Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

Nat Genet. 2020 July ; 52(7): 646–654. doi:10.1038/s41588-020-0651-0.

Privacy Challenges and Research Opportunities for Genomic Data Sharing

Luca Bonomi[1], **Yingxiang Huang**[1], **Lucila Ohno-Machado**[1,2]

[1]UCSD Health Department of Biomedical Informatics, University of California San Diego, La Jolla, CA, 92093, USA

[2]Division of Health Services Research & Development, VA San Diego Healthcare System, San Diego, CA 92161, USA

Abstract

The sharing of genomic data holds great promise for advancing precision medicine, providing personalized treatments and other types of interventions. However, there are privacy concerns, as data misuse may lead to infringement of privacy for individuals and their blood relatives. As genomic data are rapidly growing and some of these data are being made available to researchers, it is imperative to understand the current genome privacy landscape and to identify the challenges in developing effective privacy-protecting solutions. In this work, we provide an overview of major privacy threats identified by the research community and examine the privacy challenges in the context of emerging direct-to-consumer applications. We present general privacy protection techniques for genomic data sharing and their potential applications in direct-to-consumer genomic testing and forensic analyses. We discuss limitations in current privacy protection methods, highlight possible mitigation strategies, and suggest future research opportunities for advancing genomic data sharing.

Introduction

In recent years, technological improvements have significantly reduced the cost of genome sequencing, creating an unprecedented amount of genomic data that is vital for many research applications[1,2]. Several research initiatives (e.g., NIH All of Us[3]) are integrating these data with the goal of serving as the source of analyses for a wide range of studies. At the same time, there has been a significant expansion of genomic data-driven applications in the private sector, where personal genomic data are collected to provide individuals with health-related services.

lbonomi@ucsd.edu.

Contributions

LB conducted the literature review, drafted the organization of the article, and contributed the majority of the writing. YH contributed on the data sharing in direct-to-consumer genetic testing and provided helpful comments on the presentation. LO-M provided the motivation for this work, detailed edits, and critical suggestions on the organization and structure for the article.

Competing Interests

The authors declare no competing interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

There are several remarkable features that make genomic data different from other health data. For example, genomic data carry information that may be effectively used for prognosing health conditions (e.g., Alzheimer's disease[4,5]) and for designing preventive interventions. Another important property of genomic data is the presence of significant commonality among individuals who are blood relatives. Therefore, genome analysis is commonly used for susceptibly risk, paternity and relativeness testing (e.g., ancestry services), and for forensic purposes (e.g., genomic genealogy searches).

Some of these features pose significant privacy concerns when sharing genomic data[6,7]. Individual's germline genomic data provide information that can uniquely identify individuals and tend to remain relatively static over the course of life, providing excellent biometric information (i.e., genomic "fingerprint"). Lin et al.[8] showed that 75 statistically independent SNPs would suffice to uniquely identify an individual across the global population. Sharing seemingly harmless "aggregate" data (e.g., allele statistics) can also pose privacy risks[9,10].

Traditional privacy models designed for health data, provide limited protection for genomic data. An attacker may learn sensitive information about a target individual by exploiting the dependency between genomic data and other publicly available information such as: family name, demographic data, and observable features (e.g., eyes and hair color)[11–13]. As personal data are made largely available (e.g., social networks), privacy assurances from traditional methods are unlikely to be sustainable. Furthermore, the rise of direct-to-consumer companies poses new privacy and ethical concerns. These companies collect data from a growing number of individuals, some of whom may share their data without fully understanding the potential implications for themselves, existing and future blood relatives. As a notable example, the recent use of direct-to-consumer genomic data in forensic analyses has brought these privacy concerns to the attention of the general public. Privacy breaches can have serious social implications, and adverse impact on genomic-driven research, such as limiting data collection and data sharing[14,15]. Therefore, it is imperative to ensure privacy both as a fundamental right for individuals and an enabling strategy to support responsible data sharing.

Currently, healthcare organizations in US must comply to the Privacy Rule created under the Health Insurance Portability and Accountability Act (HIPAA[16]), which defines protected health information (PHI) such as name and Social Security Number, and states how such information should be protected. However, the removal of PHI cannot protect from re-identification[17,18]. For genomic data, the Genetic Information Nondiscrimination Act of 2008 (GINA[19]) provides protections against discrimination by health insurers or employers on the basis of genetic information, but it does not clearly define what information needs protection nor how such protection is carried out. The NIH Policy on Protection of Genome Data sheds some light on regulations for research personnel and contractors in terms of training and certifications (https://osp.od.nih.gov/). Data privacy regulations vary around the world. For example, the General Data Protection Regulation (GDPR[20]) law for EU members provides a strong privacy protection by regulating use and storage of the collected data, and empowering individuals with better data control (e.g., right to delete data). Some countries have more relaxed regulations based on consent or are still missing specific regulations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Clarification of the potential risks and mitigating solutions when sharing human genome data will help policy makers accept tradeoffs between data access and privacy of the individuals whose data are being shared.

In this article, we provide an overview of major privacy challenges and research opportunities for genomic data sharing. Compared to previous surveys on genome privacy[21–24], we do not aim at presenting a detailed technical review of privacy methods or general use of genomic data in the healthcare domain. Instead, we present known major privacy risks, classify privacy-preserving strategies, and contextualize the discussion by relating threats and mitigating "solutions" in light of emerging applications such as direct-to-consumer genomic testing and forensic analyses.

Privacy Risks in Sharing Human Genomic Data

In this section, we briefly illustrate known privacy attacks (Figure 1), where an adversary may leverage publicly available data (Table 1). We categorize these attacks into *identification* and *phenotype inference*.

Identification

In identification attacks, an adversary who has access to "anonymized" human genomic data successfully recovers the identity of the donors. The current practice of "anonymization" of genomic data is performed by removing protected health information (e.g., name) and quasi-identifying information (e.g., Zipcode). While this "de-identification" process may meet current privacy regulations in the USA (e.g., HIPAA[16]), it often fails to protect against identification attacks[25–29]. Sweeney et al.[25] demonstrated this vulnerability by identifying participants in the Personal Genome Project (PGP) using publicly available data. Identification attacks can also be carried out directly on the raw genomic data, as demonstrated in several studies[8,12,30]. Lippert et al.[13] showed that an adversary who has access to the whole genome data of an individual may be able to correctly predict physical features, such as: eye, hair, skin color, and facial and vocal characteristics. Additionally, once data are shared, it becomes impossible to practically track their multiple potential destinations. Attacks by foreign nations or malicious enterprises might harm individuals in ways that have not been anticipated by current regulations. Multiple authors have advocated the expansion of GINA and related regulations beyond protection of health insurance and employer discrimination[31–33]. While GDPR proactively addresses some of these risks, US regulations vary from state to state. Despite the improvement in privacy regulations, there are significant technical challenges to satisfy these new privacy standards. For example, Rocher et al.[17] showed that more than 99% of Americans would be correctly identified in any dataset using only 15 demographic attributes.

Phenotype Inference

In a phenotype inference attack, an adversary who has access to partial genomic information of a known target individual aims at inferring some sensitive phenotypes (e.g., disease). When sufficient genomic data about the target are available, the adversary may learn sensitive traits by observing the presence of characteristic genetic markers. Despite masking

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

these markers, an attacker may still restore the original genomic information via genotype imputation. For example, by exploiting the linkage disequilibrium between regions in the genome[34,35] of the target and from blood relatives (i.e., genealogical imputation). With the unprecedented amount of genomic information being collected, this type of attack poses growing privacy concerns[11,36,37]. Privacy disclosure may occur even when only aggregate statistics are released[9,10,38,39]. For example, Homer et al.[10] demonstrated that it is possible to determine the presence of a target individual in a group by comparing the target's allele frequencies with those observed in the standard population and those published for the group (Figure 2). Due to the severity of this privacy threat, the National Institutes of Health (NIH) initially responded by controlling the access to the aggregated genomic data results. However, since November 2018, given the lack of known attacks, NIH decided to broaden the access to these data (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-023.html). Broad access to genomic data is vital for collaborative medical research efforts. However, designing data sharing methods that strike the right balance between data access and privacy is challenging.

In research studies, "informed consent" processes are used to facilitate the participation of individuals, by promoting transparency and making privacy risks more explicit. *Broad* consent aims at maximizing the utility of the collected genomic data, where individuals, in addition to consent to the use of their data for primary research tasks (e.g., breast cancer study), may agree to general research use (e.g., future research). *Specific* consent reduces the use of the data to narrowly defined research tasks. *Dynamic* consent enables individuals to update their privacy preferences over time. To design effective informed consent processes, it is important to educate individuals about privacy risks and potential benefits, which may require researchers to overcome cultural and social barriers.

Studies involving genetically distinct populations may require researchers to take into account the specific concerns of these populations, which may include group privacy breaches (e.g., a population is more susceptible to a disease), transparency, data and result ownership, and oversight of the research study. To benefit these populations, it is important to address any concerns directly, by considering privacy, social, and ethical aspects[40–43].

Privacy in Emerging Genomic Data Applications

There is a privacy protection gap in direct-to-consumer (DTC) applications, as portions of current privacy regulations do not apply and there is no entity providing oversight over the data sharing process. As millions of individuals are contributing data, it is important to understand the specific privacy challenges for DTC applications. Here, we discuss how genomic data are used in the context of two emerging applications: recreational genetic testing and forensic investigations.

Data Sharing in Direct-to-Consumer Genetic Testing

DTC companies collect large volumes of personal genomic data (Table 2) that could be used for purposes beyond the services provided. In most companies' *Terms of Service*, it is not clear how data may be used and shared. Most DTC companies often refer to "appropriate" use of genetic information, without providing adequate details about what does and does not

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

qualify as “appropriate”. In fact, a recent study found that 67% of the DTC companies in that study provided consumers with insufficient information about how their genomic data would be used[44].

Most DTC companies state that additional research with the consumers’ sample may be conducted[45,46], which may include health related and non-health related research. For non-health related research, most companies aim at answering questions about human migration and population history, while some companies also conduct market research and other non-specified research[45]. For health-related research, companies provide little to no details about what such research may include. A popular DTC service states that data may be used to understand the basic causes of disease, develop drugs, design preventive measures, or predict risk of disease (for example https://www.23andme.com/about/tos/). The research conducted with consumers’ sample becomes known only when a company files for patent or releases a product. Customer reactions have not been very positive: some customers were unhappy with the company profiting off their genetic material and felt that the consent process was not sufficiently clear[47]. It is understandable that companies want to protect their intellectual property, but greater transparency is needed to enable individuals decide whether consent to the use of their data in research activities[46].

DTC companies may also share consumers’ data with third parties, but statements regarding the sharing of genetic data vary tremendously. Most companies provide, at the very least, a blanket statement about whether consumers’ data can be shared with third parties. Only a few enumerate a list of third parties that may have access to the data[45], such as healthcare professionals and research institutions, if the consumer chooses to divulge the information. Since the first paper that reported novel associations between several single-nucleotide polymorphisms and phenotypes using data from 23andMe was published in 2010, many more research studies with DTC data were conducted[48]. In addition to sharing genomic data, companies may outsource biological samples overseas where the genomes are sequenced at a lower cost. Because the data sharing process may involve entities in different countries, there is a need for better harmonized privacy policies that are coherent across countries.

Identification of Individuals in Forensic Investigations

Large datasets of DNA materials from convicted or arrested individuals, such as the US DNA Index System (NDIS) and Combined DNA Index System (CODIS), have been used for decades by law enforcement agents to search for DNA profiles that match the genomic evidence collected from a crime scene. However, we have recently witnessed a new way of using genomic data in facilitating forensic analysis: genomic materials from a crime scene can be used to query publicly available DTC genomics genealogical databases, enabling the identification of victims and perpetrators. Traditional databases of genomic material available to law enforcement agents (e.g., NDIS, CODIS) only comprise individuals with a criminal history and suspects of serious crimes, while genomic data from DTC companies enable agents to access additional individuals and perform powerful genetic searches. Agents can identify profiles that partially match the genomic information collected from the crime scene, thus significantly increasing the likelihood of successfully identifying individuals who are blood relatives of a suspected criminal (Figure 3)[30]. As a notable

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

example, in April 2018, long-range familial searches led to the arrest of the Golden State Killer who was responsible for more than 13 murders and 50 rapes in California from 1974 to 1986. Long-range familial searches are gaining increasing popularity in forensic analysis by specialized companies, leading to the resolution of more than thirty cold cases since April 2018, including the reidentification of four victims of violent crimes. A detailed summary of the cold cases solved via genetic genealogy from 2018 until late 2019 is reported in the Supplementary Table 1.

Despite the unquestionable benefits, there are privacy and ethical concerns about the use of DTC genomic data in forensic analysis[49,50]. Current regulations provide limited protection against law enforcement searches. For example, GINA only protects against genetic discrimination in health insurance and employment. DTC companies are typically not engaged in providing healthcare services, thus are not legally required to comply with HIPAA. Moreover, current regulation for protecting individual privacy from government surveillance (e.g., the Fourth Amendment of the US Constitution) does not apply to DTC genomic data, as these data are voluntarily provided[49,51]. The use of the genomic data in criminal investigations has raised the public awareness about genomic privacy. Recently, DTC companies have changed their data privacy policy, requiring individuals who uploaded their genomic data to opt in to allow law enforcement agents to access their data (for example, GEDmatch's Terms of Service and Privacy Policy - https://www.gedmatch.com/tos.htm).

Privacy-Protecting Solutions for Genomic Data

Core Privacy-Protecting Solutions for Genomic Data

Here, we briefly describe the core privacy-protecting solutions for genomic data (Table 3). For a technical review, we refer the readers to previous surveys[21–24].

Access control—*Access control* methods limit the data exposure by allowing only authorized users to access sensitive data[52]. Qualified users have to ensure that data will be appropriately stored, will not be used to identify data contributors, and may be required to file periodic reports.

Encryption—*Encryption* techniques rely on results from number theory to transform the original data (i.e., plaintext) into an encoded format (i.e., ciphertext). Homomorphic encryption (HE) is a special type of encryption that enables simple primitives (e.g., addition, multiplication) directly on the ciphertext. HE methods are used in many privacy-protecting solutions[53–56] where data are shared in the cloud[57–59] and in federated envirorments[60–62]. Privacy attacks may still be performed over homomorphically encrypted data[63].

Secure Multiparty Computation (SMC)—*Secure Multiparty Computation (SMC)* protocols are cryptography-based methods that enable a group of parties to jointly perform a task without revealing private data. Computations can be performed without the need of a trusted party, making these solutions suitable for distributed settings. For example, to perform genomic sequence comparison[64–66], secure statistical test evaluation[67,68], and GWAS[69,70].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

k-anonymity[71]—*k-anonymity*[71] ensures that, for each record, there are at least *k-1* records with the same quasi-identifiers (e.g., Zipcode) and therefore any record is hidden in a group. To achieve *k*-anonymity, the original data are transformed via suppression and generalization of attribute values (e.g., 3-digit representation for Zipcode). *k*-anonymity has been applied on quasi-identifier attributes[29,72] and at SNP-level[73,74].

Differential privacy[75]—*Differential privacy*[75] provides formal and provable privacy protection by ensuring that an adversary who observes the results cannot determine whether an individual participated in a study. Privacy is achieved via randomized mechanisms (e.g., output perturbation). Differential privacy has been deployed in GWAS studies[76–79], and specializations are recently considered in other genomic applications[80,81].

A common practice is to utilize multiple privacy techniques for combined benefits. For example, Raisaro et al.[82] proposed a solution that combines HE and differential privacy. Specifically, a central server stores the patient's genetic data (encrypted using HE), while differential privacy is used to generate summary statistics for researchers.

Privacy Solutions Applicable to DTC

Bruekers et al.[66] proposed a secure Short Tandem Repeats (STR) matching protocol, which can enable DNA-based search, paternity, and ancestry tests without revealing the identity of individuals. HE protocols are used to encrypt STR profiles and, by performing simple binary and logical operations (e.g., difference between profiles), determine whether individuals are related. Security-based methods can also be applied in forensic investigations to protect the identity of individuals whose records do not match the genomic profile of an individual of interest (e.g., suspected criminal). For example, Bohannon et al.[56] proposed an approach that encrypts the forensic DNA databases and allows the identification of an individual only when his/her genomic record is matched with the genomic profile gathered from a crime scene. These solutions prevent identity disclosure but do not protect privacy entirely. For example, when it is known that the sequence of a particular individual is included in a database, it is possible to determine non-paternity by lack of the expected partial match.

Building on cryptographic methods, Huang et al.[83] proposed GenoGuard, a tool for genomic data storage. An individual sends a password and a biological sample to a trusted certified institution in charge of sequencing the sample. Then, the sequence is encrypted with the given password and stored in a biobank, where authorized users (e.g., doctors) may retrieve and decrypt it. In the DTC setting, the biobank can store the encrypted genomic data collected by DTC companies, and an individual may access the stored genetic data to request the desired genetic test. GenoGuard relies on a novel cryptographic scheme named honey encryption[84], in which attacker who tries to decrypt the cyphertext receives an incorrect but plausible plaintext. The honey encryption provides long-lasting protection (i.e., mitigates brute-force attacks), which is suitable in the genomic setting, as the sensitivity of genomic data does not change over time.

Humbert et al.[85] proposed a data anonymization technique that enables an individual to safely publish their genomic data. Such a technique finds application in the DTC setting, where the publication of an individual genomic data may disclose sensitive information

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

about the donor and family members. The notion of health privacy is used to quantify how individual SNPs contribute in the predisposition to different diseases. Their idea is to achieve health privacy by masking SNPs and limiting the disclosure of sensitive phenotypes of the data donor or family members. The selection of the SNPs to hide is performed to satisfy the privacy preference of the individuals by considering the known correlation between genomic regions. As a result, the sanitized genomic data can be made public without compromising the privacy of the family for the genetic traits that are known today. However, given the fast pace in which genetic markers are being discovered, what is considered non-sensitive today may become sensitive in the near future. Once the genomes are disclosed, there is no backtracking. Additionally, the sanitization pattern of SNPs may reveal what type of information is being hidden.

On the policy side, recent developments in guidelines and regulations have started to bridge the privacy gap in DTC settings, but they are still in an early stage. Organizations of domain experts, such as the American College of Medical Genetics and Genomics (ACMG) and the European Society of Human Genetics (ESHG), are proposing guidelines to establish a level of transparency for all DTC companies. Since a violation of those guidelines is not sanctioned by law, companies have not been inclined to adhering to them[45]. In addition, the Food and Drug Administration (FDA) has increased its involvement and revamped its policies regarding consumer genomics. The most prominent change is the shift in the perception of consumer genetic testing from being a commercial product to a medical device product that requires regulations and restrictions under HIPAA. Although it is unclear what ethical or legal responsibilities DTC companies must carry at the moment, this is a step toward protecting consumer privacy with the same standard as patient privacy[86]. There have also been regulations set up by leading DTC companies to promote a more responsible and transparent use of the genomic data (Privacy Best Practices for Consumer Genetic Testing Services[87]). Furthermore, some companies have started to implement forms of participant consent (e.g., dynamic consent) to enable robust and transparent data sharing processes[88].

Research Opportunities

Here, we discuss some future research opportunities for improving the design of privacy-protecting approaches for genomic data.

Deployment of Privacy Solutions.

To promote the usability of privacy methods, it is crucial to build and distribute a wide range of usable privacy tools. While there are several research initiatives (e.g., International iDASH Privacy Protection Challenge, GenoPri) and publicly available privacy tools (e.g., the Harvard University Privacy Tools Project - https://privacytools.seas.harvard.edu/), most technical solutions are rarely deployed in practice. The lack of practical privacy methods is even more evident in the DTC setting, where solutions have to be compatible with a company business model. Yet, developing privacy tools is likely to provide opportunities for educating individuals, research institutions, and private companies about the beneficial impacts of privacy in genomic data sharing. Those tools could also help provide test grounds for developing new ethical and regulatory guidelines for data sharing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Measuring Privacy Risks.

The understanding of privacy risks vs. potential benefits is crucial for determining the most appropriate privacy method or policy in genomic data sharing. Modeling privacy risk is challenging, as it may depend on the available information and the power of the adversary. As research advances, it is reasonable to believe that publicly available genomic data that do not currently present privacy risks may present risks in the near future[6]. The posture adopted by some, that lack of known attacks equates lack of attacks, is naïve. The posture adopted by others, that only perfect privacy is acceptable, makes any kind of data sharing impractical. Therefore, the design of solutions that leverage technical approaches for risk assessment with appropriate regulations (e.g., data agreements, policies) may help identify concrete privacy risks and steer the development of new methods for improving data usability.

Technology for Controlling Data Flow.

In the current DTC data sharing framework, individuals can directly share their genomic data to receive health related services. However, users have often limited information and control over their data. Advances in current privacy and security domain have the potential of changing this paradigm by empowering individuals to own, track, and potentially even profit from their genomic data. Among them, blockchain technology constructs a chain of immutable blocks recording data transactions (i.e., immutable distributed ledger), which has several potential benefits in data sharing. The chain can provide a full history of what has happened to the data (i.e., data provenance), enabling individuals to track the data and achieving higher level of trust[89]. Recently, a few companies have proposed to use blockchain technology for genomic data sharing, with the goal of empowering individuals to better control their data, and potentially receiving more immediate health benefits[90].

Privacy Initiatives for Data Sharing.

Whether genomic data are shared for biomedical research or for other services, there is an increasing need for data standardization and privacy protection methods. To this end, several initiatives (e.g., GA4GH[91]) have been established to address the privacy and data harmonization challenges in collaborative research efforts. However, in DTC settings, the standards for quality and privacy are not very clear. Initiatives that aim at providing individuals with reliable, relevant and transparent information are needed to informed privacy choices[92].

Conclusion

Admirable progress has been made over recent years toward the development of privacy technologies, which are essential in broadening the sharing and collection of genomic data. In parallel with technological advances, it is crucial to further improve current regulations and guidelines. Addressing these technological, regulatory, and ethical challenges in combination may empower individuals to actively contribute to scientific research, improving genomic data sharing and benefiting medical research.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Human Genome Research Institute grant K99HG010493, National Institute of General Medical Sciences grant R01GM118609, and National Heart, Lung, and Blood Institute grant R01HL136835.

References

1. Mardis ER A decade's perspective on DNA sequencing technology. Nature 470, 198 (2011). [PubMed: 21307932]
2. Metzker ML Sequencing technologies - the next generation. Nat. Rev. Genet 11, 31–46 (2010). [PubMed: 19997069]
3. Investigators, A. of U. R. P. The "All of Us" Research Program. N. Engl. J. Med 381, 668–676 (2019). [PubMed: 31412182]
4. Green RC et al. Disclosure of APOE genotype for risk of Alzheimer's disease. N. Engl. J. Med 361, 245–254 (2009). [PubMed: 19605829]
5. Goldman JS et al. Genetic counseling and testing for Alzheimer disease: joint practice guidelines of the American College of Medical Genetics and the National Society of Genetic Counselors. Genet. Med 13, 597 (2011). [PubMed: 21577118]
6. Heeney C, Hawkins N, de Vries J, Boddington P & Kaye J Assessing the privacy risks of data sharing in genomics. Public Health Genomics 14, 17–25 (2011). [PubMed: 20339285]
7. Wang S et al. Genome privacy: challenges, technical approaches to mitigate risk, and ethical considerations in the united states. Ann. N. Y. Acad. Sci 1387, 73–83 (2017). [PubMed: 27681358]
8. Lin Z, Owen AB & Altman RB Genomic research and human subject privacy. Science (80-.) 305, 183 (2004).
9. Sankararaman S, Obozinski G, Jordan MI & Halperin E Genomic privacy and limits of individual detection in a pool. Nat. Genet 41, 965 (2009). [PubMed: 19701190]
10. Homer N et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 4, e1000167 (2008). [PubMed: 18769715]
11. Humbert M, Ayday E, Hubaux J-P & 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 1141–1152 (2013).
12. Gymrek M, McGuire AL, Golan D, Halperin E & Erlich Y Identifying personal genomes by surname inference. Science 339, 321–4 (2013). [PubMed: 23329047]
13. Lippert C et al. Identification of individuals by trait prediction using whole-genome sequencing data. Proc. Natl. Acad. Sci 114, 10166–10171 (2017). [PubMed: 28874526]
14. McGuire AL et al. To share or not to share: a randomized trial of consent for data sharing in genome research. Genet. Med 13, 948–955 (2011). [PubMed: 21785360]
15. Oliver JM et al. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. Public Health Genomics 15, 106–114 (2012). [PubMed: 22213783]
16. Health Insurance Portability and Accountability Act of 1996, 18 USC §264. (1996).
17. Rocher L, Hendrickx JM & de Montjoye Y-A Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun 10, 3069 (2019). [PubMed: 31337762]
18. Na L et al. Feasibility of Reidentifying Individuals in Large National Physical Activity Data Sets From Which Protected Health Information Has Been Removed With Use of Machine Learning. JAMA Netw. Open 1, e186040–e186040 (2018). [PubMed: 30646312]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

19. The Genetic Information Nondiscrimination Act of 2008. https://www.eeoc.gov/laws/statutes/gina.cfm (2008).
20. European Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the coucil of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general dat. Official Journal of the European Union L vol. 119 1–88 (2016).
21. Erlich Y & Narayanan A Routes for breaching and protecting genetic privacy. Nat. Rev. Genet 15, 409–21 (2014). [PubMed: 24805122]
22. Naveed M et al. Privacy in the Genomic Era. ACM Comput. Surv 48, 6:1–-6:44 (2015).
23. Mittos A, Malin B & De Cristofaro E Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective. Proc. Priv. Enhancing Technol 2019, (2019).
24. Akgün M, Bayrak AO, Ozer B & Sağıroğlu MŞ Privacy preserving processing of genomic data: A survey. J. Biomed. Inform 56, 103–111 (2015). [PubMed: 26056074]
25. Sweeney L, Abu A & Winn J Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment). 4 http://privacytools.seas.harvard.edu/files/privacytools/files/1021-1.pdf (2013).
26. 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 84, 251–258 (2009). [PubMed: 19215731]
27. Malin B Re-identification of familial database records in AMIA annual symposium proceedings vol. 2006 524 (American Medical Informatics Association, 2006).
28. Malin B & Sweeney L How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J. Biomed. Inform 37, 179–192 (2004). [PubMed: 15196482]
29. Malin B & Sweeney L Determining the identifiability of DNA database entries in Proceedings of the AMIA Symposium 537 (American Medical Informatics Association, 2000).
30. Erlich Y, Shor T, Pe'er I & Carmi S Identity inference of genomic data using long-range familial searches. Science (80-.) 362, 690–694 (2018).
31. Kahn SD On the future of genomic data. Science (80-.) 331, 728–729 (2011).
32. Areheart BA & Roberts JL GINA, Big Data, and the Future of Employee Privacy. Yale Law J. 128, 3 (2019).
33. Lee SS-J & Borgelt E Protecting posted genes: Social networking and the limits of GINA. Am. J. Bioeth 14, 32–44 (2014). [PubMed: 25325810]
34. Wheeler DA et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872 (2008). [PubMed: 18421352]
35. Nyholt DR, Yu C-E & Visscher PM On Jim Watson's APOE status: genetic information is hard to hide. Eur. J. Hum. Genet 17, 147–9 (2009). [PubMed: 18941475]
36. Humbert M, Ayday E, Hubaux J-P & Telenti A Quantifying interdependent risks in genomic privacy. ACM Trans. Priv. Secur 20, 3 (2017).
37. Ayday E & Humbert M Inference attacks against kin genomic privacy. IEEE Secur. Priv 15, 29–37 (2017).
38. Shringarpure SS & Bustamante CD Privacy risks from genomic data-sharing beacons. Am. J. Hum. Genet 97, 631–646 (2015). [PubMed: 26522470]
39. 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 534–544 (ACM, 2009).
40. James R et al. Exploring pathways to trust: a tribal perspective on data sharing. Genet. Med 16, 820–826 (2014). [PubMed: 24830328]
41. Harding A et al. Conducting research with tribal communities: Sovereignty, ethics, and data-sharing issues. Environ. Health Perspect 120, 6–10 (2012). [PubMed: 21890450]
42. Arquette M et al. Holistic risk-based environmental decision making: a Native perspective. Environ. Health Perspect 110, 259–264 (2002). [PubMed: 11929736]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

43. Mello MM & Wolf LE The Havasupai Indian tribe case—lessons for research involving stored biologic samples. N. Engl. J. Med 363, 204–207 (2010). [PubMed: 20538622]
44. Christofides E & O'Doherty K Company disclosure and consumer perceptions of the privacy implications of direct-to-consumer genetic testing. New Genet. Soc 35, 101–123 (2016).
45. Laestadius LI, Rich JR & Auer PL All your data (effectively) belong to us: data practices among direct-to-consumer genetic testing firms. Genet. Med 19, 513 (2017). [PubMed: 27657678]
46. Niemiec E & Howard HC Ethical issues in consumer genome sequencing: use of consumers' samples and data. Appl. Transl. genomics 8, 23–30 (2016).
47. Allyse M 23 and Me, We, and You: direct-to-consumer genetics, intellectual property, and informed consent. Trends Biotechnol. 31, 68–69 (2013). [PubMed: 23237855]
48. Eriksson N et al. Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS Genet 6, e1000993 (2010). [PubMed: 20585627]
49. Ram N, Guerrini CJ & McGuire AL Genealogy databases and the future of criminal investigation. Science (80-.) 360, 1078–1079 (2018).
50. Greytak EM, Kaye DH, Budowle B, Moore C & Armentrout SL Privacy and genetic genealogy data. Science (80-.) 361, 857 (2018).
51. Berkman BE, Miller WK & Grady C Is It Ethical to Use Genealogy Data to Solve Crimes? Ann. Intern. Med (2018).
52. Erlich Y et al. Redefining genomic privacy: trust and empowerment. PLoS Biol. 12, e1001983 (2014). [PubMed: 25369215]
53. Lauter K, López-Alt A & Naehrig M Private computation on encrypted genomic data. in 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy. (2014).
54. Wang S et al. HEALER: Homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS. Bioinformatics 32, 211–218 (2015). [PubMed: 26446135]
55. He D et al. Identifying genetic relatives without compromising privacy. Genome Res. 24, 664–672 (2014). [PubMed: 24614977]
56. Bohannon P, Jakobsson M & Srikwan S Cryptographic approaches to privacy in forensic DNA databases in International Workshop on Public Key Cryptography 373–390 (Springer, 2000).
57. Sousa JS et al. Efficient and secure outsourcing of genomic data storage. BMC Med. Genomics 10, 46 (2017). [PubMed: 28786363]
58. Deuber D et al. My Genome Belongs to Me: Controlling Third Party Computation on Genomic Data. Proc. Priv. Enhancing Technol 2019, 108–132 (2019).
59. Ayday E, Raisaro JL, Hubaux J-P & Rougemont J Protecting and Evaluating Genomic Privacy in Medical Tests and Personalized Medicine in Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society 95–106 (ACM, 2013). doi:10.1145/2517840.2517843.
60. Constable SD, Tang Y, Wang S, Jiang X & Chapin S Privacy-Preserving GWAS Analysis on Federated Genomic Datasets. BMC Med Inf. Decis Mak 15, S2 (2015).
61. Zhang Y, Dai W, Jiang X, Xiong H & Wang S FORESEE: Fully Outsourced secuRe gEnome Study basEd on homomorphic Encryption. BMC Med Inf. Decis Mak 15, S5 (2015).
62. Chen F et al. Princess: Privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics 33, 871–878 (2016).
63. Goodrich MT The mastermind attack on genomic data in Security and Privacy, 2009 30th IEEE Symposium on 204–218 (IEEE, 2009).
64. Atallah MJ, Kerschbaum F & Du W Secure and private sequence comparisons in Proceedings of the 2003 ACM workshop on Privacy in the electronic society 39–44 (ACM, 2003).
65. Jha S, Kruger L & Shmatikov V Towards practical privacy for genomic computation in Security and Privacy, 2008. SP 2008. IEEE Symposium on 216–230 (IEEE, 2008).
66. Bruekers F, Katzenbeisser S, Kursawe K & Tuyls P Privacy-Preserving Matching of DNA Profiles. IACR Cryptol. ePrint Arch 2008, 203 (2008).
67. Danezis G & De Cristofaro E Fast and private genomic testing for disease susceptibility in Proceedings of the 13th Workshop on Privacy in the Electronic Society 31–34 (ACM, 2014).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

68. Duverle DA, Kawasaki S, Yamada Y, Sakuma J & Tsuda K Privacy-preserving statistical analysis by exact logistic regression in Security and Privacy Workshops (SPW), 2015 IEEE 7–16 (IEEE, 2015).
69. Kamm L, Bogdanov D, Laur S & Vilo J A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics 29, 886–893 (2013). [PubMed: 23413435]
70. Cho H, Wu DJ & Berger B Secure genome-wide association analysis using multiparty computation. Nat. Biotechnol 36, 547–551 (2018). [PubMed: 29734293]
71. Sweeney L k-anonymity: A model for protecting privacy. Int. J. Uncertainty, Fuzziness Knowledge-Based Syst 10, 557–570 (2002).
72. Malin BA An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. J. Am. Med. Informatics Assoc 12, 28–34 (2005).
73. Li N, Qardaji W & Su D On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security 32–33 (ACM, 2012).
74. Malin BA Protecting genomic sequence anonymity with generalization lattices. Methods Inf. Med 44, 687–692 (2005). [PubMed: 16400377]
75. Dwork C Differential privacy. Int. Colloq. Autom. Lang. Program 4052, 1–12 (2006).
76. Simmons S & Berger B Realizing Privacy Preserving Genome-wide Association Studies. Bioinformatics (2016).
77. Johnson A & Shmatikov V Privacy-preserving data exploration in genome-wide association studies in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13 1079 (ACM Press, 2013). doi:10.1145/2487575.2487687.
78. Yu F & Ji Z Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies: An Application to iDASH Healthcare Privacy Protection Challenge. BMC Med. Inform. Decis. Mak 14, S3 (2014). [PubMed: 25521367]
79. Uhlerop C, Slavković A & Fienberg SE Privacy-preserving data sharing for genome-wide association studies. J. Priv. Confidentiality 5, 137 (2013).
80. Backes M, Berrang P, Humbert M & Manoharan P Membership privacy in MicroRNA-based studies in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security 319–330 (ACM, 2016).
81. Tramèr F, Huang Z, Hubaux J-P & Ayday E Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security 1286–1297 (ACM, 2015).
82. Raisaro JL et al. Protecting privacy and security of genomic data in I2B2 with homomorphic encryption and differential privacy. IEEE/ACM Trans. Comput. Biol. Bioinforma 15, 1413–1426 (2018).
83. Huang Z, Ayday E, Fellay J, Hubaux J-P & Juels A GenoGuard: Protecting Genomic Data against Brute-Force Attacks. in 36th IEEE Symposium on Security and Privacy (2015).
84. Juels A & Ristenpart T Honey encryption: Security beyond the brute-force bound in Annual International Conference on the Theory and Applications of Cryptographic Techniques 293–310 (Springer, 2014).
85. Humbert M, Ayday E, Hubaux J-P & Telenti A Reconciling utility with privacy in genomics in Proceedings of the 13th Workshop on Privacy in the Electronic Society 11–20 (ACM, 2014).
86. Allyse MA, Robinson DH, Ferber MJ & Sharp RR Direct-to-consumer testing 2.0: emerging models of direct-to-consumer genetic testing in Mayo Clinic Proceedings vol. 93 113–120 (Elsevier, 2018). [PubMed: 29304915]
87. Future of Privacy Forum. Privacy best practices for consumer genetic testing services. https://fpf.org/wp-content/uploads/2018/07/Privacy-Best-Practices-for-Consumer-Genetic-Testing-Services-FINAL.pdf (2018).
88. Wee R, Henaghan M & Winship I Ethics: Dynamic consent in the digital age of biology: online initiatives and regulatory considerations. J. Prim. Health Care 5, 341–347 (2013). [PubMed: 24294625]
89. Mackey TK et al. 'Fit-for-purpose?'–challenges and opportunities for applications of blockchain technology in the future of healthcare. BMC Med. 17, 68 (2019). [PubMed: 30914045]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

90. Maxmen A AI researchers embrace Bitcoin technology to share medical data. Nature 555, (2018).

91. Lawler M et al. All the world's a stage: facilitating discovery science and improved cancer care through the global alliance for genomics and health. Cancer Discov. 5, 1133–1136 (2015). [PubMed: 26526696]

92. Phillips AM Only a click away—DTC genetics for ancestry, health, love… and more: a view of the business and regulatory landscape. Appl. Transl. genomics 8, 16–22 (2016).

93. Simmons S, Sahinalp C & Berger B Enabling privacy-preserving GWASs in heterogeneous human populations. Cell Syst. 3, 54–61 (2016). [PubMed: 27453444]

94. Yu F, Fienberg SE, Slavković AB & Uhler C Scalable privacy-preserving data sharing methodology for genome-wide association studies. J. Biomed. Inform 50, 133–141 (2014). [PubMed: 24509073]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

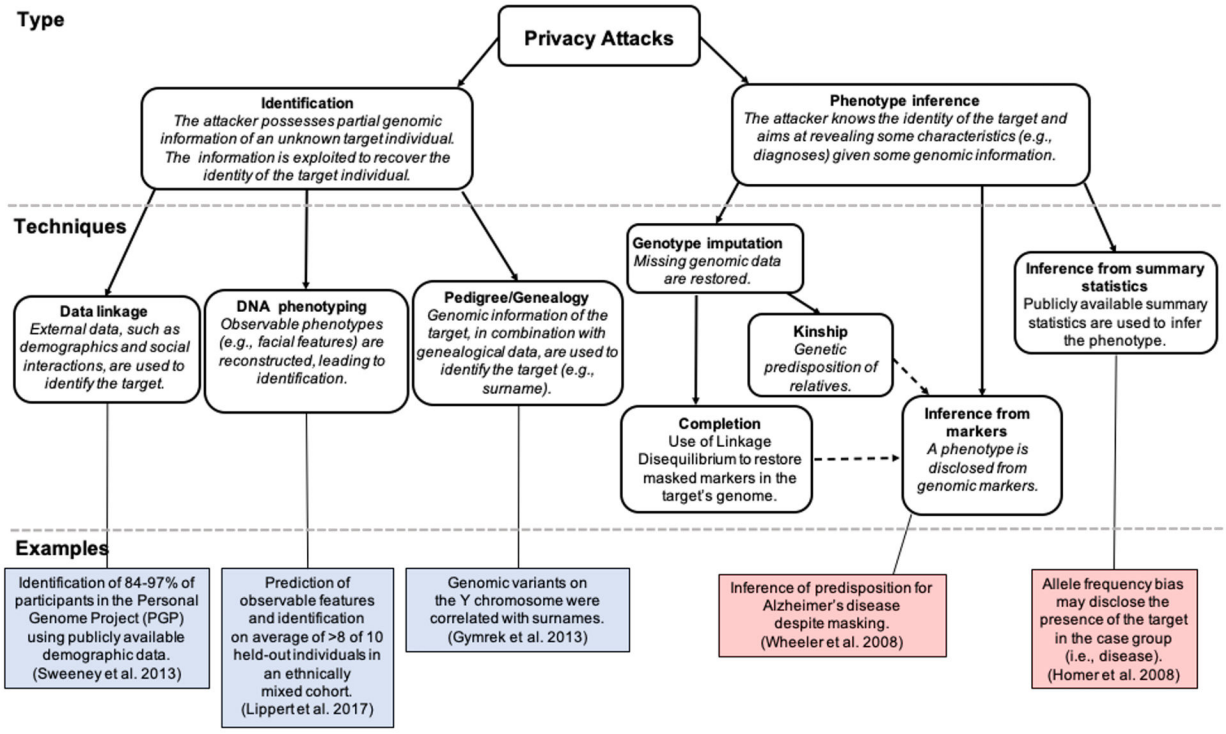

Figure 1.

Taxonomy of known privacy attacks in genomic data sharing. We differentiate between two main categories of privacy attacks: identification and phenotype inference. For each type of attack, we highlight the main known techniques and report relevant published examples.

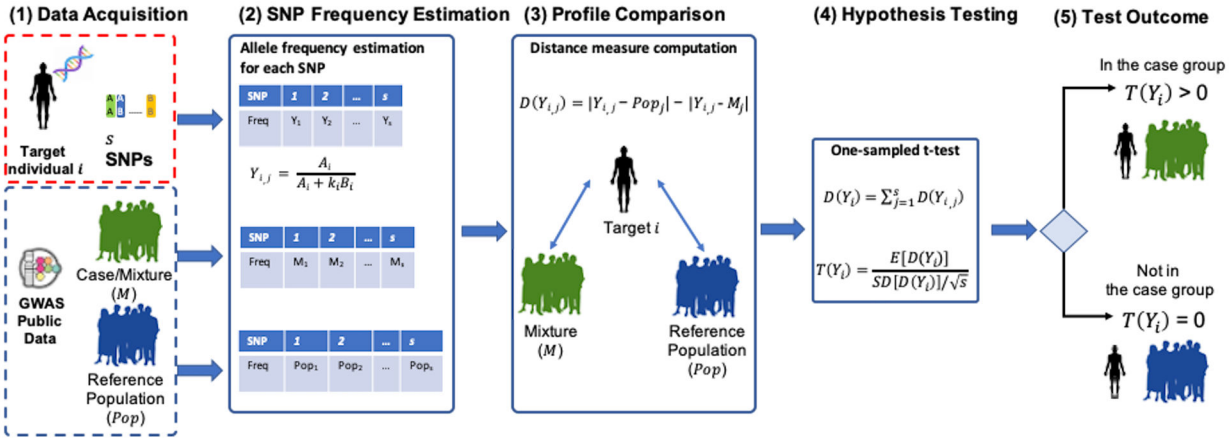

Figure 2.

Membership disclosure attack by Homer et al.,[10] where an adversary aims at determining the presence of the target in the mixture (e.g., case group). (1) Data Acquisition: the attacker has partial genomic data of a known target individual (i.e., SNPs) and he has access to publicly available summary of statistics (e.g., GWAS). (2) SNP Frequency Estimation: the attacker estimates the allele frequency for each j-SNP in the target data ($Y_{i,j}$), in the mixture (M_j), and in the reference population (Pop_j). (3) Profile Comparison: a SNP-wise distance measure ($D(Y_{i,j})$) is computed to determine how the profile of the target deviates from the reference population and mixture. Notice that $D(Y_{i,j})$ is positive when $Y_{i,j}$ is closer to M_j and negative when $Y_{i,j}$ is closer to Pop_j. Furthermore, for a sufficiently large sample, the distance $D(Y_{i,j})$ follows a normal distribution. (4) Hypothesis Testing: a one-sampled t-test is performed by the attacker to determine the likelihood of the target belonging to the mixture, where $E[\cdot]$ and $SD[\cdot]$ denote the expectation and standard deviation, respectively, and s denotes the number of SNPs. (5) Test Outcome: a positive test indicates that the target belongs to the mixture. As a result, the attacker may learn that the target individual has the phenotype that defines a “case”.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

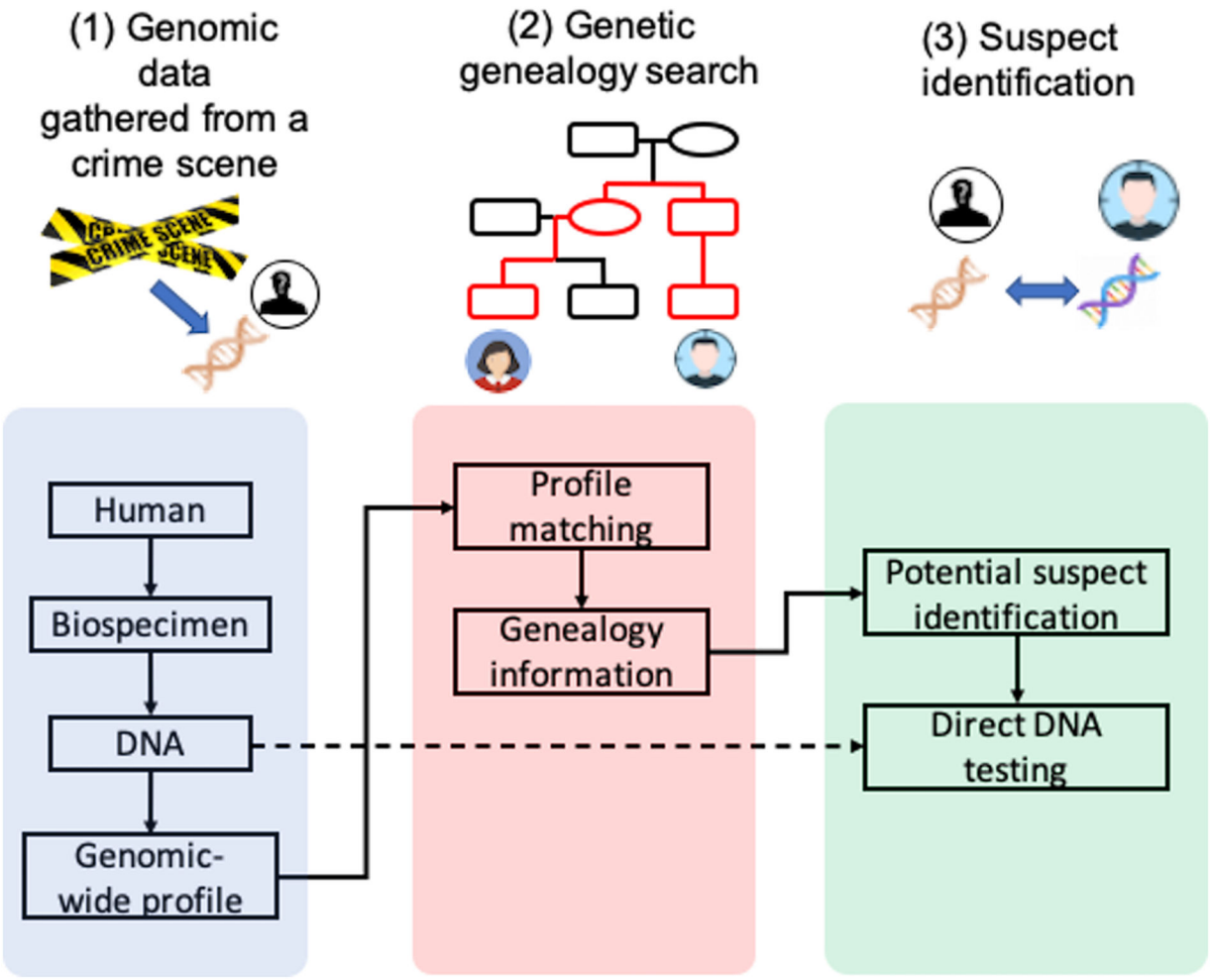

Figure 3.

Genetic Genealogy Search framework for forensics analysis. (1) Genomic data are collected from the crime scene, and a genomic-wide profile of the subject is constructed. (2) A search of matching profiles is conducted on publicly available datasets. The genomic information may lead to the identification of a match representing a relative (e.g., cousin). The genealogical information is used to narrow down the family tree, for individuals who may be suspects (e.g., living in the vicinity of the crime scene). (3) When a suitable suspect is identified, a direct DNA test is performed to confirm the match with the DNA collected from the crime scene.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Example of auxiliary information. We report some examples of auxiliary information and available data sources that can be exploited by an adversary to perform identification and phenotype inference attacks. PGP: Personal Genome Project, CEPH: Centre d'Etude du Polymorphism Humain, GTEx: Genotype-Tissue Expression, dbGaP: database of Genotypes and Phenotypes, GWAS: genome-wide association study, UK BioBank: biobank study in the United Kingdom.

Auxiliary information	Identification	Phenotype inference	Examples of data sources
Demographics, Surnames	X		Census Data (https://www.census.gov/data.html)
Pedigree, Family Tree	X	X	PGP (https://pgp.med.harvard.edu) CEPH (http://www.cephb.fr)
Gene Expression	X	X	GTEx Project (https://gtexportal.org/home/)
Genotype Data	X	X	OpenSNP (https://www.opensnp.org) 1000 Genomes Project (https://www.internationalgenome.org) dbGaP (https://www.ncbi.nlm.nih.gov/gap/)
Social Relationships	X	X	Population Registry, Social Networks
Observable Phenotypes		X	Social Networks
Clinical Data	X	X	Clinical Data Research Networks
Summary of Statistics		X	UK BioBank (https://www.ukbiobank.ac.uk)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

List of popular DTC companies (in alphabetical order) providing health-related services based on genomic data.

DTC Company	Year Founded	Number of Individuals	Main Services
23andMe (https://www.23andme.com)	2006	>10 Millions	Medical, Genealogical, Personal Ancestry
AncestryDNA (https://www.ancestry.com/dna/)	2002	>16 Millions	Genealogical, Personal Ancestry (Autosomal only)
FamilyTreeDNA (https://www.familytreedna.com)	1999	>1.1 Million	Genealogical, Personal Ancestry (Autosomal only)
GEDmatch (https://www.gedmatch.com)	2010	>1.3 Million	Genetic Genealogy Search
MyHeritage (https://www.myheritage.com)	2003	>3 Million	Genealogical, Personal Ancestry (Autosomal only)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

An overview of core techniques used for privacy-preserving genomic data.

	Goal	Techniques	Privacy Protection	Pros	Cons	Examples of Relevant Applications
Data Security	Blocking unauthorized users to access the original data	Access control, Trust-but-verify	Grant access only to authorized users	Easy to implement Allow monitoring of data usage	Vulnerable to internal attacks (e.g., a dishonest user who has access to the data)	Data Repositories such as: dbGaP (https://www.ncbi.nlm.nih.gov/gap/), EGA (https://www.ebi.ac.uk/ega/home), and the All of Us Research Program (https://www.researchallofus.org)
		Homomorphic Encryption (HE)	Data are encrypted, generating ciphertext, on which certain operations can be performed and produce the same results as when the original, non-encrypted data are used	Strong and provable security guarantees	Computationally intense	Genomic sequence matching[53,54,56], outsource computation[57–62]
		Secure Multiparty Computation (SMC)	Data are encrypted and multiple parties can jointly compute a function without learning anything about each others' private data	Strong and provable security guarantees	High communication cost	Genomic sequence comparison[64–66], secure statistical test evaluation[67,68], and GWAS[69]
Data Anonymization	Protect the identity/ presence of the individual in shared data	*k*-anonymity via generalization and suppression of SNPs	Data are transformed such that, for each record in the output, there are *k-1* other records with the same set of quasi-identifiers	Intuitive notion of privacy	Vulnerable against an informed adversary May lead to overly generalized data	"Anonymization" of DNA sequences[73,74]
		Differential privacy (adversary wants to know if target is in the database) achieved via random perturbation	Data results are perturbed to guarantee that an adversary who observes the outputs cannot determine the presence of any individual record in the data	Strong and provable privacy guarantees	Released data may have limited usability due to the noise injected	GWAS test statistics (e.g., χ^2) and SNPs highly associated with diseases of interest[77–79,93,94]