Routes for breaching and protecting genetic privacy

Yaniv Erlich[1] and Arvind Narayanan[2]

Abstract | We are entering an era of ubiquitous genetic information for research, clinical care and personal curiosity. Sharing these data sets is vital for progress in biomedical research. However, a 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, indicate the underlying assumptions, and assess their 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.

We generate genetic information for research, clinical care and personal curiosity at exponential rates. Sequencing studies that include 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 for 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]. 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 analyses, which are necessary to maximize use of such data for patients and the general public[9].

One of the key issues of broad dissemination is finding an adequate balance that ensures 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, public concerns regarding medical data privacy halted a massive plan of the National Health Service in the United Kingdom to create a centralized health care database[14]. Protecting personal identifiable information is also a demand of various regulatory statutes in the United States and the European Union[15]. Data de-identification (that is, the removal 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 maps privacy breaching techniques that are relevant to genetic information and proposes potential counter-measures. We first categorize privacy breaching strategies (FIG. 1), discuss their underlying technical concepts, and evaluate their performance and limitations (TABLE 1). We then 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 combining distinct resources to gain private information that is relevant to DNA data. Data custodians should be aware that security threats can be much broader and 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[17–18] (see Chronology of Data Breaches). We do not include these threats here, as they have been extensively discussed in the computer security field[19]. 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 partly covered by the broad privacy literature[20,21].

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

[1]*Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA.*
[2]*Department of Computer Science, Princeton University, 35 Olden Street, Princeton, New Jersey 08540, USA.*
Correspondence to Y.E.
e-mail: yaniv@wi.mit.edu
doi:10.1038/nrg3723
Published online 8 May 2014; corrected online 17 June 2014

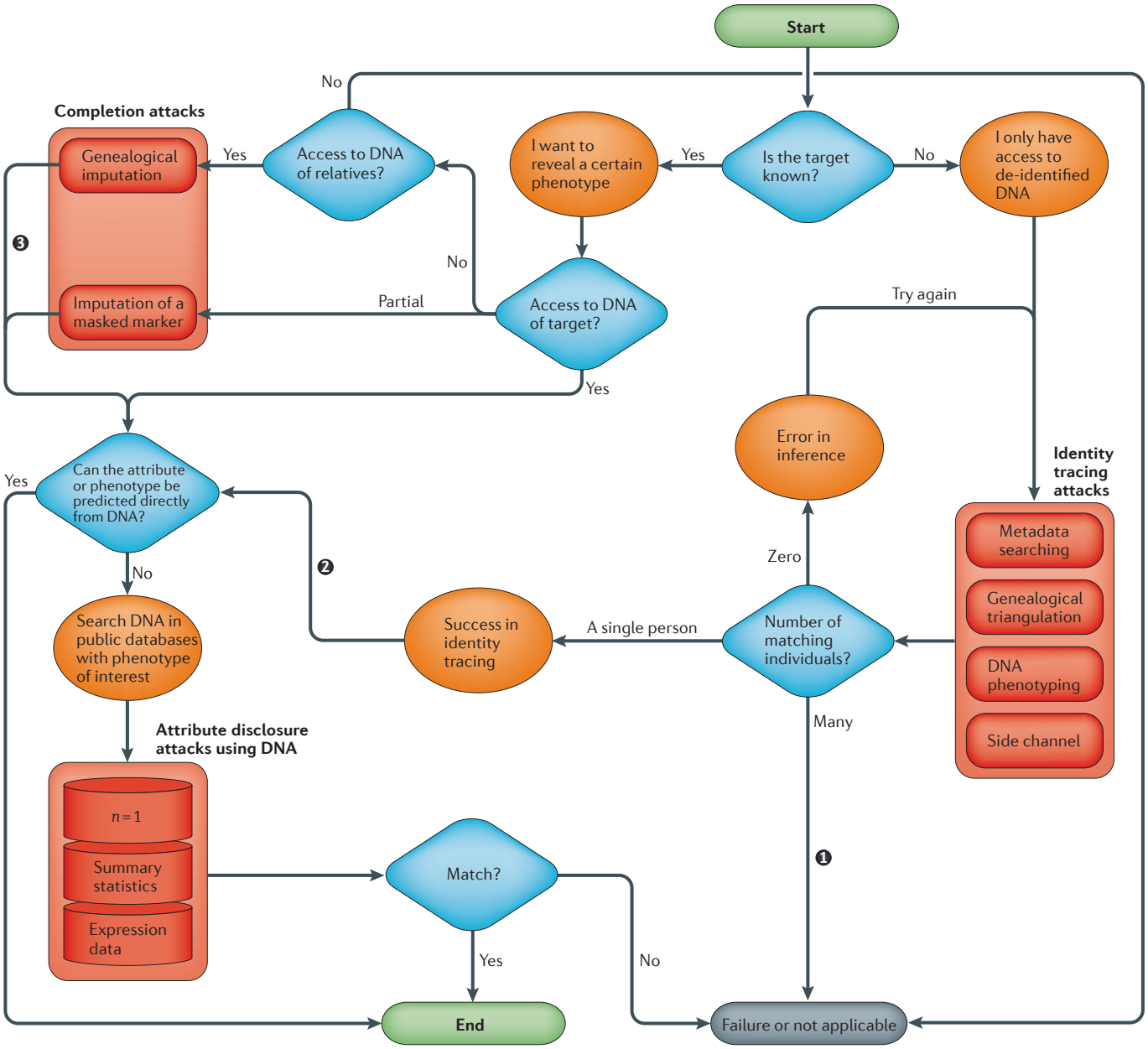


Figure 1 | **An integrative map of genetic privacy breaching techniques.** The map contrasts different scenarios, such as identifying de-identified genetic data sets, revealing an attribute from genetic data and unmasking of data. It also shows the interdependencies between the techniques and suggests potential routes to exploit further information after the completion of one attack. There are several simplifying assumptions (black circles). In certain scenarios (such as insurance decisions), uncertainty about the target's identity within a small group of people could still be considered a success (assumption 1). For certain privacy harms (such as surveillance), identity tracing can be considered a success and the end point of the process (assumption 2). The complete DNA sequence is not always necessary (assumption 3).

the data set. 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).

Tracing with metadata. Genetic data sets are typically published with additional metadata — such as basic demographic details, inclusion and exclusion criteria, pedigree structure and health conditions — that are crucial both to the study and for secondary analyses. These pieces of metadata can be exploited to trace the identity of unknown genomes.

Unrestricted demographic information conveys substantial power for identity tracing. It has been estimated that the combination of date of birth, sex and five-digit zip code uniquely identifies >60% of individuals in the

Table 1 | **Categorization of techniques for breaching genetic privacy**

Technique	Maturation level*	Technical complexity‡	Example of auxiliary information	Availability of auxiliary information§	Ref
Identity tracing attacks					
Surname inference	Level 4	Intermediate	Records of Y chromosomes and surnames	Intermediate to good	34
DNA phenotyping	Level 2	Low	Population registry of eye colour	Poor	54
Demographic identifiers	Level 4	Very low	Population registry stratified by state	Good	28
Pedigree structure	Level 3	Low	Family trees of the entire population	Poor	30
Side-channel leaks	Level 4	Intermediate	NA	Varies	25
Attribute disclosure attacks using DNA					
$n = 1$	Level 4	Low	NA	NA	60
Genotype frequencies	Level 3	Intermediate	Exome Sequencing Project	Good	62
Linkage disequilibrium	Level 2	High	1000 Genomes Project	Intermediate	133
Effect sizes	Level 2	Intermediate	NA	NA	67
Trait inference	Level 1	Low	NA	NA	68
Gene expression	Level 3	High	Genotype-Tissue Expression (GTEx) Project	Poor	75
Completion attacks					
Imputation of a masked marker	Level 4	Low	1000 Genomes Project	Good	77
Genealogical imputation of a single relative	Level 4	Low	OpenSNP and Facebook profiles	Poor	78
Genealogical imputation of multiple relatives	Level 4	High	deCODE pedigree and DNA	Poor	79

NA, not available. *Genetic privacy breaching techniques are classified into four maturation levels on the basis of the data used. For level 1, working principles are established using simulated data. Level 2 involves small-scale proof-of-concept experiments that use real data (typically only one data set) in a controlled environment, whereas level 3 involves large-scale experiments that use real data (typically more than one data set) in controlled environments. For level 4, breach of privacy was reported in a real scenario. ‡Techniques of very low complexity do not require knowledge in genetics or special tools. Low-complexity techniques require genetic knowledge, and computation can be reasonably done on a regular computer. Existing tools are available for techniques of intermediate complexity, which require genetic knowledge, intermediate-scale processing of data and/or molecular techniques. High-complexity techniques require genetic knowledge and large-scale processing of data; it may also require molecular techniques. §Availability of auxiliary information refers to the level of existing public reference databases for the US population. For identity tracing attacks, it refers to the availability of organized lists that link identities and that extract pieces of information. For attribute disclosure attacks using DNA and for completion attacks, it refers to the existence of supporting reference data sets that are necessary to complete the attacks. Poor auxiliary information has highly fragmented supporting data that are not amenable to searches. Supporting data for intermediate-level information are harmonized and searchable but require some pre-processing. Supporting data for good information are searchable using existing tools or minimal pre-processing.

United States[22,23]. 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) 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[24]. 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[25].

Since the inception of the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, dissemination of demographic identifiers has been the subject of tight regulation in the US health care system[26]. The Safe Harbor provision requires that the maximal resolution of any date field, such as hospital admission dates, is in years. In addition, the maximal resolution of a geographical subdivision is the first three digits of a zip code (for zip code areas with populations of >20,000). Statistical analyses of the census data and empirical health records have found that the Safe Harbor provision provides reasonable protection against identity tracing, assuming that the adversary has access only to demographic identifiers. The combination of sex, age, ethnic group and state of residence is unique in <0.25% of the populations of each of the US states[27,28].

Pedigree structures are another piece of metadata that is included in many genetic studies. These structures contain rich information, especially when large kinships are available[29]. 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[30]. Only the number (but not the order) of male and female individuals in each generation was available. Despite this limited information, ~30% of the pedigree structures were unique, which shows the large information content that can be obtained from such data.

Another vulnerability of pedigrees is combining demographic quasi-identifiers across records to enhance identity tracing despite HIPAA protections. For example, consider a large pedigree that shows the age and state of residence of all participants. The age and state of residence of each participant leak very minimal information, but knowing the ages of all first- and second-degree relatives of an individual markedly reduces the search space. Moreover, after a single

Safe Harbor
A standard in the US Health Insurance Portability and Accountability Act (HIPAA) rule for de-identification of protected health information by removing 18 types of quasi-identifiers.

Haplotypes
Sets of alleles along the same chromosome.

individual in a pedigree has been identified, it is easy to link the identities of other relatives with their genetic data sets. 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 number of genealogical websites. One notable exception is Israel, where the entire population registry was leaked to the web in 2006, thus allowing the construction of multigeneration family trees of all Israeli citizens[31].

Identity tracing by genealogical triangulation. Genetic genealogy attracts millions of individuals who are interested in their ancestry or in discovering distant relatives[32]. 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[33,34] (FIG. 2). In most societies, surnames are passed from father to son, which creates a transient correlation with specific Y-chromosome haplotypes[35,36]. The adversary can take advantage of the Y chromosome–surname correlation and compare the Y-chromosome haplotype of the unknown genome to haplotype records in recreational genetic genealogy databases. A close match with a fairly short time to the most common recent ancestor would indicate that the unknown genome probably 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 individual with the genome of interest. An empirical analysis estimated that 10–14% of US white male individuals from the middle and upper classes are subject to surname inference on the basis of scanning the two largest Y-chromosome genealogical websites using a built-in search engine[34]. Individual surnames are fairly rare in the population and, in most cases, a single surname is shared by <40,000 US male individuals[34], 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 used to breach genetic privacy in the past[37–40]. 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 5 successful surname inferences from Illumina data sets of 3 large families that were part of the 1000 Genomes Project, which eventually exposed the identity of nearly 50 research participants[34].

The main limitation of surname inference is that haplotype matching relies on the comparison of Y-chromosome short tandem repeats. 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[41]. Another complication is false identification of surnames and inference of surnames that are spelling variants of the original one. 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 United States. For example, 400 million people in China hold 1 of the 10 common surnames[35], and the top 100 surnames cover almost 90% of the population[42], which markedly reduces the use of surname inference for re-identification.

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 of entropy is the lowest 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 28.2 bits of entropy. He or she can then gain ~1 bit of information by inferring the individual's sex, which reduces the entropy to 27.2. Complete identification of any person is guaranteed when the entropy reaches zero. The table below lists some possible quasi-identifiers and their maximal information content expectation for the general 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 residence rarely adds new information. A second possibility is inaccurate inference of the quasi-identifier. Information theory posits 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 or she will recover close to zero bits of information.

Quasi-identifier	Expected information content (bits)
Sex*	1.0
Ethnic group*‡	1.4
Eye colour§	1.4
Blood group (ABO and Rhesus systems)‖	2.2
State of residence*	5.0
Height¶	5.0
Year of birth*	6.3
Day and month of birth#	8.5
Surname*	12.9
Zip code**	13.8

*Based on US Census data. ‡Based on self-classification field in the US Census: African American, Asian American, European American, Native American, Other race, and two or more races. §Perfect inferences of three eye colour groups (blue, brown and intermediate); data from Eye Color Distribution Percentages. ‖Based on Stanford School of Medicine Blood Center. ¶Assuming accurate measurement within 1-cm resolution and normal distribution with a standard deviation of 8 cm in the population. #Based on 400,000 births (see An analysis of the distribution of birthdays in a calendar year). **Data from ZipAtlas.

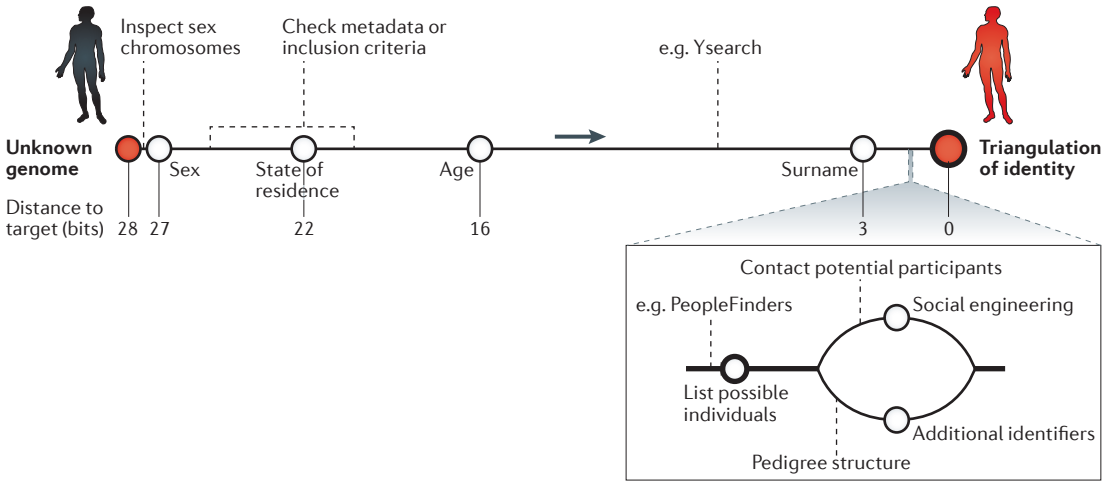


Figure 2 | **A possible route for identity tracing.** The route combines both metadata and surname inference to triangulate the identity of an unknown genome of a person in the United States (represented by the black silhouette). Without any information, there are ~300 million individuals that could match the genome, which is equivalent to 28 bits of entropy. Inferring the sex by inspecting the sex chromosomes reduces the entropy by 1 bit. The adversary then uses the metadata to find the state of residence and the age, which reduces the entropy to 16 bits. Successful surname recovery (for example, using Ysearch) leaves only ~3 bits of entropy. At this point, the adversary uses public record search engines such as PeopleFinders to generate a list of potential individuals; he or she can use social engineering or pedigree structure to triangulate the person (represented by the red silhouette).

An open research question is the use of non-Y-chromosome markers for genealogical triangulation. The Mitosearch and GEDmatch websites 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 owing to the small size of the mitochondrial genome, which means that a large number of individuals share the same mitochondrial haplotypes. In addition, matrilineal identifiers (such as surname or clan) are fairly rare in most human societies, which complicates the use of mitochondrial haplotype for identity tracing. By contrast, autosomal searches can be 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 third to fourth cousins) with fairly sufficient accuracy[43]. These hits would reduce the search space to no more than a few thousand individuals[44]. The main challenge of this approach would be to derive a list of potential people from a genealogical match. As stated above, family trees of most individuals are not publicly available; such searches are therefore demanding and would require indexing a large number 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 be used as quasi-identifiers for identity tracing[45,46]. Twin studies have estimated high heritabilities for various visible traits, for example, height[47] and facial morphology[48]. In addition, recent studies show that age prediction is possible from DNA specimens derived from blood samples[49,50], but the applicability of these DNA-derived quasi-identifiers for identity tracing has yet to be demonstrated.

A 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[51], body mass index (BMI)[52] and face morphology[53], which substantially limits their use for identification. For example, perfect knowledge about height at 1-cm resolution conveys 5 bits of information. However, as current genetic knowledge can merely provide an explanation for 10% of height variability[51], the adversary learns only 0.15 bits of information. Predictions of face morphology and BMI are much worse[8,53]. The exceptions in visible traits are eye colour[54] and age prediction[49]. Recent studies show a prediction accuracy of 75–90% of the phenotypic variability of these traits, but even these successes represent 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 are meant to be

public. A good example of such leaks is the exposure of the full names of PGP participants from filenames in the database[25]. The PGP allowed participants to upload 23andMe genotyping files to their public profile webpages. Although it seemed that these files do not contain explicit identifiers, after downloading and decompressing, the original filename — the default of which is the first and last names of the user — appeared. As 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 participants instructions to rename files before uploading and warns them of possible hidden information that can expose their identities. Generally, certain types of files, such as Microsoft Office products, can embed deleted text or hidden identifiers[55]. 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 patients, which was the aim of the contest[56]. In addition, a 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 from quasi-identifiers[57]. Some suggested the assignment of accession numbers by applying cryptographic hashing to the participants' identifiers, such as names or Social Security numbers[58]. However, this technique is vulnerable to dictionary attacks owing to the fairly small 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 using DNA

Consider the following scenario: Alice interviews Bob for a certain position. After the interview, Alice recovers Bob's DNA and uses the data to search a large genetic study of drug abuse. The study stores the DNA in an anonymous form, but a match between Bob's DNA and one of the records reveals that Bob was a drug abuser. This short story illustrates the main concepts of attribute disclosure attacks using DNA (ADAD). The adversary gains access to the DNA sample of the target and 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 genotypic data of the individual. The adversary can simply match the genotypic data that are associated with the identity of the individual to the genotypic data that are 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 1 error of 10^{-15} for most of the major populations across the globe[59]. Moreover, random subsets of ~300 common SNPs yield sufficient information to uniquely identify any person[60]. Therefore, an individual's genome is a strong identifier. In general, ADAD is a theoretical vulnerability of essentially any individual-level, DNA-derived 'omic' data set.

Genome-wide association studies (GWASs) are highly vulnerable to ADAD. To address this issue, several organizations, including the US National Institutes of Health (NIH), have adopted a two-tier access system for data sets of GWASs: 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[61]. The premise of this distinction was that summary statistics enable secondary data use for meta-GWAS analyses, although it was thought that this type of data is protected against ADAD.

The summary statistic scenario. A landmark study in 2008 reported the possibility of ADAD on GWAS data sets that only consist of the allele frequencies of the study participants[62]. 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 with the allele frequencies of the general population. A good illustration of this concept is the case of an extremely rare variation in the subject's genome. Positive 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 carry out ADAD with the common variations that are analysed in GWASs.

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[63–65]. Under the assumption of common SNPs in linkage equilibrium, the improved test statistic is mathematically guaranteed to yield maximal power for any specificity level (BOX 2). Another group went beyond allele frequencies and showed that it is possible to exploit local linkage disequilibrium (LD) structures for ADAD[66]. The power of this approach stems from searching 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 GWASs that involve quantitative traits to detect the presence of the target[67]. A powerful development of this study is exploiting GWASs that use 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 enhance 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[68].

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
Approaches to reverse cryptographic hashing by scanning only highly probable inputs.

Alice
A common generic name in computer security to denote party A.

Bob
A common generic name in computer security to denote party B.

Type 1 error
The probability of obtaining a positive answer from a negative item.

Linkage equilibrium
Absence of correlation between the alleles at two loci.

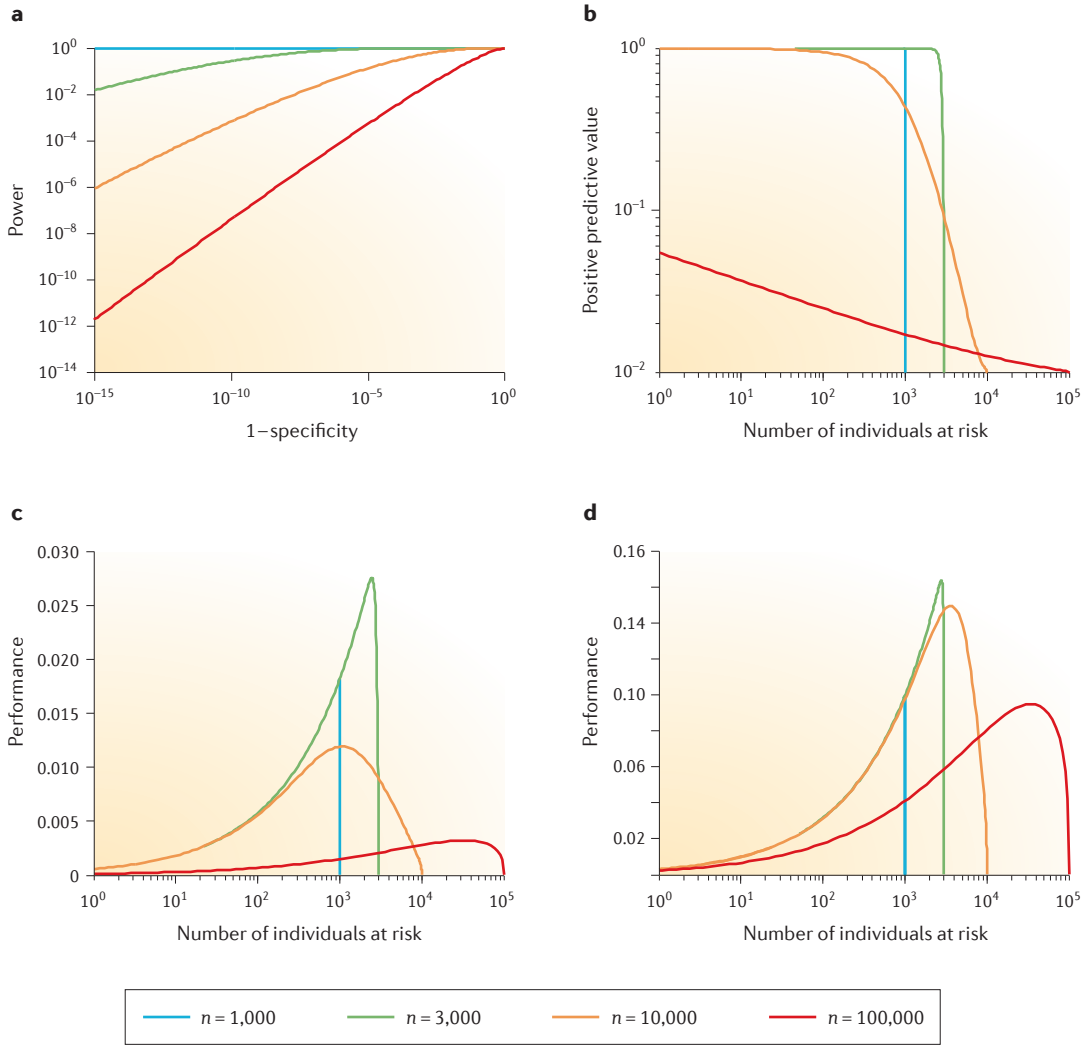
Power
The probability of obtaining a positive answer for a positive item.

Specificity
The probability of obtaining a negative answer for a negative item.

Linkage disequilibrium
(LD). The correlation between alleles at two loci.

Effect sizes
The contributions of alleles to the values of particular traits.

Box 2 | **The performance of ADAD using allele frequencies**



The theoretical performance of attribute disclosure attacks using DNA (ADAD) with summary statistics is a complex function of the size of the study and the prior knowledge of the adversary[124,125]. To illustrate this point, consider an adversary that has access to the allele frequencies of a genome-wide association study (GWAS) of schizophrenia, which has a prevalence of 1%, in the United States[126]. Without any other prior knowledge, the adversary randomly obtains DNA of people from the US population and attempts to infer their schizophrenia status. When the study size (*n*) is small, the adversary enjoys higher power and specificity to discriminate between participants and non-participants than with larger study samples (see the figure, part **a**). However, with smaller studies, the adversary almost never encounters individuals that were part of the study. He or she keeps consuming resources to carry out the attack, only to implicate fairly 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. Part **b** of the figure depicts the positive predictive value as a function of the number of individuals at risk when the only prior knowledge of the adversary is that the participants are in the United States. 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 on the size of the study. The ADAD performance (that is, Matthews correlation coefficient between truth and disease prediction) is shown as a function of the number of individuals at risk when the prior knowledge of the adversary is that participants are in the United States (see the figure, part **c**). Compare this with the case in which the prior knowledge of the adversary is that participants are sampled from a US subpopulation of 10 million people (for example, the adversary knows that a schizophrenia study enrolled only adults with Hispanic ancestry that live in California) (see the figure, part **d**). Restricting the ADAD efforts to this specific demographic group increases the accuracy for all study sizes but in different proportions. As a rule of thumb, ADAD performs best when the adversary can narrow down the base population from which participants were sampled (such as with studies of ethnic minorities or in a specific geographical region) or when detailed inclusion criteria are given.

Positive predictive value
The probability that a positive answer belongs to a true positive.

The actual risk of ADAD has been the subject of intense debate. Following the original study in 2008 (REF. 62), the NIH and other data custodians moved their GWAS summary statistic data from public databases to access-controlled databases, such as the database of Genotypes and Phenotypes (dbGaP)[69]. A retrospective analysis found that significantly fewer GWASs publicly released their summary statistic data after the discovery of this attack[70]. Currently, most of the studies publish summary statistic data on 10–500 SNPs, which is compatible with one suggested guideline to manage risk[68]. However, some researchers have warned that these policies are too harsh[71]. 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[72], and accurate matching between the target ancestries and those listed in the reference database[73]. Failure to address any of these prerequisites can severely affect the performance of the ADAD. In addition, for a range of GWASs, the associated attributes are neither sensitive nor 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[74].

The gene expression scenario. Databases such as the NIH's Gene Expression Omnibus (GEO) publicly hold hundreds of thousands of gene expression profiles from individuals that are linked to a range of medical attributes. A recent study proposed a potential route to exploit these expression profiles for ADAD[75]. The method starts with a training step that uses a standard expression quantitative trait locus (eQTL) analysis with a reference data set. The goal of this step is to identify several hundred strong eQTLs and to learn the distributions of expression level 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, then the algorithm links the identity of the target to the medical attribute in the gene expression experiment. This ADAD technique has the potential for high accuracy in ideal conditions. On the basis of large-scale simulations, the authors predicted that the method can reach a type 1 error of 10^{-5} with a power of 85% when tested on an expression database of the entire US population[75].

There are several practical limitations to ADAD with expression data. Although the training and inference steps can work with expression profiles from different tissues, the method reaches its maximal power when the training and inference use eQTLs from the same tissue. Additionally, there is substantial loss of accuracy when the expression data in the training phase and those in the inference phase are collected using different technologies. Another complication is that, to fully execute the technique on a large database such as the GEO, the adversary will need to manage and process substantial amounts of expression data. 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 and is known as genotype imputation[76]. This method takes advantage of the LD between markers and uses reference panels with complete genetic information to restore missing genotypic values in the data of interest. The very same strategies enable the adversary to expose certain regions of interest when 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 James Watson's predisposition for Alzheimer's disease from the apolipoprotein E (*APOE*) locus despite masking of this gene[77]. As a result of this study, a 2-Mb 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 data set of a known individual. He or she then exploits this information to estimate genetic predispositions for relatives whose genetic information is inaccessible. A recent study showed the feasibility of this attack by taking advantage of self-identified genetic data sets from openSNP[78], which is an Internet platform for public sharing of genetic information. Using Facebook searches, the research team was able to find relatives of the individuals that self-identified their genetic data sets. Next, the team predicted the genotypes of these relatives and estimated their genetic predisposition to Alzheimer's disease using a Bayesian approach.

In the advanced setting, the adversary has access to the genealogical and genetic information of several relatives of the target[79]. The algorithm finds relatives of the target who donated their DNA to the reference panel and who 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 or her DNA. This approach is more accurate than the basic setting and enables genotypes of more distant relatives to be inferred. In Iceland, deCODE genetics took advantage of their large reference panel and genealogical information to infer genetic variants of an additional 200,000 living individuals who never donated their DNA[80]. 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.

Expression quantitative trait locus
(eQTL). A genetic variant associated with variability in gene expression.

Genotype imputation
A class of statistical techniques to predict a genotype from information on surrounding genotypes.

Mitigation techniques

Most of the genetic privacy breaching approaches 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[81,82]. In this view, an appropriate mitigation strategy is to simply remove obvious identifiers from the data sets 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[83]. Technologies only get better with time, and what is technically challenging but possible now will be much easier in the future. Known in cryptography as Shannon's maxim[84], this school of thought assumes that the adversary exists and is equipped with the knowledge and means to execute the breach. Robust data protection is therefore achieved by explicit design of the data access protocol rather than by relying on the small chances of a breach[85].

Access control. Privacy risks are both amplified and more uncertain when data are 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. After approval has been granted, 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 use of such data and any adverse events. This approach is the 'cornerstone' of dbGAP[61,86]. On the basis of periodic reports by users, a retrospective analysis of dbGAP access control has identified 8 data management incidents of ~750 studies, most of which involved non-adherence to the technical regulations, and there was no report of breaching the privacy of participants[87].

Despite the absence of privacy breaches so far, some have criticized the lack of real oversight once the data are in the hand of the applicants[88]. An alternative model uses a trust-but-verify approach, in which users cannot download the data without restriction but, on the basis of their privileges, may execute certain types of queries, which are recorded and audited by the system[89,90]. 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 behaviour to detect adversaries[91]. It has yet to be proved that such methods can reliably distinguish between legitimate and malicious use of genetic data. Auditing also requires that any interaction with the genetic data sets is done using a standard set of application programming interface (API) calls that can be analysed. 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 analyses[92,93].

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[94,95]. This model centres on dynamic consent based on ongoing 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 of such an effort is Platform for Engaging Everyone Responsibly (PEER). In this setting, Private Access 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 grants access to data on the basis of participants' decisions. Participant-based access control is still a fairly new method. As data custodians gain more experience with such a framework, a better picture will emerge regarding its use 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 extent that no record in the database has a unique combination of quasi-identifiers. One heuristic is *k*-anonymity[96], 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 use of the data for subsequent analyses, 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 use of the data. As a rule of thumb, $k=5$ is commonly used in practice[97]. Most of the *k*-anonymity work centres on protecting demographic identifiers. For genetic data, one study suggested a two-anonymity protocol by generalizing the four nucleotides in DNA sequences into broader types of biochemical groups such as pyrimidine and purines[98]. However, the use of such data for broad genetic applications is unclear. Furthermore, *k*-anonymity is vulnerable to ADAD when the adversary has prior knowledge about the presence of the target in the database[99,100]. Thus, although this heuristic is easy to comprehend, its privacy properties and relevance to genomic studies are in question.

Differential privacy is an emerging methodology for privacy-preserving reporting of results, primarily of summary statistics[101] (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

Application programming interface
(API). A set of commands that specify the interface with a data set or software applications.

χ^2-statistic
A measure of association in case–control genome-wide association studies.

Read mapping
A computationally intensive step in the analysis of high-throughput sequencing to find the location of a short DNA sequence (string) in the genome.

Edit distance
The total number of insertions, deletions and substitutions between two strings.

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 data set. This way, an adversary with any type of prior knowledge can never be sure whether a specific individual was part of the original data set because the data release process produces results that are almost exactly the same if the individual was not included. Owing to its theoretical guarantees and tractable computational demands, differential privacy has become an active research area in computer science and statistics. In perhaps the best-known large-scale implementation, the US Census Bureau uses this technique for privacy-preserving release of data in the online OnTheMap tool[102].

In the context of genetic privacy, several studies have explored the 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[103,104]) or shifting the original locations of variants[105]. 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 (see Supplementary information S1 (figure)). 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 advances to data dissemination beyond the traditional use 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 discussed in the previous section, cryptographic solutions enable computing exact answers of the protected data sets.

One line of cryptographic work considers the problem of privacy-preserving approaches that 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 range of online genetic interpretation applications, such as Promethease, which increases the chance of a privacy breach. Recent cryptographic work has suggested homomorphic encryption (BOX 4) for secure genetic interpretation[106]. 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; therefore, it cannot read the plain genotypic values. Instead, the interpretation service executes the risk prediction algorithm on the encrypted genotypes. Owing to the special mathematical properties of the homomorphic 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 who are concerned about privacy. Preliminary results have highlighted the potential feasibility of this scheme[107]. A proof-of-concept study encrypted the variants of an individual in the 1000 Genomes Project and simulated a secure inference of heart disease risk on the basis of 23 SNPs and 17 environmental factors[107]. The total size of the encrypted genome was 51 gigabytes, 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 amenable to future improvements.

Cryptographic studies have also considered the task of outsourcing read mapping without revealing any genetic information to the service provider[108–110]. The basis of some of these protocols is secure multiparty computation (SMC). SMC allows two or more entities, each of which has 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[111]. Earlier studies suggested SMC versions for edit distance-based mapping of DNA sequences that does not reveal their content[108,109]. A more recent study proposed a privacy-preserving version of the popular seed-and-extend algorithm[110], which serves as the basis of several high-throughput alignment tools[110,112]. The privacy-preserving version is a hybrid: the seeding part is securely outsourced to a cloud, in which a cryptographic hashing hides the actual DNA sequences while permitting string matching. The cloud results are streamed to a local trusted computer that carries out the extension part. By adjusting the underlying parameters of the seed-and-extend algorithm, this method puts most of the computation burden on the cloud. Experiments with real sequencing

Box 3 | **A 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 only have a minimal effect on the output, then 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 data set altogether, rather than the effect of their attributes.

Differential privacy randomizes the released data. Let *D* be the original data set and *D′* be the data set with any single user record removed. Differential privacy requires that the output distributions that correspond to *D* and *D′* are close throughout the output space[101]. A privacy parameter (ε) quantifies the difference of the distributions and hence the level of information leakage. Low values of ε such that $e^{\varepsilon} \cong 1 + \varepsilon$ are considered more secure, but they typically come at the expense of data utility. Practical values of ε are still in question, but several models have been proposed[127,128].

A simple addition of 'noise' or randomness to the true output satisfies the requirement above. Let $t(D)$ be the summary statistic function that operates on the input data set, such as mean, median or the number of individuals with a specific property. $f(D) = t(D) + z$ is called ε-differentially private if *z* is randomly drawn from a Laplace distribution with a mean of zero and a scale of S/ε, where *S* is the sensitivity (that is, a bound on how much a single record can affect the output of *t*)[129]. For example, the mean of a binary attribute has sensitivity of $1/n$, where *n* is the number of records in *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 that has 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 to do so. Bob is a skilled goldsmith but has an unclear reputation. Using homomorphic encryption, Alice sets up a securely locked glovebox with the raw gold. Bob uses the gloves to construct the jewellery without unlocking the box. After that, Alice receives the glovebox and opens the lock with her key. The raw gold can be thought of as genotypes, Bob as an interpretation service and the necklace as disease risk status.

Homomorphic encryption creates the 'glovebox' by adding additional mathematical properties besides the basic encryption and decryption operations in traditional cryptographic protocols. This property takes a regular function (y) that operates on plaintext (that is, genotypes) — for example, $y(M_1,M_2) = M_1 + M_2$, where M_1 and M_2 are two integers — and maps it to a secure function, $y'(X_1,X_2)$, that carries out 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 is $D(y'(X_1,X_2)) = M_1 + M_2$ in this example. In 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[130], 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 owing to the additive properties of genetic predispositions[131]. A breakthrough in 2009 established a fully homomorphic encryption scheme that supports calculating arbitrary functions on the plaintext[132]. This innovation is not yet efficient in terms of computational time, but further developments can complete the collection of secure functions in genetic epidemiology.

data showed that the cloud performs >95% of the computation efforts. In addition, the secure algorithm takes only 3.5× longer than a similar insecure implementation, which suggests 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 the other way round[113]. 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 genotypic information (for example, from a crime scene) can reveal the identity but not vice versa. 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[114,115]. A smart-phone-based implementation was presented for one version of this algorithm[116]. A recent study suggested a scalable approach for finding relatives using genome-wide data without disclosing the raw genotypes to a third party or to other participants[117]. First, users collectively decide the minimal degree of relatedness they wish to accept. Next, each user posts a secure version of his or her genome to a public repository using a fuzzy encryption scheme. Users then 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 using experiments with hundreds of individuals from the 1000 Genomes Project showed that second-degree relatives can reliably find each other[117].

A major open question is whether cryptographic protocols can facilitate data sharing for research purposes. Cryptographic schemes have so far focused on developing protocols for GWAS analyses without the need to reveal individual-level genetic data. One study presented a scheme in which genetic data and computation of GWAS contingency tables are securely outsourced through homomorphic encryption to external data centres[118]. A trusted party (for example, the NIH) acts as a gateway that accepts requests from researchers in the community, instructs the data centres to carry out 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 centres[119]. Another study evaluated the outsourcing of GWAS analyses to a commercially available tamper-resistant hardware[120]. Different from the schemes above[118,119], the individual-level genotypes are decrypted as part of the GWAS summary statistic 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. So far, cryptography has yet to devise a comprehensive data sharing solution for GWASs.

Conclusions

In the past few years, a large number of studies have suggested that a motivated, technically sophisticated adversary is capable of exploiting a wide range of genetic data. On the one hand, with the constant innovation in genetics and the rapid accumulation of online information, we can expect that new privacy breaching techniques will be discovered in the next few years and that technical barriers to existing attacks will diminish. On the other hand, privacy-preserving strategies for data dissemination are an active area of research. Rapid progress has been made, and powerful frameworks such as differential privacy and homomorphic encryption are now part of the mitigation strategy. At least for certain tasks in genetics, there are protocols that preserve the privacy of individuals. However, protecting privacy is only one aspect 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[121,122]. We have yet to fully achieve this aim.

In addition, successful balancing of privacy demands and data sharing is not restricted to technical means[123]. Balanced informed consent that outlines both benefits and risks is a key element 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 programmes to reduce the chance of misuse of genetic data regardless of the ability to identify data sets.

1. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2013).
2. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
3. Roberts, J. P. Million veterans sequenced. *Nature Biotech.* **31**, 470–470 (2013).
4. Drmanac, R. Medicine. The ultimate genetic test. *Science* **336**, 1110–1112 (2012).
5. Burn, J. Should we sequence everyone's genome? Yes. *BMJ* **346**, f3133 (2013).
6. Kaye, J., Heeney, C., Hawkins, N., de Vries, J. & Boddington, P. Data sharing in genomics — re-shaping scientific practice. *Nature Rev. Genet.* **10**, 331–335 (2009).
7. Park, J. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet.* **42**, 570–575 (2010).
8. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nature Genet.* **45**, 400–405 (2013).
9. Friend, S. H. & Norman, T. C. Metcalfe's law and the biology information commons. *Nature Biotech.* **31**, 297–303 (2013).
10. Rodriguez, L. L., Brooks, L. D., Greenberg, J. H. & Green, E. D. The complexities of genomic identifiability. *Science* **339**, 275–276 (2013).
11. Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care. *Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary* (National Academies Press (US), 2010).
12. McGuire, A. L. *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).
13. Oliver, J. M. *et al.* Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Publ. Health Genom.* **15**, 106–114 (2012).
14. Careless.data. *Nature* **507**, 7 (2014).
15. Schwartz, P. M. & Solove, D. J. Reconciling personal information in the United States and European Union. *102 California Law Rev.* http://dx.doi.org/10.2139/ssrn.2271442 (2013).
16. El Emam, K. Heuristics for de-identifying health data. *IEEE Secur. Priv.* **6**, 58–61 (2008).
17. Lunshof, J. E., Chadwick, R., Vorhaus, D. B. & Church, G. M. From genetic privacy to open consent. *Nature Rev. Genet.* **9**, 406–411 (2008).
18. Brenner, S. E. Be prepared for the big genome leak. *Nature* **498**, 139 (2013).
19. McClure, S., Scambray, J. & Kurtz, G. *Hacking Exposed 7: Network Security Secrets and Solutions* (McGraw Hill, 2012).
20. Solve, D. J. A taxonomy of privacy. *Univ. Pennsylvania Law Rev.* **154**, 477 (2006).
This work organizes various concepts of privacy violations from a legal perspective.
21. Ohm, P. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Rev.* **57**, 1701 (2010).
22. Golle, P. Revisiting the uniqueness of simple demographics in the US population. *Proc. 5th ACM Workshop Privacy in Electron. Soc.* 77–80 (2006).
23. Sweeney, L. A. *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon Univ. Data Privacy Working Paper 3 (2000).
24. Sweeney, L. Testimony of Latanya Sweeney before the Privacy and Integrity Advisory Committee of the Department of Homeland Security. *US Homeland Security* [online], http://www.dhs.gov/xlibrary/assets/privacy/privacy_advcom_06-2005_testimony_sweeney.pdf (2005).
25. Sweeney, L. A., Abu, A. & Winn, J. Identifying participants in the personal genome project by name. *Data Privacy Lab* [online], http://dataprivacylab.org/projects/pgp/1021-1.pdf (2013).
This study shows identity tracing of PGP participants using metadata and side-channel techniques.
26. Code of Federal Regulations Title 45 Section 164.514 (US Federal Register, 2002).
27. Benitez, K. & Malin, B. Evaluating re-identification risks with respect to the HIPAA Privacy Rule. *J. Am. Med. Informat. Associ.* **17**, 169–177 (2010).
28. Kwok, P., Davern, M., Hair, E. & Lafky, D. *Harder Than You Think: a Case Study of Re-identification Risk of HIPAA-Compliant Records*. NORC at The University of Chicago Abstract 302255 (2011).
29. Bennett, R. L. *et al.* Recommendations for standardized human pedigree nomenclature. Pedigree standardization task force of the national society of genetic counselors. *Am. J. Hum. Genet.* **56**, 745–752 (1995).
30. Malin, B. Re-identification of familial database records. *AMIA Annu. Symp. Proc.* **2006**, 524–528 (2006).
31. Israel v. N. Bilik and others 24441-05-12 [online], http://www.law.co.il/media/computer-law/bilik2.pdf (in Hebrew) (2013).
32. Khan, R. & Mittelman, D. Rumors of the death of consumer genomics are greatly exaggerated. *Genome Biol.* **14**, 139 (2013).
33. 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).
34. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
This paper reports end-to-end identity tracing of anonymous research participants from DNA information and Internet searches, and a risk assessment for the US population.
35. King, T. E. & Jobling, M. A. What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *Trends Genet.* **25**, 351–360 (2009).
36. King, T. E. & Jobling, M. A. Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames. *Mol. Biol. Evol.* **26**, 1093–1102 (2009).
37. Motluk, A. Anonymous sperm donor traced on internet. *New Scientist* 2 (3 Nov 2005).
This article discusses the first public case of identity tracing using genealogical triangulation.
38. Stein, R. Found on the web, with DNA: a boy's father. *Washington Post* A09 (13 Nov 2005).
39. Naik, G. Family secrets: an adopted man's 26-year quest for his father. *The Wall Street Journal* (2 May 2009).
40. Lehmann-Haupt, R. Are sperm donors really anonymous anymore? *Slate* (1 Mar 2010).
41. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
42. China News Network. Ministry of Public Security statistics: "King" into the most common surname in China has 9288 million. *Eastday* [online], http://news.eastday.com/c/20070424/u1a2791347.html (in Chinese) (2007).
43. Huff, C. D. *et al.* Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res.* **21**, 768–774 (2011).
44. Henn, B. M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS ONE* **7**, e34267 (2012).
45. Lowrance, W. W. & Collins, F. S. Identifiability in genomic research. *Science* **317**, 600–602 (2007).
46. Kayser, M. & de Knijff, P. Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Rev. Genet.* **12**, 179–192 (2011).
This is a comprehensive review of methods to predict phenotypes from DNA information.
47. Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res.* **6**, 399–408 (2003).
48. Kohn, L. A. P. The role of genetics in craniofacial morphology and growth. *Annu. Rev. Anthropol.* **20**, 261–278 (1991).
49. Zubakov, D. *et al.* Estimating human age from T-cell DNA rearrangements. *Curr. Biol.* **20**, R970–R971 (2010).
50. Ou, X. L. *et al.* Predicting human age with bloodstains by sjTREC quantification. *PLoS ONE* **7**, e42412 (2012).
51. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
52. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature Genet.* **44**, 659–669 (2012).
53. Liu, F. *et al.* A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet.* **8**, e1002932 (2012).
54. Walsh, S. *et al.* IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forens. Sci. Int. Genet.* **5**, 170–180 (2011).
55. Byers, S. Information leakage caused by hidden data in published documents. *IEEE Secur. Priv.* **2**, 23–27 (2004).
56. Kaufman, S., Rosset, S. & Perlich, C. Leakage in data mining: formulation, detection, and avoidance. *Proc. 17th ACM SIGKDD Int. Conf. Knowledge Discov. Data Mining* 556–563 (2011).
57. Acquisti, A. & Gross, R. Predicting Social Security numbers from public data. *Proc. Natl Acad. Sci. USA* **106**, 10975–10980 (2009).
58. Noumeir, R., Lemay, A. & Lina, J. M. Pseudonymization of radiology data for research purposes. *J. Digital Imag.* **20**, 284–295 (2007).
59. Pakstis, A. J. *et al.* SNPs for a universal individual identification panel. *Hum. Genet.* **127**, 315–324 (2010).
60. Lin, Z., Owen, A. B. & Altman, R. B. Genomic research and human subject privacy. *Science* **305**, 183 (2004).
61. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–1186 (2007).
62. 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).
This is the first study to show an ADAD from summary statistic data.
63. Jacobs, K. B. *et al.* A new statistic and its power to infer membership in a genome-wide association study using genotype frequencies. *Nature Genet.* **41**, 1253–1257 (2009).
64. Visscher, P. M. & Hill, W. G. The limits of individual identification from sample allele frequencies: theory and statistical analysis. *PLoS Genet.* **5**, e1000628 (2009).
65. Sankararaman, S., Obozinski, G., Jordan, M. I. & Halperin, E. Genomic privacy and limits of individual detection in a pool. *Nature Genet.* **41**, 965–967 (2009).
References 64–65 provide excellent mathematical analyses of ADAD using allele frequency data.
66. Wang, R., Li, Y. F., Wang, X., Haixu, T. & Zhou, X. Learning your identity and disease from research papers: information leaks in genome wide association study. *Proc. 16th ACM Conf. Comput. Commun. Security* 534–544 (2009).
67. Im, H. K., Gamazon, E. R., Nicolae, D. L. & Cox, N. J. On sharing quantitative trait, GWAS results in an era of multiple-omics data and the limits of genomic privacy. *Am. J. Hum. Genet.* **90**, 591–598 (2012).
68. Lumley, T. Potential for revealing individual-level information in genome-wide association studies. *JAMA* **303**, 659 (2010).

69. Zerhouni, E. A. & Nabel, E. G. Protecting aggregate genomic data. *Science* **322**, 44 (2008).
70. Johnson, A. D., Leslie, R. & O'Donnell, C. J. Temporal trends in results availability from genome-wide association studies. *PLoS Genet.* **7**, e1002269 (2011).
71. Gilbert, N. Researchers criticize genetic data restrictions. *Nature* http://dx.doi.org/10.1038/news.2008.1083 (2008).
72. Malin, B., Karp, D. & Scheuermann, R. H. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J. Investig. Med.* **58**, 11–18 (2010).
73. Clayton, D. On inferring presence of an individual in a mixture: a Bayesian approach. *Biostatistics* **11**, 661–673 (2010).
74. Report on the workshop on establishing a central resource of data from genome sequencing projects. *National Genome Research Institute* [online], http://www.genome.gov/Pages/Research/DER/GVP/Data_Aggregation_Workshop_Summary.pdf (2012).
75. Schadt, E. E., Woo, S. & Hao, K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genet.* **44**, 603–608 (2012).
76. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature Rev. Genet.* **11**, 499–511 (2010).
77. Nyholt, D. R., Yu, C. E. & Visscher, P. M. On Jim Watson's *APOE* status: genetic information is hard to hide. *Eur. J. Hum. Genet.* **17**, 147–149 (2009).
This study clearly shows the limited use of masking sensitive DNA areas.
78. Humbert, M., Ayday, E., Hubaux, J.-P. & Telenti, A. Addressing the concerns of the Lacks family: quantification of kin genomic privacy. *Proc. 2013 ACM SIGSAC Conf. Comput. Commun. Secur.* 1141–1152 (2013).
79. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genet.* **40**, 1068–1075 (2008).
80. Kaiser, J. Agency nixes deCODE's new data-mining plan. *Science* **340**, 1388–1389 (2013).
81. Bambauer, J. R. Tragedy of the data commons. *Harvard J. Law Technol.* http://dx.doi.org/10.2139/ssrn.1789749 (2011).
82. Hartzog, W. & Stutzman, F. The case for online obscurity. *Calif. Law Rev.* **101**, 1 (2013).
83. Taleb, N. N. *The Black Swan: the Impact of the Highly Improbable* (Random House, 2007).
84. Shannon, C. Communication theory of secrecy systems. *Bell System Techn. J.* **28**, 656–715 (1949).
85. Cavoukian, A. Privacy by design. *Information and Privacy Commissioner, Ontario, Canada* [online], http://www.ipc.on.ca/images/Resources/privacybydesign.pdf%3E (2009).
86. Tryka, K. A. *et al.* NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res.* **42**, D975–D979 (2014).
87. Ramos, E. M. *et al.* A mechanism for controlled access to GWAS data: experience of the GAIN Data Access Committee. *Am. J. Hum. Genet.* **92**, 479–488 (2013).
88. Church, G. *et al.* Public access to genome-wide data: five views on balancing research with privacy and protection. *PLoS Genet.* **5**, e1000665 (2009).
89. Agrawal, R., Kiernan, J., Srikant, R. & Xu, Y. Hippocratic databases. *Proc. 28th Int. Conf. Very Large Databases* 143–154 (2002).
90. Agrawal, R. *et al.* Auditing compliance with a hippocratic database. *Proc. 30th Int. Conf. Very Large Databases* 516–527 (2004).
91. Venter, H. S., Olivier, M. S. & Eloff, J. H. PIDS: a privacy intrusion detection system. *Internet Res.* **14**, 360–365 (2004).
92. Creating a global alliance to enable responsible sharing of genomic and clincal data. [online], http://www.broadinstitute.org/files/news/pdfs/GAWhitePaperJune3.pdf (2013).
93. Bafna, V. *et al.* Abstractions for genomics. *Commun. ACM* **56**, 83–93 (2013).
94. Terry, S. F. & Terry, P. F. Power to the people: participant ownership of clinical trial data. *Sci. Transl Med.* **3**, 69cm3 (2011).
95. Kaye, J. *et al.* From patients to partners: participant-centric initiatives in biomedical research. *Nature Rev. Genet.* **13**, 371–376 (2012).
96. Sweeney, L. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzz.* **10**, 557–570 (2002).
97. El Emam, K. & Dankar, F. K. Protecting privacy using *k*-anonymity. *J. Am. Med. Informat. Associ.* **15**, 627–637 (2008).
98. Malin, B. A. Protecting genomic sequence anonymity with generalization lattices. *Methods Inform. Med.* **44**, 687–692 (2005).
99. Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkitasubramaniam, M. *L*-diversity: privacy beyond *k*-anonymity. *ACM Trans. Knowl. Discov. Data* **1**, 3 (2007).
100. Li, N., Li, T. & Venkatasubramanian, S. *t*-closeness: privacy beyond *k*-anonymity and *L*-diversity. *IEEE 23rd Int. Conf. Data Eng.* 106–115 (2007).
101. Dwork, C. Differential privacy. *Automata, Languages and Programming* 1–12 (Springer Verlag, 2006).
102. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. & Vilhuber, L. Privacy: theory meets practice on the map. *IEEE 24th Int. Conf. Data Eng.* 277–286 (2008).
103. Uhler, C., Slavkovic, A. B. & Fienberg, S. E. Privacy-preserving data sharing for genome-wide association studies. *arXiv* 1205.0739 (2012).
104. Yu, F., Fienberg, S. E., Slavkovic, A. & Uhler, C. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *arXiv* 1401.5193 (2014).
105. Johnson, A. & Shmatikov, V. Privacy-preserving data exploration in genome-wide association studies. *Proc. 19th ACM SIGKDD Int. Conf. Knowledge Discov. Data Mining* 1079–1087 (2013).
106. Ayday, E., Raisaro, J. L. & Hubaux, J. P. Privacy-enhancing technologies for medical tests using genomic data. *Ecole Polytechnique Federale de Lausanne* [online], http://infoscience.epfl.ch/record/182897/files/CS_version_technical_report.pdf (2013).
107. Ayday, E., Raisaro, J. L., McLaren, P. J., Fellay, J. & Hubaux, J.-P. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. *Proc. USENIX Security Workshop Health Inf. Technol.* http://citeseerx.ist.psu.edu/viewdoc/summary?doi = 10.1.1.309.1513 (2013).
This pioneering work shows the use of homomorphic encryption for privacy-preserving genetic risk predictions.
108. Atallah, M. J., Kerschbaum, F. & Du, W. Secure and private sequence comparisons. *Proc. 2003 ACM Workshop Privacy in Electron. Soc.* 39–44 (2003).
109. Jha, S., Kruger, L. & Shmatikov, V. Towards practical privacy for genomic computation. *IEEE Symp. Security and Privacy* 216–230 (2008).
110. Chen, Y., Peng, B., Wang, X. & Tang, H. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. *Proc. 19th Annu. Netw. Distributed Syst. Security Symp.*(2013).
The paper presents an interesting concept of privacy-preserving alignment of high-throughput sequencing data that allows the use of untrusted cloud providers.
111. Yao, A. C.-C. Protocols for secure computations. *23rd Annu. Symp. Found. Comput. Sci.* 160–164 (1982).
112. Li, H. & Homer, N. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform.* **11**, 473–483 (2010).
113. Bohannon, P., Jakobsson, M. & Srikwan, S. in *Public Key Cryptography* (eds Imai, H. & Zheng, Y.) 373–390 (Springer, 2000).
114. Fons, B., Stefan, K., Klaus, K. & Pim, T. Privacy-preserving matching of DNA profiles. *Cryptology ePrint Archive* **2008**, 203 (2008).
115. Baldi, P., Baronio, R., Cristofaro, E. D., Gasti, P. & Tsudik, G. Countering GATTACA: efficient and secure testing of fully-sequenced human genomes. *Proc. 18th ACM Conf. Comput. Commun. Security* 691–702 (2011).
116. De Cristofaro, E., Faber, S., Gasti, P. & Tsudik, G. Genodroid: are privacy-preserving genomic tests ready for prime time? *Proc. 2012 ACM Workshop Privacy in Electron. Soc.* 97–108 (2012).
117. He, D. *et al.* Identifying genetic relatives without compromising privacy. *Genome Res.* **24**, 664–672 (2014).
118. Kantarcioglu, M., Jiang, W., Liu, Y. & Malin, B. A cryptographic approach to securely share and query genomic sequences. *IEEE Trans. Inf. Technol. Biomed.* **12**, 606–617 (2008).
119. 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).
120. Canim, M., Kantarcioglu, M. & Malin, B. Secure management of biomedical data with cryptographic hardware. *IEEE Trans. Inf. Technol. Biomed.* **16**, 166–175 (2012).
121. Narayanan, A. What happened to the crypto dream? *IEEE Secur. Priv.* **11**, 75–76 (2013).
122. Ayday, E., De Cristofaro, E. Hubaux, J.-P. & Tsudik, G. The chills and thrills of whole genome sequencing. *Computer* http://doi.ieeecomputersociety.org/10.1109/MC.2013.333 (2013).
This is a good overview of cryptographic work for protecting genetic data and of open questions in the area.
123. Presidential Commission for the Study of Bioethical Issues. *Privacy and Progress in Whole Genome Sequencing* (2012).
124. Craig, D. W. *et al.* Assessing and managing risk when sharing aggregate genetic variant data. *Nature Rev. Genet.* **12**, 730–736 (2011).
125. Braun, R., Rowe, W., Schaefer, C., Zhang, J. & Buetow, K. Needles in the haystack: identifying individuals present in pooled genomic data. *PLoS Genet.* **5**, e1000668 (2009).
This is a critical assessment of the performance of ADAD with allele frequency data.
126. Kendler, K. S., Gallagher, T. J., Abelson, J. M. & Kessler, R. C. Lifetime prevalence, demographic risk factors, and diagnostic validity of nonaffective psychosis as assessed in a US community sample: the National Comorbidity Survey. *Arch. Gen. Psychiatry* **53**, 1022–1031 (1996).
127. Lee, J. & Clifton, C. in *Information Security* 325–340 (Springer, 2011).
128. Hsu, J. *et al.* Differential privacy: an economic method for choosing epsilon. *arXiv* 1402.3329 (2014).
129. Dwork, C., McSherry, F., Nissim, K. & Smith, A. in *Theory of Cryptography* 265–284 (Springer, 2006).
130. Paillier, P. in *Advances in Cryptology — EUROCRYPT '99* (ed. Stern, J.) 223–238 (Springer, 1999).
131. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
132. Gentry, C. Fully homomorphic encryption using ideal lattices. *Proc. 41st Annu. ACM Symp. Theory of Comput.* 169–178 (2009).
133. Wang, R., Li, Y. F., Wang, X. F., Tang, H. & Zhou, X. Learning your identity and disease from research papers: information leaks in genome wide association study. *Proc. 16th ACM Conf. Comput. Commun. Security* 534–544 (2009).

Acknowledgements
Y.E. 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 in part by a US National Human Genome Research Institute grant R21HG006167, and by a gift from C. Stone and J. Stone. The authors thank D. Zielinski and M. Gymrek for comments.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
An analysis of the distribution of birthdays in a calendar year: www.panix.com/~murphy/bday.html
Chronology of Data Breaches: http://www.privacyrights.org/data-breach
dbGaP: http://www.ncbi.nlm.nih.gov/gap
Eye Color Distribution Percentages: http://www.statisticbrain.com/eye-color-distribution-percentages/
GEDmatch: http://gedmatch.com
Mitosearch: http://www.mitosearch.org
OpenSNP: https://opensnp.org
PeopleFinders: http://www.peoplefinders.com
Platform for Engaging Everyone Responsibly: http://www.geneticalliance.org/programs/biotrust/peer
Promethease: promethease.com
Stanford School of Medicine Blood Center: http://bloodcenter.stanford.edu/education/blood_types.html
Ysearch: http://www.ysearch.org
ZipAtlas: http://zipatlas.com

SUPPLEMENTARY INFORMATION
See online article: S1 (figure)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF