Annual Review of Biomedical Data Science

Genome Privacy and Trust

Gamze Gürsoy[1,2]

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA;
email: gamze.gursoy@columbia.edu

[2]New York Genome Center, New York, NY, USA

Annu. Rev. Biomed. Data Sci. 2022. 5:163–81

First published as a Review in Advance on May 4, 2022

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-122120-021311

Keywords

genome privacy, information leakage, data sharing, cryptography

Abstract

Genomics data are important for advancing biomedical research, improving clinical care, and informing other disciplines such as forensics and genealogy. However, privacy concerns arise when genomic data are shared. In particular, the identifying nature of genetic information, its direct relationship to health status, and the potential financial harm and stigmatization posed to individuals and their blood relatives call for a survey of the privacy issues related to sharing genetic and related data and potential solutions to overcome these issues. In this work, we provide an overview of the importance of genomic privacy, the information gleaned from genomics data, the sources of potential private information leakages in genomics, and ways to preserve privacy while utilizing the genetic information in research. We discuss the relationship between trust in the scientific community and protecting privacy, illuminating a future roadmap for data sharing and study participation.

GENOME PRIVACY AS A GLOBAL CHALLENGE

In 1890, Samuel D. Warren and Louis Brandeis defined privacy as the “right to be let alone” in their *Harvard Law Review* article entitled “The Right to Privacy” as a response to the increase in printing technologies such as newspapers and photographs (1). Although the meaning of “let alone” was debated extensively, the most common interpretation was the right of a person to choose seclusion. As vague as it is, this concept was used as a basis for the design of the broad legal framework for privacy that has developed over the years. As part of this concept, genetic privacy concerns the protection of one’s genetic information.

Increase in Genomic Data Brings in New Privacy Challenges

The human genome research field has come a long way since the first sequencing of the human genome. There has been a rapid increase in technological advances, which has resulted in not only a decrease in the cost of sequencing but also an increase in high-throughput biochemical assays (e.g., transcriptomic and epigenomic assays) that produce large amounts of data based on next-generation sequencing (NGS) (2, 3). The resulting data are extremely valuable in research settings where scientists can link genome variations to function. With more genomes being sequenced from patients, we can make better associations between genetic information and diseases and traits. The increase in data enables researchers to use the genetic information in clinical settings and enables funding agencies to initiate large-scale research projects such as the All of Us Research Program to integrate genetic data with other health observations from participants (4). In the past decade, more initiatives, funded publicly or privately, are generating a vast amount of genetic and related data in order to mine the genetic determinants of diseases, traits, and conditions (5–9). Moreover, direct-to-consumer (DTC) genetic testing companies are collecting a wealth of genetic data in order to provide health- or ancestry-related information to their customers. As genome scientists, we live in this exciting era of genomics data science, where we can characterize the human anatomy at different molecular levels using a mountain of genetic and related data. However, this immense personal data collection presents unique challenges, some of which are understudied and must be addressed. The three big challenges we face are the analysis, storage, and privacy of the collected data. The most obvious challenge is how we are going to analyze this vast amount of data to make meaningful biological and medical inferences. With the introduction of artificial intelligence and other computational advances in biology and healthcare research, we are increasingly getting better at addressing the analysis challenge. However, the slow pace in addressing the challenges related to secure storage and privacy of genetic and related data is threatening our progress in the analysis space.

As genomics becomes a bigger part of clinical practice, we will see more data being stored in public repositories that can be trained and used for surveillance purposes. Moreover, we already know that there is bias in the current genomic data ecosystem. Although the National Institutes of Health (NIH) have established guidelines for the diversity of collected genomic data, there are still large disparities in current data collection efforts. The amount of genomic data we have from certain demographic groups overwhelmingly outnumbers that from other demographics. This issue also partly relates to the privacy concerns of different populations. For example, due to the scarcity of genomic data from Indigenous and Native American individuals, the broad sharing of their genetic data will be more invasive to privacy. Moreover, a vast amount of genomic data from European individuals have been collected and broadly shared, rendering this population more prone to reidentification through forensic investigations conducted by law enforcement (10).

In the United States, one health privacy protection, the Health Insurance Portability and Accountability Act (HIPAA), was established in the late 1990s. HIPAA regulates the privacy of medical health records, including the anonymization of genetic testing. As genetic testing has become more widespread, the US Congress passed the Genetic Information Nondiscrimination Act (GINA) in 2008. GINA aims to protect individuals from discrimination by insurance companies and employers based on their genetic information. With the recent debate on whether genetic information can truly be anonymized based on HIPAA guidelines, genome privacy has increasingly become a major focus among ethicists, legal scholars, geneticists, and computer scientists.

What is Genetic Information?

Genetic information is highly personal and identifying. Although *Homo sapiens* share a large portion of their genome, the small number of differences are extremely unique and identifying (11). These changes can take different forms (**Figure 1**). For example, a single-nucleotide polymorphism (SNP) is a single base change in an individual's genome. Insertions and deletions (indels) are single or a few base pair insertions or deletions in the genome. Most SNPs and indels are common and shared across individuals. However, some can be rare or unique (i.e., de novo SNPs), which makes them extremely identifying. In addition, some can be population specific (e.g., a variation specific to European or African genomes). SNPs and indels (or combinations of different SNPs or indels) can be biomarkers for many human traits such as eye color, hair color, height, and

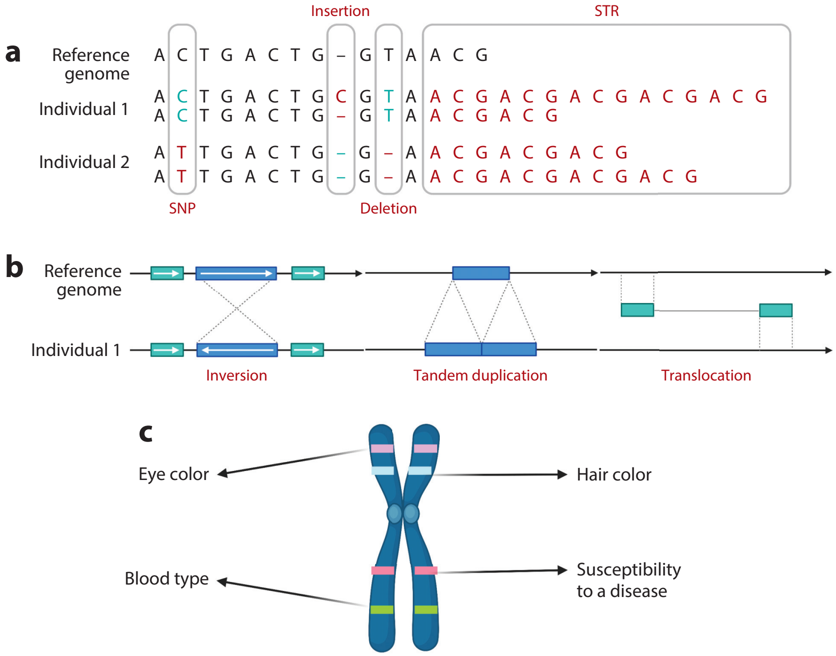
a
Insertion
STR
Reference genome
A C T G A C T G – G T A A C G
Individual 1
A C T G A C T G C G T A A C G A C G A C G A C G A C G
A C T G A C T G – G T A A C G A C G
Individual 2
A T T G A C T G – G – A A C G A C G A C G
A T T G A C T G – G – A A C G A C G A C G A C G
SNP
Deletion
b
Reference genome
Individual 1
Inversion
Tandem duplication
Translocation
c
Eye color
Hair color
Blood type
Susceptibility to a disease

Figure 1

(*a*) Illustration of the most abundant and identifying genomic variations: single-nucleotide polymorphisms (SNPs), insertions and deletions (indels), and short tandem repeats (STRs). Every individual carries two copies of DNA, one from maternal and one from paternal genomes (called haplotypes). The difference on the DNA sequence with respect to a reference genome and whether it is on the maternal copy only, the paternal copy only, or both are genomic variations that make every genome unique and, thus, identifying. (*b*) Examples of structural variants. (*c*) Examples of genomic variations on a chromosome and which trait they might affect. Genomic variation on a specific locus or combinations of variations on different specific loci might be responsible for different traits and diseases. Figure adapted from images created with BioRender.com.

susceptibility to disease. The genome also contains short tandem repeats (STRs), which are short units of DNA sequences repeated at different times in different individuals. These are used by law enforcement agencies for DNA profiling and sometimes in paternity testing. Structural variants (SVs) are larger changes in the genome, such as large insertions, deletions, and inversions. SVs are less abundant and less commonly used for identification or DNA profiling but might be disease related. The combination of all of these variations in an individual's genome can be analogous to that person's Social Security number (SSN): They might be meaningless sequences of letters by themselves but can reveal a dangerous wealth of private information specific to the individual. However, one can easily change their SSN in case of a data breach but cannot change their genetics. Moreover, genetic information is largely shared with close family members. Therefore, the personal genetic data implicate not only the immediate owner but many third-party relatives as well. Such power of genetic information has led to its different uses in forensics, creating additional concerns over privacy (12).

Interplay Between Policy and Technology to Enable Genomic Data Sharing

Data sharing among researchers, institutions, and even continents is essential for biomedical research. In particular, genetic data sharing can increase the statistical power for rare genetic disorders and can save lives (13). Indeed, the international scientific community recognizes that broad data sharing is a prerequisite for many important integrative analyses. However, the extremely identifying nature of genetic data brings global challenges. For example, the regulations that govern the protection of genetic data radically differ by jurisdiction, thus severely hampering data sharing. Even fundamental expectations and definitions of privacy differ between US frameworks (14) and European frameworks such as the General Data Protection Regulation (15). In compliance with this patchwork of regulations, genetic data are increasingly siloed into more restrictive archives that may constrain data access to the physical boundaries of a jurisdiction, even though the data are generated without regard to jurisdiction. However, there are serious privacy-related concerns that form the basis of these different jurisdictions and policies, which we expand upon throughout this review. Studies have shown that anonymized data can be linked to the identities of the donors; hence, sensitive phenotypes of known individuals can be revealed (16). Without protective legislation, this issue will have monetary consequences, such as discrimination during employment. It can also create harm through the stigmatization of medical conditions and health statuses. The use of genetic data can also lead to racial profiling and bias (17). Moreover, improved technology, increased data availability, and a deeper understanding of genotype-to-phenotype associations make it easier to identify links between datasets, thus demanding a careful and dynamic reevaluation of our understanding of genetic privacy.

Data privacy and security are traditionally studied under the umbrella of computer science, law, and ethics. However, maintaining the privacy and security of biological data while preserving their utility differs from that of other data types, requiring a deep understanding of how biological systems work, how biotechnological advances (e.g., omics data generation) produce new data, and how existing bioinformatics file formats and tools operate on the data. Recent advances in biotechnology, increases in scalable computing, and the immense amount of large-scale biological data at the population level provide the perfect opportunity for integrating multidimensional information to study human variation and disease. However, these efforts must be complemented with interdisciplinary privacy research fueled by an understanding of the biology behind the data in order to develop methods and file formats with high utility. Given the interconnected nature of data privacy and utility, bridging these diverse fields will be imperative to maximizing the utility of biological data while preserving patient privacy.

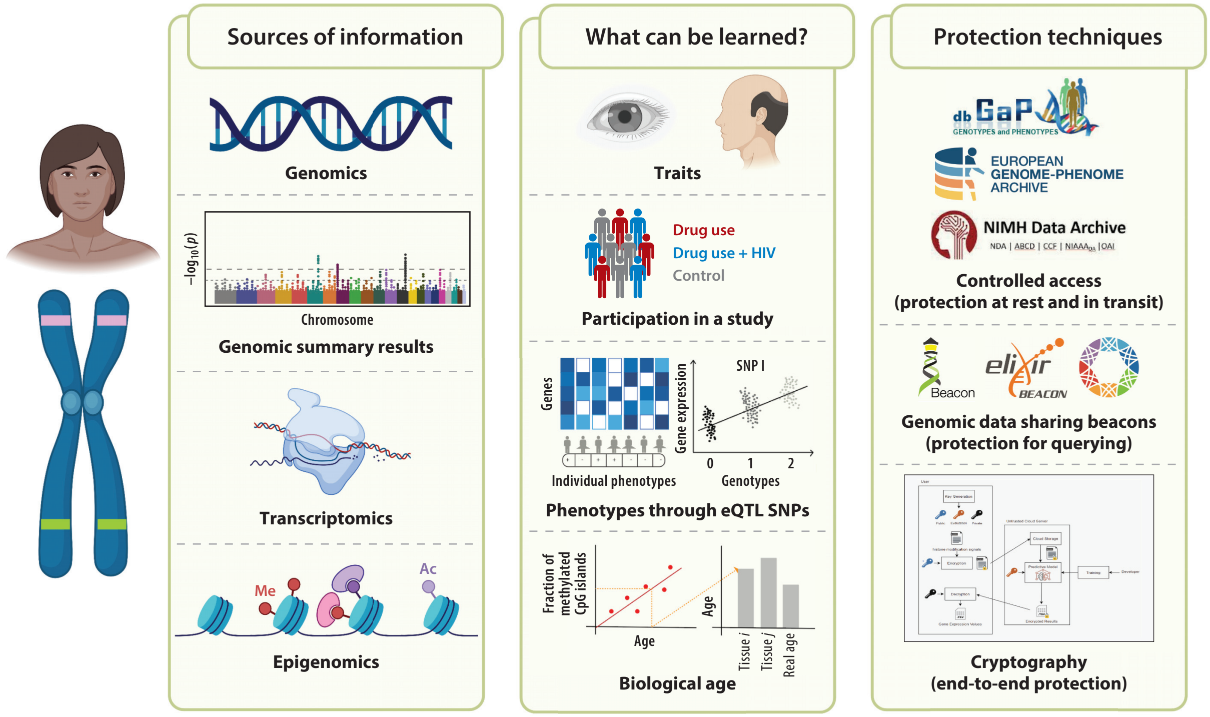

Figure 2

Where genetic information can be obtained, what can be learned from it, and how we can protect it. Abbreviations: eQTL, expression quantitative trait locus; SNP, single-nucleotide polymorphism. Parts of figure adapted from images created with BioRender.com; middle column plots adapted from Reference 62.

As "privacy" is often an ill-defined term itself, this review is organized as a walk-through of the important aspects of genetic privacy to give the full picture on why we do and should care about privacy in biomedical research (**Figure 2**). We start by describing what kind of information can be revealed about an individual from their genetic data. We then talk about the obvious and hidden sources of genetic data leakages in research, clinical, and commercial settings. We discuss how we can prevent this leakage using technology and policy and argue that the current one-size-fits-all types of protections are too stringent for some data types and too lenient for others. Finally, we conclude our review by describing the relationship between privacy protection and public trust, sketching a roadmap of practical solutions that might simultaneously benefit open science and restore the public trust in research participation.

WHAT CAN BE LEARNED FROM AND WITH GENETIC DATA?

When it comes to the information that can be gleaned from genetic data, various parties are relevant to the discussion. There are legal responsibilities defined by laws and policies, ethical considerations that take into account the risk–benefit balance in data sharing, and technical ways to both breach privacy and prevent these breaches. These three aspects of the information inferred from genetic data do not necessarily overlap. In this review, we focus on this problem from a technical perspective and direct readers to other valuable studies for the ethical and legal considerations (18–27).

Data deidentification, such as the removal of identifiers mandated by HIPAA (28), have been shown not to be effective for genetic data (14, 16). There are several ways to breach the privacy of individuals using genetic information, namely, completion attacks, identity tracing attacks, and attribute disclosure attacks (14, 16). In a nutshell, a completion attack relies on using a partial genetic variant list of a known individual to impute more genetic variants using the nonrandom association between the alleles of multiple genetic loci in a given population, also known as linkage disequilibrium (LD) (29). Identity tracing attacks start with an unknown genome and aim to identify the person to which the genome belongs using various auxiliary information such as genealogical information (10, 30). In attribute disclosure attacks, one can use partial, full, or summary-level genetic information about a known individual to discover an attribute about that person (e.g., participation in a drug abuse study). These attacks and countermeasures have been extensively surveyed by other genetic privacy perspectives and review papers (16, 31).

Since the jargon about these attacks and how they penetrate privacy can be confusing for researchers outside this field, here we delineate the information that can be learned through genetic data leakage in categories for clarity (**Figure 2**).

Full Genetic Information

The genome of an organism is like a manual. It contains the information for the cells and tissues to grow and develop. As a result, our traits and phenotypes have direct or indirect links to our genomes. Although there are environmental factors at play, there is a wealth of information that can be garnered from having access to an individual's genome. For example, ancestry, disposition to genetic diseases, and physical traits such as height and eye color are among the information that can be directly or indirectly inferred from genetic data.

Access to a known person's full genetic information (if not voluntary) is the most obvious information leakage. The leak of such data can cause monetary harm such as the denial of life insurance. Note that discrimination in health insurance coverage and employment based on genetic information has been protected against by GINA in the United States (32). The leak of these data might also cause harm in the form of stigmatizing people with dispositions to certain health conditions such as mental health problems. Moreover, there could be indirect inferencing of stigmatizing phenotypes that are not directly related to one's genetic information. For example, a patient's genome might be sequenced under a large-scale "anonymized" HIV database. If an adversary has access to the identity and the genome of an individual, they can easily search the genome in this database and learn whether this person has HIV (**Figure 3*a***).

With the current protections in place, it is fairly difficult to access genetic information of a known individual. However, studies have shown that "anonymized" genomes, i.e., genomes of unknown individuals, can be easily deanonymized if the individual deposited their genome into a genealogy database (30). It was also shown that the metadata of the anonymized genomes can be overlapped with other datasets (such as the US Census) to deanonymize and reveal private information about the individuals (33). Moreover, unknown genomes sequenced from the evidence in crime scenes can be used to search for genetic relatives of the culprit in genealogical databases, such as in the famous case of the Golden State Killer (12). In short, the DNA sample obtained from a victim was sequenced and searched against public and private genome databases to find a matching relative of the culprit. Furthermore, recent discussions have centered on whether facial images can be reconstructed using genomes, sparking further privacy concerns (34–36).

Partial Genetic Information

Leakage of partial genetic information of a known individual can be as dangerous as leakage of full genetic information (**Figure 3*b***). This is mostly because there are strong correlations between the

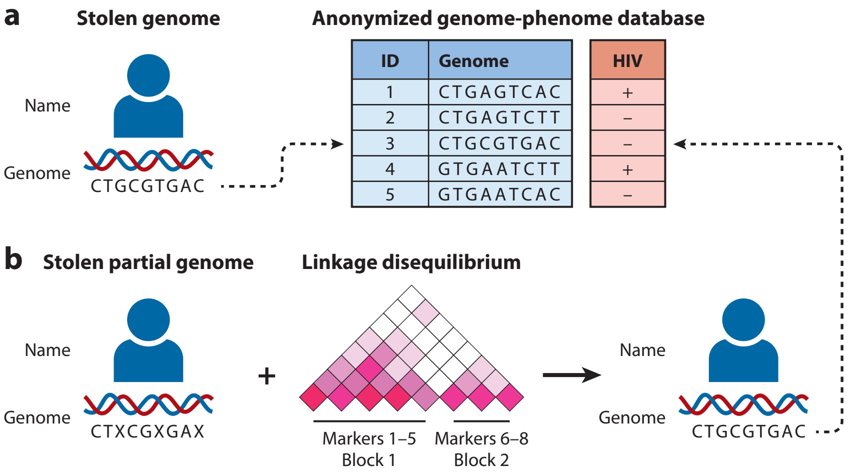

Figure 3

(*a*) An anonymized dataset can be deanonymized by using the genome of an individual, which can serve as a unique identifier. (*b*) Partial genomes can be completed using linkage disequilibrium information inferred from genomes of the population to which the individual belongs. Completed genomes can be further used to deanonymize sensitive datasets. Figure adapted from images created with BioRender.com.

genetic variants residing on the same haplotype block. A haplotype block is a region in the genome with little evidence of recombination; thus, the genetic variants on the same block are typically inherited together (37). Alleles in a haplotype block are in LD. Given partial genetic information of an individual, one can use the principles of genotype imputation (38–41) and complete the full genome using many available and highly accurate imputation tools (42–45). For example, in 2008, Dr. James Watson's genome was sequenced and released in a public database (46). Per Dr. Watson's request, all gene information related to apolipoprotein E (encoded by *APOE*, a gene implicated in Alzheimer's disease and cardiovascular disease) was masked from the sequence. However, it was later communicated that the risk status of this gene can easily be inferred using available LD information between the *APOE* risk alleles and other alleles present in the released genome (47). Moreover, just like in the case of full genetic information, partial genetic information can be used to query datasets to either deanonymize the genomes or mine private information about a known individual.

Summary-Level Genetic Information

Genomic summary results (GSRs) are by far the most controversial data type in terms of broad sharing. The privacy discussions surrounding GSR data have resulted in the NIH changing their data sharing policy multiple times in the past 20 years (48–50). GSRs are the results of genome-wide association studies (GWAS) performed on genomes from a cohort of individuals within a single study or across multiple studies. In particular, GSRs represent two types of information: the allele frequency of a variant in the cohort and the statistical association between genotype and phenotype, such as p-values, beta values in regression, odds ratios, and effect sizes (51).

At first glance, one might think that GSRs do not contain any individual-level data; therefore, sharing them would not violate the privacy of the cohort individuals. However, several studies have shown that GSRs can be combined with genotypes from known individuals to reveal participation in potentially stigmatizing cohort studies (52–55). For example, in their seminal work, Homer et al. (52) showed that one can accurately predict whether a genome of an individual is present in a mixture of genomes by comparing the allele frequencies of SNPs in a reference population and in the mixture. It is important to point out that this study was intended for applications in forensic science, but it raised the issue of potential inferring whether an individual participated in a GWAS. Furthermore, Sankararaman et al. (53) developed a statistical quantification method that challenges the limits of detection of individuals in GWAS in an effort to provide GSR sharing guidelines for researchers. In 2012, another study developed a statistical membership inference method to show that *p*-values and regression coefficients that are typically shared in GSRs can reveal the participation of an individual in a study and their private phenotypes (55).

Recently, Global Alliance for Genomics and Health's genomic data sharing beacons (56) have been proposed as another way of sharing summary-level genetic data, which have been shown to leak private information about the cohort individuals (56–58). After the seminal work by Homer et al. (52) in 2008, the NIH and other institutions changed their GSR data sharing policies to place the data behind controlled access. Although it is still heavily debated whether the risk to privacy in sharing GSRs broadly is theoretical or practical, the NIH revised their policy in 2018, releasing the aggregate data from controlled access. This decision was not made lightly, but considered findings from public workshops, stakeholder comments, and accumulated public input over the years. In the meantime, researchers have developed a myriad of technological solutions using cryptographic techniques for performing GWAS and releasing GSRs.

Machine learning models that are trained on genetic information from research participants or patients are another type of summary-level data that might leak genetic information. Although the model parameters do not directly summarize the data, as with GSRs, studies have shown that they carry enough information to determine whether data from an individual are in the training set (59, 60). Moreover, these attacks can be utilized to infer information about the genomes in a training dataset of generative models used for creating synthetic genomic data (61).

Proxy-Level Genetic Information

By "proxy-level genetic information," we refer to the molecular measurements that have associations with or relations to genetic variants. These measurements are not necessarily made for the purpose of inferring genetic variants, but because they are correlated with genetic variants, they might contain cryptic private information about the individuals from which the measurements are taken. Proxy-level genetic information is probably the hardest information leakage to pinpoint and sometimes requires different levels of sophisticated statistical inference. It is also not clear if the theoretical risk to privacy can or will translate into practical risk moving forward. Therefore, this level of privacy risk is also the least studied. Functional genomics data, such as from RNA sequencing (RNA-seq) and chromatin immunoprecipitation and sequencing (ChIP-seq), are the most studied molecular measurements known to leak private information and have been extensively reviewed (62).

There are various ways that these molecular measurements leak private information. NGS-based functional genomics assays such as RNA-seq or ChIP-seq contain direct genotype information of individuals (63, 64), and, therefore, we often share the raw data behind controlled access (**Figure 4**). One of the most common ways to analyze NGS-based functional genomics measurements is to examine the signal tracks and identify highly active regions with peaks (i.e., regions that overlap with sequenced reads). Signal tracks are often deemed to be safe to share in terms of private information leakage, but recent work has shown that deletions can be easily inferred by considering the lack of signal in these tracks (65).

Some of the most summarized forms of RNA-seq data are gene expression values, which are broadly shared. However, it has been shown that using the correlation between the publicly

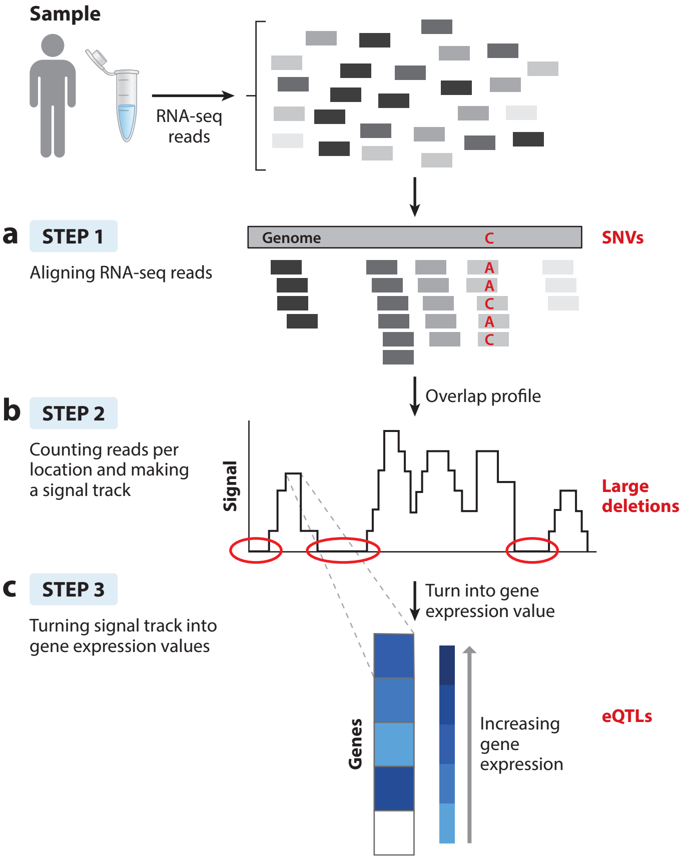

Figure 4

Progressive summarization of RNA-seq data and the leakage of genetic variants at each summarization level. RNA-seq data can be a proxy for genetic variants through (*a*) NGS reads obtained from the sequences, (*b*) signal profiles obtained from the sequencing coverage per gene, and (*c*) inferences of eQTLs that belong to the sample. Figure adapted from Reference 62. Abbreviations: eQTLs, expression quantitative trait loci; NGS, next-generation sequencing; RNA-seq, RNA sequencing; SNVs, single-nucleotide variants.

available expression quantitative trait loci, genetic variations whose genotype correlates with expression of the gene, and the expression values of an individual's genes, one can easily predict the private genotypes of the individual (66, 67) (**Figure 4**). Moreover, even lists of allele-specific genes have been shown to leak various genetic variants of the research participants (68). Summary-level DNA methylation data have also been shown to leak genetic variants or private disease information (69, 70). In general, a rule of thumb in assessing these molecular measurements is that if at least 20 independent SNP genotypes can be inferred from the measurements, then it is likely that one can use these measurements to reidentify individuals or infer private phenotypes about them (11).

Summary

It is almost impossible to precisely quantify the amount of genetic information leakage in a dataset. This is because there can always be a correlation between an observation and genetics that has not yet been discovered. Another reason, and probably the most problematic, is that multiple datasets can be combined to infer more private information about an individual, and we have very little idea of what kind of data types, datasets, or measurements will be available in the future. A dataset that is privacy-attack-proof today could be revealing too much genetic information tomorrow. Overall, the inferred genetic information can be used in many ways that penetrate privacy, such as tracing the identity of an individual, recovering an individual's full genome, or recovering phenotypic information about an individual (16).

WHAT ARE THE SOURCES OF GENETIC DATA LEAKAGES?

Access to genomics data by scientists is immensely important and should be made as easy as possible. However, if this access infrastructure is not secure enough, then the data can be obtained by parties for purposes that violate and harm the patients or the research participants. Here, we describe how such adversarial access can be possible.

Forensic Databases

All over the world, there are DNA databases that are maintained by governments. In the United States, all states are required to collect DNA samples from individuals who are convicted of certain crimes (71), and many states and the federal government collect DNA samples from all individuals arrested (not convicted!) for various offenses. Some of these collected DNA data are incorporated into the national Combined DNA Index System (CODIS) database. Although the CODIS database is composed of 13 STR loci that are very unique and identifying, there are also correlations between these STRs and more medically meaningful genetic variants (72, 73). This means that access to forensic databases not only will result in identity tracing but also will reveal potentially private or stigmatizing phenotypes about the individuals. Moreover, due to the selective bias in collecting the samples for these databases, there is more risk of privacy breaches for certain communities, which start with racial profiling and lead to even more racial profiling (74–76).

Genealogy Databases

In the past 20 years, genetic genealogy has become increasingly popular as the cost of sequencing has decreased. Several genetic genealogy services (e.g., GEDmatch and FamilyTreeDNA) enable matchmaking among users' uploaded genomic data to identify genetic relatives. Oftentimes, the uploaded genetic data are accompanied by the identities and sometimes the addresses of the users. As was the case with the famous Golden State Killer case (12), law enforcement agencies have increasingly been using these services to identify the relatives of culprits (77). It was indeed shown that about 60% of searches for individuals of European ancestry in the United States through these services will return a third-cousin or closer match (10). The privacy implications of law enforcement using these searches to solve cold cases have been extensively discussed (78, 79). Moreover, some of these genetic genealogy servers such as GEDmatch have insecure underlying infrastructure and information dissemination systems such that falsified uploads can be used to obtain private genetic information about the individuals in the database (80).

Direct-to-Consumer Genetic Services

As the interest in genetic genealogy increases, an increasing number of companies are offering DTC tests for a variety of purposes ranging from genetic ancestry inference to linking genetic variants to disease risk. The way that these services work is by collecting biosamples (most commonly saliva) from customers and returning a written report through a web application in return for a fee. There are many privacy-related issues concerning how DTC services work. Similar to genetic genealogy services, most of these companies allow law enforcement to use their databases for searching genetic relatives of suspects (81). Another issue arises regarding data ownership. Since the companies own the data they generate from customers' DNA, they also hold the right to sell the data as they see profitable. Some of these companies also biobank the collected specimen for a long time, risking the physical security of the samples and creating an unknown sense of the future for the privacy of the data. The legal problems in providing privacy and how solutions fall short of providing required protections have also been discussed extensively (82).

Cloud Computing and Shared Servers

Unintentional data leakages might happen while researchers compute on the data in cloud computing settings or in shared servers. First and foremost, the genetic data that are not deidentified should not be stored and computed on cloud servers or in high-performance computing settings without a trusted system between the user and service provider in place. However, considering how impossible it is to fully deidentify genetic data, genetic data are vulnerable in cloud and shared server settings even in a deidentified form.

Attacks on Healthcare and Related Data

There is always risk for adversaries to attack databases in institutions or hospitals that contain genetic data for monetary gain. Many of these attacks take advantage of vulnerabilities such as network endpoints (i.e., laptops or tablets that are not properly secured). These attacks can especially be easier for the data in transit, as moving data from one place to another creates more opportunities for adversaries to access the data through unsecure endpoints if proper protections are not in place (83).

Surreptitious DNA Sequencing

When an adversary has access to a discarded item that contains a known individual's genetic material without the consent of that individual and sequences the genetic information on the sample, they can obtain almost the entire personal DNA sequence of the individual depending on the sample (63). Such information can be used for malicious purposes, and currently there are no laws protecting against surreptitious DNA sequencing in the United States (22).

HOW CAN WE PROTECT GENETIC DATA?

Just like any other data, genetic and related data can take three forms: data in transit, data at rest, and data in use. Many of the privacy policies surrounding genetic and related data focus on the protection of data in transit and data at rest and do not provide guidelines for data in use. This is because it is a much harder problem to solve. Moreover, traditionally it was expected that researchers would compute private data locally without using shared servers or cloud services. However, this expectation is no longer realistic, as the amount of data and the resources needed for the complexity of the analyses can no longer be addressed with local computational power.

Shortcomings of the Current System

In research settings, we face three genetic data sharing models (84): (*a*) Generated data cannot be shared due to a lack of consent or proprietary reasons, (*b*) the data behind a firewall can be shared to authorized users due to terms of the consent (controlled access), and (*c*) the data can be broadly shared without any limits.

The first model raises important questions related to the reliability of the published results, reproducibility, and democratized access to research results. If there is no consent by the research participant or the patient, the data, of course, should not be shared with any parties. But if the data are not shared due to proprietary reasons, but consent from the participants is provided, then there are ways to compute the data without breaching confidentiality using cryptographic techniques (85).

The second model, also known as controlled access, is an audit-based protection. The users apply for access with a justification on how they will use the data. They agree not to use the data for purposes other than the justification they promised. A data-access committee reviews these applications and grants access to the users. The NIH Database of Genotypes and Phenotypes (86), National Institute of Mental Health's Data Archive (87), and the European Genome-Phenome Archive (88) are commonly used controlled-access databases. Historically, the expectation from the authorized users is to download the encrypted data using secured channels provided by these institutions, decrypt the data with the provided key, compute the data locally, and then delete the sensitive data from the local computers. As the analyses become more computationally intensive and data become larger, funding agencies are moving toward cloud-based solutions. For example, a wealth of controlled-access genetic data can be accessed through the National Human Genome Research Institute's Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL), which is a cloud-based infrastructure for genomic data access, sharing, and computing across large genomic and genomic-related datasets, as long as the user is authorized to access the dataset of interest. There are two important issues related to this system from both a utility and a privacy perspective. A controlled-access model puts a burden on researchers and delays scientific progress. First of all, researchers must choose the best data types for their analysis, mostly based on poorly developed metadata that often do not follow a standardized structure. Once they identify a dataset, they draft an application and submit it to the data use committee. This is often done via signing officials at the institutions. This application step can take a significant amount of time, as small institutions with limited resources may not have an adequate number of signing officials. After a months-long application period, the researchers wait for the data use committee to review their applications and grant access, which also can take up to several months depending on the access-granting institution.

After access is granted, the data can be downloaded locally for analysis or the computation can take place in a dedicated cloud computing environment such as AnVIL. This is when privacy problems arise. Research labs usually have a certified secure partition within their local high-performance computing system to store the sensitive data for which they obtained authorization. However, oftentimes, junior members of these labs are not adequately trained on how to handle sensitive data, and they might unintentionally misplace the data into servers or computers (89). Although there are serious consequences if identified, these behaviors often go unnoticed because there are no reliable audit mechanisms in place once the data leave the access-granting institution. There are no imposed actions but rather best practices to safeguard the sensitive data to which researchers were granted access (90). The issues with storing and analyzing the controlled-access data in the cloud environment are, as described above, the control of the data by third parties and the potential collusion of the information during data analysis.

Innovative Solutions

It is about time that the human genetics community recognizes the sensitive nature of genetic data and gets inspired by the industry-scale cybersecurity tagline *Never decrypt!* Genetic data need to be protected at all times: at rest, in transit, and in use. However, cryptographic frameworks that can provide such properties have been deemed theoretical and not practical or scalable. With increasing advances in algorithm designs and computing power, this view has started to change in other sectors and recently in genomics. Many of these emerging technologies and their applications in genomics have been discussed in other reviews (62, 89, 91).

Funding agencies such as the National Human Genome Research Institute have recognized this need and have been funding initiatives such as the iDASH Genomic Privacy Challenges (92), where interdisciplinary scientists come together to solve problems in genomics using cryptographic techniques. In these challenges, researchers have demonstrated that differential privacy can be implemented to protect queries on genomics data in a scalable manner (93). Efforts have also shown that secure GWAS is possible using a technique called homomorphic encryption, which enables the data to be encrypted in transit, in analysis, and at rest (94). Many other studies have also shown that secure GWAS is possible even when the data are distributed across different sites (95–97). A traditional genomics tool, genotype imputation, was also shown to be feasible using different cryptographic techniques (98–100). Recently, iDASH challenges have started to include tracks related to blockchain-based solutions, which are the first applications of blockchain technology in genome privacy and security (101–105).

TRUST

As scientists, an important principle in designing a research study is to balance the risk and benefit to the research participants (106). However, this ideally requires a quantitative way of measuring the risks and benefits. When it comes to collecting genetic information from participants, the risk in terms of privacy cannot be objectively quantified for many reasons: (*a*) We do not have a way of knowing if the participants have given their genetic and related data to other parties, (*b*) we cannot oversee what kind of potential data leakages will arise in the future, and (*c*) we cannot anticipate potential policy and legislative changes in data sharing and protection in the future. Moreover, disentangling risk from its historical and political context will also result in underestimating the harm (107). We can help biomedical science move forward and save lives only if we can recruit participants to the studies. However, since we cannot quantify the risk, the system must be built on trust. Trust can only be gained if we actively seek new technologies and innovations to minimize the risk instead of pushing the status quo because it is more convenient. Trust will also stem from active efforts to bring the benefits to everyone (**Figure 5**).

Research participants are humans, not just subjects. In 1948, the United Nations issued the Universal Declaration of Human Rights (108), which defines privacy as a human right (109). In working in a field that seeks to help human life, it is dangerous for biomedical researchers to feel entitled to data, considering the harm is almost always disproportionately distributed toward disadvantaged groups. We can never anticipate how such data will be used in the future; therefore, we must take extra precautions to protect research participants' data. For example, recent news that China has used publicly and privately collected genetic data to trace minorities is a great example of an unanticipated yet harmful use of genetic data (110). Moreover, it is extremely important to think about genetic data sharing in the historical context. Trust is not given; it is gained. Exploitative and discriminatory practices in the history of biomedical research will always be importantly relevant when individuals from marginalized groups decide to participate in genetic research (111).

The path forward in protecting genetic privacy is a legal, ethical, and technological ecosystem that involves all of the stakeholders. If the goal in biosciences is to reduce human suffering, then this ecosystem cannot be only based on what is convenient for the scientists. If history has taught us anything related to discrimination, then we must start taking proactive rather than reactive measures to protect genetic privacy given the sensitive nature of genetic information. This involves

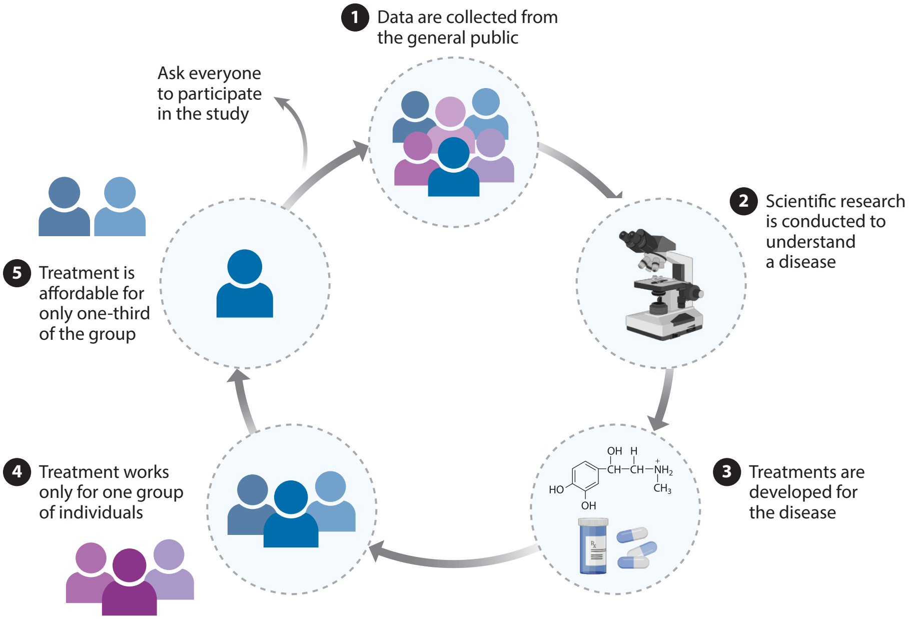
1 Data are collected from the general public
2 Scientific research is conducted to understand a disease
3 Treatments are developed for the disease
4 Treatment works only for one group of individuals
5 Treatment is affordable for only one-third of the group
Ask everyone to participate in the study
OH H
CH CH NH2
HO
CH3
OH

Figure 5

Studies involving genetic data collection efforts pose significant privacy risks to participants. However, the benefits of these studies go to a tiny fraction of the participants. Keeping this system in place increasingly risks the recruitment of participants to future studies. This cycle can be broken by placing appropriate precautions to secure the data and actively working to remove biases in the healthcare system so that diagnosis and treatment options work for and are affordable to everyone. Figure adapted from images created with BioRender.com.

not only the creation of a legal and ethical protection system that takes into account historical context, but also the promotion of genetic privacy studies, the implementation of technological solutions, the training of the next generation of genetic privacy researchers, and the creation of funding opportunities for genetic privacy studies that are different from traditional hypothesis-driven science. None of these endeavors should be perceived as in competition with advances in biomedical research but should be deemed as important complementary efforts.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This study was supported by US National Institutes of Health grant R00 HG010909 to G.G. I thank Dr. Mark Gerstein for fruitful discussions on genome privacy over the years.

LITERATURE CITED

1. Warren S, Brandeis L. 1890. The right to privacy. *Harvard Law Rev.* 4(5):193–220
2. Werner T. 2010. Next generation sequencing in functional genomics. *Brief. Bioinform.* 11(5):499–511
3. Hirst M, Marra MA. 2010. Next generation sequencing based approaches to epigenomics. *Brief. Funct. Genom.* 9(5–6):455–65

4. All Us Res. Progr. Investig., Denny JC, Rutter JL, Goldstein DB, Philippakis A, et al. 2019. The "All of Us" Research Program. *N. Engl. J. Med.* 381(7):668–76
5. Cancer Genome Atlas Res. Netw., Weinstein JN, Collisson EA, Mills GB, Shaw KRM, et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45(10):1113–20
6. Wellcome Trust, MRC (Med. Res. Counc.), UK Dep. Health. 2002. *The UK Biobank: A Study of Genes, Environment and Health*. London: Wellcome Trust
7. Ponting CP. 2019. The Human Cell Atlas: making "cell space" for disease. *Dis. Model. Mech.* 12(2):dmm037622
8. GTEx Consort. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* 45(6):580–85
9. ENCODE Proj. Consort. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74
10. Erlich Y, Shor T, Pe'er I, Carmi S. 2018. Identity inference of genomic data using long-range familial searches. *Science* 362(6415):690–94
11. Lin Z. 2004. Genomic research and human subject privacy. *Science* 305(5681):183
12. Wickenheiser RA. 2019. Forensic genealogy, bioethics and the Golden State Killer case. *Forensic. Sci. Int. Synerg.* 1:114–25
13. Gürsoy G. 2020. Criticality of data sharing in genomic research and public views of genomic data sharing. In *Responsible Genomic Data Sharing: Challenges and Approaches*, ed. X Jiang, H Tang, pp. 3–18. London: Elsevier
14. Arellano AM, Dai W, Wang S, Jiang X, Ohno-Machado L. 2018. Privacy policy and technology in biomedical data science. *Annu. Rev. Biomed. Data Sci.* 1:115–29
15. Harbord K. 2019. Genetic data privacy solutions in the GDPR. *Tex. A&M Law Rev.* 7(1):269–97
16. Erlich Y, Narayanan A. 2014. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* 15(6):409–21
17. Boylan M. 2008. *Racial profiling and genetic privacy*. Tech. Rep., Cent. Am. Progr., Washington, DC. **https://www.americanprogress.org/article/racial-profiling-and-genetic-privacy/**
18. Knoppers BM, Beauvais MJS. 2021. Three decades of genetic privacy: a metaphoric journey. *Hum. Mol. Genet.* 30(R2):R156–60
19. Robinson JC. 2004. Ethics and genetic privacy. *Online J. Health Ethics* 1(1):1
20. DeCew JW. 2004. Privacy and policy for genetic research. *Ethics Inf. Technol.* 6:5–14
21. Troy ESF. 1997. The Genetic Privacy Act: an analysis of privacy and research concerns. *J. Law Med. Ethics* 25:256–72
22. Strand NK. 2016. Shedding privacy along with our genetic material: What constitutes adequate legal protection against surreptitious genetic testing? *AMA J. Ethics* 18(3):264–71
23. Paillier F. 2018. About consumer genomics, genetic data privacy and ethics. *J. Bioanal. Biomed.* 10:132–33
24. Anderlik MA, Rothstein MA. 2001. Privacy and confidentiality of genetic information: what rules for the new science? *Annu. Rev. Genom. Hum. Genet.* 2:401–33
25. Springer JA, Beever J, Morar N, Sprague JE, Kane MD. 2013. Ethics, privacy, and the future of genetic information in healthcare information assurance and security. In *Bioinformatics: Concepts, Methodologies, Tools, and Applications*, pp. 1405–23. Hershey, PA: IPI Global
26. O'Neill M. 2021. Genetic information, social justice, and risk-sharing institutions. *J. Med. Ethics* 47:473–79
27. Kaan T, Ho CW-L, eds. 2013. *Genetic Privacy: An Evaluation of the Ethical and Legal Landscape*. Hackensack, NJ: Imp. Coll. Press
28. Tovino SA. 2021. HIPAA compliance. In *The Cambridge Handbook of Compliance*, ed. B van Rooij, DD Sokol, pp. 895–908. Cambridge, UK: Cambridge Univ. Press
29. Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A. 2003. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* 12:771–76
30. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* 339(6117):321–24
31. Mittos A, Malin B, De Cristofaro E. 2019. Systematizing genome privacy research: a privacy-enhancing technologies perspective. *Proc. Priv. Enhanc. Technol.* 2019(1):87–107

32. EEOC (Equal Employ. Oppor. Comm.). 2014. *Genetic Information Nondiscrimination Act*. Fact Sheet, EEOC, Washington, DC. **https://www.eeoc.gov/laws/guidance/fact-sheet-genetic-information-nondiscrimination-act**
33. Sweeney L, Abu A, Winn J. 2013. *Identifying participants in the personal genome project by name*. White Pap., Data Priv. Lab, IQSS, Harvard Univ., Cambridge, MA
34. Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, et al. 2017. Identification of individuals by trait prediction using whole-genome sequencing data. *PNAS* 114(38):10166–71
35. Venkatesaramani R, Malin BA, Vorobeychik Y. 2021. Re-identification of individuals in genomic datasets using public face images. *Sci. Adv.* 7(47):eabg3296
36. Erlich Y. 2017. Major flaws in "Identification of individuals by trait prediction using whole-genome sequencing data." bioRxiv 10.1101/185330. **https://doi.org/10.1101/185330**
37. Zhu X, Zhang S, Kan D, Cooper R. 2003. Haplotype block definition and its application. *Pac. Symp. Biocomput.* 2004:152–63
38. Naj A. 2019. Genotype imputation in genome-wide association studies. *Curr. Protoc. Hum. Genet.* 102(1):e84
39. Rubinacci S. 2020. *Genotype imputation methods for next generation datasets*. PhD Thesis, Univ. Oxford, Oxford, UK
40. Roshyara NR. 2020. *Genome-wide genotype imputation-aspects of quality, performance and practical implementation*. PhD Thesis, Univ. Leipzig, Leipzig, Ger.
41. Sherman MA. 2021. Paving the path toward genomic privacy with secure imputation. *Cell Syst.* 12(10):950–52
42. Browning BL, Zhou Y, Browning SR. 2018. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103(3):338–48
43. van Leeuwen EM, Kanterakis A, Deelen P, Kattenberg MV, Genome Neth. Consort., et al. 2015. Population-specific genotype imputations using minimac or IMPUTE2. *Nat. Protoc.* 10(9):1285–96
44. Davies RW, Kucka M, Su D, Shi S, Flanagan M, et al. 2021. Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* 53:1104–11
45. Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genet*. 5(6):e1000529
46. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452(7189):872–76
47. Nyholt DR, Yu C-E, Visscher PM. 2009. On Jim Watson's *APOE* status: Genetic information is hard to hide. *Eur. J. Hum. Genet.* 17(2):147–49
48. Paltoo DN, Rodriguez LL, Feolo M, Gillanders E, Ramos EM, et al. 2014. Data use under the NIH GWAS Data Sharing Policy and future directions. *Nat. Genet.* 46(9):934–38
49. Krawczak M, Goebel JW, Cooper DN. 2010. Is the NIH policy for sharing GWAS data running the risk of being counterproductive? *Investig. Genet.* 1:3
50. NIH (Natl. Inst. Health). 2018. *Update to NIH management of genomic summary results access*. Public Not. NOT-OD-19-023, NIH, US Dept. Health Hum. Serv., Washington, DC. **https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html**
51. Tsunoda T, Tanaka T, Nakamura Y, eds. 2019. *Genome-Wide Association Studies*. Singapore: Springer Nature
52. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genet*. 4(8):e1000167
53. Sankararaman S, Obozinski G, Jordan MI, Halperin E. 2009. Genomic privacy and limits of individual detection in a pool. *Nat. Genet.* 41:965–67
54. Lumley T, Rice K. 2010. Potential for revealing individual-level information in genome-wide association studies. *JAMA* 303(7):659–60
55. Im HK, Gamazon ER, Nicolae DL, Cox NJ. 2012. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* 90(4):591–98
56. Fiume M, Cupak M, Keenan S, Rambla J, de la Torre S, et al. 2019. Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.* 37(3):220–24

57. Shringarpure SS, Bustamante CD. 2015. Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* 97(5):631–46
58. Ayoz K, Ayday E, Cicek AE. 2021. Genome reconstruction attacks against genomic data-sharing beacons. *Proc. Priv. Enhancing Technol.* 2021(3):28–48
59. De Cristofaro E. 2021. A critical overview of privacy in machine learning. *IEEE Secur. Priv.* 19:19–27
60. Shokri R, Stronati M, Song C, Shmatikov V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. Los Alamitos, CA: IEEE Comput. Soc.
61. Oprisanu B, Ganev G, De Cristofaro E. 2021. On utility and privacy in synthetic genomic data. arXiv:2102.03314 [q-bio.GN]. **https://arxiv.org/abs/2102.03314**
62. Gürsoy G, Li T, Liu S, Ni E, Brannon CM, Gerstein MB. 2022. Functional genomics data: privacy risk assessment and technological mitigation. *Nat. Rev. Genet.* 23:245–58
63. Gürsoy G, Emani P, Brannon CM, Jolanki OA, Harmanci A, et al. 2020. Data sanitization to reduce private information leakage from functional genomics. *Cell* 183(4):905–17.e16
64. Gürsoy G, Brannon CM, Navarro FCP, Gerstein M. 2020. FANCY: fast estimation of privacy risk in functional genomics data. *Bioinformatics* 36(21):5145–50
65. Harmanci A, Gerstein M. 2018. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. *Nat. Commun.* 9:2453
66. Harmanci A, Gerstein M. 2016. Quantification of private information leakage from phenotype–genotype data: linking attacks. *Nat. Methods* 13(3):251–56
67. Schadt EE, Woo S, Hao K. 2012. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.* 44(5):603–8
68. Gürsoy G, Lu N, Wagner S, Gerstein M. 2021. Recovering genotypes and phenotypes using allele-specific genes. *Genome Biol.* 22:263
69. Backes M, Berrang P, Bieg M, Eils R, Herrmann C, et al. 2017. Identifying personal DNA methylation profiles by genotype inference. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 957–76. Los Alamitos, CA: IEEE Comput. Soc.
70. Philibert RA, Terry N, Erwin C, Philibert WJ, Beach SR, Brody GH. 2014. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. *Clin. Epigenet.* 6(1):28
71. DNAresource.com. 2011. *State DNA database laws: qualifying offenses*. Web Resour., DNAresource.com, Tacoma, WA. **https://www.dnaresource.com/documents/statequalifyingoffenses2011.pdf**
72. Kim J, Edge MD, Algee-Hewitt BFB, Li JZ, Rosenberg NA. 2018. Statistical detection of relatives typed with disjoint forensic and biomedical loci. *Cell* 175(3):848–58.e6
73. Edge MD, Algee-Hewitt BFB, Pemberton TJ, Li JZ, Rosenberg NA. 2017. Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *PNAS* 114(22):5671–76
74. Birzer ML. 2012. *Racial Profiling: They Stopped Me Because I'm ———!* Boca Raton, FL: CRC
75. Ramirez D, McDevitt J, Farrell A. 2014. *A resource guide on racial profiling data collection systems: promising practices and lessons learned*. Resour. Guide, US Dep. Justice, Washington, DC
76. Lynch MJ, Patterson EB, Childs KK, eds. 2008. *Racial Divide: Racial and Ethnic Bias in the Criminal Justice System*. Monsey, NY: Crim. Justice Press
77. Greytak EM, Moore C, Armentrout SL. 2019. Genetic genealogy for cold case and active investigations. *Forensic Sci. Int.* 299:103–13
78. Syndercombe Court D. 2018. Forensic genealogy: some serious concerns. *Forensic Sci. Int. Genet.* 36:203–4
79. Ram N, Guerrini CJ, McGuire AL. 2018. Genealogy databases and the future of criminal investigation. *Science* 360(6393):1078–79
80. Edge MD, Coop G. 2020. Attacks on genetic privacy via uploads to genealogical databases. *eLife* 9:e51810
81. Kennett D. 2019. Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. *Forensic Sci. Int.* 301:107–17
82. Chung J, Kaufman A, Rauenzahn B. 2021. Privacy problems in the genetic testing industry. *The Regulatory Review*, Jan. 23. **https://www.theregreview.org/2021/01/23/saturday-seminar-privacy-problems-genetic-testing/**

83. Albugmi A, Alassafi MO, Walters R, Wills G. 2016. Data security in cloud computing. In *2016 Fifth International Conference on Future Generation Communication Technologies (FGCT)*, pp. 55–59. New York: IEEE
84. Byrd JB, Greene AC, Prasad DV, Jiang X, Greene CS. 2020. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* 21(10):615–29
85. Hie B, Cho H, Berger B. 2018. Realizing private and practical pharmacological collaboration. *Science* 362(6412):347–50
86. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, et al. 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res*. 42:D975–79
87. Lee SM, Majumder MA. 2021. National Institutes of Mental Health Data Archive: privacy, consent, and diversity considerations and options for improvement. *AJOB Neurosci*. 13(1):3–9
88. Fernandez-Orth D, Lloret-Villas A, Rambla de Argila J. 2019. European Genome-phenome Archive (EGA)—granular solutions for the next 10 years. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 4–6. New York: IEEE
89. Berger B, Cho H. 2019. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol*. 20(1):128
90. NIH (Natl. Inst. Health). 2021. *NIH security best practices for controlled-access data subject to the NIH Genomic Data Sharing (GDS) Policy*. Web Resour., NIH, Washington, DC. **https://osp.od.nih.gov/wp-content/uploads/NIH_Best_Practices_for_Controlled-Access_Data_Subject_to_the_NIH_GDS_Policy.pdf**
91. Bonomi L, Huang Y, Ohno-Machado L. 2020. Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* 52(7):646–54
92. Tang H, Jiang X, Wang X, Wang S, Sofia H, et al. 2016. Protecting genomic data analytics in the cloud: state of the art and opportunities. *BMC Med. Genom.* 9:63
93. Wang S, Mohammed N, Chen R. 2014. Differentially private genome data dissemination through top-down specialization. *BMC Med. Inform. Decis. Mak.* 14(Suppl. 1):S2
94. Lu W-J, Yamada Y, Sakuma J. 2015. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. *BMC Med. Inform. Decis. Mak.* 15(Suppl. 5):S1
95. Cho H, Wu DJ, Berger B. 2018. Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* 36(6):547–51
96. Constable SD, Tang Y, Wang S, Jiang X, Chapin S. 2015. Privacy-preserving GWAS analysis on federated genomic datasets. *BMC Med. Inform. Decis. Mak.* 15(Suppl. 5):S2
97. Kockan C, Zhu K, Dokmai N, Karpov N, Kulekci MO, et al. 2020. Sketching algorithms for genomic data analysis and querying in a secure enclave. *Nat. Methods* 17(3):295–301
98. Kim M, Harmanci AO, Bossuat J-P, Carpov S, Cheon JH, et al. 2021. Ultrafast homomorphic encryption models enable secure outsourcing of genotype imputation. *Cell Syst.* 12(1):1108–20.e4
99. Dokmai N, Kockan C, Zhu K, Wang X, Sahinalp SC, Cho H. 2021. Privacy-preserving genotype imputation in a trusted execution environment. *Cell Syst.* 12(1):983–93.e7
100. Gürsoy G, Chielle E, Brannon CM, Maniatakos M, Gerstein M. 2021. Privacy-preserving genotype imputation with fully homomorphic encryption. *Cell Syst*. 13(12):173–82.e3
101. Gürsoy G, Bjornson R, Green ME, Gerstein M. 2020. Using blockchain to log genome dataset access: efficient storage and query. *BMC Med. Genom.* 13(Suppl. 7):78
102. Ma S, Cao Y, Xiong L. 2020. Efficient logging and querying for blockchain-based cross-site genomic dataset access audit. *BMC Med. Genom.* 13(Suppl. 7):91
103. Pattengale ND, Hudson CM. 2020. Decentralized genomics audit logging via permissioned blockchain ledgering. *BMC Med. Genom.* 13(Suppl. 7):102
104. Ozdayi MS, Kantarcioglu M, Malin B. 2020. Leveraging blockchain for immutable logging and querying across multiple sites. *BMC Med. Genom.* 13(Suppl. 7):82
105. Kuo T-T, Bath T, Ma S, Pattengale N, Yang M, et al. 2021. Benchmarking blockchain-based gene-drug interaction data sharing methods: a case study from the iDASH 2019 secure genome analysis competition blockchain track. *Int. J. Med. Inform.* 154:104559
106. Gefenas E. 2006. The concept of risk and responsible conduct of research. *Sci. Eng. Ethics* 12(1):75–83

107. Tsosie KS, Yracheta JM, Dickenson D. 2019. Overvaluing individual consent ignores risks to tribal participants. *Nat. Rev. Genet.* 20(9):497–98
108. U.N. 1948. *Universal declaration of human rights.* U.N. Declar., U.N. Gen. Assem., Paris
109. Grant S. 2019. *Privacy is a human right—It can't be bought or sold*. Blog Post, Consum. Fed. Am., Dec. 19. **https://consumerfed.org/privacy-is-a-human-right-it-cant-be-bought-or-sold/**
110. Wee S-L. 2019. China uses DNA to track its people, with the help of American expertise. *The New York Times*, Feb. 21
111. Fox K. 2020. The illusion of inclusion—the "All of Us" research program and Indigenous peoples' DNA. *N. Engl. J. Med.* 383(5):411–13