may not be experts in genomic data privacy, with essential background knowledge about the privacy vulnerabilities associated with genetic data. This understanding will help them identify privacy-critical aspects and serve as a starting point for conducting risk assessments on genetic data sets. Notably, the reidentification risks associated with data that complement genetic data (eg, clinical data and demographic data) as well as aspects of the data environment (access and governance) are crucial for a comprehensive risk assessment [57], but these aspects are not in the scope of this research. From our literature review, we synthesized 9 features that are both inherent to any genetic data and informative about privacy risk (Figure 2). The features are not mutually exclusive. Instead, they represent different "views" on genetic data and highlight various aspects that should be considered in a privacy risk assessment. For each feature, we lay out why this feature is associated with privacy risk by summarizing the relevant evidence in the scientific literature, and we assess the criticality of these attacks. In addition, we provide guiding questions that help to assess the risk of a given data set. The features can be divided into three groups:

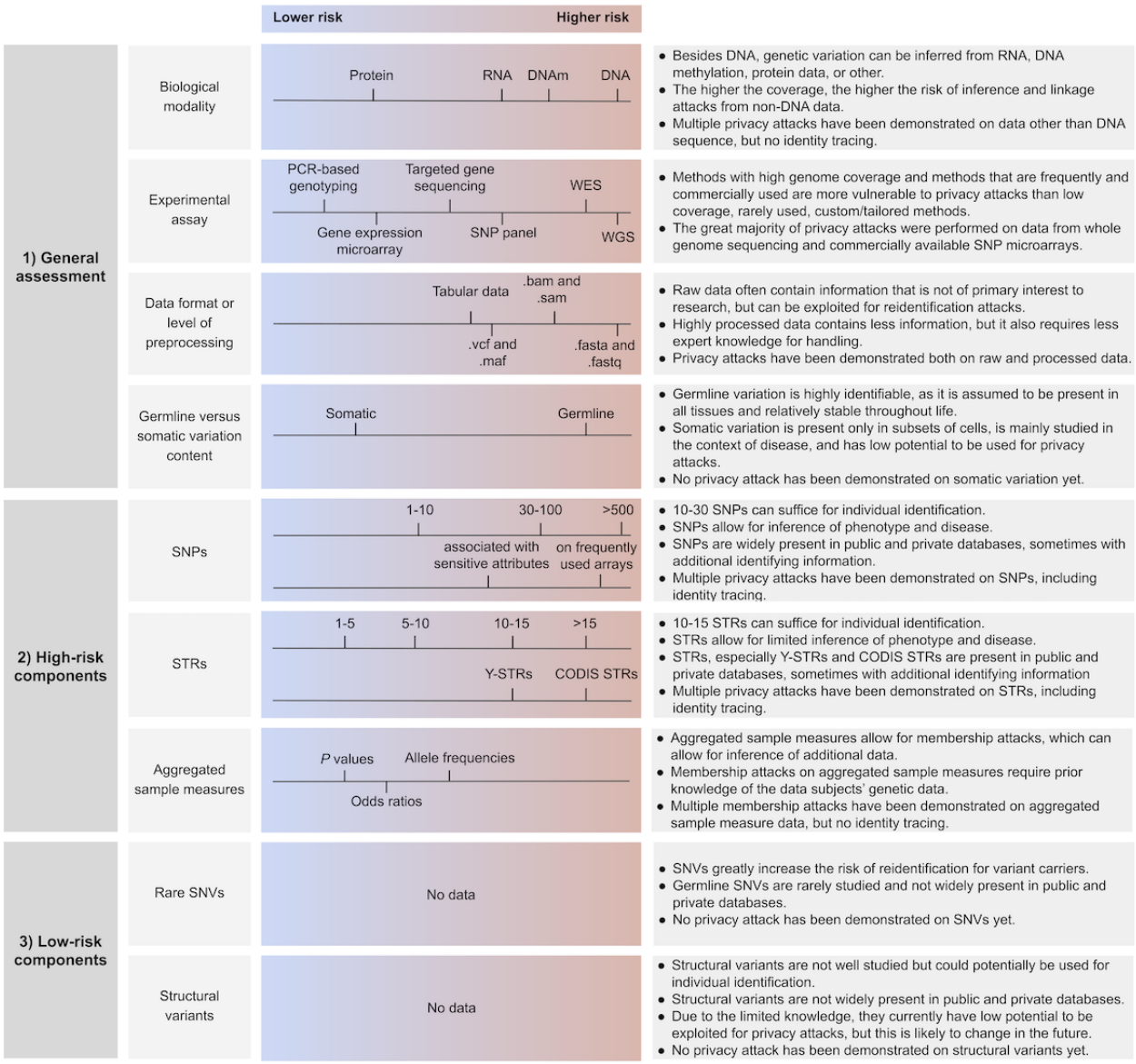
1. The first 4 features are general categorizations of the genomic data set and serve as a very rough estimate of the amount of privacy-critical information in the data.
2. The next 3 features are specific genomic features that are known to be a high risk for privacy. Their assessment is critical for estimating the reidentification risk.
3. The last 2 features are genomic features that have not been exploited for privacy attacks yet but should still be considered and could present a risk if they are present to a high degree in the data.

We summarize our findings in an overview figure, which lists the 9 features and their relevance for privacy. While it is challenging to define clear risk thresholds, there is a recognized need for practical guidance and orientation. To address this, we

provide a scale that ranges from lower to higher risk and offer illustrative examples derived from the overview of privacy attack studies. These scales and examples serve as the initial guidance for risk assessment, emphasizing their purpose as guiding principles rather than exact measurements. The assessment of each individual feature is intricate and thoroughly explained in the corresponding sections. In addition, while the scales offer a framework to compare and assess different features, it is crucial to consider all features comprehensively to arrive at a conclusive assessment. Furthermore, the text sections highlight important interactions that arise from the comprehensive evaluation of these features.

Table S1 in Multimedia Appendix 1 presents a detailed description of the original attack studies.

Figure 2. Overview of the privacy-critical features of genetic data sets, with exemplary values and key points to consider for risk assessment. CODIS: Combined DNA Index System; SNP: single nucleotide polymorphism; SNV: single nucleotide variant; STR: short tandem repeat; WES: whole exome sequencing; WGS: whole genome sequencing; Y-STR: short tandem repeat on the Y chromosome.



Evidence of Privacy Risks in Genetic Data

Part 1. General Assessment

Biological Modality

While most privacy attacks have been demonstrated on DNA sequence data, other types of molecular data (eg, DNA methylation data or data derived from RNA) are also considered genetic data under General Data Protection Regulation, can also be identifiable, and have also been exploited for attacks [58-67]. Attacks on these types of data are performed mainly by 3 mechanisms. The first mechanism is direct extraction of DNA sequence from raw or low-processed data. This is possible, because even if not of primary interest, DNA sequence information is often a by-product of gene expression or DNA methylation studies [68-70]. For example, Gürsoy et al [70] demonstrated how genetic variants can be called from raw RNA sequencing data. The second mechanism is inference of DNA sequence, for example, through known associations of genetic sequence and gene expression or other modalities. For example,

Schadt et al [65] used gene expression data of individuals (40,000 transcript counts) to infer genetic variants (1000 SNPs), which allowed them to determine with high certainty whether individuals with known SNPs were members of a gene expression study cohort (N=378). They also assessed the success rate of matching gene expression records to SNP records in a simulated cohort of 300 million individuals and correctly matched 97.1% of the records, demonstrating the feasibility of cross-linking these data types, which since then has been confirmed in additional studies [60,62,63]. Less literature has been published on other types of data, such as protein or epigenetic data (eg, DNA methylation), but similar proof of concept of cross-linkage to SNP data has been demonstrated in prior studies [58-60,63,64,66,67,71]. In the third mechanism, sensitive information such as disease phenotypes, demographic information, and behavioral traits is inferred from gene expression, protein levels, or other modalities (eg, age [72], cigarette smoking, and alcohol consumption [59] from DNA methylation).

However, such inference and linkage are not error free. For example, in the study by Schadt et al [65], the accuracy of the imputed SNPs from gene expression data was low (average Pearson correlation coefficient was 0.35 between true and inferred genotype). It is not clear whether such imputed data could be used for privacy attacks in the real world, such as in an identity tracing attack (eg, via upload of the imputed genetic data to GEDmatch or other). Considering that previous successful identity tracing attacks have used >500,000 SNPs [6], the inference of 1000 SNPs (with errors) may not be sufficient for such an attack. If the reconstruction of a larger set of SNPs were attempted, it is likely that the initial imputation error would propagate and thereby reduce the probability of a successful identity tracing attack. Furthermore, Schadt et al [65] reported much lower matching performance if training and test data stem from different array manufacturers, a scenario that is likely to occur in real-world data. Finally, although biological associations between genomic variants and gene expression are publicly accessible, substantial expert knowledge is still required for accessing this information and implementing the attack. Similar limitations apply to all the aforementioned studies. Altogether, data sets of RNA, protein, or epigenetic data, especially if they are large (eg, genome-wide), do allow for linkage and inference attacks. However, true reidentification would require matching the inferred genetic or phenotypic information to databases with identifying or quasi-identifying information in a next step, and no such full identity tracing attack starting with data other than DNA sequence has been demonstrated yet.

The guiding questions in this context are as follows:

- Do the data contain DNA sequence information directly (eg, DNA sequencing reads)? If yes, could the data be processed such that sequence information is no longer available (eg, report DNA methylation levels in percentage instead of providing raw sequencing read files)?
- Could DNA sequence information be inferred from the data (eg, via biological correlations such as expression or methylation quantitative trait loci)?
- What sensitive information could be inferred from the data (eg, age, sex, diseases, or physical traits)?

Experimental Assay

Knowing the experimental assay that was used to generate the data can already provide a first estimate of its information content and linkability. For example, sequencing-based assays generally produce very rich data (eg, high genome coverage and high precision, such as whole genome DNA sequencing), whereas polymerase chain reaction–based genotyping assays provide more sparse data (eg, information on only 1 nucleotide of the DNA sequence). However, genome coverage alone (ie, the percentage of all base pairs or loci of the genome covered by the method) is not a reliable proxy for privacy risk. In some circumstances, a data set with only 10 sequenced positions of the DNA could in fact be more critical than a data set containing hundreds of positions, if those 10 positions are in highly identifiable loci. However, as a very rough indicator of information content, we believe it is still valuable to consider the genome coverage of the data as one of many factors in the risk assessment. In many cases, the rule of thumb that more sequence information equals higher information content and hence risk of cross-linking, inference, and reidentification is true. Nevertheless, these aspects need to be carefully evaluated together with the biological modality of the data, the level of processing, and the specific content of the data.

It is also important to consider that data produced with frequently used methods, such as commercially available kits (eg, SNP microarrays), often target the same genetic variants that are also interrogated by DTC-GT companies and genome-wide disease association studies and can thus more easily be linked to public data and exploited for privacy attacks than data generated with tailor-made, targeted analysis methods (refer to the study by Lu et al [73] for an overview of genotyping arrays commonly used by direct-to-consumer companies). Finally, as nearby variants are more likely to be correlated, it is also important to consider how the genetic information in the data is spread across the genome. A targeted assay that reads all SNPs within a specific gene likely carries less information than an assay that interrogates the same number of SNPs distributed across the full genome, as nearby SNPs are more likely to be correlated [74]. In line with these arguments, the great majority of published privacy attacks were performed on data obtained from whole genome sequencing and commercially available SNP microarrays (ie, rich, genome-wide data in the order of hundreds of thousands of SNP loci from a commercial assay).

The guiding questions in this context are as follows:

- Which method was used to generate the data? Does this method produce rich or sparse data? (What percentage of all base pairs or loci of the genome are covered by the method?)
- How do the data produced with this method cover the genome (ie, genome-wide vs targeted approach)?
- How likely is it that data generated with the same method are present in publicly available databases (ie, commercial assay vs custom)?

Data Format or Level of Processing

The format of the data gives some indication on its processing level and can thus help to estimate its information content. Genetic data processing consists of cleaning, filtering, normalizing, and reducing raw data to a version that contains only the information that is relevant for its intended use. Important standard formats in genomic sequencing experiments sorted from raw to processed are *.fasta* and *.fastq* (raw nucleobase reads); *.bed*, *.bam*, and *.sam* (reads aligned to reference genome); *.vcf* and *.maf* files (deviations from the reference genome only), whereas highly processed data are often represented in tabular (*.csv* and *.tsv*) or otherwise structured form (*.json*, *.xml*, or other) containing only variants or regions of interest. Raw or low-processed data (*.fasta*, *.fastq*, *.bed*, *.bam*, or *.sam*) often contain information that is not of primary interest to research but can be exploited for reidentification attacks (eg, raw read files from gene expression studies can contain genomic variant information [63]). While the possibilities for privacy attacks are greater in raw data, it is important to note that the required effort and expert knowledge for handling these data are usually higher than those for processed data, where genetic variants such as SNPs do not need to be extracted.

The guiding question in this context is as follows:

- If the data are in a raw or semiprocessed format, do the data contain any information that is not directly relevant for their intended use?

Germline Versus Somatic Variation Content

Genetic variants found in an individual's genome can be categorized into germline and somatic variants. This categorization is specific to individuals and depends on the heritability of the variant (ergo, its presence in the individual's reproductive tissues). Heritable variants are categorized as germline (ie, present in germ and usually also in somatic cells) and nonheritable variants are categorized as somatic (ie, present in somatic cells only). In the context of genetic privacy, it is important to understand that germline variation comprises all variants that can be assumed to be present in every cell of the body, are not expected to change much throughout the lifetime of an individual, are inherited from parental DNA, and are expected to be passed to the offspring. Such variation can inform about identity, ancestry, and kinship and is, therefore, used by DTC-GT providers, forensics, and genetic genealogy services. The most prominent example for germline variation are SNPs, as variation found at known SNP loci is generally assumed to be germline. (However, the terms germline variants and SNPs cannot be used interchangeably, as they refer to different concepts: germline describes the heritability, and SNP describes the type of variant and its frequency in the population.) Overall, germline variants are not only highly relevant for individual identification because of their stability and omnipresence across tissues but are also of great interest for scientific research. Associations of germline variants to disease, physical traits, or other biomedical modalities are well studied, with results being publicly accessible. As such, germline variants are vulnerable to identity, inference, and linkage attacks, and indeed, all the reviewed privacy attacks targeted germline variants.

In contrast, somatic variants are acquired during life (after fertilization) and are usually present only in specific, nonreproductive tissues or even only in single cells or cell populations. They are intensively studied in the context of diseases (eg, cancer), and as they are often found to be associated with diseases, data on somatic variants could be used to infer sensitive attributes about data subjects. However, their low association with identity and use limited to clinical diagnostics and scientific research makes it very difficult to cross-link them to databases with identifying or quasi-identifying information. DTC-GT companies, forensics services, or genetic genealogy services do not use somatic variants to determine identity, familial relations, or ancestry, as somatic variation is neither stable nor present in all tissues and cells (usually found only in a fraction of cells analyzed in a sample). A linkage attack based on somatic variation would require a matching data record of the same tissue, ideally taken at a similar time in life, which is unlikely to exist for most cases (as somatic variant patterns can change rapidly, eg, in cancer tissue). No identity tracing, inference, or membership attack based on somatic variation data has been published yet, and considering its low potential for identifiability, somatic variation data can currently be considered a low risk for reidentification attacks.

To determine whether a variant is germline or somatic, one would ideally analyze multiple samples from one individual to determine whether the variant is present in germ cells or only in specific somatic cells. In practice, experts can assess the status of a variant from its sequencing read signal (determining whether it is present in all cells of the sample or only in a few), genomic location, and type alone by comparing it to public knowledge of known loci of germline and somatic variation or through computational approaches [75]. In processed genetic data, variants which are with high certainty germline have often already been identified and are indicated as such (eg, SNPs are identified by a specific reference SNP cluster ID, such as "rs343543"), whereas somatic variants are described by standard mutation nomenclature (eg, single nucleotide variants [SNVs] are described by the Human Genome Variation Society nomenclature, containing the reference genome used; the genomic location of the variant; the nucleotide in the reference sequence; and the detected nucleotide, such as "NC_000023.9:g.32317682G>A"). Furthermore, the type of tissue that was used to generate genetic data, most importantly whether samples were taken from healthy or tumor tissue, can also give some indication on the amount of germline variation included in the data. When analyzing tumor tissue data, germline variations such as SNPs are typically removed during processing, as the focus is on studying somatic variation. However, especially if the data are raw and unfiltered, they often contain germline variants irrespective of whether they were taken from healthy or tumor tissue and must hence be considered a higher risk for reidentification. Therefore, while data that are both derived from tumor tissues and highly processed are often a low privacy risk, the amount of information on germline variation that is contained in the data needs to be assessed case by case. Public databases (eg, dbSNP, hosted by the National Institutes of Health's National Center for Biotechnology Information) store information about the genomic locations and population frequencies of SNPs and can

be used to search data for this important type of germline variation.

The guiding questions in this context are as follows:

- Was germline or somatic variation of primary interest when generating or processing the data?
- If somatic variation was of primary interest, was germline variation removed from the data?

Part 2. High-Risk Components

SNPs

SNPs are germline SNVs that are present in >1% of the population. They are highly relevant features for individual reidentification and the most privacy-critical component of genetic data sets. Because SNPs usually have 2 different states (ie, a common or reference and a rare nucleotide) and human somatic cells have 2 DNA copies (ie, are diploid), an individual usually has 1 of 3 different states at a SNP locus, often represented as 0,1, and 2 (0 represents 2 copies of the common variant [ie, homozygous for major allele], 1 represents 1 copy of the common variant and 1 copy of the rare variant [heterozygous], and 2 represents 2 copies of the rare variant [homozygous for minor allele]). Knowing an individual's state at 30 to 80 statistically independent SNPs (or a random set of approximately 300 SNPs) can suffice for individual identification [76-79], yet commonly used SNP or genome sequencing assays often read hundreds of thousands of SNPs at once. As germline variation, SNPs are assumed to be stable and present in every cell of the body, signifying that they can identify individuals across samples taken at different times or from different tissues. As they are heritable, DTC-GT providers and forensic institutes compare SNP patterns of individuals to determine familial relations and ancestry [80]. Furthermore, SNPs are associated with physiological traits (eg, skin, hair and eye color [2,3], facial features [81], BMI [82], and height [4]), ethnicity [1], and susceptibility to diseases [5], making them central to research and genetic testing (refer to the study by Dabas et al [83] for a review of association of SNPs with externally visible characteristics).

SNP data can be directly used for reidentification by matching it to publicly accessible databases, as demonstrated in the reidentification attack by Erlich et al [6], who uploaded SNP data (700,000 SNPs) from an anonymous study participant to the genetic genealogy website GEDmatch and identified the participant's surname through matches with relatives. Such identity tracing attacks are possible because millions of people send their DNA to DTC-GT companies such as AncestryDNA, 23andMe, FamilyTreeDNA, or MyHeritage [84], and many also decide to share their genetic data on publicly accessible websites, such as GEDmatch, the Personal Genome Project [85], or OpenSNP [86]. Enabling individuals to identify and contact relatives, learn about their ancestry, disease predispositions, and contribute their data to research, these platforms often contain genetic data accompanied by information about an individual's diseases and traits or even personal data such as place of residence, age, sex, surname, or phone number. In addition, there is a wealth of publicly accessible knowledge on associations of SNPs with physical features, diseases, other genetic variants or genetic modalities (eg, gene expression and DNA methylation; eg, dbSNP database [87], the GWAS catalog [5], the International Genome Sample Resource from the 1000 Genomes Project [88], and data from the HapMap project [89]), which can and have been exploited for completion and inference attacks (eg, inference of additional genetic variation in genomic regions that were not studied originally, other biomedical modalities such as gene expression and DNA methylation, or physical attributes [9,90-96]). For example, Humbert et al [92] predicted phenotypic traits (eye, hair and skin color, blood type, and more) of individuals from their SNP data (20 SNPs) using publicly available knowledge on SNP-phenotype associations from the public database SNPedia and used this information to cross-link individuals between genetic and phenotypic data sets. In addition, Humbert et al [92] inferred additional and sensitive information (eg, susceptibility to Alzheimer disease) from the SNP data. However, this linkage attack had a success rate of only 5% (ie, proportion of correctly matched individuals) in a data set of 80 individuals and is likely to perform worse in more realistic scenarios with larger data sets. Nyholt et al [9] imputed the status of multiple risk variants for Alzheimer disease in the published genome of Dr James Watson [94] from SNPs in nearby genomic regions, although the respective gene had been masked. Edge et al [90] cross-linked individuals in SNP and short tandem repeat (STR) data sets, a highly identifiable type of genetic variation that is used in forensics, by imputing STR from SNP data (642,563 loci). In a highly debated study, Lippert et al [93] developed a model to predict phenotypic traits (facial structure, voice, eye color, skin color, age, sex, height, and BMI) from whole genome sequencing (WGS) data containing >6 million SNPs and used it to cross-link high-resolution face photographs of individuals to their genetic data in a cohort of 1061 study participants. In a real-life scenario, photos and personal data from social media could be exploited for such an attack and matched to the inferred phenotype. However, it has been argued that the predictive power in this study stems mainly from the estimation of the participant's ancestry and sex [97] and that the attack is unlikely to be successful in the real world and with more realistic, lower-quality images [98]. Furthermore, large, genome-wide association studies indicate that the currently known associations between SNPs and facial structure, voice, height, and BMI are too small to be useful for accurate phenotype prediction on an individual level; however, this will likely improve in the future. Nevertheless, other characteristics, such as ancestry, eye, hair color, and skin color, can be inferred from specific SNPs with high accuracy, and corresponding DNA phenotyping kits are already commercially available and used in forensics today [99]. As a small number of SNPs can already uniquely identify an individual and SNPs are widely available in public databases together with identifying and quasi-identifying information, SNPs must be considered a high risk for privacy and data sanitization efforts (eg, as proposed by Emani et al [100]) should be used in any genetic data set containing >20 SNPs.

The guiding questions in this context are as follows:

- How many SNPs do the data contain (directly or indirectly)?
- Are the SNPs in close proximity or spread across the genome (nearby SNPs are more likely to be correlated and

thus often contain less information than statistically independent SNPs)?
- Are the interrogated SNPs frequently assessed in research or by DTC-GT providers (ie, how likely is it that they can be linked to publicly available, identifying data sets)? The study by Lu et al [73] presents an overview of genotyping arrays commonly used by direct-to-consumer companies.
- Are all SNPs relevant to the intended use of the data or could some be removed from the data?
- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with STRs or other)?

STRs

The human genome contains more than half a million regions of repetitive units of 2 to 6 bases, the so-called STRs or microsatellites [101]. The number of repeats in these regions is highly variable across individuals and can affect protein function or expression or be linked to medical conditions or physical traits [102]. Knowing the repeat numbers of as little as 10 to 30 STRs can suffice for individual identification. Because of their high identifiability, STRs are used to determine identity and kinship in forensics, law enforcement, paternity testing, and genetic genealogy. For example, the Combined DNA Index System (CODIS; a set of 20 STRs) is used to connect suspects to crime scenes or establish identity of missing persons. While CODIS STRs are usually not of interest in research studies or genetic genealogy, STRs on the Y chromosome (ie, Y-STRs, only present in male individuals) are included in several DTC-GT kits, where they are used to identify relatives along the paternal ancestry line (eg, "Y-STR Testing" by FamilyTreeDNA). Consequently, several large databases of STR loci with accompanying identifying and quasi-identifying information exist (eg, mitoYDNA from mitoYDNA Ltd). In addition, the CODIS forensic database and analysis software contains genetic data and identifying information from >14 million individuals in the United States alone [103].

Several studies demonstrate reidentification attacks on Y-STRs. Gitschier et al [104] provided first evidence for surname inference from Y-STRs by matching genetic STR profiles of anonymous study participants from the international HapMap project [89] to 2 genetic genealogy databases (Ysearch and Sorenson Molecular Genealogy Foundation [SGMF]). Later, Gymrek et al [105] demonstrated that it is not only possible to infer surnames from STR data (eg, 34 Y-STR loci extracted from WGS data) but also to triangulate the actual identity of data subjects with high probability using publicly accessible genealogy databases, record search engines, obituaries, and genealogical websites. The authors attempted this for 10 study participants of the 1000 Genomes Project and correctly identified 5 out of 10 individuals. It is important to note that STR data can also be fortuitously included in genetic data derived from targeted gene or WGS, even if they are not of primary interest for the study. Moreover, STR markers can be imputed from genetic data sets that do not even cover STR regions by exploiting known associations between SNPs and STRs [90]. While the authors of this study report a low imputation accuracy for STRs from SNPs (likely too low to reliably impute full STR profiles even from large SNP data), they did demonstrate the ability to cross-link records across SNP and STR databases. In detail, they correctly matched 90% to 98% of paired SNP (642,563 loci) and STR data records (13 STRs) to each other, and such successful linkage has also been demonstrated elsewhere [106].

Due to the high association of STRs with identity, any genetic data that directly (eg, repeat numbers for specific STR regions) or indirectly (eg, WGS data covering STR regions) contain >10 STR regions could be considered identifiable. However, the actual risk of reidentification depends on the availability of STR databases with identifying and quasi-identifying information and the ability to cross-link records. It is important to note that the databases used in the seminal study by Gymrek et al [105] (ie, Ysearch and SGMF) are no longer available (Ysearch, belonging to FamilyTreeDNA, closed in 2018, and SGMF, belonging to Ancestry, was shut down in 2015), and access to the CODIS database is restricted to criminal justice agencies for law enforcement identification purposes. However, databases from DTC-GT providers (eg, FamilyTreeDNA) and public platforms (eg, mitoYDNA) are still available and allow uploading results from third-party providers; therefore, an attacker could fabricate a genetic testing result from STR data [107,108] and reproduce the demonstrated surname inference attacks. From information about possible surnames, sex, and residence inferred from matches on the platform, the triangulation of identity could be possible with the help of additional publicly available resources [105,109]. However, such an attack would only be possible on male data records (ie, Y chromosome based) and is not guaranteed to find matches that allow surname inference; the success rate in the demonstrated attack was 11.9% (109/911 cases), and the 2 previous studies used >30 STR loci (all located in close vicinity of each other and on the Y chromosome). Furthermore, the know-how and effort necessary for such an attack is high. Finally, even if genetic matches or surnames are identified, the reconstruction of identity from surname is not trivial and can take months to complete, as others have pointed out [110]. Still, because of their high identifiability potential and their use in DTC-GT, paternity testing, and forensics, STRs should be removed from genetic data if they are not of primary interest and otherwise considered a high risk for privacy.

The guiding questions in this context are as follows:

- Do the data directly or indirectly (eg, STRs in raw data and STRs imputable from SNPs) contain >10 STR loci?
- Are these STR loci either (1) part of the CODIS system or (2) on the Y chromosome (ie, high linkability)?
- Could additional DNA sequence information be inferred from the data (eg, association with SNPs or other)?

Aggregated Sample Measures

Aggregated sample measures, that is, variables that are the result of aggregating genetic data across multiple samples can also be exploited for privacy attacks (reviewed by Craig et al [111]). The most prominent examples are summary statistics from association studies such as SNP frequencies, odds ratios, or correlation coefficients. However, the limited information content in these summary statistics usually only allows for

membership attacks, that is, assessing whether an individual of known genetic background is part of a study group or database or not [112-114]. Multiple studies demonstrate such an attack [113,115-119], although Homer et al [114] were the first to explain how membership of an individual in a mixture can be predicted from the reported SNP allele frequencies (ie, if SNPs of that individual are known, in this case >10,000 SNPs). The authors accomplished this by comparing the reported study allele frequencies to allele frequencies in a reference cohort of similar ancestry (obtained from public resources) and detecting the bias introduced by the sample of interest. Their method performed well even if the individual's contribution to the mixture was <1%, and this method can easily be extended to predicting membership from aggregated data from a study cohort. In response to that, the US National Institutes of Health has restricted the publication of aggregate GWAS results in their databases [120]; however, the feasibility of the attack has been critically discussed. Its power depends on the size and quality of the actual and reference cohorts, the number of reported SNP allele frequencies, prior knowledge of the attacker, and the fulfillment of several underlying assumptions, many of which are likely not fulfilled in practice [115,116,121,122]. Aside from membership attacks, it was also shown that aggregate results, such as linear models that have been fitted to study data or polygenic risk scores, can be exploited to predict sensitive attributes and genotypes via model inversion [28,123]. However, this attack required background information on the data subject and on the distribution of variables in the study data. Furthermore, its performance is limited by the predictive power and complexity of the fitted model. Membership and attribute inference attacks on aggregate data can reveal demographic, genetic, and phenotypic information (such as country or place of residence due to participation in a local study, ethnicity, disease, age group, or presence of specific genetic variants due to descriptions of inclusion or exclusion criteria in the cohort) and can thus facilitate linkage and identity tracing attacks, which is why they can be a risk for privacy. However, no identity tracing attack based on aggregate data has been demonstrated yet.

The guiding question in this context is as follows:

- What sensitive information could an attacker gain from ascertaining the membership of an individual to the data set (eg, geographic information, sex, disease, and age)?

Part 3. Low-Risk Components

No privacy attack has been demonstrated on these components, but due to their high association with identifying and sensitive attributes, we recommend including them in the risk assessment.

Rare SNVs

Rare SNVs are single nucleotide substitutions that are present in <1% of the population. They may be somatic or germline and can be associated with pathological conditions and thus reveal sensitive information. Furthermore, while less informative than common SNVs (ie, SNPs) from an information theoretical standpoint, rare variants greatly increase the risk of reidentification for the small subpopulation of variant carriers. However, because of their low frequency in the population, germline SNVs are rarely the target of large scientific studies (eg, for phenotype or disease association) and have very limited use for ancestry and disease susceptibility analysis. Therefore, most DTC-GT providers and research studies specifically target regions of common genetic variation (eg, SNPs) and either use assays that do not detect SNVs or remove them during preprocessing, making it very unlikely that a set of SNVs could be linked to any database with quasi-identifying information. No identity tracing, completion, or inference attack has been published on SNVs yet; therefore, they can currently be viewed as a low risk for reidentification, despite their high theoretical potential for identifiability.

The guiding questions in this context are as follows:

- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with SNPs or other)?
- Are there any databases that could be used to cross-link the data to identifiable data, and how accessible are the databases?

Structural Variants

The study of structural variants (SVs) in the human genome is in its early stages, but it is already clear that it accounts for even more individual variation than SNPs [124,125]. The best-studied type of SVs is copy number variation (CNV), that is, deletions and duplications of regions larger than 50 base pairs. CNVs can be used as measures of relatedness and identifiers of population origin [126], have a strong impact on gene expression [127], and could allow for the inference of physical features [128] and pathological conditions [129], thereby revealing sensitive information of data subjects. However, CNVs are still not well studied, and sequencing technologies have only recently progressed to a level that allows to capture their full scope in the human genome (reviewed by Mahmoud et al [124]). Most importantly, human CNV databases are very scarce in comparison to databases of SNVs (refer to the study by Ho et al [130] for an overview of the available human SV reference sets), and they are currently not used for genetic genealogy analyses, making it difficult to link CNVs across databases to obtain identifying information. A privacy attack based on CNVs or any other type of SV yet remains to be demonstrated. Finally, it is important to note that many SVs that are assessed in medical and research studies are somatic, that is, nonhereditary, not present in all cells of the body, not stable, and thus not strongly associated with identity. For example, tumor tissue is characterized by frequent and dynamic changes in SVs (eg, CNVs in tumor tissue, also referred to as CNAs), which are likely neither directly nor indirectly identifiable. Therefore, the risk of reidentification from SVs can currently be considered low, but the growth of public databases and their use in genealogical or clinical research should be monitored. The same holds true for common SVs, such as CNVs that occur in >1% of the population and are hence classified as polymorphisms (ie, CNPs). Little is known about the population frequencies of CNVs, and while public databases are growing, no privacy attack based on CNPs has been demonstrated yet. Due to the limited knowledge about CNPs or other common SVs in the

population, their presence in genetic data is difficult to assess, and they can be considered a low risk for reidentification at the current time.

The guiding questions in this context are as follows:

- What sensitive information could be inferred from the data (eg, diseases and physical traits)?
- Could additional DNA sequence information be inferred from the data (eg, association with SNPs or other)?
- Are there any databases that could be used to cross-link the data to identifiable data, and how accessible are the databases?

Discussion

Limitations

It is important to acknowledge some key limitations of our review. First, it is possible that we may have missed relevant studies. This is particularly true for recent research, as our search was confined to original studies referenced in existing reviews. While the search strategy was designed to retrieve the most pertinent studies, it carries the risk of overlooking lesser-known or very recent studies. Therefore, we recommend conducting periodic reviews to stay updated with scientific advancements and changes in the availability of public genetic data that may contain (indirectly) identifying information susceptible to identity tracing attacks. Second, even under the assumption that all relevant literature was considered, it is still possible that we may have overlooked certain vulnerabilities. This is known as the "proof of nonexistence fallacy"—the absence of evidence for risk does not imply the absence of those risks. Finally, it was necessary to balance our aim of providing a comprehensive and evidence-based overview of genetic privacy vulnerabilities with our aim of providing practical and useful guidance. Therefore, we provide both a detailed assessment (refer to the *Results* section and Table S1 in Multimedia Appendix 1) as well as a simplified overview (Figure 2). However, this trade-off necessitated compromises in practical utility on one hand and scientific exhaustiveness on the other hand.

Conclusions

On the basis of the findings of this review, it can be argued that the privacy risks of genetic data vary greatly between data sets. Considering all genetic data at all times as information relating to an identifiable natural person is not correct, and it is becoming apparent that reidentification risk in genetic data must be assessed on a case-by-case basis and under the consideration of all the means reasonably likely to be used [131]. However, while efforts are underway [132], no practical guidelines or recommendations for performing such a reidentification risk assessment on genetic data have been proposed yet. On the basis of a review of the scientific literature on privacy attacks on genetic data, we provide an overview of genetic data privacy risks that can guide data processors in risk assessment by providing the necessary background knowledge and an overview of the existing evidence. We believe that a careful examination of the 9 described features in the data set at hand (biological modality or type of data, experimental assay, data format or level of processing, germline vs somatic variation content, content of SNPs, STRs, aggregated sample measures, rare SNVs, and SVs) provides a strong foundation for a data risk assessment. While completely eliminating the possibility of reidentification is rarely achievable, a more practical approach of risk minimization is warranted [133,134], accompanied by organizational and technical measures to safeguard genetic data from reidentification attack attempts and a transparent communication of the remaining risks to data subjects.

Acknowledgments

The authors would like to thank Florian Schneider for reviewing the manuscript.

Conflicts of Interest

NM was contracted by Roche Diagnostics GmbH, Penzberg, Germany, while contributing to this work. All other authors declare no other conflicts of interest.

Multimedia Appendix 1

List of identified reviews and a table with the description and evaluation of original privacy attack studies.
[DOCX File , 36 KB-Multimedia Appendix 1]

Multimedia Appendix 2

PRISMA checklist.
[PDF File (Adobe PDF File), 85 KB-Multimedia Appendix 2]

References

1. Huang T, Shu Y, Cai YD. Genetic differences among ethnic groups. BMC Genomics. Dec 21, 2015;16:1093. [FREE Full text] [doi: 10.1186/s12864-015-2328-0] [Medline: 26690364]
2. Simcoe M, Valdes A, Liu F, Furlotte NA, Evans DM, Hemani G, et al. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. Sci Adv. Mar 2021;7(11):eabd1239. [FREE Full text] [doi: 10.1126/sciadv.abd1239] [Medline: 33692100]

3. Pavan WJ, Sturm RA. The genetics of human skin and hair pigmentation. Annu Rev Genomics Hum Genet. Aug 31, 2019;20:41-72. [doi: 10.1146/annurev-genom-083118-015230] [Medline: 31100995]
4. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. Oct 2022;610(7933):704-712. [doi: 10.1038/s41586-022-05275-y] [Medline: 36224396]
5. Buniello A, MacArthur JA, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. Jan 08, 2019;47(D1):D1005-D1012. [FREE Full text] [doi: 10.1093/nar/gky1120] [Medline: 30445434]
6. Erlich Y, Shor T, Pe'er I, Carmi S. Identity inference of genomic data using long-range familial searches. Science. Nov 09, 2018;362(6415):690-694. [FREE Full text] [doi: 10.1126/science.aau4832] [Medline: 30309907]
7. Greytak EM, Moore C, Armentrout SL. Genetic genealogy for cold case and active investigations. Forensic Sci Int. Jun 2019;299:103-113. [doi: 10.1016/j.forsciint.2019.03.039] [Medline: 30991209]
8. Kennett D. Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. Forensic Sci Int. Aug 2019;301:107-117. [doi: 10.1016/j.forsciint.2019.05.016] [Medline: 31153988]
9. Nyholt DR, Yu CE, Visscher PM. On Jim Watson's APOE status: genetic information is hard to hide. Eur J Hum Genet. Feb 2009;17(2):147-149. [FREE Full text] [doi: 10.1038/ejhg.2008.198] [Medline: 18941475]
10. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. Jun 2014;15(6):409-421. [FREE Full text] [doi: 10.1038/nrg3723] [Medline: 24805122]
11. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One. 2011;6(12):e28071. [FREE Full text] [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
12. Mohammed Yakubu A, Chen YP. Ensuring privacy and security of genomic data and functionalities. Brief Bioinform. Mar 23, 2020;21(2):511-526. [doi: 10.1093/bib/bbz013] [Medline: 30759195]
13. Joly Y, Dalpe G. Genetic discrimination still casts a large shadow in 2022. Eur J Hum Genet. Dec 2022;30(12):1320-1322. [FREE Full text] [doi: 10.1038/s41431-022-01194-8] [Medline: 36163420]
14. Tiller J, Lacaze P. Australians can be denied life insurance based on genetic test results, and there is little protection. The Conversation. Aug 24, 2017. URL: https://theconversation.com/australians-can-be-denied-life-insurance-based-on-genetic-test-results-and-there-is-little-protection-81335 [accessed 2024-04-05]
15. Humbert M, Ayday E, Hubaux JP, Telenti A. 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. Presented at: CCS '13; November 4-8, 2013:1141-1152; Berlin, Germany. URL: https://dl.acm.org/doi/10.1145/2508859.2516707 [doi: 10.1145/2508859.2516707]
16. Deznabi I, Mobayen M, Jafari N, Tastan O, Ayday E. An inference attack on genomic data using kinship, complex correlations, and phenotype information. IEEE/ACM Trans Comput Biol Bioinform. 2018;15(4):1333-1343. [doi: 10.1109/TCBB.2017.2709740] [Medline: 30102600]
17. Trost B, Loureiro LO, Scherer SW. Discovery of genomic variation across a generation. Hum Mol Genet. Oct 01, 2021;30(R2):R174-R186. [FREE Full text] [doi: 10.1093/hmg/ddab209] [Medline: 34296264]
18. EDPB documenton response to the request from the European commission for clarifications on the consistent application of the GDPR, focusing on health research. European Data Protection Board. 2021. URL: https://edpb.europa.eu/sites/default/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf [accessed 2024-04-05]
19. Global Alliance for Genomics and Health. GENOMICS. A federated ecosystem for sharing genomic, clinical data. Science. Jun 10, 2016;352(6291):1278-1280. [doi: 10.1126/science.aaf6162] [Medline: 27284183]
20. Shringarpure SS, Bustamante CD. Privacy risks from genomic data-sharing beacons. Am J Hum Genet. Nov 05, 2015;97(5):631-646. [FREE Full text] [doi: 10.1016/j.ajhg.2015.09.010] [Medline: 26522470]
21. von Thenen N, Ayday E, Cicek AE. Re-identification of individuals in genomic data-sharing beacons via allele inference. Bioinformatics. Feb 01, 2019;35(3):365-371. [doi: 10.1093/bioinformatics/bty643] [Medline: 30052749]
22. Raisaro JL, Tramèr F, Ji Z, Bu D, Zhao Y, Carey K, et al. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks. J Am Med Inform Assoc. Jul 01, 2017;24(4):799-805. [FREE Full text] [doi: 10.1093/jamia/ocw167] [Medline: 28339683]
23. Ayoz K, Ayday E, Cicek AE. Genome reconstruction attacks against genomic data-sharing beacons. Proc Priv Enhanc Technol. 2021;2021(3):28-48. [FREE Full text] [doi: 10.2478/popets-2021-0036] [Medline: 34746296]
24. Fienberg SE, Slavkovic A, Uhler C. Privacy preserving GWAS data sharing. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops. 2011. Presented at: ICDMW '11; December 11, 2011:628-635; Washington, DC. URL: https://dl.acm.org/doi/10.1109/ICDMW.2011.140 [doi: 10.1109/icdmw.2011.140]
25. Huang Z, Ayday E, Fellay J, Hubaux JP, Juels A. GenoGuard: protecting genomic data against brute-force attacks. In: Proceedings of the 2015 IEEE Symposium on Security and Privacy. 2015. Presented at: SP '15; May 17-21, 2015:447-462; San Jose, CA. URL: https://ieeexplore.ieee.org/document/7163041 [doi: 10.1109/sp.2015.34]
26. Wang Y, Wen J, Wu X, Shi X. Infringement of individual privacy via mining differentially private GWAS statistics. In: Proceedings of the 2nd International Conference on Big Data Computing and Communications. 2016. Presented at: BigCom

'16; July 29-31, 2016:29-31; Shenyang, China. URL: https://link.springer.com/chapter/10.1007/978-3-319-42553-5_30 [doi: 10.1007/978-3-319-42553-5_30]
27. Cavallaro L, Kinder J, Domingo-Ferrer J, Oprisanu B, Dessimoz C, Cristofaro ED. How much does Genoguard really "guard"?: an empirical analysis of long-term security for genomic data. In: Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society. 2019. Presented at: WPES '19; November 11, 2019:93-105; London, UK. URL: https://tinyurl.com/4w8sxk6f [doi: 10.1145/3338498.3358641]
28. Fredrikson M, Lantz E, Jha S, Lin S, Page D, Ristenpart T. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. Proc USENIX Secur Symp. Aug 2014;2014:17-32. [FREE Full text] [Medline: 27077138]
29. Oprisanu B, Ganev G, Cristofaro ED. On utility and privacy in synthetic genomic data. In: Proceedings of the 2022 Network and Distributed Systems Security. 2022. Presented at: NDSS '22; April 24-28, 2022:1-18; San Diego, CA. URL: https://www.ndss-symposium.org/wp-content/uploads/2022-92-paper.pdf [doi: 10.14722/ndss.2022.24092]
30. Mittos A, Malin B, Cristofaro ED. Systematizing genome privacy research: a privacy-enhancing technologies perspective. Proc Priv Enhancing Technol. 2019;(1):87-107. [FREE Full text] [doi: 10.2478/popets-2019-0006]
31. Martinez C, Jonker E. A practical path towards genetic privacy in the United States. Future of Privacy Forum. 2020. URL: https://fpf.org/wp-content/uploads/2020/04/APracticalPathTowardGeneticPrivacy_April2020.pdf [accessed 2022-10-31]
32. Bernier A, Liu H, Knoppers BM. Computational tools for genomic data de-identification: facilitating data protection law compliance. Nat Commun. Nov 29, 2021;12(1):6949. [FREE Full text] [doi: 10.1038/s41467-021-27219-2] [Medline: 34845213]
33. The GDPR and genomic data - the impact of the GDPR and DPA 2018 on genomic healthcare and research. PHG Foundation. 2020. URL: https://tinyurl.com/dfk7e3xs [accessed 2024-04-05]
34. Wagner I. Evaluating the strength of genomic privacy metrics. ACM Trans Priv Secur. Jan 09, 2017;20(1):1-34. [doi: 10.1145/3020003]
35. Abinaya B, Santhi S. A survey on genomic data by privacy-preserving techniques perspective. Comput Biol Chem. Aug 2021;93:107538. [doi: 10.1016/j.compbiolchem.2021.107538]
36. Azencott CA. Machine learning and genomics: precision medicine versus patient privacy. Philos Trans A Math Phys Eng Sci. Sep 13, 2018;376(2128):20170350. [doi: 10.1098/rsta.2017.0350] [Medline: 30082298]
37. Ayday E, Humbert M. Inference attacks against kin genomic privacy. IEEE Secur Privacy. 2017;15(5):29-37. [doi: 10.1109/msp.2017.3681052]
38. Aziz MM, Sadat MN, Alhadidi D, Wang S, Jiang X, Brown CL, et al. Privacy-preserving techniques of genomic data-a survey. Brief Bioinform. May 21, 2019;20(3):887-895. [FREE Full text] [doi: 10.1093/bib/bbx139] [Medline: 29121240]
39. Berger B, Cho H. Emerging technologies towards enhancing privacy in genomic data sharing. Genome Biol. Jul 02, 2019;20(1):128. [FREE Full text] [doi: 10.1186/s13059-019-1741-0] [Medline: 31262363]
40. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. Nat Genet. Jul 29, 2020;52(7):646-654. [FREE Full text] [doi: 10.1038/s41588-020-0651-0] [Medline: 32601475]
41. Carter AB. Considerations for genomic data privacy and security when working in the cloud. J Mol Diagn. Jul 2019;21(4):542-552. [FREE Full text] [doi: 10.1016/j.jmoldx.2018.07.009] [Medline: 30703562]
42. Clayton EW, Halverson CM, Sathe NA, Malin BA. A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. PLoS One. Oct 31, 2018;13(10):e0204417. [FREE Full text] [doi: 10.1371/journal.pone.0204417] [Medline: 30379944]
43. Gürsoy G. Genome privacy and trust. Annu Rev Biomed Data Sci. Aug 10, 2022;5(1):163-181. [doi: 10.1146/annurev-biodatasci-122120-021311] [Medline: 35508070]
44. Knoppers BM, Beauvais MJ. Three decades of genetic privacy: a metaphoric journey. Hum Mol Genet. Oct 01, 2021;30(R2):R156-R160. [FREE Full text] [doi: 10.1093/hmg/ddab164] [Medline: 34155499]
45. May T. Sociogenetic risks — ancestry DNA testing, third-party identity, and protection of privacy. N Engl J Med. Aug 02, 2018;379(5):410-412. [doi: 10.1056/nejmp1805870]
46. Oestreich M, Chen D, Schultze JL, Fritz M, Becker M. Privacy considerations for sharing genomics data. EXCLI J. 2021;20:1243-1260. [FREE Full text] [doi: 10.17179/excli2021-4002] [Medline: 34345236]
47. Schwab AP, Luu HS, Wang J, Park JY. Genomic privacy. Clin Chem. Dec 2018;64(12):1696-1703. [doi: 10.1373/clinchem.2018.289512] [Medline: 29991478]
48. Shen H, Ma J. Privacy challenges of genomic big data. Adv Exp Med Biol. 2017;1028:139-148. [doi: 10.1007/978-981-10-6041-0_8] [Medline: 29058220]
49. Shi X, Wu X. An overview of human genetic privacy. Ann N Y Acad Sci. Jan 14, 2017;1387(1):61-72. [FREE Full text] [doi: 10.1111/nyas.13211] [Medline: 27626905]
50. Stiles D, Appelbaum PS. Cases in precision medicine: concerns about privacy and discrimination after genomic sequencing. Ann Intern Med. May 07, 2019;170(10):717. [doi: 10.7326/m18-2666]
51. Wan Z, Hazel JW, Clayton EW, Vorobeychik Y, Kantarcioglu M, Malin BA. Sociotechnical safeguards for genomic data privacy. Nat Rev Genet. Jul 04, 2022;23(7):429-445. [FREE Full text] [doi: 10.1038/s41576-022-00455-y] [Medline: 35246669]

52. Wang S, Jiang X, Singh S, Marmor R, Bonomi L, Fox D, et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the United States. Ann N Y Acad Sci. Jan 28, 2017;1387(1):73-83. [FREE Full text] [doi: 10.1111/nyas.13259] [Medline: 27681358]
53. Belani S, Tiarks GC, Mookerjee N, Rajput V. "I agree to disagree": comparative ethical and legal analysis of big data and genomics for privacy, consent, and ownership. Cureus. Oct 2021;13(10):e18736. [FREE Full text] [doi: 10.7759/cureus.18736] [Medline: 34796049]
54. Du L, Wang M. Genetic privacy and data protection: a review of Chinese direct-to-consumer genetic test services. Front Genet. Apr 28, 2020;11:416. [FREE Full text] [doi: 10.3389/fgene.2020.00416] [Medline: 32425986]
55. Dugan T, Zou X. Privacy-preserving evaluation techniques and their application in genetic tests. Smart Health. Jun 2017;1-2:2-17. [doi: 10.1016/j.smhl.2017.03.003]
56. Lu D, Zhang Y, Zhang L, Wang H, Weng W, Li L, et al. Methods of privacy-preserving genomic sequencing data alignments. Brief Bioinform. Nov 05, 2021;22(6):bbab151. [doi: 10.1093/bib/bbab151] [Medline: 34021302]
57. Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J. Assessing the privacy risks of data sharing in genomics. Public Health Genomics. 2011;14(1):17-25. [FREE Full text] [doi: 10.1159/000294150] [Medline: 20339285]
58. Backes M, Berrang P, Bieg M, Eils R, Herrrnann C, Humbert M, et al. Identifying personal DNA methylation profiles by genotype inference. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy. 2017. Presented at: SP '17; May 22-26, 2017:957-976; San Jose, CA. URL: https://ieeexplore.ieee.org/document/7958619 [doi: 10.1109/sp.2017.21]
59. Philibert RA, Terry N, Erwin C, Philibert WJ, Beach SR, Brody GH. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. Clin Epigenetics. 2014;6(1):28. [FREE Full text] [doi: 10.1186/1868-7083-6-28] [Medline: 25859287]
60. Gürsoy G, Lu N, Wagner S, Gerstein M. Recovering genotypes and phenotypes using allele-specific genes. Genome Biol. Sep 07, 2021;22(1):263. [FREE Full text] [doi: 10.1186/s13059-021-02477-x] [Medline: 34493313]
61. Hagestedt I, Zhang Y, Humbert M, Berrang P, Tang H, Wang X, et al. MBeacon: privacy-preserving beacons for DNA methylation data. In: Proceedings of the 2019 Network and Distributed Systems Security Symposium. 2019. Presented at: NDSS '19; February 24-27, 2019:1-15; San Diego, CA. URL: https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03A-2_Hagestedt_paper.pdf [doi: 10.14722/ndss.2019.23064]
62. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods. Mar 2016;13(3):251-256. [FREE Full text] [doi: 10.1038/nmeth.3746] [Medline: 26828419]
63. Harmanci A, Gerstein M. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. Nat Commun. Jun 22, 2018;9(1):2453. [FREE Full text] [doi: 10.1038/s41467-018-04875-5] [Medline: 29934598]
64. Boonen K, Hens K, Menschaert G, Baggerman G, Valkenborg D, Ertaylan G. Beyond genes: re-identifiability of proteomic data and its implications for personalized medicine. Genes (Basel). Sep 05, 2019;10(9):682. [FREE Full text] [doi: 10.3390/genes10090682] [Medline: 31492022]
65. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nat Genet. May 2012;44(5):603-608. [doi: 10.1038/ng.2248] [Medline: 22484626]
66. Dyke SO, Cheung WA, Joly Y, Ammerpohl O, Lutsik P, Rothstein MA, et al. Epigenome data release: a participant-centered approach to privacy protection. Genome Biol. Jul 17, 2015;16(1):142. [FREE Full text] [doi: 10.1186/s13059-015-0723-0] [Medline: 26185018]
67. Berrang P, Humbert M, Zhang Y, Lehmann I, Eils R, Backes M. Dissecting privacy risks in biomedical data. In: Proceedings of the 2018 IEEE European Symposium on Security and Privacy. 2018. Presented at: EuroS&P'18; April 24-26, 2018:62-76; London, UK. URL: https://ieeexplore.ieee.org/document/8406591/similar#similar [doi: 10.1109/eurosp.2018.00013]
68. Zhao Y, Wang K, Wang W, Yin T, Dong W, Xu C. A high-throughput SNP discovery strategy for RNA-seq data. BMC Genomics. Feb 27, 2019;20(1):160. [FREE Full text] [doi: 10.1186/s12864-019-5533-4] [Medline: 30813897]
69. Gürsoy G, Li T, Liu S, Ni E, Brannon CM, Gerstein MB. Functional genomics data: privacy risk assessment and technological mitigation. Nat Rev Genet. Apr 2022;23(4):245-258. [doi: 10.1038/s41576-021-00428-7] [Medline: 34759381]
70. Gürsoy G, Emani P, Brannon CM, Jolanki OA, Harmanci A, Strattan JS, et al. Data sanitization to reduce private information leakage from functional genomics. Cell. Nov 12, 2020;183(4):905-917. [FREE Full text] [doi: 10.1016/j.cell.2020.09.036] [Medline: 33186529]
71. Li S, Bandeira N, Wang X, Tang H. On the privacy risks of sharing clinical proteomics data. AMIA Jt Summits Transl Sci Proc. 2016;2016:122-131. [FREE Full text] [Medline: 27595046]
72. Dupras C, Beck S, Rothstein MA, Berner A, Saulnier KM, Pinkesz M, et al. Potential (mis) use of epigenetic age estimators by private companies and public agencies: human rights law should provide ethical guidance. Environ Epigenet. 2019;5(3):dvz018. [doi: 10.1093/eep/dvz018]
73. Lu C, Greshake Tzovaras B, Gough J. A survey of direct-to-consumer genotype data, and quality control tool (GenomePrep) for research. Comput Struct Biotechnol J. Jun 27, 2021;19:3747-3754. [FREE Full text] [doi: 10.1016/j.csbj.2021.06.040] [Medline: 34285776]
74. Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci U S A. Dec 21, 1999;96(26):15173-15177. [FREE Full text] [doi: 10.1073/pnas.96.26.15173] [Medline: 10611357]

75. Sun JX, He Y, Sanford E, Montesion M, Frampton GM, Vignot S, et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. PLoS Comput Biol. Feb 2018;14(2):e1005965. [FREE Full text] [doi: 10.1371/journal.pcbi.1005965] [Medline: 29415044]
76. Yousefi S, Abbassi-Daloii T, Kraaijenbrink T, Vermaat M, Mei H, van 't Hof P, et al. A SNP panel for identification of DNA and RNA specimens. BMC Genomics. Jan 25, 2018;19(1):90. [FREE Full text] [doi: 10.1186/s12864-018-4482-7] [Medline: 29370748]
77. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science. Jul 09, 2004;305(5681):183. [doi: 10.1126/science.1095019] [Medline: 15247459]
78. Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, et al. A multiplex assay with 52 single nucleotide polymorphisms for human identification. Electrophoresis. May 2006;27(9):1713-1724. [doi: 10.1002/elps.200500671] [Medline: 16586411]
79. Pakstis AJ, Speed WC, Fang R, Hyland FC, Furtado MR, Kidd JR, et al. SNPs for a universal individual identification panel. Hum Genet. Mar 2010;127(3):315-324. [doi: 10.1007/s00439-009-0771-1] [Medline: 19937056]
80. Kling D, Phillips C, Kennett D, Tillmar A. Investigative genetic genealogy: current methods, knowledge and practice. Forensic Sci Int Genet. May 2021;52:102474. [FREE Full text] [doi: 10.1016/j.fsigen.2021.102474] [Medline: 33592389]
81. White JD, Indencleef K, Naqvi S, Eller RJ, Hoskens H, Roosenboom J, et al. Insights into the genetic architecture of the human face. Nat Genet. Jan 2021;53(1):45-53. [FREE Full text] [doi: 10.1038/s41588-020-00741-7] [Medline: 33288918]
82. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. Feb 12, 2015;518(7538):197-206. [FREE Full text] [doi: 10.1038/nature14177] [Medline: 25673413]
83. Dabas P, Jain S, Khajuria H, Nayak BP. Forensic DNA phenotyping: inferring phenotypic traits from crime scene DNA. J Forensic Leg Med. May 2022;88:102351. [doi: 10.1016/j.jflm.2022.102351] [Medline: 35427851]
84. Regalado A. More than 26 million people have taken an at-home ancestry test. MIT Technology Review. Feb 11, 2019. URL: https://www.technologyreview.com/2019/02/11/103446/more-than-26-million-people-have-taken-an-at-home-ancestry-test/ [accessed 2024-04-05]
85. Church GM. The personal genome project. Mol Syst Biol. 2005;1:2005.0030. [FREE Full text] [doi: 10.1038/msb4100040] [Medline: 16729065]
86. Greshake B, Bayer PE, Rausch H, Reda J. openSNP--a crowdsourced web resource for personal genomics. PLoS One. 2014;9(3):e89204. [FREE Full text] [doi: 10.1371/journal.pone.0089204] [Medline: 24647222]
87. Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res. Jan 01, 2000;28(1):352-355. [FREE Full text] [doi: 10.1093/nar/28.1.352] [Medline: 10592272]
88. Fairley S, Lowy-Gallego E, Perry E, Flicek P. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. Nucleic Acids Res. Jan 08, 2020;48(D1):D941-D947. [FREE Full text] [doi: 10.1093/nar/gkz836] [Medline: 31584097]
89. International HapMap Consortium. A haplotype map of the human genome. Nature. Oct 27, 2005;437(7063):1299-1320. [FREE Full text] [doi: 10.1038/nature04226] [Medline: 16255080]
90. Edge MD, Algee-Hewitt BF, Pemberton TJ, Li JZ, Rosenberg NA. Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. Proc Natl Acad Sci U S A. May 30, 2017;114(22):5671-5676. [FREE Full text] [doi: 10.1073/pnas.1619944114] [Medline: 28507140]
91. He Z, Yu J, Li J, Han Q, Luo G, Li Y. Inference attacks and controls on genotypes and phenotypes for individual genomic data. IEEE/ACM Trans Comput Biol Bioinform. 2020;17(3):930-937. [doi: 10.1109/TCBB.2018.2810180] [Medline: 29994587]
92. Humbert M, Huguenin K, Hugonot J, Ayday E, Hubaux JP. De-anonymizing genomic databases using phenotypic traits. Proc Priv Enhanc Technol. 2015;2015:99-114. [FREE Full text] [doi: 10.1515/popets-2015-0020]
93. Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan O, et al. Identification of individuals by trait prediction using whole-genome sequencing data. Proc Natl Acad Sci U S A. Sep 19, 2017;114(38):10166-10171. [FREE Full text] [doi: 10.1073/pnas.1711125114] [Medline: 28874526]
94. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. Apr 17, 2008;452(7189):872-876. [doi: 10.1038/nature06884] [Medline: 18421352]
95. Sero D, Zaidi A, Li J, White JD, Zarzar TB, Marazita ML, et al. Facial recognition from DNA using face-to-DNA classifiers. Nat Commun. Jun 11, 2019;10(1):2557. [FREE Full text] [doi: 10.1038/s41467-019-10617-y] [Medline: 31186421]
96. Wang Y, Wu X, Shi X. Using aggregate human genome data for individual identification. In: Proceedings of the 2013 IEEE International Conference on Bioinformatics and Biomedicine. 2013. Presented at: BIBM '13; December 18-21, 2013:410-415; Shanghai, China. URL: https://ieeexplore.ieee.org/abstract/document/6732527
97. Erlich Y. Major flaws in “identification of individuals by trait prediction using whole-genome sequencing data”. bioRxiv. Preprint posted online September 7, 2017. [FREE Full text] [doi: 10.1101/185330]
98. Venkatesaramani R, Malin BA, Vorobeychik Y. Re-identification of individuals in genomic datasets using public face images. Sci Adv. Nov 19, 2021;7(47):eabg3296. [FREE Full text] [doi: 10.1126/sciadv.abg3296] [Medline: 34788101]

99. Schneider PM, Prainsack B, Kayser M. The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry. Dtsch Arztebl Int. Dec 23, 2019;51-52(51-52):873-880. [FREE Full text] [doi: 10.3238/arztebl.2019.0873] [Medline: 31941575]
100. Emani PS, Gürsoy G, Miranker A, Gerstein MB. Assessing and mitigating privacy risk of sparse, noisy genotypes by local alignment to haplotype databases. bioRxiv. Preprint posted online August 30, 2022. [FREE Full text] [doi: 10.1101/2021.07.18.452853]
101. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. Feb 15, 2001;409(6822):860-921. [doi: 10.1038/35057062] [Medline: 11237011]
102. Wyner N, Barash M, McNevin D. Forensic autosomal short tandem repeats and their potential association with phenotype. Front Genet. 2020;11:884. [FREE Full text] [doi: 10.3389/fgene.2020.00884] [Medline: 32849844]
103. FBI. US Law Enforcement Resources: Biometrics and Fingerprints. Combined DNA Index System (CODIS). URL: https://tinyurl.com/3by74dhj [accessed 2023-03-29]
104. Gitschier J. Inferential genotyping of Y chromosomes in latter-day saints founders and comparison to Utah samples in the HapMap project. Am J Hum Genet. Feb 2009;84(2):251-258. [FREE Full text] [doi: 10.1016/j.ajhg.2009.01.018] [Medline: 19215731]
105. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. Jan 18, 2013;339(6117):321-324. [doi: 10.1126/science.1229566] [Medline: 23329047]
106. Kim J, Edge MD, Algee-Hewitt BF, Li JZ, Rosenberg NA. Statistical detection of relatives typed with disjoint forensic and biomedical loci. Cell. Oct 18, 2018;175(3):848-58.e6. [FREE Full text] [doi: 10.1016/j.cell.2018.09.008] [Medline: 30318150]
107. Edge MD, Coop G. Attacks on genetic privacy via uploads to genealogical databases. Elife. Jan 07, 2020;9:e51810. [FREE Full text] [doi: 10.7554/eLife.51810] [Medline: 31908268]
108. Ney P, Ceze L, Kohno T. Genotype extraction and false relative attacks: security risks to third-party genetic genealogy services beyond identity inference. In: Proceedings of the 2020 Network and Distributed Systems Security (NDSS) Symposium. 2020. Presented at: NDSS '20; February 23-26, 2020:1-15; San Diego, CA. URL: https://www.ndss-symposium.org/wp-content/uploads/2020/02/23049.pdf [doi: 10.14722/ndss.2020.23049]
109. Sweeney L, Abu A, Winn J. Identifying participants in the personal genome project by name. SSRN Journal. Preprint posted online April 29, 2013. [FREE Full text] [doi: 10.2139/ssrn.2257732]
110. Guerrini CJ, Wickenheiser RA, Bettinger B, McGuire AL, Fullerton SM. Four misconceptions about investigative genetic genealogy. J Law Biosci. 2021;8(1):lsab001. [FREE Full text] [doi: 10.1093/jlb/lsab001] [Medline: 33880184]
111. Craig DW, Goor RM, Wang Z, Paschall J, Ostell J, Feolo M, et al. Assessing and managing risk when sharing aggregate genetic variant data. Nat Rev Genet. Sep 16, 2011;12(10):730-736. [FREE Full text] [doi: 10.1038/nrg3067] [Medline: 21921928]
112. Im HK, Gamazon ER, Nicolae DL, Cox NJ. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. Am J Hum Genet. Apr 06, 2012;90(4):591-598. [FREE Full text] [doi: 10.1016/j.ajhg.2012.02.008] [Medline: 22463877]
113. Wang R, Li YF, Wang X, Tang H, Zhou 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. 2009. Presented at: CCS '09; November 9-13, 2009:534-544; Chicago, IL. URL: https://dl.acm.org/doi/10.1145/1653662.1653726 [doi: 10.1145/1653662.1653726]
114. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. Aug 29, 2008;4(8):e1000167. [FREE Full text] [doi: 10.1371/journal.pgen.1000167] [Medline: 18769715]
115. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. Nat Genet. Sep 2009;41(9):965-967. [doi: 10.1038/ng.436] [Medline: 19701190]
116. Braun R, Rowe W, Schaefer C, Zhang J, Buetow K. Needles in the haystack: identifying individuals present in pooled genomic data. PLoS Genet. Oct 2009;5(10):e1000668. [FREE Full text] [doi: 10.1371/journal.pgen.1000668] [Medline: 19798441]
117. Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, et al. Deterministic identification of specific individuals from GWAS results. Bioinformatics. Jun 01, 2015;31(11):1701-1707. [FREE Full text] [doi: 10.1093/bioinformatics/btv018] [Medline: 25630377]
118. Bu D, Wang X, Tang H. Haplotype-based membership inference from summary genomic data. Bioinformatics. Jul 12, 2021;37(Suppl_1):i161-i168. [FREE Full text] [doi: 10.1093/bioinformatics/btab305] [Medline: 34252973]
119. Almadhoun N, Ayday E, Ulusoy Ö. Inference attacks against differentially private query results from genomic datasets including dependent tuples. Bioinformatics. Jul 01, 2020;36(Suppl_1):i136-i145. [FREE Full text] [doi: 10.1093/bioinformatics/btaa475] [Medline: 32657411]
120. Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science. Oct 03, 2008;322(5898):44. [doi: 10.1126/science.322.5898.44b] [Medline: 18772394]

121. Visscher PM, Hill WG. The limits of individual identification from sample allele frequencies: theory and statistical analysis. PLoS Genet. Oct 2009;5(10):e1000628. [FREE Full text] [doi: 10.1371/journal.pgen.1000628] [Medline: 19798439]
122. Masca N, Burton PR, Sheehan NA. Participant identification in genetic association studies: improved methods and practical implications. Int J Epidemiol. Dec 2011;40(6):1629-1642. [FREE Full text] [doi: 10.1093/ije/dyr149] [Medline: 22158671]
123. Pardo R, Rafnsson W, Steinhorn G, Lavrov D, Lumley T, Probst C, et al. Privacy with good taste: a case study in quantifying privacy risks in genetic scores. arXiv. Preprint posted online August 26, 2022. [FREE Full text] [doi: 10.1007/978-3-031-25734-6_7]
124. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. Nov 20, 2019;20(1):246. [FREE Full text] [doi: 10.1186/s13059-019-1828-7] [Medline: 31747936]
125. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010;11(5):R52. [FREE Full text] [doi: 10.1186/gb-2010-11-5-r52] [Medline: 20482838]
126. Chen W, Hayward C, Wright AF, Hicks AA, Vitart V, Knott S, et al. Copy number variation across European populations. PLoS One. 2011;6(8):e23087. [FREE Full text] [doi: 10.1371/journal.pone.0023087] [Medline: 21829696]
127. Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, et al. The impact of structural variation on human gene expression. Nat Genet. May 3, 2017;49(5):692-699. [FREE Full text] [doi: 10.1038/ng.3834] [Medline: 28369037]
128. Ueki M, Takeshita H, Fujihara J, Kimura-Kataoka K, Iida R, Yasuda T. Simple screening method for copy number variations associated with physical features. Leg Med (Tokyo). Mar 2017;25:71-74. [doi: 10.1016/j.legalmed.2017.01.006] [Medline: 28457514]
129. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. Feb 2013;14(2):125-138. [doi: 10.1038/nrg3373] [Medline: 23329113]
130. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. Nat Rev Genet. Mar 2020;21(3):171-189. [FREE Full text] [doi: 10.1038/s41576-019-0180-9] [Medline: 31729472]
131. Shabani M, Marelli L. Re-identifiability of genomic data and the GDPR: assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. EMBO Rep. Jun 2019;20(6):e48316. [FREE Full text] [doi: 10.15252/embr.201948316] [Medline: 31126909]
132. Molnár-Gábor F, Korbel JO. Genomic data sharing in Europe is stumbling-Could a code of conduct prevent its fall? EMBO Mol Med. Mar 06, 2020;12(3):e11421. [FREE Full text] [doi: 10.15252/emmm.201911421] [Medline: 32072760]
133. Martinez-Martin N, Magnus D. Privacy and ethical challenges in next-generation sequencing. Expert Rev Precis Med Drug Dev. 2019;4(2):95-104. [FREE Full text] [doi: 10.1080/23808993.2019.1599685] [Medline: 32775691]
134. Clayton EW, Evans BJ, Hazel J, Rothstein MA. The law of genetic privacy: applications, implications, and limitations. J Law Biosci. 2019:1-36. [doi: 10.2139/ssrn.3384321]

Abbreviations

CNV: copy number variation
CODIS: Combined DNA Index System
DTC-GT: direct-to-consumer genetic testing
SGMF: Sorenson Molecular Genealogy Foundation
SNP: single nucleotide polymorphism
SNV: single nucleotide variant
STR: short tandem repeat
SV: structural variant
Y-STR: short tandem repeat on the Y chromosome

Edited by Z Yue; submitted 06.11.23; peer-reviewed by L Guo, J Lai; comments to author 01.01.24; revised version received 26.03.24; accepted 29.03.24; published 27.05.24

Please cite as:
Thomas M, Mackes N, Preuss-Dodhy A, Wieland T, Bundschus M
Assessing Privacy Vulnerabilities in Genetic Data Sets: Scoping Review
JMIR Bioinform Biotech 2024;5:e54332
URL: https://bioinform.jmir.org/2024/1/e54332
doi: 10.2196/54332
PMID: 38935957