Published in final edited form as:
Nat Rev Genet. 2014 June ; 15(6): 409–421. doi:10.1038/nrg3723.

Routes for breaching and protecting genetic privacy

Yaniv Erlich[1,*] and **Arvind Narayanan**[2]

[1]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA USA 02142

[2]Department of Computer Science, Princeton University, 35 Olden Street, Princeton, NJ USA 08540

Abstract

We are entering an era of ubiquitous genetic information for research, clinical care and personal curiosity. Sharing these datasets is vital for progress in biomedical research. However, one growing concern is the ability to protect the genetic privacy of the data originators. Here, we present an overview of genetic privacy breaching strategies. We outline the principles of each technique, point to the underlying assumptions, and assess its technological complexity and maturation. We then review potential mitigation methods for privacy-preserving dissemination of sensitive data and highlight different cases that are relevant to genetic applications.

Introduction

We produce genetic information for research, clinical care and out of personal curiosity at exponential rates. Sequencing studies including thousands of individuals have become a reality[1,2], and new projects aim to sequence hundreds of thousands to millions of individuals[3]. Some geneticists envision whole genome sequencing of every person as part of routine health care[4,5].

Sharing genetic findings is vital for accelerating the pace of biomedical discoveries and fully realizing the promises of the genetic revolution[6]. Recent studies suggest that robust predictions of genetic predispositions to complex traits from genetic data will require the analysis of millions of samples[7,8]. Clearly, collecting cohorts at such scales is typically beyond the reach of individual investigators and cannot be achieved without combining different sources. In addition, broad dissemination of genetic data promotes serendipitous discoveries through secondary analysis, which is necessary to maximize its utility for patients and the general public[9].

One of the key issues of broad dissemination is an adequate balance of data privacy[10]. Prospective participants of scientific studies have ranked privacy of sensitive information as one of their top concerns and a major determinant of participation in a study[11–13]. Recently,

*Correspondence to: Y.E (yaniv@wi.mit.edu).

Competing interests statement
None.

public concerns regarding medical data privacy halted a massive plan of the National Health Service in the UK to create a centralized health-care database[14]. In addition, protecting personal identifiable information is also a demand of an array of regulatory statutes in the USA and in the European Union[15]. Data de-identification, the removing of personal identifiers, has been suggested as a potential path to reconcile data sharing and privacy demands[16]. But is this approach technically feasible for genetic data?

This review categorizes privacy breaching techniques that are relevant to genetic information and maps potential counter-measures. We first categorize privacy-breaching strategies (Figure 1), discuss their underlying technical concepts, and evaluate their performance and limitations (Table 1). Then, we present privacy-preserving technologies, group them according to their methodological approaches, and discuss their relevance to genetic information. As a general theme, we focus only on breaching techniques that involve data mining and fusing distinct resources to gain private information relevant to DNA data. Data custodians should be aware that security threats can be much broader. They can include cracking weak database passwords, classic techniques of hacking the server that holds the data, stealing of storage devices due to poor physical security, and intentional misconduct of data custodians[17–19]. We do not include these threats since they have been extensively discussed in the computer security field[20]. In addition, this review does not cover the potential implications of loss of privacy, which heavily depend on cultural, legal and socio-economical context and have been covered in part by the broad privacy literature[21,22].

Identity Tracing attacks

The goal of identity tracing attacks is to uniquely identify an anonymous DNA sample using quasi-identifiers – residual pieces of information that are embedded in the dataset. The success of the attack depends on the information content that the adversary can obtain from these quasi-identifiers relative to the size of the base population (Box 1).

Searching with meta-data

Genetic datasets are typically published with additional metadata, such as basic demographic details, inclusion and exclusion criteria, pedigree structure, as well as health conditions that are critical to the study and for secondary analysis. These pieces of metadata can be exploited to trace the identity of the unknown genome.

Unrestricted demographic information conveys substantial power for identity tracing. It has been estimated that the combination of date of birth, sex, and 5-digit zip code uniquely identifies more than 60% of US individuals[23,24]. In addition, there are extensive public resources with broad population coverage and search interfaces that link demographic quasi-identifiers to individuals, including voter registries, public record search engines (such as PeopleFinders.com) and social media. An initial study reported the successful tracing of the medical record of the Governor of Massachusetts using demographic identifiers in hospital discharge information[25]. Another study reported the identification of 30% of Personal Genome Project (PGP) participants by demographic profiling that included zip code and exact birthdates found in PGP profiles[26].

Since the inception of the Health Insurance Portability and Accountability Act (HIPAA) Privacy rule, dissemination of demographic identifiers have been the subject of tight regulation in the US health care system[27]. The safe harbor provision requires that the maximal resolution of any date field, such as hospital admissions, will be in years. In addition, the maximal resolution of a geographical subdivision is the first three digits of a zip code (for zip codes of populations of greater than 20,000). Statistical analyses of the census data and empirical health records have found that the Safe Harbor provision provides reasonable immunity against identity tracing assuming that the adversary has access only to demographic identifiers. The combination of sex, age, ethnic group, and state is unique in less than 0.25%of the populations of each of the states[28,29].

Pedigree structures are another piece of metadata that are included in many genetic studies. These structures contain rich information, especially when large kinships are available[30]. A systematic study analysed the distribution of 2,500 two-generation family pedigrees that were sampled from obituaries from a US town of 60,000 individuals[31]. Only the number (but not the order) of male and female individuals in each generation was available. Despite this limited information, about 30% of the pedigree structures were unique, demonstrating the large information content that can be obtained from such data.

Another vulnerability of pedigrees is combining demographic quasi-identifiers across records to boost identity tracing despite HIPAA protections. For example, consider a large pedigree that states the age and state of all participants. The age and state of each participant leaks very minimal information, but knowing the ages of all first and second-degree relatives of an individual dramatically reduces the search space. Moreover, once a single individual in a pedigree is identified, it is easy to link between the identities of other relatives and their genetic datasets. The main limitation of identity tracing using pedigree structures alone is their low searchability. Family trees of most individuals are not publicly available, and their analysis requires indexing a large spectrum of genealogical websites. One notable exception is Israel, where the entire population registry was leaked to the web in 2006, allowing the construction of multi-generation family trees of all Israeli citizens[32].

Identity tracing by genealogical triangulation

Genetic genealogy attracts millions of individuals interested in their ancestry or in discovering distant relatives[33]. To that end, the community has developed impressive online platforms to search for genetic matches, which can be exploited by identity tracers. One potential route of identity tracing is surname inference from Y-chromosome data[34,35] (Figure 2). In most societies, surnames are passed from father to son, creating a transient correlation with specific Y chromosome haplotypes[36,37]. The adversary can take advantage of the Y chromosome–surname correlation and compare the Y haplotype of the unknown genome to haplotype records in recreational genetic genealogy databases. A close match with a relatively short time to the most common recent ancestor (MRCA) would signal that the unknown genome likely has the same surname as the record in the database.

The power of surname inference stems from exploiting information from distant patrilineal relatives of the unknown's genome. Empirical analysis estimated that 10–14% of US white male individuals from the middle and upper classes are subject to surname inference based

on scanning the two largest Y-chromosome genealogical websites with a built-in search engine[35]. Individual surnames are relatively rare in the population, and in most cases a single surname is shared by less than 40,000 US male individuals[35], which is equivalent to 13 bits of information (Box 1). In terms of identification, successful surname recovery is nearly as powerful as finding one's zip code. Another feature of surname inference is that surnames are highly searchable. From public record search engines to social networks, numerous online resources offer query interfaces that generate a list of individuals with a specific surname. Surname inference has been utilized to breach genetic privacy in the past[38–41]. Several sperm donor conceived individuals and adoptees successfully used this technique on their own DNA to trace their biological families. In the context of research samples, a recent study reported five successful surname inferences from Illumina datasets of three large families that were part of the 1000 Genomes project, which eventually exposed the identity of nearly fifty research participants[35].

The main limitation of surname inference is that haplotype matching relies on comparing Y chromosome Short Tandem Repeats (Y-STRs). Currently, most sequencing studies do not routinely report these markers, and the adversary would have to process large-scale raw sequencing files with a specialized tool[42]. Another complication is false identification of surnames and inference of surnames with spelling variants compared to the original surname. Eliminating incorrect surname hits necessitates access to additional quasi-identifiers such as pedigree structure and typically requires a few hours of manual work. Finally, in certain societies, a surname is not a strong identifier and its inference does not provide the same power for re-identification as in the USA. For example, 400 million people in China hold one of the ten common surnames[36], and the top hundred surnames cover almost 90% of the population[43], dramatically reducing the utility of surname inference for re-identification.

An open research question is the utility of non Y chromosome markers for genealogical triangulation. Websites such as Mitosearch.org and GedMatch.com run open searchable databases for matching mitochondrial and autosomal genotypes, respectively. Our expectation is that mitochondrial data will not be very informative for tracing identities. The resolution of mitochondrial searches is low due to the small size of the mitochondrial genome, meaning that a large number of individuals share the same mitochondrial haplotypes. In addition, matrilineal identifiers such as surname or clan are relatively rare in most human societies, complicating the usage of mitochondria haplotype for identity tracing. Autosomal searches on the other hand can be quite powerful. Genetic genealogy companies have started to market services for dense genome-wide arrays that enable the identification of distant relatives (on the order of 3rd to 4th cousins) with fairly sufficient accuracy[44]. These hits would reduce the search space to no more than a few thousand individuals[45]. The main challenge of this approach would be to derive a list of potential people from a genealogical match. As we stated earlier, family trees of most individuals are not publicly available, making such searches a very demanding task that would require indexing a large spectrum of genealogical websites. With the growing interest in genealogy, this technique might be easier in the future and should be taken into consideration.

Identity tracing by phenotypic prediction

Several reports on genetic privacy have envisioned that predictions of visible phenotypes from genetic data could serve as quasi-identifiers for identity tracing[46,47]. Twin studies have estimated high heritabilities for various visible traits such as height[48] and facial morphology[49]. In addition, recent studies show that age prediction is possible from DNA specimens derived from blood samples[50,51]. But the applicability of these DNA-derived quasi-identifiers for identity tracing has yet to be demonstrated.

The major limitation of phenotypic prediction is the fast decay of the identification power with small inference errors (Box 1). Current genetic knowledge explains only a small extent of the phenotypic variability of most visible traits, such as height[52], body mass index (BMI)[53], and face morphology[54], substantially limiting their utility for identification. For example, perfect knowledge about height at one-centimeter resolution conveys 5 bits of information. However, with current genetic knowledge that explains 10% of height variability[52], the adversary learns only 0.15 bits of information. Predictions of face morphology and BMI are much worse[8,54]. The exceptions in visible traits are eye colour[55] and age prediction[50]. Recent studies show a prediction accuracy of 75–90% of the phenotypic variability of these traits. But even these successes translate to no more than 3–4 bits of information. Another challenge for phenotypic prediction is the low searchability of some of these traits. We are not aware of population-wide registries of height, eye colour or face morphology that are publicly accessible and searchable. However, future developments in social media might circumvent this barrier.

Identity tracing by side-channel leaks

Side-channel attacks exploit quasi-identifiers that are unintentionally encoded in the database building blocks and structure rather than the actual data that is meant to be public. A good example for such leaks is the exposure of the full names of PGP participants from filenames in the database[26]. The PGP allowed participants to upload 23andMe genotyping files to their public profile webpages. While it seemed that these do not contain explicit identifiers, after downloading and decompressing the 23andMe file, the original filename, whose default is the first and last name of the user, appeared. Since most of the users did not change the default naming convention, it was possible to trace the identity of a large number of PGP profiles. The PGP now offers instructions to participants how to rename files before uploading and warns them that the file may contain hidden information that can expose their identities. Generally, certain types of files, such as Microsoft Office products, can embed deleted text or hidden identifiers[56]. Data custodians should be aware that mere scanning of the file content might not always be sufficient to ensure that all identifiers have been removed.

The mechanism to generate database accession numbers can also leak personal information. For example, in a top medical data mining contest, the accession numbers revealed the disease status of the patient, which was the aim of the contest[57]. In addition, pattern analysis of a large amount of public data revealed temporal and spatial commonalities in the assignment system that allowed predictions of US social security numbers (SSNs) from quasi-identifiers[58]. Some suggested the assignment of accession numbers by applying

cryptographic hashing to the participant identifiers, such as name or SSN[59]. However, this technique is extremely vulnerable to dictionary attacks due to the relatively low search space of the input. In general, it is advisable to add some sort of randomization to procedures that generate accession numbers.

Attribute disclosure attacks via DNA (ADAD)

Consider the following scenario Alice interviews Bob for a certain position

After the interview, Alice recovers Bob's DNA and uses this data to search a large genetic study of drug abuse. The study stores the DNA in anonymous form, but a match between Bob's DNA and one of the records reveals that Bob was a drug abuser. While the short story above has some practical limitations, it illustrates the main concepts of ADAD attack. The adversary gains access to the DNA sample of the target. He or she uses the identified DNA to search genetic databases with sensitive attributes (for example, drug abuse). A match between the identified DNA and the database links the person and the attribute.

The n=1 scenario—The simplest scenario of ADAD is when the sensitive attribute is associated with the genotype data of the individual. The adversary can simply match the genotype data that is associated with the identity of the individual and the genotype data that is associated with the attribute. Such an attack requires only a small number of autosomal single nucleotide polymorphisms (SNPs). Empirical data showed that a carefully chosen set of 45 SNPs is sufficient to provide matches with a type I error of 10^{-15} for most of the major populations across the globe[60]. Moreover, random subsets of ~300 common SNPs yield sufficient information to uniquely identify any person[61]. As such, an individual's genome is a strong identifier. In general, ADAD is a theoretical vulnerability of virtually any individual level DNA-derived omics dataset such as RNA-seq and personal proteomics.

Genome-wide association studies (GWAS) are highly vulnerable to ADAD. In order to address this issue, several organizations, including the NIH, have adopted a two-tier access system for GWAS datasets: a restricted access area that stores individual level genotypes and phenotypes and a public access area for high level data summary statistics of allele frequencies for all cases and controls[62]. The premise of this distinction was that summary statistics enable secondary data usage for meta-GWAS analysis while it was thought that this type of data is immune to ADAD.

The summary statistic scenario—A landmark study in 2008 reported the possibility of ADAD on GWAS datasets that only consist of the allele frequencies of the study participants[63]. The underlying concept of this approach is that, with the target genotypes in the case group, the allele frequencies will be positively biased towards the target genotypes compared to the allele frequencies of the general population. A good illustration of this concept is considering an extremely rare variation in the subject's genome. Non-zero allele frequency of this variation in a small-scale study increases the likelihood that the target was part of the study, whereas zero allele frequency strongly reduces this likelihood. By integrating the slight biases in the allele frequencies over a large number of SNPs, it is also possible to conduct ADAD with the common variations that are analysed in GWAS.

Subsequent studies extended the range of vulnerabilities for summary statistics. One line of studies improved the test statistic in the original work and analysed its mathematical properties[64–66]. Under the assumption of common SNPs in linkage-equilibrium (LD), the improved test statistic is mathematically guaranteed to yield maximal power for any specificity level (Box 2). Another group went beyond allele frequencies and demonstrated that it is possible to exploit local LD structures for ADAD[67]. The power of this approach stems from scavenging for the co-occurrence of two relatively uncommon alleles in different haplotype blocks that together create a rare event. Another study developed a method to exploit the effect sizes of GWAS involving quantitative traits to detect the presence of the target[68]. A powerful development of this study is exploiting GWAS studies that utilize the same cohort for multiple phenotypes. The adversary repeats the identification process of the target with the effect sizes of each phenotype and integrates them to boost the identification performance. After determining the presence of the target in a quantitative trait study, the adversary can further exploit the GWAS data to predict the phenotypes with high accuracy[69].

The actual risk of ADAD has been the subject of intense debate. Following the original 2008 study[63], the NIH and other data custodians moved their GWAS summary statistics data from public databases to access-controlled databases such as dbGAP[70]. A retrospective analysis found that significantly fewer GWAS studies publicly released their summary statistics data after the discovery of this attack[71]. As of now, most of the studies publish summary statistic data on 10–500 SNPs, which is compatible with one suggested guideline to manage risk[69]. However, some researchers have warned that these policies are too harsh[72]. There are several practical complications that the adversary needs to overcome to launch a successful attack, such as access to the target's DNA data[73] and accurate matching between the target ancestries and those listed in the reference database[74]. Failure to address any of these prerequisites can severely impact the performance of the ADAD. In addition, for a range of GWAS studies, the associated attributes are not sensitive or private (for example, height). Thus, even if ADAD occurs, the impact on the participant should be minimal. A recent NIH workshop has proposed the release of summary statistics as the default policy and the development of an exemption mechanism for studies with increased risk due to the sensitivity of the attribute or the vulnerability level of the summary data[75].

The gene expression scenario—Databases such as the NIH's Gene Expression Omnibus (GEO) publicly hold hundreds of thousands of gene expression profiles from human that are linked to a range of medical attributes. A recent study proposed a potential route to exploit these profiles for ADAD[76]. The method starts with a training step that employs a standard expression quantitative trait loci (eQTL) analysis with a reference dataset. The goal of this step is to identify several hundred strong eQTLs and to learn the expression level distributions for each genotype. Next, the algorithm scans the public expression profiles. For each eQTL, it uses a Bayesian approach to calculate the probability distributions of the genotypes given the expression data. Last, the algorithm matches the target's genotype with the inferred allelic distributions of each expression profile and tests the hypothesis that the match is random. If the null hypothesis is rejected, the algorithm links the identity of the target to the medical attribute in the gene expression experiment.

This ADAD technique has the potential for relatively high accuracy in ideal conditions. Based on large-scale simulations, the authors predicted that the method can reach a type I error of 1×10^{-5} with a power of 85% when tested on an expression database of the entire US population.

There are several practical limitations to ADAD via expression data. While the training and inference steps are capable of working with expression profiles from different tissues, the method reaches its maximal power when the training and inference utilize eQTL from the same tissue. Additionally, there is a substantial loss of accuracy when the expression data in the training phase is collected using a different technology than the expression data in the inference phase. Another complication is that in order to fully execute the technique on a large database such as GEO, the adversary will need to manage and process substantial amounts of expression data. Due to the technical complexities, the NIH did not issue any changes to their policies regarding sharing expression data from human subjects.

Completion attacks

Completion of genetic information from partial data is a well-studied task in genetic studies, called genotype imputation[77]. This method takes advantage of the linkage disequilibrium between markers and uses reference panels with complete genetic information to restore missing genotype values in the data of interest. The very same strategies enable the adversary to expose certain regions of interest where only partial access to the DNA data is available. In a famous example of a completion attack, a recent study showed that it is possible to infer Jim Watson's predisposition for Alzheimer's disease from the ApoE locus despite masking of this gene[78]. As a result of the study, a 2Mb segment around the ApoE gene was removed from Watson's published genome.

In some cases, completion techniques also enable the prediction of genomic information when there is no access to the DNA of the target. This technique is possible when genealogical information is available in addition to genetic data. In the basic setting, the adversary obtains access to a single genetic dataset of a known individual. He then exploits this information to estimate genetic predispositions for relatives whose genetic information is inaccessible. A recent study demonstrated the feasibility of this attack by taking advantage of self-identified genetic datasets from OpenSNP.org, an internet platform for public sharing of genetic information[79]. Using Facebook searches, the research team was able to find relatives of the individuals that self-identified their genetic datasets. Next, the team predicted the genotypes of these relatives and estimated their genetic predisposition to Alzheimer's using a Bayesian approach.

In the advanced setting, the adversary has access to the genealogical and genetic information of multiple relatives of the target[80]. The algorithm finds relatives of the target that donated their DNA to the reference panel and that reside on a unique genealogical path that includes the target, for example, a pair of half-first cousins when the target is their grandfather. A shared DNA segment between the relatives indicates that the target has the same segment. By scanning more pairs of relatives that are connected through the target, it is possible to infer the two copies of autosomal loci and collect more genomic information on the target

without any access to his DNA. This approach is more accurate than the basic setting and enables to infer genotypes of more distant relatives. In Iceland, decode genetics leveraged their large reference panel and genealogical information to infer genetic variants of an additional 200,000 living individuals who never donated their DNA[81]. In May 2013, Iceland's Data Protection Authority prohibited the use of this technique until consent is obtained from the individuals who are not part of the original reference panel.

Mitigation techniques

Most of the genetic privacy breaches presented above require a background in genetics and statistics and – importantly – a motivated adversary. One school of thought posits that these practical complexities markedly diminish the probability of an adverse event[82,83]. In this view, an appropriate mitigation strategy is to simply remove obvious identifiers from the datasets before publicly sharing the information. In the field of computer security, this risk management strategy is called security by obscurity. The opponents of security by obscurity posit that risk management schemes based on the probability of an adverse event are fragile and short lasting. Technologies only get better with time and what is technically challenging but possible today will be much easier in the future. In other words, it is impossible to estimate future risks of adverse events[84]. Known in cryptography as Shannon's maxim[85], this school of thought assumes that the adversary exists and is equipped with the knowledge and means to execute the breach. Robust data protection, therefore, is achieved by explicit design of the data access protocol rather than by relying on the small chances of a breach[86].

Access control

Privacy risks are both amplified and more uncertain when data is shared publicly with no record of who accesses it. An alternative is to place sensitive data in a secure location and to screen the legitimacy of the applicants and their research projects by specialized committees. Once approval is made, the applicants are allowed to download the data under the conditions that they will store it in a secure location and will not attempt to identify individuals. In addition, the applicants are required to file periodic reports about the data usage and any adverse events. This approach is the cornerstone of dbGAP[62,87]. Based on periodic reports by users, a retrospective analysis of dbGAP access control has identified 8 data management incidents out of close to 750 studies, mostly involving non-adherence to the technical regulations, with no reports of breaching the privacy of participants[88].

Despite the absence of privacy breaches thus far, some have criticized the lack of real oversight once the data is in the hand of the applicant[89]. An alternative model uses a trust-but-verify approach, where users cannot download the data without restriction but, based on their privileges, may execute certain types of queries, which are recorded and audited by the system[90,91]. Supporters of this model state that monitoring has the potential to deter malicious users and to facilitate early detection of adverse events. One technological challenge is that audit systems usually rely on anomalous behavior to detect adversaries[92]. It is yet to be proven that such methods can reliably distinguish between legitimate and malicious use of genetic data. Auditing also requires that any interaction with the genetic datasets is done using a standard set of API calls that can be analyzed. By contrast, most of

the genomic formats currently operate using more liberal text parsing approaches, but several efforts in the community have been made to standardize genomic analysis[93,94].

Another model of access control is allowing the original participants to grant access to their data instead of delegating this responsibility to a data access committee[95,96]. This model centers on dynamic consent based on on-going communication between researchers and participants regarding data access. Supporters of this model state that this approach streamlines the consent process, enables participants to modify their preferences throughout their lifetimes, and can promote greater transparency, higher levels of participant engagement, and oversight. An example for such an effort is PEER (Platform for Engaging Everyone Responsibly). In this setting, Private Access Inc. operates a service that manages the access rights and mediates the communication between researchers and participants, without revealing the identity of the participants. A trusted agent, Genetic Alliance, holds the participants health data, offers stewardship regarding privacy preferences, and grant access to data based on participants' decisions. Participant-based access control is still a relatively new method. As data custodians gain more experience with such a framework, a better picture will emerge regarding its utility as an alternative for risk-benefit management compared to traditional access control methodologies.

Data anonymization

The premise of anonymity is the ability to be 'lost in the crowd'. One line of studies suggested restoring anonymity by restricting the granularity of quasi-identifiers to the point that no record in the database has a unique combination of quasi-identifiers. One heuristic is k-anonymity[97], in which attribute values are generalized or suppressed such that for each record there are at least 'k-1' records with the same combination of quasi-identifiers. To maximize the utility of the data for subsequent analysis, the generalization process is adaptive. Certain records will have a lower resolution depending on the distribution of the other records and certain data categories that are too unique are suppressed entirely. There is a strong trade-off in the selection of the value of k; high values better protect privacy but at the same time reduce the utility of the data. As a rule of thumb, k=5 is commonly used in practice[98]. Most of the k-anonymity work centers on protecting demographic identifiers. For genetic data, one study suggested a 2-anonymity protocol by generalizing the 4 nucleotides in DNA sequences into broader types of biochemical groups such as pyrimidine and purines[99]. However, the utility of such data for broad genetic applications is unclear. Furthermore, k-anonymity is vulnerable to attribute disclosure attacks when the adversary has prior knowledge about the presence of the target in the database[100,101]. Thus, while this heuristic is easy to comprehend, its privacy properties as well as its relevance to genomic studies are in question.

Differential privacy is an emerging methodology for privacy-preserving reporting of results, primarily of summary statistics[102] (Box 3). In contrast to k-anonymity, this method guarantees privacy against an adversary with arbitrary prior knowledge. Differential privacy operates by adding noise to the results before their release. The algorithm tunes the amount of noise such that the reported results will be statistically indistinguishable from similar reported results that would have been obtained if a single record had been removed from the

original dataset. This way, an adversary with any type of prior knowledge can never be sure whether a specific individual was part of the original dataset because the data release process produces results that are almost exactly the same if the individual was not included. Due to its theoretical guarantees and tractable computational demands, differential privacy has become a vibrant research area in computer science and statistics. In perhaps the best-known large-scale implementation, the US Census Bureau utilizes this technique for privacy-preserving release of data in the online OnTheMap tool[103].

In the context of genetic privacy, several studies have explored differential private release of common summary statistics of GWAS data, such as the allele frequencies of cases and controls, χ^2-statistic, and p-values[104,105] or shifting the original locations of variants[106]. Currently, these techniques require a large amount of noise even for the release of a GWAS statistics from a small number of SNPs, which renders these measures impractical. It is unclear whether there is a perturbation mechanism that can add much smaller amounts of noise to GWAS results while satisfying the differential privacy requirement, or whether perturbation can be shown to be effective for privacy preservation under a different theoretical model.

Cryptographic solutions

Modern cryptography brought new advancements to data dissemination beyond the traditional usage of encrypting sensitive information and distributing the key to authorized users. These solutions enable well-defined usability of data while blocking unauthorized operations. Different from solutions in the previous section, the underlying data is not perturbed within the authorized usability.

One line of cryptographic work considers the problem of privacy-preserving approaches to outsource computation on genetic information to third parties. For example, with the advent of ubiquitous genetic data, patients (or their physicians) will interact throughout their lives with a variety of online genetic interpretation applications, such as promethease.com, increasing the chance of a privacy breach. Recent cryptographic work has suggested homomorphic encryption (Box 4) for secure genetic interpretation[107]. In this method, users send encrypted versions of their genomes to the cloud. The interpretation service can access the cloud data but does not have the key and therefore cannot read the plain genotype values. Instead, the interpretation service executes the risk prediction algorithm on the encrypted genotypes. Due to the special mathematical properties of the underlying cryptosystem, the user simply decrypts the results given by the interpretation service to obtain the risk prediction. This way, the user does not expose genotypes or disease susceptibility to the service provider and interpretation companies can offer their service to users concerned with privacy. Preliminary results have highlighted the potential feasibility of this scheme[108]. A proof-of-concept study encrypted the variants of a 1000 Genomes individual and simulated a secure inference of heart disease risk based on 23 SNPs and 17 environmental factors. The total size of the encrypted genome was 51 Gbyte and the risk calculation took 6 minutes on a standard computer. The current scope of risk prediction models is still narrow but this approach might be quite amenable to future improvements.

Cryptographic studies have also considered the task of outsourcing read mapping without revealing any genetic information to the service provider[109–111]. The basis of some of these protocols is Secure Multiparty Computation (SMC). SMC allows two or more entities who each have some private data to execute a computation on these private inputs without revealing the input to each other or disclosing it to a third party. In one classic example of SMC, two individuals can determine who is richer without either one revealing their actual wealth to the other[112]. Earlier studies suggested SMC versions for edit distance-based mapping of DNA sequences that does not reveal their content[109,110]. However, regular (unsecure) edit distance-based mapping is too slow to handle the volumes of high-throughput sequencing reads, narrowing the applications for the much-slower secure version. A more recent study proposed a privacy-preserving version of the popular seed-and-extend algorithm[111], which serves as the basis of several high-throughput alignment tools[111,113]. The privacy-preserving version is a hybrid: the seeding part is securely outsourced to a cloud where a cryptographic hashing hides the actual DNA sequences while permitting string matching. The cloud results are streamed to a local trusted computer that performs the extension part. By tuning the underlying parameters of the seed-and-extend algorithm, this method puts most of the computation burden on the cloud. Experiments with real sequencing data showed that the cloud performs >95% of the computation efforts. In addition, the secure algorithm takes only 3.5× longer than a similar unsecure implementation, suggesting a tractable price tag to maintain privacy.

Beyond outsourcing of computation, several studies designed cryptographically secure algorithms for searching genetic databases. One study suggested searchable genetic databases for forensic purposes that allow only going from genetic data to identity but not from identity to genetic data[114]. The forensic database stores the individuals' names and contact information in an encrypted form. The key for each entry is the corresponding individual's genotypes. This way, knowing genotype information (for example, from a crime scene) can reveal the identity but not the opposite. In addition, to tolerate genotyping errors or missing data, the study suggested a fuzzy encryption scheme in which a decryption key can approximately match the original key. Another cryptographic protocol proposed matching genetic profiles between two parties for paternity tests or carrier screening without exposing the actual genetic data[115,116]. A smartphone-based implementation was presented for one version of this algorithm[117]. A recent study suggested a scalable approach for finding relatives using genome-wide data without disclosing the raw genotypes to a third party or other participants[118]. First, users collectively decide the minimal degree of relatedness they wish to accept. Next, each user posts a secure version of her genome to a public repository using a fuzzy encryption scheme. Then, users compare their own secure genome to the secure genomes of other users. Comparison of two encrypted genomes reveals no information if the genomes are farther than the threshold degree of relatedness; otherwise, it reveals the exact genetic distance. An evaluation of the efficacy of this approach via experiments with hundreds of individuals from the 1000 Genomes Project showed that even second-degree relatives can reliably find each other[118].

A major open question is whether cryptographic protocols can facilitate data sharing for research purposes. So far, cryptographic schemes have focused on developing protocols for GWAS analysis without the need to reveal individual-level genetic data. One study

presented a scheme where genetic data and computation of GWAS contingency tables are securely outsourced via homomorphic encryption to external data centers[119]. A trusted party (for example, the NIH) acts as a gateway that accepts requests from researchers in the community, instructs the data centres to perform computation on the encrypted data, and decrypts and disseminates the GWAS results back to the researchers. A more recent study tested a scheme to generate GWAS summary statistics without a trusted party using only SMC between the data centeres[120]. Another study evaluated the outsourcing of GWAS analysis to a commercially available tamper-resistant hardware[121]. Different from the schemes above[119,120], the individual-level genotypes are decrypted as part of the GWAS summary statistics computation but the exposure occurs for a short amount of time in a secure hardware environment, which prevents any leakage. All of the cryptographic GWAS schemes above suffer from one common drawback: the protocols produce summary statistics, which are theoretically amenable to ADAD methods. As of today, cryptography has yet to devise a comprehensive data sharing solution for GWAS studies.

Conclusions

In the last few years, a torrent of studies has suggested that a motivated, technically sophisticated adversary is capable of exploiting a wide range of genetic data. With the constant innovation in genetics and the explosion of online information, we can expect that new privacy breaching techniques will be discovered in the next few years and that technical barriers to conducting existing attacks will diminish. On the other hand, privacy-preserving strategies for data dissemination are a vibrant area of research. Rapid progress has been made, and powerful frameworks such as differential privacy and homomorphic encryption are now part of the mitigation arsenal. At least for certain tasks in genetics, there are protocols that preserve the privacy of individuals. However, protecting privacy is only one facet of the solution. Lessons from computer security have highlighted that usability is a key component for the wide adoption of secure protocols. Successful implementations should hide unnecessary technical details from the users, minimize the computational overhead, and enable legitimate research[122,123]. We have yet to fully achieve this aim.

In addition, successful balancing of privacy demands and data sharing is not restricted to technical means[124]. Balanced informed consent outlining both benefits and risks are key ingredients for maintaining long-lasting credibility in genetic research. With the active engagements of a wide range of stakeholders from the broad genetics community and the general public, we as a society can facilitate the development of social and ethical norms, legal frameworks, and educational programs to reduce the chance of misuse of genetic data regardless of the ability to identify datasets.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

YE is an Andria and Paul Heafy Family Fellow and holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. This study was supported by a National Human Genome Research Institute grant

R21HG006167 and by a gift from Cathy and Jim Stone. The authors thank Dina Zielinski and Melissa Gymrek for useful comments.

Biographies

Yaniv Erlich is a Fellow at the Whitehead Institute for Biomedical Research. Erlich received his Ph.D. from Cold Spring Harbor Laboratory in 2010 and B.Sc. from Tel-Aviv University in 2006. Prior to that, Erlich worked in computer security and was responsible for conducting penetration tests on financial institutes and commercial companies. Dr. Erlich's research involves developing new algorithms for computational human genetics.

Arvind Narayanan is an Assistant Professor in the Department of Computer Science and the Center for Information Technology and Policy at Princeton. He studies information privacy and security. His research has shown that data anonymization is broken in fundamental ways, for which he jointly received the 2008 Privacy Enhancing Technologies Award. His current research interests include building a platform for privacy-preserving data sharing.

Glossary

SAFE HARBOR	A standard in the HIPAA Rule for de-identification of protected health information by removing 18 types of quasi-identifiers
HAPLOTYPES	A set of alleles along the same chromosome
CRYPTOGRAPHIC HASHING	A procedure that yields a fixed length output from any size of input in a way that is hard to determine the input from the output
DICTIONARY ATTACKS	A brute force approach to reverse cryptographic hashing by scanning the relatively small input space
ALICE AND BOB	Common generic names in computer security to denote party A and party B
TYPE I ERROR	The probability to obtain a positive answer from a negative item
LINKAGE EQUILIBRIUM	Absence of correlation between the alleles in two loci
POWER	The probability to obtain a positive answer for a positive item
SPECIFICITY	The probability to obtain a negative answer for a negative item
EFFECT SIZES	The contribution of an allele to the value of the trait
POSITIVE PREDICTIVE VALUE	The probability that a positive answer belongs to a true positive

EXPRESSION QUANTITATIVE TRAIT LOCI	Genetic variants associated with variability in gene expression
GENOTYPE IMPUTATION	A class of statistical techniques to predict a genotype from information on surrounding genotypes
LINKAGE DISEQUILIBRIUM	The correlation between alleles in two loci
API	A set of commands that specify the interface with a dataset
χ^2 STATISTIC	A measure of association in case-control GWAS studies
READ MAPPING	A computational intensive step in high throughput sequencing to find the location of a DNA strings in the genome
EDIT DISTANCE	The total number of insertions, deletions, and substitution between two strings

References

1. Fu W, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013; 493:216–220.10.1038/nature11690 [PubMed: 23201682]
2. Genomes Project, C. et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65.10.1038/nature11632 [PubMed: 23128226]
3. Roberts JP. Million veterans sequenced. Nat Biotech. 2013; 31:470–470.10.1038/nbt0613-470
4. Drmanac R. Medicine. The ultimate genetic test. Science. 2012; 336:1110–1112.10.1126/science.1221037 [PubMed: 22654043]
5. Burn J. Should we sequence everyone's genome? Yes Bmj. 2013; 346:f3133.10.1136/bmj.f3133
6. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics–re-shaping scientific practice. Nat Rev Genet. 2009; 10:331–335.10.1038/nrg2573 [PubMed: 19308065]
7. Park JH, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet. 2010; 42:570–575.10.1038/ng.610 [PubMed: 20562874]
8. Chatterjee N, et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013; 45:400–405. 405e401–403.10.1038/ng.2579 [PubMed: 23455638]
9. Friend SH, Norman TC. Metcalfe's law and the biology information commons. Nature biotechnology. 2013; 31:297–303.10.1038/nbt.2555
10. Rodriguez LL, Brooks LD, Greenberg JH, Green ED. Research ethics. The complexities of genomic identifiability. Science. 2013; 339:275–276.10.1126/science.1234593 [PubMed: 23329035]
11. Care, I. o. M. U. R. o. V. S.-D. H. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary The National Academies Collection: Reports funded by National Institutes of Health. 2010
12. McGuire AL, et al. To share or not to share: a randomized trial of consent for data sharing in genome research. Genetics in medicine : official journal of the American College of Medical Genetics. 2011; 13:948–955.10.1097/GIM.0b013e3182227589 [PubMed: 21785360]

13. Oliver JM, et al. Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. Public Health Genomics. 2012; 15:106–114.10.1159/000334718 [PubMed: 22213783]
14. Careless.data. Nature. 2014; 507:7. [PubMed: 24605371]
15. Schwartz PM, Solove DJ. Reconciling Personal Information in the United States and European Union. SSRN Electronic Journal. 201310.2139/ssrn.2271442
16. El Emam K. Heuristics for De-identifying Health Data. Security & Privacy, IEEE. 2008; 6:58–61.10.1109/MSP.2008.84
17. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. From genetic privacy to open consent. Nat Rev Genet. 2008; 9:406–411.10.1038/nrg2360 [PubMed: 18379574]
18. Brenner SE. Be prepared for the big genome leak. Nature. 2013; 498:139.10.1038/498139a [PubMed: 23765454]
19. http://www.privacyrights.org/data-breach
20. McClure S, Scambray J, Kurtz G. Hacking exposed 7 : network security secrets & solutions. 2012
21. Solove DJ. A Taxonomy of Privacy. University of Pennsylvania Law Review. 2006; 154:477.
22. Ohm P. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review. 2010; 57
23. Golle, P. Proceedings of the 5th ACM workshop on Privacy in electronic society. ACM, Alexandria; Virginia, USA: 2006. p. 77-80.
24. Sweeney LA. Simple Demographics Often Identify People Uniquely. 2000
25. Sweeney L. Testimony of Latanya Sweeney before the Privacy and Integrity Advisory Committee of the Department of Homeland Security. 2005
26. Sweeney, LA.; Abu, A.; Winn, J. Identifying Participants in the Personal Genome Project by Name. 2013. <http://dataprivacylab.org/projects/pgp/1021-1.pdf>
27. United States. General Accounting Office. & United States. U.S. General Accounting Office; Washington, D.C.: 2002.
28. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. Journal of the American Medical Informatics Association : JAMIA. 2010; 17:169–177.10.1136/jamia.2009.000026 [PubMed: 20190059]
29. Kwok, P.; Davern, M.; Hair, E.; Lafky, D. NORC at The University of Chicago. Chicago: 2011.
30. Bennett RL, et al. Recommendations for standardized human pedigree nomenclature. Pedigree Standardization Task Force of the National Society of Genetic Counselors. Am J Hum Genet. 1995; 56:745–752. [PubMed: 7887430]
31. Malin B. Re-identification of familial database records. AMIA … Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:524–528. [PubMed: 17238396]
32. Israel v. N. Bilik and others 24441-05-12. 2013. [online], (in Hebrew)
33. Khan R, Mittelman D. Rumors of the death of consumer genomics are greatly exaggerated. Genome biology. 2013; 14:139. [PubMed: 24280313]
34. 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. 2009; 84:251–258. doi:S0002-9297(09)00025-1 [pii] 10.1016/j.ajhg.2009.01.018. [PubMed: 19215731]
35. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013; 339:321–324.10.1126/science.1229566 [PubMed: 23329047]
36. King TE, Jobling MA. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. Trends Genet. 2009; 25:351–360. doi:S0168-9525(09)00133-4 [pii] 10.1016/j.tig.2009.06.003. [PubMed: 19665817]
37. King TE, Jobling MA. Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. Mol Biol Evol. 2009; 26:1093–1102. doi:msp022[pii] 10.1093/molbev/msp022. [PubMed: 19204044]
38. Motluk A. Anonymous sperm donor traced on internet. New Sci. 2005; 188:2.
39. Stein R. Found on the Web, With DNA: a Boy's Father. Washington Post. 2005; 1

40. Naik G. Family Secrets: An Adopted Man's 26-Year Quest for His Father. Wall Street Journal. 2009
41. Lehmann-Haupt R. Are Sperm Donors Really Anonymous Anymore? Slate. 2010
42. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome research. 2012 doi:
43. Network CN. China East Day. 2007
44. Huff CD, et al. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research. 2011; 21:768–774.10.1101/gr.115972.110 [PubMed: 21324875]
45. Henn BM, et al. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One. 2012; 7:e34267.10.1371/journal.pone.0034267 [PubMed: 22509285]
46. Lowrance WW, Collins FS. Ethics. Identifiability in genomic research. Science. 2007; 317:600–602.10.1126/science.1147699 [PubMed: 17673640]
47. Kayser M, de Knijff P. Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet. 2011; 12:179–192.10.1038/nrg2952 [PubMed: 21331090]
48. Silventoinen K, et al. Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin research : the official journal of the International Society for Twin Studies. 2003; 6:399–408.10.1375/136905203770326402 [PubMed: 14624724]
49. Kohn LAP. The Role of Genetics in Craniofacial Morphology and Growth. Annu Rev Anthropol. 1991; 20:261–278.
50. Zubakov D, et al. Estimating human age from T-cell DNA rearrangements. Curr Biol. 2010; 20:R970–971.10.1016/j.cub.2010.10.022 [PubMed: 21093786]
51. Ou XL, et al. Predicting human age with bloodstains by sjTREC quantification. PLoS One. 2012; 7:e42412.10.1371/journal.pone.0042412 [PubMed: 22879970]
52. Lango Allen H, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838.10.1038/nature09410 [PubMed: 20881960]
53. Manning AK, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012; 44:659–669.10.1038/ng.2274 [PubMed: 22581228]
54. Liu F, et al. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. PLoS Genet. 2012; 8:e1002932. doi:papers2://publication/doi/10.1371/journal.pgen.1002932. [PubMed: 23028347]
55. Walsh S, et al. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet. 2011; 5:170–180.10.1016/j.fsigen.2010.02.004 [PubMed: 20457092]
56. Byers S. Information leakage caused by hidden data in published documents. Security & Privacy, IEEE. 2004; 2:23–27.10.1109/MSECP.2004.1281241
57. Kaufman, S.; Rosset, S.; Perlich, C. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; San Diego, California, USA: 2011. p. 556-563.
58. Acquisti A, Gross R. Predicting Social Security numbers from public data. Proc Natl Acad Sci U S A. 2009; 106:10975–10980.10.1073/pnas.0904891106 [PubMed: 19581585]
59. Noumeir R, Lemay A, Lina JM. Pseudonymization of radiology data for research purposes. Journal of digital imaging. 2007; 20:284–295.10.1007/s10278-006-1051-4 [PubMed: 17191099]
60. Pakstis AJ, et al. SNPs for a universal individual identification panel. Hum Genet. 2010; 127:315–324.10.1007/s00439-009-0771-1 [PubMed: 19937056]
61. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science. 2004; 305:183.10.1126/science.1095019 [PubMed: 15247459]
62. Mailman MD, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet. 2007; 39:1181–1186.10.1038/ng1007-1181 [PubMed: 17898773]
63. Homer N, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008; 4:e1000167.10.1371/journal.pgen.1000167 [PubMed: 18769715]

64. Jacobs KB, et al. A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. Nat Genet. 2009; 41:1253–1257. doi:ng.455 [pii] 10.1038/ng.455. [PubMed: 19801980]
65. Visscher PM, Hill WG. The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis. PLoS Genet. 2009; 5:e1000628. doi:papers2://publication/doi/10.1371/journal.pgen.1000628. [PubMed: 19798439]
66. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. Nat Genet. 2009; 41:965–967.10.1038/ng.436 [PubMed: 19701190]
67. Wang, R.; Li, YF.; Wang, X.; Haixu, T.; Zhou, X. CCS'09. Chicago, IL, USA: 2009.
68. Im HK, Gamazon ER, Nicolae DL, Cox NJ. On Sharing Quantitative Trait GWAS Results in an Era of Multiple-omics Data and the Limits of Genomic Privacy. Am J Hum Genet. 2012; 90:591–598. doi:S0002-9297(12)00093-6 [pii] 10.1016/j.ajhg.2012.02.008. [PubMed: 22463877]
69. Lumley T. Potential for Revealing Individual-Level Information in Genome-wide Association Studies. JAMA. 2010; 303:659. doi:papers2://publication/doi/10.1001/jama.2010.120. [PubMed: 20159874]
70. Zerhouni EA, Nabel EG. Protecting aggregate genomic data. Science. 2008; 322:44.10.1126/science.1165490 [PubMed: 18772394]
71. Johnson AD, Leslie R, O'Donnell CJ. Temporal trends in results availability from genome-wide association studies. PLoS Genet. 2011; 7:e1002269.10.1371/journal.pgen.1002269 [PubMed: 21931563]
72. Gilbert N. Researchers criticize genetic data restrictions. Nature. 200810.1038/news.2008.1083
73. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. Journal of investigative medicine : the official publication of the American Federation for Clinical Research. 2010; 58:11–18.10.231/JIM.0b013e3181c9b2ea [PubMed: 20051768]
74. Clayton D. On inferring presence of an individual in a mixture: a Bayesian approach. Biostatistics. 2010; 11:661–673. doi:papers2://publication/doi/10.1093/biostatistics/kxq035. [PubMed: 20522729]
75. Workshop on Establishing a Central Resource of Data from Genome Sequencing Projects. 2012. <http://www.genome.gov/Pages/Research/DER/GVP/Data_Aggregation_Workshop_Summary.pdf>
76. Schadt EE, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. Nat Genet. 2012; 44:603–608.10.1038/ng.2248 [PubMed: 22484626]
77. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010; 11:499–511.10.1038/nrg2796 [PubMed: 20517342]
78. Nyholt DR, Yu CE, Visscher PM. On Jim Watson's APOE status: genetic information is hard to hide. European journal of human genetics : EJHG. 2009; 17:147–149.10.1038/ejhg.2008.198 [PubMed: 18941475]
79. Humbert, M.; Ayday, E.; Hubaux, JP.; Telenti, A. Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. ACM; p. 1141-1152.
80. Kong A, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet. 2008; 40:1068–1075.10.1038/ng.216 [PubMed: 19165921]
81. Kaiser J. Human genetics. Agency nixes deCODE's new data-mining plan. Science. 2013; 340:1388–1389.10.1126/science.340.6139.1388 [PubMed: 23788773]
82. Bambauer JR. Tragedy of the Data Commons. Harvard Journal of Law and Technology. 2011; 25 doi:http://dx.doi.org/10.2139/ssrn.1789749.
83. Hartzog W, Stutzman F. The Case for Online Obscurity. California Law Review. 2013; 101:1. doi:http://dx.doi.org/10.2139/ssrn.159774.
84. Taleb, NN. The black swan : the impact of the highly improbable. Random House; 2007.
85. Shannon C. Communication Theory of Secrecy Systems. Bell System Technical Journal. 1949; 28:656–715.
86. Cavoukian, A. Privacy by Design. 2009. <http://www.ipc.on.ca/images/Resources/privacybydesign.pdf>

87. Tryka KA, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014; 42:D975–D979. [PubMed: 24297256]
88. Ramos EM, et al. A mechanism for controlled access to GWAS data: experience of the GAIN Data Access Committee. Am J Hum Genet. 2013; 92:479–488.10.1016/j.ajhg.2012.08.034 [PubMed: 23561843]
89. Church G, et al. Public access to genome-wide data: five views on balancing research with privacy and protection. PLoS Genet. 2009; 5:e1000665.10.1371/journal.pgen.1000665 [PubMed: 19798440]
90. Agrawal, R.; Kiernan, J.; Srikant, R.; Xu, Y. Proceedings of the 28th international conference on Very Large Data Bases. VLDB Endowment; p. 143-154.
91. Agrawal, R., et al. Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. VLDB Endowment; p. 516-527.
92. Venter HS, Olivier MS, Eloff JH. PIDS: a privacy intrusion detection system. Internet Research. 2004; 14:360–365.
93. Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clincal Data. 2013. <https://www.broadinstitute.org/files/news/pdfs/GAWhitePaperJune3.pdf>
94. Bafna V, et al. Abstractions for genomics. Communications of the ACM. 2013; 56:83–93.
95. Terry SF, Terry PF. Power to the people: participant ownership of clinical trial data. Science Translational Medicine. 2011; 3:69cm63–69cm63.
96. Kaye J, et al. From patients to partners: participant-centric initiatives in biomedical research. Nature Reviews Genetics. 2012; 13:371–376.
97. Sweeney L. k-anonymity: a model for protecting privacy. International journal of uncertainty, fuzziness, and knowledge-based systems. 2002; 10:557–570.
98. El Emam K, Dankar FK. Protecting privacy using k-anonymity. Journal of the American Medical Informatics Association : JAMIA. 2008; 15:627–637.10.1197/jamia.M2716 [PubMed: 18579830]
99. Malin BA. Protecting genomic sequence anonymity with generalization lattices. Methods of information in medicine. 2005; 44:687–692. [PubMed: 16400377]
100. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity. ACM Trans Knowl Discov Data. 2007; 1:3-es. doi:papers2://publication/doi/10.1145/1217299.1217302.
101. Ninghui L, Tiancheng L, Venkatasubramanian S. Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. :106–115.
102. Dwork C. ICALP. :1–12.
103. Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L. Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. :277–286.
104. Uhler C, Slavkovic AB, Fienberg SE. Privacy-preserving data sharing for genome-wide association studies. arXiv preprint. 2012 arXiv:1205.0739.
105. Yu F, Fienberg SE, Slavković A, Uhler C. Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies. arXiv preprint. 2014 arXiv:1401.5193.
106. Johnson, A.; Shmatikov, V. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; Chicago, Illinois, USA: 2013. p. 1079-1087.
107. Ayday, E.; Raisaro, JL.; Hubaux, JP. Privacy-Enhancing Technologies for Medical Tests Using Genomic Data. Technical Report. 2013. <http://infoscience.epfl.ch/record/182897/files/CS_version_technical_report.pdf>
108. Hubaux, JP., et al. Proceedings of USENIX Security Workshop on Health Information Technologies (HealthTech" 13).
109. Atallah, MJ.; Kerschbaum, F.; Du, W. Proceedings of the 2003 ACM workshop on Privacy in the electronic society. ACM; p. 39-44.
110. Jha, S.; Kruger, L.; Shmatikov, V. Security and Privacy, 2008 SP 2008 IEEE Symposium on. IEEE; p. 216-230.
111. Chen, Y.; Peng, B.; Wang, X.; Tang, H. Proceeding of the 19th network & distributed system security symposium.
112. Yao, AC-C. FOCS. p. 160-164.

113. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010; 11:473–483. [PubMed: 20460430]
114. Bohannon, P.; Jakobsson, M.; Srikwan, S. Public Key Cryptography. Imai, Hideki; Zheng, Yuliang, editors. Vol. 1751. Springer; Berlin Heidelberg: 2000. p. 373-390.Lecture Notes in Computer ScienceCh. 25
115. Fons B, Stefan K, Klaus K, Pim T. Privacy-Preserving Matching of DNA Profiles. 2008; 2008
116. Baldi, P.; Baronio, R.; Cristofaro, ED.; Gasti, P.; Tsudik, G. Proceedings of the 18th ACM conference on Computer and communications security. ACM; Chicago, Illinois, USA: 2011. p. 691-702.
117. Cristofaro, ED.; Faber, S.; Gasti, P.; Tsudik, G. Proceedings of the 2012 ACM workshop on Privacy in the electronic society. ACM; Raleigh, North Carolina, USA: 2012. p. 97-108.
118. He, Dan, et al. Identifying Genetic Relatives without Compromising Privacy. Genome research. 2014
119. Kantarcioglu M, Jiang W, Liu Y, Malin B. A cryptographic approach to securely share and query genomic sequences. Information Technology in Biomedicine, IEEE Transactions on. 2008; 12:606–617.
120. Kamm L, Bogdanov D, Laur S, Vilo J. A new way to protect privacy in large-scale genome-wide association studies. Bioinformatics. 2013; 29:886–893.10.1093/bioinformatics/btt066 [PubMed: 23413435]
121. Canim M, Kantarcioglu M, Malin B. Secure management of biomedical data with cryptographic hardware. Information Technology in Biomedicine, IEEE Transactions on. 2012; 16:166–175.
122. Narayanan A. What Happend to the Crypto Dream? Security & Privacy, IEEE. 2013; 11:75–76.
123. Hubaux, JP.; Tsudik, G.; De Cristofaro, E.; Ayday, E. The Chills and Thrills of Whole Genome Sequencing. 2013.
124. Presidential Commission for the Study of Bioethical Issues. Privacy and Progress in Whole Genome Sequencing. 2012
125. Craig DW, et al. Assessing and managing risk when sharing aggregate genetic variant data. Nat Rev Genet. 2011; 12:730–736. doi:10.1038/nrg3067 nrg3067 [pii]. [PubMed: 21921928]
126. Braun R, Rowe W, Schaefer C, Zhang J, Buetow K. Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data. PLoS Genet. 2009; 5:e1000668.10.1371/journal.pgen.1000668 [PubMed: 19798441]
127. Kendler KS, Gallagher TJ, Abelson JM, Kessler RC. Lifetime prevalence, demographic risk factors, and diagnostic validity of nonaffective psychosis as assessed in a US community sample: the National Comorbidity Survey. Archives of General Psychiatry. 1996; 53:1022–1031. [PubMed: 8911225]
128. Lee, J.; Clifton, C. Information Security. Springer; 2011. p. 325-340.
129. Hsu J, et al. Differential Privacy: An Economic Method for Choosing Epsilon. arXiv preprint. 2014 arXiv:1402.3329.
130. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Theory of Cryptography. Springer; 2006. p. 265-284.
131. Paillier, P. Advances in Cryptology — EUROCRYPT ’99. Stern, Jacques, editor. Vol. 1592. Springer; Berlin Heidelberg: 1999. p. 223-238.Lecture Notes in Computer ScienceCh. 16
132. Hill WG, Goddard ME, Visscher PM. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 2008; 4:e1000008. [PubMed: 18454194]
133. Gentry C. Fully Homomorphic Encryption Using Ideal Lattices. Acm S Theory Comput. 2009:169–178.

Box 1

Entropy and the contribution of quasi-identifiers

Entropy measures the degree of uncertainty in the outcome of a random variable. One bit of entropy is equivalent to the uncertainty of tossing a fair coin. Two bits are equivalent to two independent tosses of a fair coin and so on. Zero bits is the lowest entropy level and implies that there is no uncertainty. The reciprocal measure of entropy is information content, which quantifies the expected contribution of a new piece of data in reducing the entropy level.

Information content captures the average usefulness of quasi-identifiers for identity tracing. Consider an anonymous individual's record in a study that randomly samples subjects from the US population. A priori, the adversary has 310 million equiprobable possibilities of a match, which translates to 28.2bits of entropy. He can then gain ~1 bit of information by inferring the individual's sex, reducing the entropy to 27.2. Complete identification of any person is guaranteed when the entropy reaches zero. The table below lists possible quasi-identifiers and their maximal information content expectation for the US population.

Several factors reduce the expected information content of quasi-identifiers from the maximal level. One possibility is that two quasi-identifiers are correlated. For example, after inference of a US zip code, obtaining the state of residency rarely adds new information. A second possibility is inaccurate inference of the quasi-identifier. Information theory dictates a rapid decline of information content with deviations of the inferred quasi-identifier from the truth. Another possibility is low-searchability of the quasi-identifier. For example, in the case that the adversary can only access a height registry of 100 random US individuals, even with perfect knowledge of height, he will recover close to zero bits of information.

Table

Information content of quasi-identifiers for the general US population

Quasi-identifier	Expected information content (bits)
Sex[1]	1.0
Ethnic group[1,2]	1.4
Eye color[3]	1.4
Blood group (ABO/Rh)[4]	2.2
State[1]	5.0
Height[5]	5.0
Year of birth[1]	6.3
Day and month of birth[6]	8.5
Surname[1]	12.9
Zip code[7]	13.8

[1] Based on US Census data.

[2] Based on self-classification field in the US census: African American, Asian American, European American, Native American, Other race, and two or more races.

[3] Perfect inferences of three eye color groups (blue, brown, intermediate). Data from www.statisticbrain.com/eye-color-distribution-percentages/

[4] Data is based on Stanford School of Medicine Blood Center (bloodcenter.stanford.edu/about_blood/blood_types.html)

[5] Assuming accurate measurement within 1cm resolution and normal distribution with standard deviation of 8cm in the population

[6] Data is based on 400,000 births (www.panix.com/~murphy/bday.html)

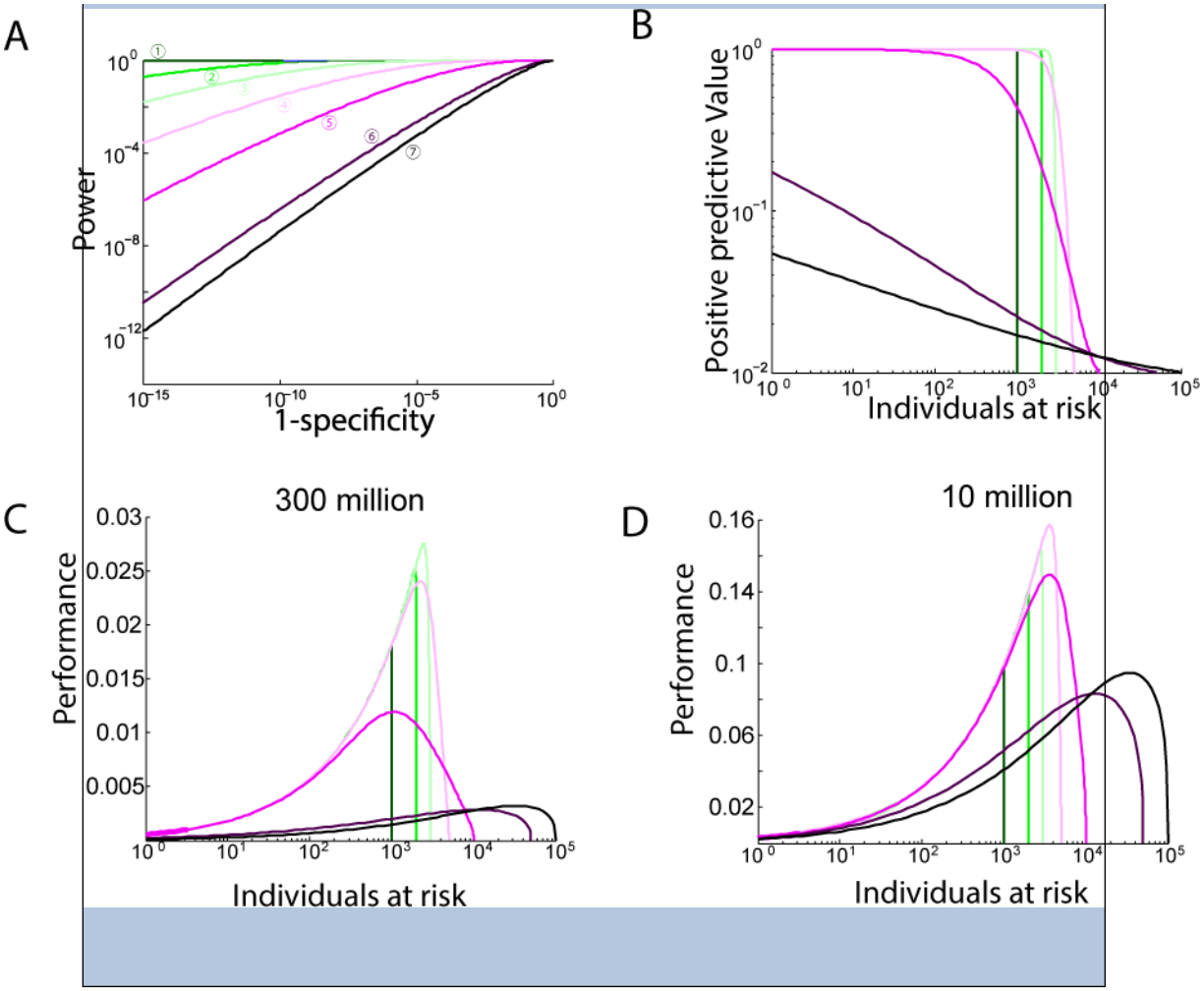
[7] Data is from zipatals.com

Box 2

The performance of ADAD attacks using allele frequencies

The theoretical performance of ADAD with summary statistics is a complex function of the size of the study and the prior knowledge of the adversary[125,126]. To illustrate this point, consider an adversary that has access to the allele frequencies of a GWAS study of schizophrenia in the US, a disease with 1% prevalence[127]. Without any other prior knowledge, the adversary randomly meets with people from the US population and attempts to infer their schizophrenia status. When the study size is small, the adversary enjoys higher power and specificity to discriminate between participants and non-participants than with larger study samples (**Left panel;** cyan – GWAS with 1,000 participants, green – 3,000, yellow – 10,000, purple – 100,000). On the other hand, with smaller studies, the adversary almost never encounters individuals that were part of the study. He keeps consuming resources to conduct the attack, just to implicate relatively few people. Moreover, attacks on non-participants can result in false positives and lower the positive predictive value of the attack. The adversary can compensate by increasing the specificity, but this will further reduce the number of people that can be implicated in the attack. The middle panel depicts the positive predictive value as a function of individuals at risk when the prior knowledge of the adversary is that participants are in the USA. Intermediate sized studies place risk on the largest number of individuals for most of the positive predictive values.

The overall performance trade-off depends on prior knowledge of the adversary and the size of the study. The **right panel** shows the ADAD performance (Matthews correlation coefficient between truth and disease prediction) as a function of individuals at risk when the prior knowledge of the adversary is that participants are in the USA versus when the prior knowledge of the adversary is that participants are sampled from a US subpopulation of 10 million people (say that the adversary knows that a schizophrenia study enrolled only adults with Hispanic ancestry that live in California). Restricting the ADAD efforts to this specific demographic group boosts the accuracy for all study sizes but with different proportions. As a rule of thumb, ADAD performs best when the adversary can narrow down the base population from which participants where sampled, such as with studies of ethnic minorities, a specific geographical region, or when detailed inclusion criteria are given.



BOX 3

Mathematical introduction to differential privacy

Differential privacy seeks to ensure that no single individual's attributes can affect the output of the data release mechanism too much. If an individual's attributes have only a minimal impact on the output, the adversary cannot use the output to accurately infer those inputs. It is necessary and sufficient to consider the impact of adding or dropping an individual from the dataset altogether, rather than the effect of their attributes.

Differential privacy randomizes the released data. Let D be the original dataset and D' be the dataset with any single user record removed. Differential privacy requires that the output distributions corresponding to D and D' are close throughout the output space[102]. A privacy parameter ϵ quantifies the difference of the distributions, and hence the level of information leakage. Low values of ϵ such that $e^{\epsilon} \approxeq 1+\epsilon$ are considered more secure but they typically come at the expense of data utility. Practical values of ϵ is still an open question but several models have been proposed[128,129].

A simple addition of "noise" or randomness to the true output satisfies the requirement above. Let $t(\boldsymbol{D})$ be the summary statistic function that operates on the input dataset, such as mean, median, or counting the number of individuals with a specific property. $f(\boldsymbol{D}) = t(\boldsymbol{D}) + z$ is called ϵ-differentially private if z is randomly drawn from Laplace distribution with mean 0 and a scale of S/ϵ; where S, called sensitivity, is a bound on how much a single record can affect the output of t[130]. For example, the mean of a binary attribute has sensitivity of $1/n$ where n is the number of records in $\boldsymbol{D}$. Thus, by analysing the summary statistic function and a desired privacy level (ϵ), the data custodian can add the appropriate level of noise.

BOX 4

Homomorphic encryption

Homomorphic encryption is an area of cryptography with great potential for certain types of privacy-preserving computation. It is best explained by the following analogy: Alice possesses raw gold and wants to create a necklace, but she is not equipped with the knowledge or tools. Bob is a skillful goldsmith but with an unclear reputation. Using homomorphic encryption, Alice sets up a securely locked glove box with the raw gold. Bob uses the gloves to construct the jewelry without unlocking the box. After that, Alice receives the glove box and opens the lock with her key. Genotypes can be thought of as the raw gold, Bob can be an interpretation service, and the necklace is disease risk status.

Homomorphic encryption creates the glove box by adding additional mathematical properties besides the basic encryption and decryption operations in traditional cryptographic protocols. This property takes a regular function that operates on plaintext (genotypes), say $y(M_1,M_2)=M_1+M_2$, and maps it to a secure function, $y'(X_1,X_2)$ that performs the same computation on the ciphertext. Decrypting $y'(X_1,X_2)$ yields exactly the same answer as calculating the original function with the corresponding plaintext, which in our example is $D(y'(X_1,X_2)) = M_1+M_2$. This way, Bob can compute secure functions on the ciphertext and Alice can decrypt his answer to obtain the result.

Until recently, cryptographic studies achieved encrypted versions of very basic algebraic operations. One example is the Paillier Cryptosystem[131], which supports the addition of plaintexts and multiplication by a constant to be carried out on ciphertexts. Such narrow designs are called Partially Homomorphic Encryption. They operate relatively fast, and despite their limitations, might prove sufficient for a wide range of computations on genotypes due to the additive properties of genetic predispositions[132]. A breakthrough in 2009 established a Fully Homomorphic Encryption scheme that supports calculating arbitrary functions on the plaintext[133]. This innovation is not yet efficient in terms of computational time but further developments can complete the arsenal of secure functions in genetic epidemiology.

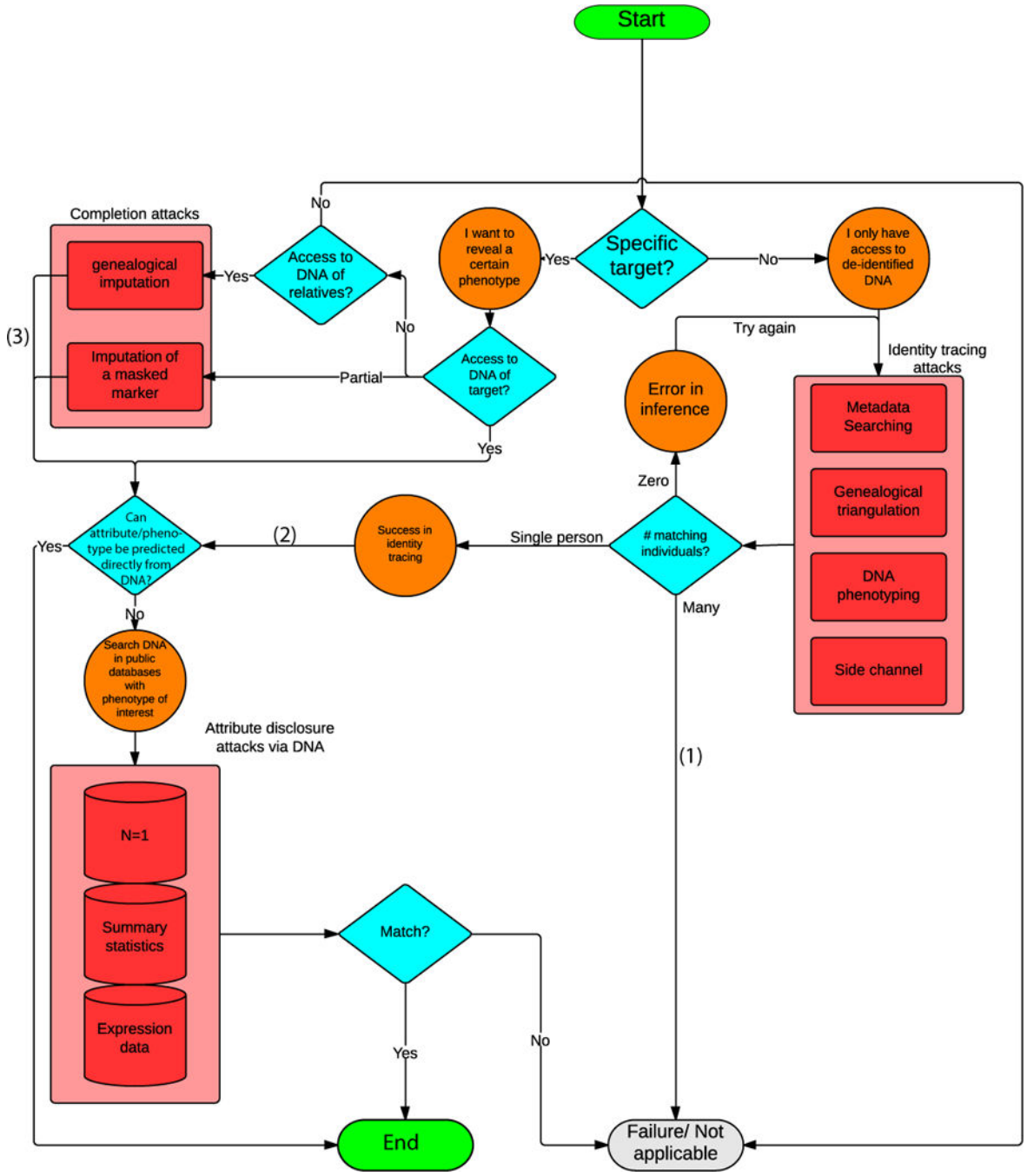


Figure 1. An integrative map of genetic privacy breaching techniques

The map contrasts different scenarios such as identifying de-identified genetic datasets, revealing an attribute from genetic data, and unmasking of data. It also shows the interdependencies between the techniques and suggests potential routes to exploit further information after the completion of one attack. We made several simplifying assumptions [corresponding to numbering in the figure]: (1) in certain scenarios, such as insurance decisions, uncertainty about the identity within a small group of people could still be considered a success (2) for certain privacy harms such as surveillance, identity tracing can

be considered a success and the end point of the process (3) complete DNA sequence is not always necessary.

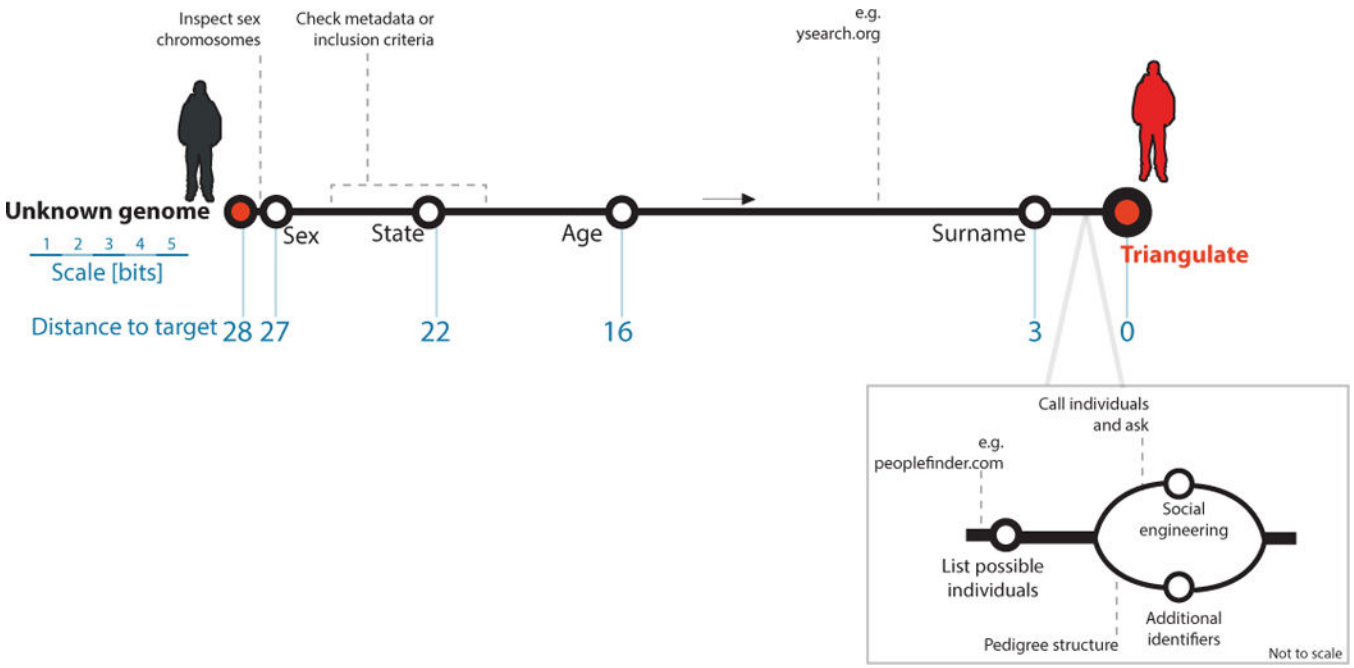


Figure 2. A possible route for identity tracing

The route combines both metadata and surname inference to triangulate the identity of an unknown male genome of a US person. Without any information, there are ~300 million individuals that could match the genome, which is equivalent to 28 bits of entropy (black silhouette). Inferring the sex by inspecting the sex chromosomes reduces the entropy by a bit. The adversary than uses the metadata to find the state and the age, which reduces the entropy to 16bits. Successful surname recovery leaves only ~3bits. At this point, the adversary uses public record search engines such as PeopleFinders.com to generate a list of potential individuals, he can use social engineering or pedigree structure to triangulate the person (red silhouette).

Nat Rev Genet. Author manuscript; available in PMC 2014 September 02.

Table 1

Categorization of techniques for breaching genetic privacy

Technique	Maturation Level	Technical complexity	Example of auxiliary information	Availability of auxiliary information	Example of a reference
Identity Tracing					
Surname Inference	★★★★	●●●	Records of Y-chromosome and surnames	Intermediate-Good	35
DNA Phenotyping	★★	●●	Population registry of eye color	Poor	55
Demographic identifiers	★★★★	●	Population registry stratified by state	Good	29
Pedigree structure	★★★	●●	Family trees of the entire population	Poor	31
Side channel leakage	★★★★	●●●	–	Varies	26
Attribute Disclosure Attacks via DNA (ADAD)					
N=1	★★★★	●●	n/a	n/a	61
Genotype frequencies	★★★	●●●	Exome Sequencing Project	Good	63
Linkage disequilibrium	★★	●●●●	1000 Genomes	Intermediate	67
Effect sizes	★★	●●●	n/a	n/a	68
Trait inference	★	●●	n/a	n/a	69
Gene expression	★★★	●●●●	GTEx project	Poor	76
Completion Attacks					
Imputation of a masked marker	★★★★	●●	1000 Genomes	Good	78
Genealogical imputation (single relative)	★★★★	●●	OpenSNP and Facebook profiles	Poor	79
Genealogical imputation (multiple relatives)	★★★★	●●●●	deCode pedigree and DNA	Poor	80

Maturation level:

★Working principles established with simulated data.

★★Small scale proof of concept with real data in a controlled environment (typically only one dataset).

★★★Large scale experiments in controlled environments with real data (typically more than one dataset).

★★★★Breach of privacy was reported in a real scenario.

Technical complexity:

●knowledge in genetics or special tools are not required.

●● Require genetic knowledge; computation can reasonably be done on a regular computer. Existing tools are available

●●● Require genetic knowledge, intermediate scale processing of data and/or molecular techniques.

●●●● Require genetic knowledge; large scale processing of data is a prerequisite; may also require molecular techniques.

Auxiliary information: this column refers to the level of existing public reference databases for the US population. For identity tracing, it refers to the availability of organized lists that link identities and extract pieces of information. For ADAD and completion techniques, it refers to the existence of supporting reference datasets that are necessary to complete the attack. Poor – supporting data is highly fragmented and not amenable to searches. Intermediate – supporting data is harmonized and searchable but requires some pre-processing. Good – supporting data is searchable using existing tools or minimal pre-processing.

Table

Information content of quasi-identifiers for the general US population

Quasi-identifier	Expected information content (bits)
Sex[1]	1.0
Ethnic group[1,2]	1.4
Eye color[3]	1.4
Blood group (ABO/Rh)[4]	2.2
State[1]	5.0
Height[5]	5.0
Year of birth[1]	6.3
Day and month of birth[6]	8.5
Surname[1]	12.9
Zip code[7]	13.8

[1]Based on US Census data.

[2]Based on self-classification field in the US census: African American, Asian American, European American, Native American, Other race, and two or more races.

[3]Perfect inferences of three eye color groups (blue, brown, intermediate). Data from www.statisticbrain.com/eye-color-distribution-percentages/

[4]Data is based on Stanford School of Medicine Blood Center (bloodcenter.stanford.edu/about_blood/blood_types.html)

[5]Assuming accurate measurement within 1cm resolution and normal distribution with standard deviation of 8cm in the population

[6]Data is based on 400,000 births (www.panix.com/~murphy/bday.html)

[7]Data is from zipatals.com