Functional genomics data: privacy risk assessment and technological mitigation

Gamze Gürsoy, Tianxiao Li, Susanna Liu, Eric Ni, Charlotte M. Brannon and Mark B. Gerstein

Abstract | The generation of functional genomics data by next-generation sequencing has increased greatly in the past decade. Broad sharing of these data is essential for research advancement but poses notable privacy challenges, some of which are analogous to those that occur when sharing genetic variant data. However, there are also unique privacy challenges that arise from cryptic information leakage during the processing and summarization of functional genomics data from raw reads to derived quantities, such as gene expression values. Here, we review these challenges and present potential solutions for mitigating privacy risks while allowing broad data dissemination and analysis.

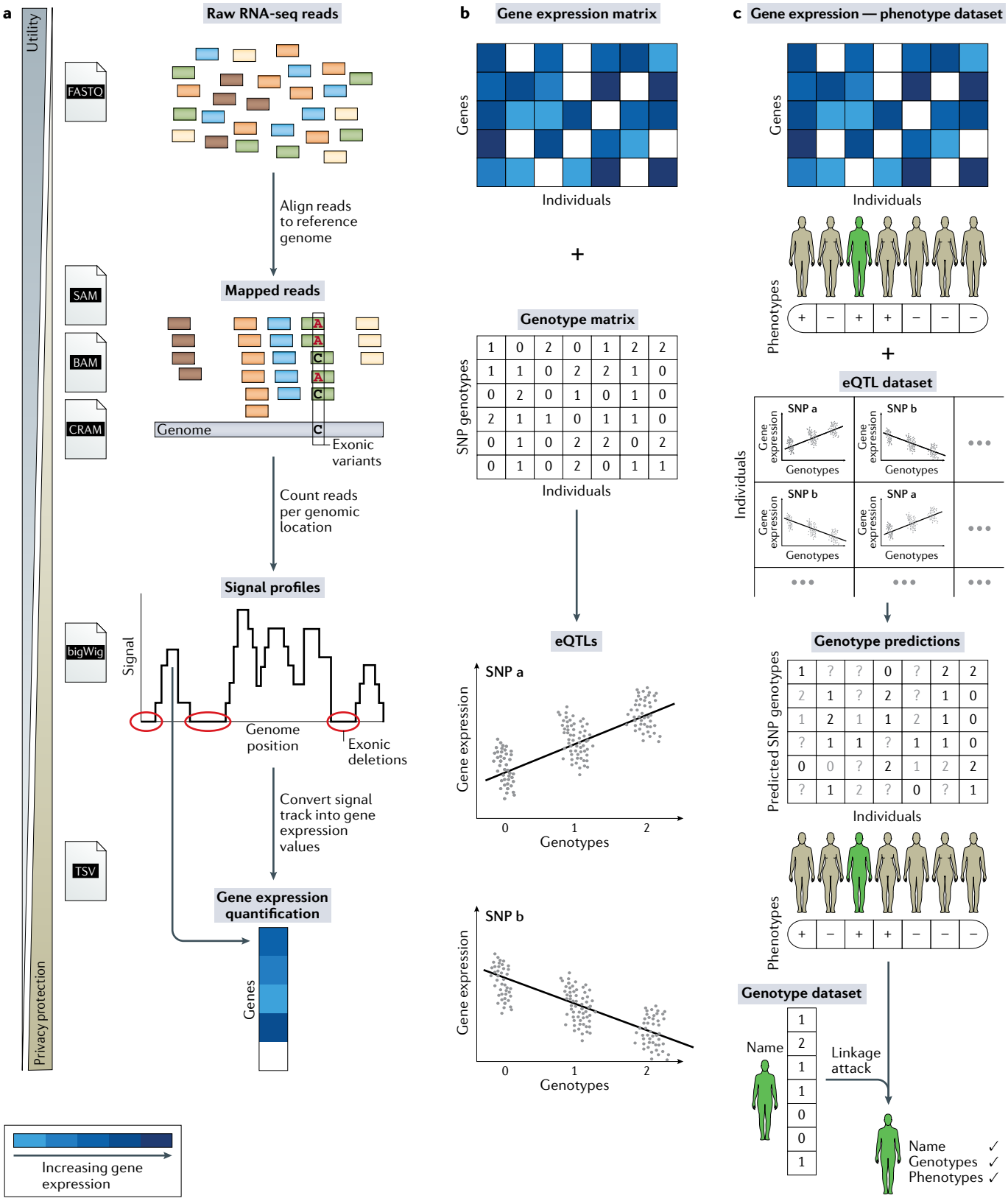
Thirty years ago, sequencing a single individual's genome was a substantial challenge. Today, scientists are launching projects that involve not only sequencing thousands of genomes, but also assaying other aspects of the genome, such as how genes are differentially expressed, which transcription factors are bound, or how chromatin is formed and organized[1–3]. These activities of the genome are studied with the aid of next-generation sequencing (NGS) under the umbrella of functional genomics[1,2]. Functional genomics experiments provide a wealth of information on genomic activities related to developmental stages or diseases, which is essential for personalized medicine[3] and can be medically actionable[4]. These studies use large-scale, high-throughput assays to quantify transcription (RNA sequencing (RNA-seq))[5], epigenetic regulation (chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq))[6], chromatin accessibility (DNase digestion and high-throughput sequencing (DNase-seq) or assay for transposase-accessible chromatin using sequencing (ATAC-seq))[7,8] or the 3D organization of the genome (Hi-C)[9] under different conditions (for example, samples from patients and healthy individuals). In addition, the field is rapidly evolving, and new biochemical techniques based on NGS or other technologies are continuously being developed[10]. Many consortia, such as the Genotype-Tissue Expression project[11] (GTEx), Encyclopedia of DNA Elements (ENCODE)[12], gEUVADIS[13] and The Cancer Genome Atlas (TCGA)[14], generate large-scale functional genomics datasets for many individuals. Although mining these data is essential to understand human biology in health and disease[3,15], these new modalities bring challenges in protecting the privacy of individuals that differ from those of traditional DNA sequencing data. Importantly, although there has been an increase in risk assessment methods and studies tackling the privacy issues related to gene expression data[16–19], the privacy issues of other functional genomics data types have not been studied as widely. Moreover, the rapid increase in the number of functional genomics assays creates further privacy issues to be accommodated. Therefore, current privacy solutions need to account for the unique aspects of functional genomics data.

Inferring biological information from functional genomics experiments is a multistep procedure, in which data are progressively summarized from raw sequencing reads to gene expression quantifications[20–22] (FIG. 1), transcription factor binding peaks[23–25] or chromatin interaction matrices[26]. The private information that can be inferred from functional genomics data can be broadly categorized into two groups: information related to genetic variants and information not related to genetic variants. The first category includes noisy genotypes that can be directly observed from the reads[16,27], similar to DNA sequencing data, and cryptic variant information hidden in the derived data[17–19,28] (for example, gene expression quantifications or transcription factor binding enrichment). The second category comprises phenotypic information that can be inferred from genome activity (for example, differential gene expression associated with disease[15]) and characterizing information that can be observed from small amounts of exogenous reads[29] (for example, the microbiome is a better predictor for disease phenotype[30] than genotype and can even reveal the location of an individual[31]).

In this Perspective, we first introduce general privacy issues with genomic technologies, which are applicable to both DNA sequencing and functional genomics data. We then discuss privacy problems specific to functional genomics data. We consider the unique characteristics of functional genomics data and privacy issues related to the sharing of these data, and discuss how the response should differ from current sharing practices for DNA sequencing data. We describe the genotypic and phenotypic information leakage from various different functional genomics data types as well as from various data summarization steps. Lastly, we discuss various techniques that will enable the broad sharing and/or confidential analysis of functional genomics data.

Privacy and genomics

In the United States, one of the most important health privacy protections was established in the late 1990s with the enactment of the Health Insurance Portability and Accountability Act (HIPAA)[32]. HIPAA requires regulations on the privacy of medical health records, which includes the anonymization of 'genetic testing'. Because the technologies in genomics have improved tremendously, Congress passed the Genetic Information Nondiscrimination Act (GINA) in 2008



to protect individuals from discrimination by insurance companies and employers based on their genomes[33]. The latest data sharing policy of the National Institutes of Health (NIH) highlights the importance of maintaining participant privacy and describes various approaches that provide privacy while meeting data-sharing expectations. Of note, the landscape of privacy rules surrounding genetic data differs globally; with the introduction of the European General Data Protection Regulation (GDPR) in 2016, the use of personal health and genetic data has been regulated substantially in Europe[34].

◀ Fig. 1 | **Private information leakage in functional genomics data. a** | Different layers of data are produced from functional genomics experiments, as exemplified by the stepwise procedure from functional genomics reads to gene expression values in RNA sequencing (RNA-seq). The processing of RNA-seq data starts with aligning a sample's raw reads obtained from the sequencer, which are stored as FASTQ files, to the reference genome. After the mapping, these reads reveal the genetic variants of the individual from whom the sample is taken. These are the data types that leak the greatest amount of sensitive information, while also possessing the highest utility. Once the mapping is complete, a signal track is created by counting the reads at each location. The signal track reveals large deletions in the genome. Different signal levels from the signal track can then be averaged and turned into gene expression values in the case of RNA-seq or used to call peaks in chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq) or assay for transposase-accessible chromatin using sequencing (ATAC-seq), for example. Moving downwards through the different stages, there is less privacy concern but largely reduced utility and amount of data. **b** | For expression quantitative trait locus (eQTL) mapping, the gene expression matrix is composed of the expression values of all genes from a cohort of individuals. The genotype matrix comprises each individual's single nucleotide polymorphism (SNP) genotype in this cohort. An eQTL is observed by regressing the gene expression values of a gene across different genotypes for a given SNP. **c** | Schematic representation of a linkage attack using functional genomics data. Public gene expression values of a cohort of individuals, along with their phenotypes, can be combined with publicly available eQTL data to predict SNP genotypes for these individuals (note that these predicted SNPs are noisy (denoted by grey values) and incomplete (denoted by question marks)). We can then overlap the genotypes of a known individual to link the individual to the gene expression cohort and hence to their phenotypes. BAM, binary alignment map; CRAM, compressive alignment map; SAM, sequence alignment map.

Genomics has emerged as a major focus of studies on privacy, not only among ethicists and legal scholars but also among geneticists and computer scientists[35–47]. This focus can be attributed primarily to the advancement of technologies for high-throughput data acquisition, which have led to a surge in datasets[48,49]. Owing to steep declines in sequencing costs, a growing number of companies now offer to collect, analyse and return genomic information directly to the public. In addition to commercial entities, several research organizations have harnessed technological developments to collect and process thousands of genomic datasets for research[14,50–52].

Genomic information is the genetic variation within each genome, such as single nucleotide polymorphisms (SNPs), small insertions and deletions (indels) and other large-scale complex rearrangements such as structural variation (SV) data. In contrast to standard medical data, this information is inherently personally identifiable information. By its very definition, raw genomic data identify the owner, as illustrated by a number of high-profile cases reported in the news, for example, a case in which the DNA on a stamp was sufficient to charge an individual with a crime[53]. Moreover, genomic data are largely shared by close family members. Therefore, even for instances in which patients have provided broad consent for their genomic information to be accessed and used, care must be taken to preserve patient privacy as the data implicate not only the immediate owner of the genome sequence but many third-party relatives as well.

There are also issues related to the privacy of groups, which is important from the perspective of disparities in health care and biomedical research[54,55]. For example, a vast amount of genomic data from individuals of European ancestry have been collected and broadly shared, rendering this population more prone to re-identification through forensic investigations conducted by law enforcement[56]. By contrast, owing to the scarcity of genomic data from Indigenous and Native American individuals, the broad sharing of functional genomics data derived from their biosamples will be more invasive to privacy as it is easier to find rare, and thus more informative, cryptic relationships between genomic and functional genomics data. These two competing issues need to be taken into account at the stages of informed consent, genomic data collection and sharing. When the NIH research programme All of Us was launched in 2008 with the aim of using health and DNA data collected from Native Americans to build a precision medicine database, this effort received heavy criticism from Indigenous geneticists for bypassing the consent of tribes[54,57,58].

Several privacy issues in genomics are newly emerging. For example, human genome sequencing can result in reads that are sequenced from exogenous species potentially belonging to the microbiome of an individual, which can reveal disease conditions or phenotypes[29,30,59]. These exogenous sequences can be the result of functional genomics assays, such as RNA-seq or ChIP–seq, or can be obtained using other sequencing-based assays, such as whole-genome or exome sequencing. In addition, lower barriers to sequencing have enabled a wider range of people, including citizen scientists, to access genomic and functional genomic data. Easy-to-use and portable sequencing technologies, such as the Oxford Nanopore MinION (which is used for both genome sequencing and functional genomics assays), and access to large databases via programmes such as All of Us are fuelling the 'do-it-yourself' science movement[60].

Early genomic privacy studies focused on the identification of individuals in a mixture, such as in a genome-wide association study (GWAS) case– cohort, by using phenotype–genotype association[8,61,62]. These studies showed that private information of an individual, such as participation in a drug-abuse study, can be revealed by using properties of genetic data such as allele frequencies observed in GWAS summary statistics. Researchers have further quantified the privacy risk associated with de-identified genomes by showing that multiple datasets can be linked to infer sensitive information such as participants' surnames[63] or addresses[64]. Such cross-referencing is performed through linkage attacks (BOX 1; FIG. 1c) and relies on quasi-identifiers, which are pieces of information that are not unique identifiers by themselves but correlate well with unique identifiers[65]. Multiple quasi-identifiers can be combined to obtain a unique identifier[64].

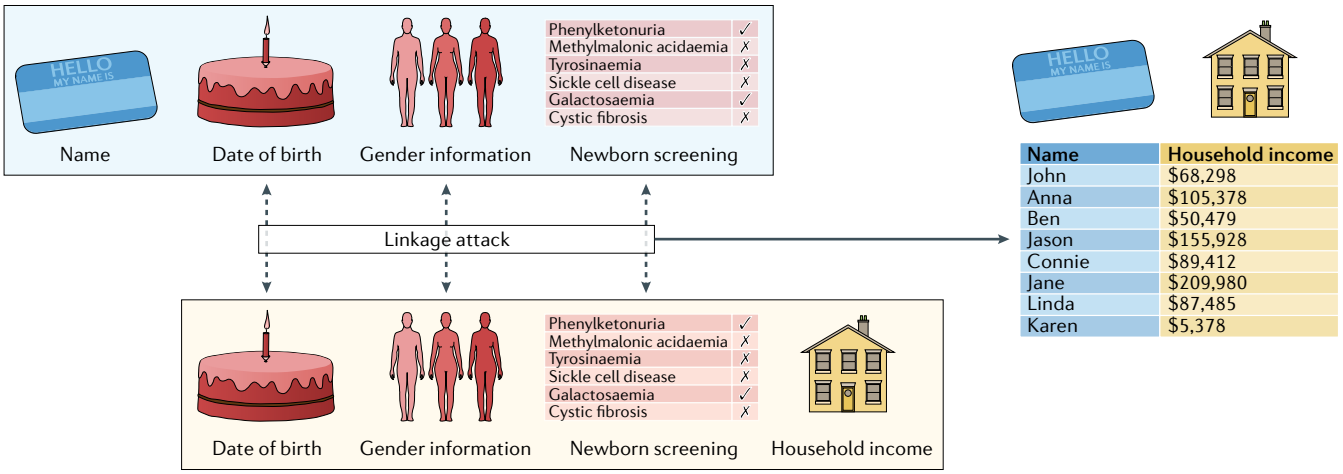
Functional genomics data obtained through NGS, such as RNA-seq and ChIP–seq data, bring about similar privacy issues in relation to the observation of private genetic variants of patients and research participants. Nonetheless, functional genomics data also differ from traditional genome sequencing data in two aspects: first, genetic variants are only a by-product of these experiments and are often not needed; second, potentially more invasive private information such as phenotypes and lifestyle can be gleaned from these data. Below, we focus on the private information leakage from functional genomics data; we refer readers to other reviews for more detail on the privacy of genetic data[66,67] and general biomedical data[68,69].

Privacy and functional genomics

Functional genomics data analysis starts with the generation of DNA or RNA sequencing reads that are stored in a file format called FASTQ[70] (FIG. 1a). These files are then mapped to the human reference genome and stored as compressed binary file types called binary alignment maps (BAMs) and/or compressive alignment maps

Box 1 | **Linkage attacks**

A linkage attack is a privacy breach that aims to de-anonymize an anonymized dataset. Such attacks are based on overlapping two datasets — one anonymized, one known — to learn sensitive information about an individual who is present in both datasets (see the figure). Consider two datasets A and B containing different kinds of information about the same group of people. Dataset A contains individuals' names along with their date of birth, zip code, gender information and the genetic screening that was done at the time of birth. Dataset B contains some of this information, such as gender, date of birth and newborn screening along with a new, sensitive piece of information, such as household income. The sensitive information is supposedly 'anonymized'; it is stored in dataset B without identifying information. However, an adversary interested in revealing sensitive information can simply link the information in dataset B, the 'quasi-identifiers', to that in dataset A to find out the household income of the named individuals in dataset A. This breach of privacy can also be used to de-anonymize functional genomics datasets (FIG. 1c).



(CRAMs), which are derived from sequence alignment map (SAM) files in text format[71]. Further summarization of the mapped reads (such as signal profiles, which depict the cumulative number of reads mapping to a genomic region, or gene expression quantification) allows researchers to make accurate biological conclusions while providing an additional 20-fold reduction in data size. In particular, read alignment files (SAM, BAM or CRAM) are of great interest owing to their large amount of biological data, and these files constitute the most important input in most genome annotation pipelines. However, these files contain sequence information of individuals and therefore leak sensitive data.

To provide solutions for maintaining privacy while sharing functional genomics data, one must comprehensively quantify the amount of private information at every data summarization level of functional genomics data processing (BOX 2). There are different ways in which private information can be leaked from functional genomics data, which can be categorized based on whether the data are related to genetic variants.

Leakage based on genetic variants

Direct genotyping from reads. The first and most obvious source of privacy leakage from functional genomics data is based on direct genotyping from reads (FIG. 1a). One must consider the following question: if we obtain raw functional genomics data from known individuals, can we recover their genotypes without using any other datasets? In contrast to DNA sequencing data, owing to the targeted nature of functional genomics assays (which focus on exons or transcription factor binding sites, for example), sequences present in functional genomics reads often cover only a small portion of the genome and are subject to various biases and base changes due to, for example, RNA editing or methylation. Therefore, genotypes inferred from an individual functional genomics dataset are noisy and incomplete and, hence, are typically not readily used for building personal genomes. However, sequencing from multiple functional genomics datasets can be combined to call more variants, as some assays target different regions in the genome (FIG. 2a). The called variants can be used further to impute, that is, statistically infer, genotypes for missing variants[72,73]; by combining assays and performing genotype imputation, we can infer almost as many genotypes as inferred from 30× whole-genome sequencing data (FIG. 2a). Although partial, the genotypes observed in raw reads can be used to identify individuals[16] (FIG. 2b). If these reads are part of an anonymized database that contains phenotypic information, with the help of linkage attacks one can quantify the privacy risk by inferring phenotypes of a known individual with a known genome[16]. Using overall sequencing statistics (for example, read-depth distribution) and supervised learning, the number of genotypes, and hence the private information leakage, from functional genomics reads could be predicted before the release of such data[74].

Cryptic genotyping from signal profiles and expression values. Another source of leakage is signal profiles in functional genomics data (FIG. 1a), in which small and large deletions can be inferred by linking signal profile data with SV data[28]. In this situation, if there is a deletion in an individual's genome, no reads will map to the location of the deletion, which yields an absence of signal when the depth of sequencing is calculated. If these signal profiles are part of an anonymized database that contains phenotypic information, one can quantify the privacy risk via linkage attacks that can help infer phenotypes of a known individual with known genomic deletions. In more elaborate settings, one can predict possible genotypes of an individual from their gene expression profiles via expression quantitative trait loci (eQTLs)[17,19] (FIG. 1b,c). This privacy risk quantification is done by inverting the genotype–molecular phenotype relationship. Because eQTLs are determined by the slope of the gene expression versus genotype curve for a given SNP (FIG. 1b), one can predict the genotype for that SNP for an individual by comparing the individual's gene expression level with those of a cohort. A set of SNP genotypes

can then be accumulated by looking at SNPs that are found to be eQTLs, which are publicly available. These genotypes can easily be cross-referenced with the genetic variants of a known individual, which allows for linking of a known individual to their gene expression data. Since gene expression levels are often linked to a phenotype, this linking allows for inference of potentially sensitive and stigmatizing phenotypes of known individuals.

Similar approaches can potentially be applied by using chromatin accessibility QTLs or splicing QTLs to predict genotypes for an individual. Recent work has demonstrated that given access to a list of an individual's allele-specific genes, the risk of private information leakage can be quantified by making inferences about the individual's heterozygous variants; one can then link these variants to full genotypes or phenotypes[18]. The noisy, cryptic genotypic information present in summary-level data (gene expression or chromatin accessibility information) requires statistical analysis for privacy leakage quantification. Such quantification is typically performed via linkage attacks[17] (FIG. 1c). For example, predicted eQTL genotypes can be overlapped with genotypes gathered from an illicitly obtained DNA sample, such as from a coffee cup, to connect the donor to a functional genomics data cohort involving the phenotype of the individual[16,17]. Note that, in this scenario, the anonymized functional genomics cohort contains information about the phenotypes of the samples. The genotypes from 'coffee cups' and functional genomics data are used as quasi-identifiers to connect the identity of an individual to a phenotype. In addition, it has been shown that summarized DNA methylation (whole-genome or array-based) data can leak identifying genotypes or private health information[75,76].

Leakage without genetic variants

Direct phenotyping. The activities of genes and proteins often correlate highly with phenotypes of tissues or individuals. For instance, associations between the expression values of certain genes and the cancer status of patients can be derived using publicly available TCGA data[77]. The gene expression values of a known individual can then be compared with the expected values from a cancer type to infer the disease status of the individual; a diagnosis of leukaemia, for example, can be predicted on the basis of gene expression values derived from blood RNA-seq[78]. Moreover, genome-wide DNA methylation patterns can be used to infer age-related diseases[79] (FIG. 3a), as sets of DNA methylation biomarkers can predict the biological age of any tissue during the human lifespan[80–83]. If methylation data were obtained from a biosample of a known individual, one could easily calculate a biological age that is older or younger than the chronological age of the individual (FIG. 3a). This information could be used by insurance companies or employers in a discriminatory way. Although the definition of genetic testing under GINA is the "analysis of human DNA, RNA, chromosomes, proteins, or metabolites that detect genotypes, mutations, or chromosomal changes"[84], which technically covers functional genomics data, it is unclear whether summary-level functional genomics data for direct phenotyping is covered under GINA. The privacy risk of sharing DNA methylation data in terms of age prediction has yet to be studied.

Incidental phenotyping. In addition to recovering the genomes of individuals or inferring sensitive phenotypes via linkage attacks, functional genomic data can potentially be even more intrusive, as they can reflect privately held life choices, for example, diet and residence, based on microbiome inferences.

Box 2 | **Private information leakage during summarization of functional genomics data**

To demonstrate the extent of the genetic variant leakage in functional genomics data, we reviewed all known sources of genetic variants at different stages of the data summarization process (see the table). Variants in the 1000 Genomes Project panel[52] were overlapped with exons to calculate the number of potential leaking variants. The number of variants that can be genotyped from a typical RNA sequencing (RNA-seq) experiment was calculated as the number of accessible variants. Multiplying the number of accessible variants with the average leakage per variant quantifies the total amount of leakage. This procedure was repeated to calculate the leakage from other sources such as signal profiles and gene expression quantifications.

The most obvious leakage occurs directly from the reads and can be largely avoided by converting alignment files into privacy-preserving counterparts[16]. The next source of leakage is from the signal profiles[28]. Additional leakage can come from further summarization of the signal profiles, such as the quantification of gene expression values[17] in RNA-seq or peak calling in other assays such as chromatin immunoprecipitation followed by high-throughput sequencing (ChIP–seq) and assay for transposase-accessible chromatin using sequencing (ATAC-seq).

At the read level, we can potentially observe all single nucleotide variants (SNVs) on the exons; however, only a fraction are accessible through RNA-seq depending on which gene is expressed in which cell line or type. This level applies to other next-generation sequencing (NGS)-based functional genomics data types such as ChIP–seq or ATAC-seq.

At the signal profile level, we can potentially observe deletions on the exons. However, only a fraction are accessible to the experiment depending on the expressed transcripts in the given cell line or type. Here, we show the number of accessible deletions that can be genotyped using the signal profile of the poly(A) RNA-seq experiment of individual NA12878, one of the individuals sequenced as part of the 1000 Genomes Project[16,28]. This level applies to other NGS-based one-dimensional functional genomics data types such as ChIP–seq or ATAC-seq. It is also equivalent to interaction matrices in two-dimensional functional genomics data types such as Hi-C or Hi-ChIP[28]. One potential countermeasure is to create pseudo-signals that follow the signal distribution of the data[28].

At the gene expression quantification level, the potential number of variants that can be observed are all expression quantitative trait loci (eQTLs) connected to the genes. We show the number of accessible variants through gene expression quantification as the average number of eQTLs per individual[16,17]. Although the privacy issues related to summary-level data of other functional genomics data have not yet been studied, this level corresponds to binding peaks in ChIP–seq, accessible elements in ATAC-seq or DNase-seq, and topologically associating domains in Hi-C or Hi-ChIP data.

Leakage source	Leaking variants	Number of potential variants	Average leakage per variant (bits)	Maximum leakage per variant (bits)	Number of accessible variants	Total leakage (bits)
Mapped reads	Exonic variants	2,772,064	0.10 ± 0.28	9.88 ± 2.12	221,293	22,129
Signal profiles	Exonic deletions	51,408	0.28 ± 0.45	7.97 ± 2.42	1,067	299
Gene expression quantification (summary-level data)	eQTLs	3,175	1.19 ± 0.36	4.00 ± 1.92	158	188

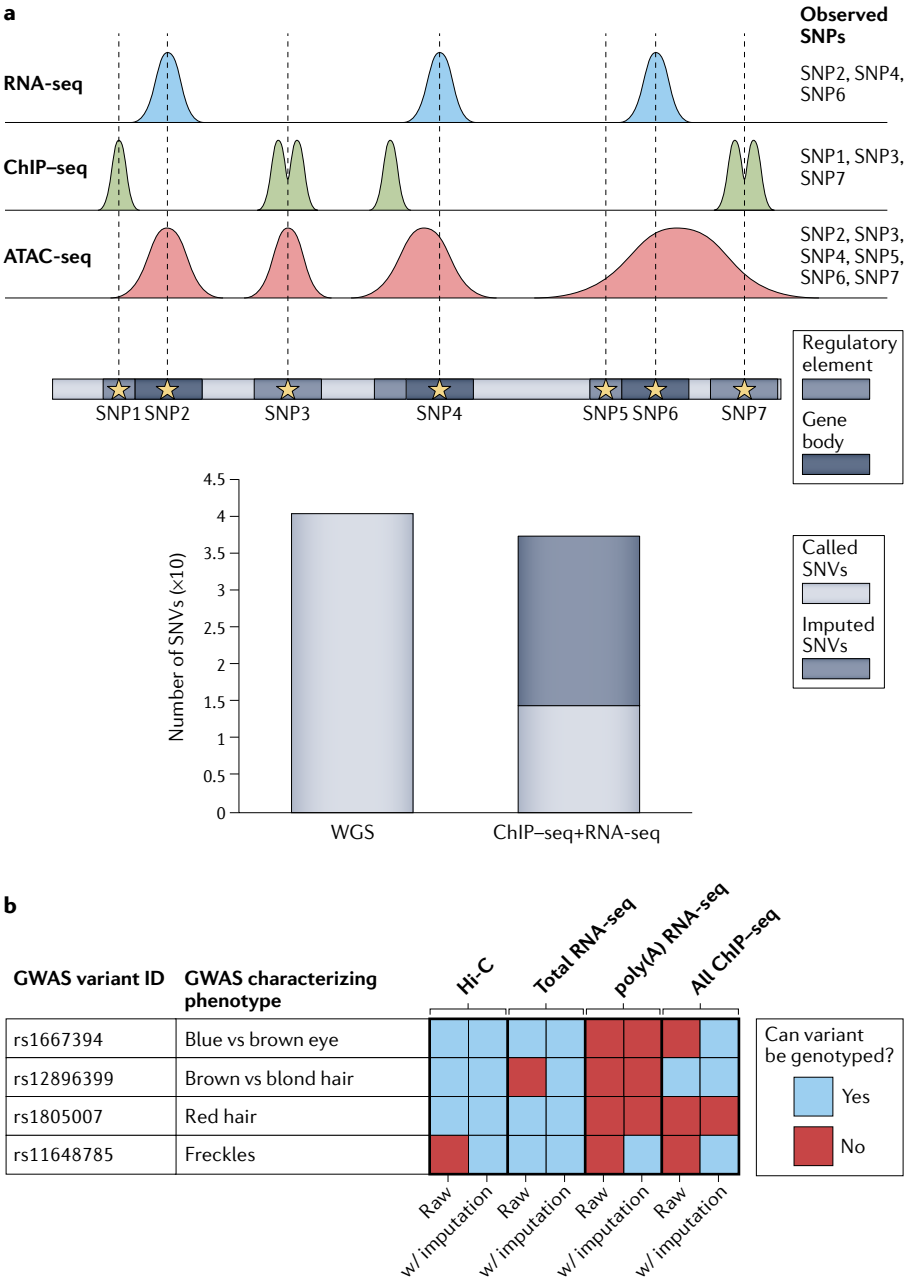


Fig. 2 | **Genetic characterization of an individual through functional genomics. a** | Different functional genomics assays target different locations in the genome; therefore, next-generation sequencing (NGS) reads from different assays can be combined to boost genotyping power. For example, the genetic variants observed by combining all chromatin immunoprecipitation followed by sequencing (ChIP–seq) and RNA sequencing (RNA-seq) reads of the individual NA12878 can be used for genotype imputation (see graph). Functional genomics datasets used in this calculation can be accessed from the ENCODE[12] data portal using the experiment matrix in this link: https://www.encodeproject.org/matrix/?type=Experiment&control_type%21=%2A&status=released&biosample_ontology.term_name=GM12878&assay_title=Histone+ChIP-seq&assay_title=polyA+plus+RNA-seq&assay_title=-total+RNA-seq. Genotyping based on whole-genome sequencing (WGS) data was obtained from the 1000 Genomes Project[52] (https://www.internationalgenome.org/data-portal/sample/NA12878). The total number of genotypes obtained by combining different functional genomics assays and performing imputation almost approximates to the total number of genotypes obtained from a high-depth WGS assay. **b** | Characterizing an individual by investigating genome-wide association study (GWAS) single nucleotide polymorphisms (SNPs) in functional genomics data. We identified GWAS variants that characterize phenotypes in Hi-C, RNA-seq and ChIP–seq data of NA12878 before and after genotype imputation. That is, owing to different coverage of different functional genomics data, one can look for GWAS variants associated with individual characterizing information in the genotypes inferred from functional genomics data.

Recent studies have shown a strong connection between the microbiome ecosystem of an individual and their phenotypes and behaviour[30,31]. These findings suggest that knowledge of the microbial communities of a person can help make predictions about their disease status, which can be stigmatizing in certain situations (for example, in the case of sexually transmitted viruses or bacteria). The non-human reads present in collected data can be analysed to calculate the abundance of microbial species (FIG. 3b). Evidence in the literature may suggest associations between these species and disease phenotypes, smoking status or even previous whereabouts of an individual[85]. Although not as extensively studied as genomic variants, the privacy risk of sharing the abundances of various microbial species and strains has been quantified for re-identification and characterization purposes[86].

Although functional genomics experiments are not intended to collect information from the microbiome of a tissue, research has shown that a small amount of exogenous reads might be collected by virtue of the experimental techniques[59]. This was especially studied in RNA-seq data, in which the fragments of RNA from the microbiome were used as a proxy to calculate the abundances of the species[29]. However, as the amounts of these reads are probably less than that of a typical microbiome sequencing assay, it might be difficult to perform a linkage attack with these abundances. That is, such microbial inference from functional genomics data can be used to characterize an individual but may not be as readily used for re-identification purposes.

Secure sharing of functional genomics data

With the availability of more advanced techniques, such as single-cell RNA-seq and ATAC-seq, we can now assay millions of cells[87–89]. This increased resolution is bringing a surge of new data from large cohorts of individuals, which will surpass the number of available sequenced genomes. We can sequence the DNA of an individual once, but numerous functional genomics assays can be performed on a single sample. When considering the privacy of DNA sequencing data, we deal with one data file per individual; when considering the privacy of functional genomics data, we must deal with many more data files per individual. This difference will soon create substantial hurdles in terms of movement and

storage of these private data in centralized controlled-access servers (FIG. 4).

Traditional data-sharing models, such as controlled-access models, can be challenging: access to private data requires complex, often overly bureaucratic, user agreements[90,91] (FIG. 4a). Unfortunately, the amount of broadly shared functional genomics data is much smaller than that stored behind controlled access (FIG. 4b), which affects the utility of these data. Therefore, funding agencies are increasingly supporting new means of broad data sharing and new requirements for making data publicly available while preserving participants' privacy[92]. It is important to find ways to honour privacy protection without losing practical data utility, as large-scale mining of functional genomics data will allow researchers to genetically and environmentally characterize disease states and susceptibility in detail. Moreover, there is an increasing push towards moving genomic analysis and data sharing to cloud computing platforms because of cost-effectiveness. This creates further privacy issues related to sharing private data with third-party cloud service providers. It also challenges the current controlled-access data-sharing protocols by moving computation on sensitive data from local computers to third-party cloud services. Although the utility of functional genomics data is different from the utility of DNA sequencing data, currently, raw functional genomics data are shared by following protocols based on traditional DNA sequencing data, that is, raw reads are shared using controlled access, whereas summary-level data are shared broadly.

DNA sequencing seeks to determine the genetic variants of individuals; by contrast, genetic variants are not the focus of functional genomics assays. When pipelines proceed from sequencing reads to final quantifications, the genetic variants in these reads are not used in key calculations such as gene expression quantification or transcription factor binding peak enrichment. However, other features in the reference genome mapped reads, such as the location in the genome to which they are mapped, read length and mapping quality of the reads, are important quantities. Hence, although raw reads are an important component in functional genomics data processing, the genetic variants in these data are irrelevant in most of the calculations. If one can disseminate data without the variants in the reads, the data can still be largely used by the community, although we note that cryptic variant information still potentially exists in summary-level data. Nevertheless, they require a different mode of data sharing that must shift away from traditional controlled access for better practical utility[93].

Solutions to data sharing due to privacy concerns cannot be separated from data utility. Unlike privacy, it is often difficult to formally define utility, as a single dataset can be used in different ways depending on the situation. Moreover, future use cases could arise that may change the definition of data utility in a manner that we cannot anticipate. Therefore, utility can be divided into two categories: practical utility and rigorous (mathematical) utility with theoretical guarantees. For example, the practical utility of data is often measured by the number of citations or downloads of a dataset, whereas the mathematical utility of data can be based on whether a defined function can be precisely calculated using the data. Of note, defining a unified utility metric for functional genomics data sharing may not be mathematically possible. Because different functional data types from the data summarization steps are used for different purposes, the utility

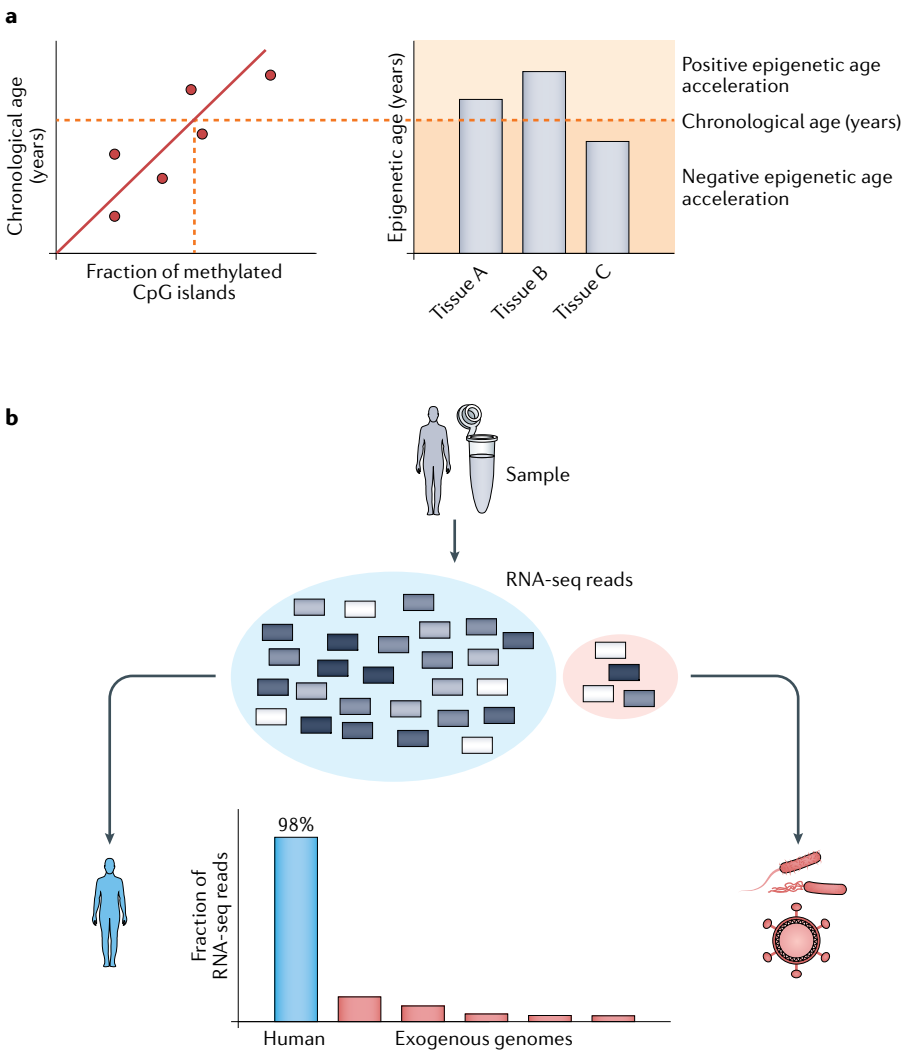
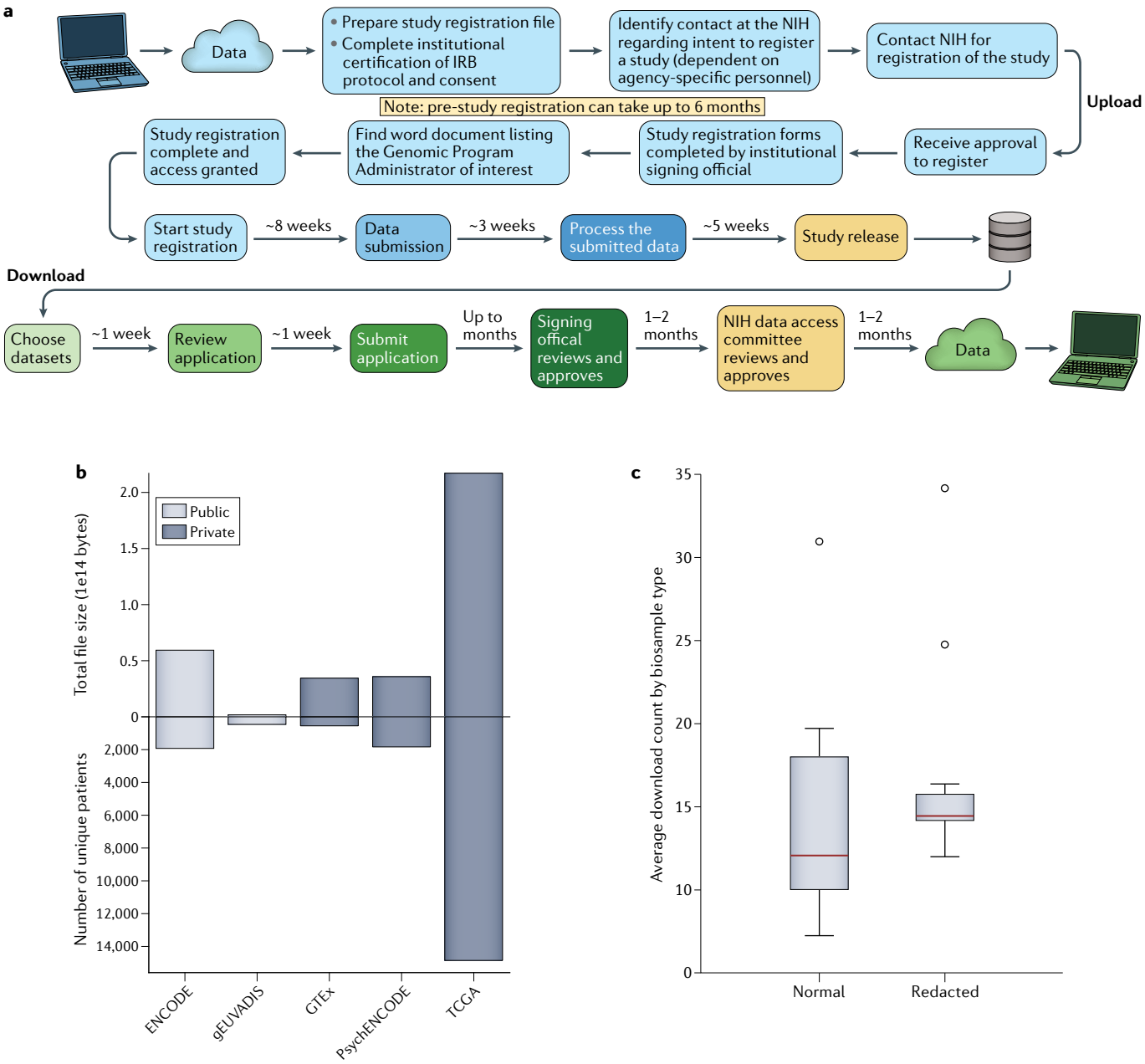


Fig. 3 | **Cryptic private information in functional genomics data. a** | The DNA methylation levels of a genome correlate with a person's chronological age, that is, the time since birth (left graph). The methylation states of specific sets of CpG dinucleotides can be coupled with a mathematical algorithm to determine the 'epigenetic age' of a DNA source, such as cells, tissues or organs, exemplified by tissues A–C (right graph). This estimated age reflects not only chronological age but also the biological age of the DNA source, thus representing a molecular readout of its physiological state. Positive epigenetic age acceleration, seen in tissue A and tissue B, might be associated with age-related diseases. This information could be exploited, for example, by insurance companies. **b** | RNA sequencing (RNA-seq) reads in human-derived samples can originate from exogenous species, such as bacteria or viruses. It was found that 1–2% of RNA-seq reads from tissues originate from the microbiome of the individual[60].


a
Data
• Prepare study registration file
• Complete institutional certification of IRB protocol and consent
Identify contact at the NIH regarding intent to register a study (dependent on agency-specific personnel)
Contact NIH for registration of the study
Upload
Note: pre-study registration can take up to 6 months
Receive approval to register
Study registration forms completed by institutional signing official
Find word document listing the Genomic Program Administrator of interest
Study registration complete and access granted
Start study registration
~8 weeks
Data submission
~3 weeks
Process the submitted data
~5 weeks
Study release
Download
Choose datasets
~1 week
Review application
~1 week
Submit application
Up to months
Signing offical reviews and approves
1–2 months
NIH data access committee reviews and approves
1–2 months
Data
b
Public
Private
Total file size (1e14 bytes)
Number of unique patients
2.0
1.5
1.0
0.5
0
2,000
4,000
6,000
8,000
10,000
12,000
14,000
ENCODE
gEUVADIS
GTEx
PsychENCODE
TCGA
c
Average download count by biosample type
35
30
25
20
15
10
0
Normal
Redacted
d
Reference genome
ACTGCGTAACTGTTTGGTTCACTGTGTACATGCTGATGCAAACCAGTTGCGTACAGTGCAGTGTTGCAACCA
BAM read
ACTGCGTAACGGTTTGGTTCACTGTGT
TGATGCAAA CAGTTGCGTACAGTGCA
Redacted BAM read
ACTGCGTAACTGTTTGGTTCACTGTGT
TGATGCAAACCAGTTGCGTACAGTGC
Add deletion back
Remove from end to keep read length fixed
Reference genome
GTGCGTACATGTGCAAACATGTGCGCTGACATGCAGCATGCAACGTGTCGTGACACGTGCAGTACGGACACGTGCAGT
C
C
BAM read
GTGCGTACATGTGCAAACATGTGCGCT
TGCAACGTGTCGTGACAC TGCAGTAC
Redacted BAM read
GTGCGTACATGTGCAAACATGTGCGCTG
TGCAACGTGTCGTGACACGTGCAGTAC
Remove insertion from the read
Add at the end to keep read length fixed
Remove insertion from the read
Add deletion back

◀ Fig. 4 | **Challenges in accessing private functional genomics data and potential solutions from data sanitization. a** | The steps and approximate amount of time required to upload and download a controlled-access dataset to/from the database of Genotypes and Phenotypes (dbGaP). **b** | The amount of functional genomics data from Encyclopedia of DNA Elements (ENCODE)[12], gEUVADIS[13], Genotype-Tissue Expression project (GTEx)[11], PsychENCODE[145] and The Cancer Genome Atlas (TCGA)[14] that is behind controlled access (dark grey) and that is broadly shared (light grey) on top; the total number of unique individuals in these datasets is shown at the bottom. **c** | Comparison of the download frequency between redacted and normal binary alignment map (BAM) files for histone chromatin immunoprecipitation followed by sequencing (ChIP–seq) experiments on ENCODE (from 29 March 2019 to 31 December 2020). Redacted BAM files are those labelled as 'redacted alignments' or 'redacted unfiltered alignments'. Each data point is the average download frequency of all files for one biosample type. We found that redacted BAMs have a higher number of downloads from the ENCODE data portal than BAMs for comparable datasets. Most likely, users could not find these datasets in the Gene Expression Omnibus (GEO), because these data were in the dbGaP, therefore were downloaded from the ENCODE data portal. **d** | A schematic depiction of reads in normal and redacted BAM files.

of functional genomics data generally falls under practical utility, whereas the utility for a defined function can be categorized as mathematical utility. For example, sharing raw functional genomics data in a privacy-preserving manner provides general practical utility for biomedical research for multiple different use cases. However, if we want to define mathematical utility of the privacy-preserving form of the data, we can often specifically identify a purpose for the data, for example, gene expression quantification or eQTL inference, and define the mathematical utility for this specific purpose. For example, the general purpose of RNA-seq is to quantify gene expression levels; therefore, one can define the mathematical utility of raw RNA-seq data as the accuracy of the resulting gene expression values. The mathematical utility of transcription factor ChIP–seq data can be formalized as the accuracy of the inference of binding peaks.

Redacted BAMs

Open sharing of raw functional genomics data is particularly important for developing analysis methods and discovering novel mechanisms related to the human genome. Privacy-preserving BAMs (pBAMs) were recently proposed as an alternative to the current controlled-access functional genomics data-sharing mechanism[16]. pBAMs were developed to allow the public sharing of read alignments of functional genomics experiments while protecting sensitive information and minimizing the amount of private data that requires special access and storage. This aim is achieved by differentiating the public components of the data, that is, the genetic variants that are often unnecessary for most downstream calculations, and storing the data in small, binarized '.diff' files. This differentiation is performed in a systematic way such that, first, the utility loss due to removal of the private component can be quantified in advance, and secondly, original BAM files can be recovered by combining the '.diff', pBAM and human reference genome sequence files. The privacy of the pBAM is provided by masking the information from BAM attributes that might be indicative of the presence of genetic variants (FIG. 4d).

Studies have shown that the utility loss from pBAM conversion is less than the difference between two biological replicates of a functional genomics sample[16]. The pBAM has also been shown to work with single-cell functional genomics data[16]. pBAMs currently provide support for commonly generated functional genomics data types, including single-cell and bulk RNA-seq, ChIP–seq and ATAC-seq. The pBAM file format is based on the existing file format system (SAM, BAM or CRAM). This allows users to use the existing functional genomics pipelines with pBAMs as input, without exposing sensitive genotype information. Users can also treat pBAM files as BAMs and use existing tools to parse and analyse pBAMs. Download statistics of ENCODE data to determine whether researchers prefer working with these pBAMs when the original BAMs are not accessible clearly reveal community interest in pBAMs (FIG. 4c).

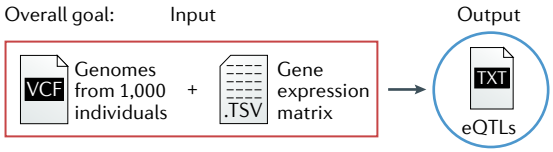
Cryptographic approaches

Traditional cryptography offers a solution for open sharing by allowing private data to be encrypted and uploaded inexpensively on public clouds, where only privileged users with a private decryption key can access the data after downloading. However, this approach still requires large operating costs, as data must be transferred to, decrypted and computed on local systems to ensure privacy. Owing to these challenges, the genomic privacy field is moving towards developing efficient and scalable newer cryptographic solutions that can be used in the genomic ecosystem, that is, allowing computations on private data in the cloud[94,95]. Instead of creating secure ways of sharing data, these methods instead allow confidential computing on the data, which can be an alternative solution to the data-sharing problem.

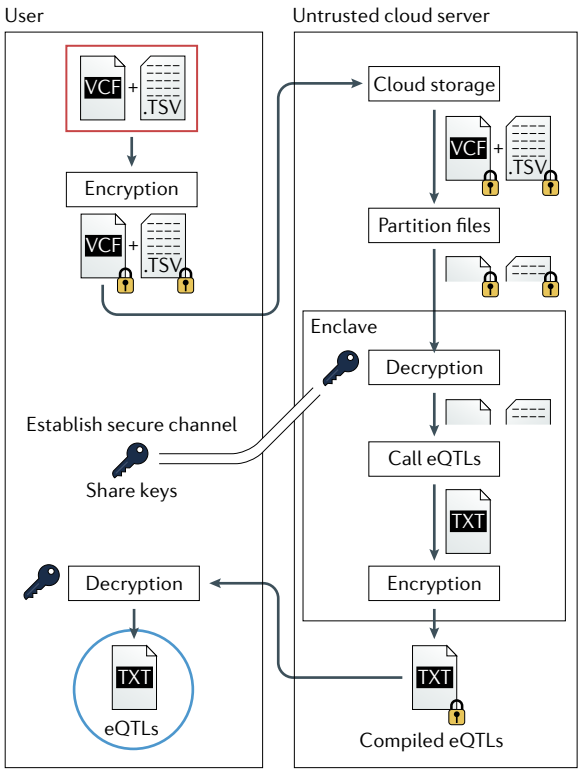
There are a variety of privacy-enhancing techniques developed by the cryptography community, each of which is suitable for a different problem (FIG. 5). For example, homomorphic encryption[96,97] and trusted execution environments, such as software guard extension (SGX)[98], are usually deployed to tackle the problem of the confidentiality of input and output data, whereas secure multiparty computation (SMC)[99] and federated learning[100] can be used when private data are distributed across different sites. (Note, however, that multi-key homomorphic encryption can also be used for private data that are distributed across different sites by allowing the usage of multiple secret keys[101]. Similarly, SGX can also be used for distributed data).

Of note, functional genomics data analysis has not been the focus of recent cryptographic studies. However, the privacy of gene expression data has previously been addressed with various techniques[102–104]. Indeed, the NIH-funded iDASH secure genome analysis competition[94,95,105–108] has been incorporating challenges related to the privacy of gene expression data solved with cryptographic techniques. Moreover, federated techniques for privacy preservation have been applied to differential gene expression analysis when the gene expression data are distributed across different sites[103]. Researchers have proposed a blockchain-based privacy-preserving machine learning model, called swarm learning, to predict disease types using transcriptomics data as features[109]. We can envision similar techniques being used to solve problems related to other types of functional genomics data processing in the future.

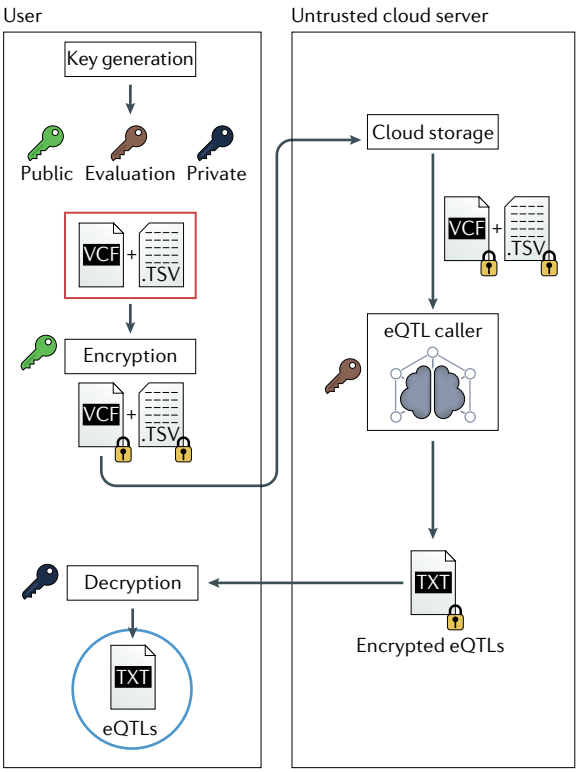
In general, all cryptographic technologies require specific technical knowledge, rendering the use of these technologies difficult for most genome scientists. Misuse of these technologies, whether intentional or not, can lead to privacy leakages. In addition to reducing the performance overhead, simpler software libraries, generic protocols and established guidelines may be necessary before widespread adoption can occur. Below, we provide high-level explanations for these techniques, with specific examples in FIG. 5. Note that these examples are somewhat aspirational based on the recent developments in cryptographic



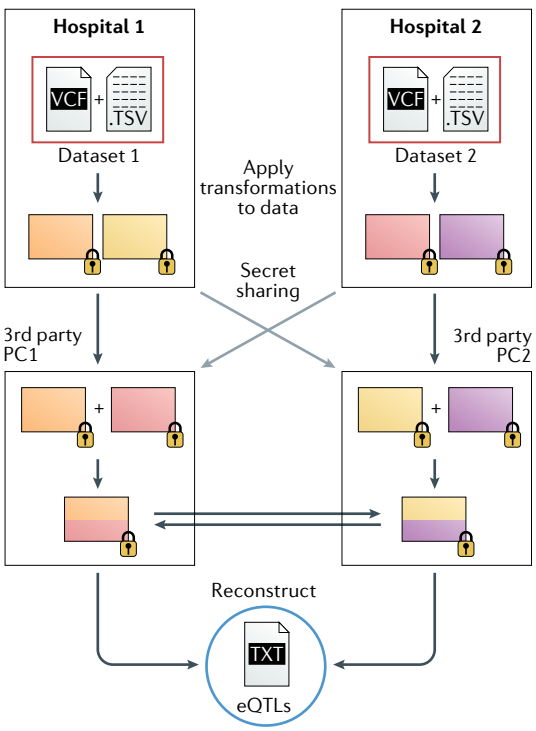
a Trusted execution environments



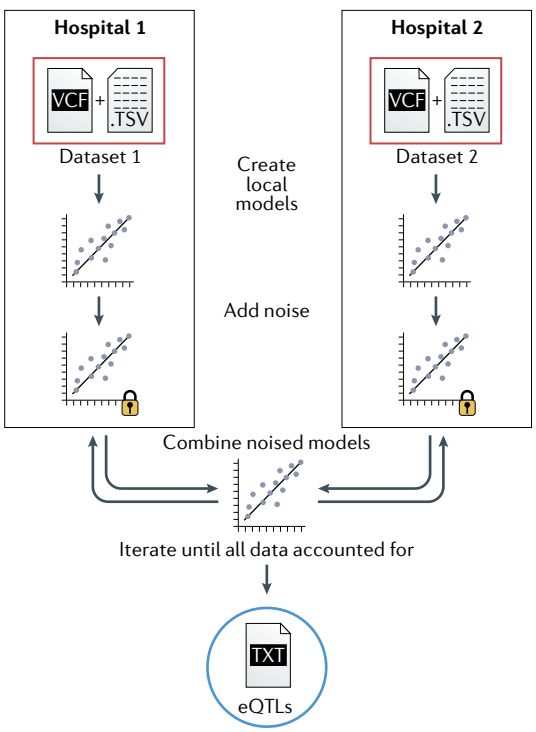
b Homomorphic encryption



c Secure multiparty computation



d Federated learning



◀ Fig. 5 | **Cryptographic techniques to perform confidential eQTL mapping.** This figure outlines a hypothetical scenario for the use of functional genomics data and the various cryptographic techniques that can be used. Note that all solutions discussed here are hypothetical and may not scale with real-world requirements. Let us assume that we want to take single nucleotide polymorphism (SNP) genotypes and gene expression values from a cohort of individuals and map the expression quantitative trait loci (eQTLs)[146–148]. **a** | eQTL mapping using trusted executive environments, such as Intel SGX (software guard extensions). The user-encrypted data are uploaded to a computer server with SGX-enabled hardware. The user application creates an enclave, with which the user can establish a secure channel and exchange keys. In the enclave, the data are decrypted, computed and the resulting eQTLs are re-encrypted to be sent back to the user. Of note, SGX has memory limitations, so one must devise an algorithm that can work on pieces of the data. **b** | eQTL mapping using homomorphic encryption (HE). Similar to the first scenario, eQTL mapping is outsourced to cloud services without decrypting the data. We accomplish this using HE, which allows computation on ciphertexts. The data owner encrypts the data with their public key and uploads the data to the Cloud, where it is computed, and they receive encrypted results, which can then be decrypted with the data owner's private key. **c** | eQTL mapping with secure multiparty computation (SMC). In this scenario, eQTLs are mapped collectively for data residing at two separate institutions. Using SMC, the institutions can share a transformed version of their data with a third-party compute node, where these inputs are combined. The transformations are done such that the real inputs are not visible, but when combined they result in the correct output. Note that the addition of the third server in this figure is optional in SMC. **d** | eQTL mapping with federated learning. Similar to part **c**, this approach is used when data are located at different sites that are not allowed to share data with each other. One can think of an eQTL mapping problem as a classical regression problem, in which the gene expression of each gene is regressed across each SNP genotype in the cohort. A traditional regression model can work iteratively without the need for data sharing; each site can create a mini model using private SNP data and exchange the model parameters back and forth until it converges to a model that represents data at all sites. Here, we added another step, that is, the addition of noise to the exchanged parameters for differential privacy, to prevent possible snooping into training data via model parameters.

solutions for genomic data analysis, such as privacy-preserving GWAS[110–112], and are included to give an idea of how these techniques could be used for functional genomics data.

Homomorphic encryption. Homomorphic encryption is useful when the data need to be protected while in use (FIG. 5a). It enables direct computation on encrypted data within the public cloud[113–116]. The computed results are also encrypted and can be downloaded and decrypted via a private key. In some cases, two or more groups wish to contribute data on the same project and compute the pooled data jointly, while keeping their individual inputs private[117]. There are several classes of homomorphic encryption, designated as partially, somewhat or fully homomorphic, which make various trade-offs in accuracy, performance and the number of operations that they handle. Thus far, homomorphic encryption is only suited for arithmetical tasks that involve addition and multiplication and cannot be extended to more complex algorithms. Additionally, both the storage size of the encrypted plaintexts, known as ciphertexts, and the computation times for homomorphic encryption are several orders of magnitude greater than those for the original plaintexts. Because of these challenges, homomorphic encryption has yet to be adopted for large-scale functional genomics data analysis.

Trusted execution environments. Hardware-based technologies reduce the computational overhead by basing trust in the processing unit. Notably, one example is Intel SGX, which isolates the parts of a user application containing private data and code within a protected unit inside the central processing unit (CPU) called an enclave. The enclave has its own cache that stores private code and encrypts or decrypts data as they are transferred in and out, protecting the data from the non-private parts of the application as well as any other processes (FIG. 5b). In this way, Intel SGX allows secure computation with much lower computational complexity and thus better performance compared with homomorphic encryption. However, the small enclave cache size can be limiting, requiring developers to partition their data into workable chunks[111,118].

Secure multiparty computation. SMC techniques have been developed to solve the issue of exchanging private data between sites without relying on another mutually trusted party[110,119] (FIG. 5c); however, these techniques have some disadvantages. On top of the additional computational overhead, SMC incurs a heavy network overhead for the cost of communication between parties. Thus, this approach is impractical for raw sequencing data from functional genomics experiments, as transfer of and computation on large datasets cannot be performed in a reasonable amount of time.

Federated learning. In machine learning, when access to data that are stored at different sites is prohibited owing to privacy issues, federated learning can be applied. In federated learning, different sites train their own model locally and merge the predictive models at a third site[120] (FIG. 5d). Note that this approach does not involve a cryptographic solution and raises issues related to privacy leakages from the model. Therefore, in privacy-preserving settings, federated learning is often coupled with techniques such as differential privacy[121,122]. Previous work has shown that federated learning is useful for biomedical data privacy[123].

Differential privacy. Differential privacy is useful when summary-level information about a dataset is shared without compromising the privacy of the data points in the dataset[124]. In summary, an algorithm is considered differentially private if the output cannot differentiate the dataset with and without a data point. This is achieved by adding a principled noise to the dataset. One of the uses of differential privacy in functional genomics is as an addition to federated learning. If a machine learning model is developed using, for example, transcriptomics data from multiple sites, then the model parameters can be exchanged in a differentially private manner such that the privacy of the transcriptomic data can be preserved.

Blockchain. Recently, there has been increasing interest in blockchain technology for use in genome privacy and security[125–137] and recently for the use of transcriptomic data in artificial intelligence (AI) models[109]. Blockchain has several key properties, including a decentralized, distributed architecture and cryptographic protocols that yield immutability, that is, data integrity and security[138]. It is important to note that the security and integrity of the data, although related, are different from the privacy of the data. Integrity ensures the reliability and accuracy of the data in its entire life cycle; security protects the data from unwanted actions and unauthorized users; and privacy guarantees the proper handling of the sensitive data. Of note, there are applications of the blockchains that can be beneficial for the security and potentially the privacy of functional genomics data. For example, functional genomics data access by authorized users can be logged and stored in a blockchain[125–127,135]. This allows reliable and immutable auditing of the users accessing sensitive data and early prevention of

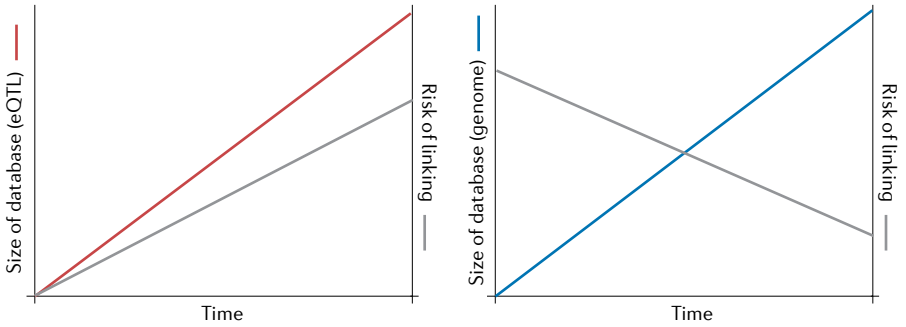


Fig. 6 | **Future perspectives for the privacy of functional genomics data.** As we perform more functional genomics assays, we will infer more relationships between the genotypes and phenotypes of individuals, which will create unique privacy challenges. However, the direction of privacy concerns will depend on the type of data. For example, as we obtain more relationships between single nucleotide polymorphisms (SNPs) and gene expression, a large number of expression quantitative trait loci (eQTLs) can be obtained that can be used to infer genotypes for individuals (left graph). However, if the databases that contain direct or indirect genotypic information from individuals also increase, then the relative overlap between eQTLs and the databases will either stay the same or become smaller (right graph), which might reduce the risk of linking individuals to their identity.

mishandling of sensitive data. Already today, there are multiple personalized medicine start-ups that aim to use blockchain to improve genomic data storage[138]. Similar to other cryptography-based solutions, blockchain comes with its own overhead, and the field is currently in its infancy. As we obtain more personalized multiomics data and as computing resources expand, blockchain technology might be adopted for resolving security and privacy issues related to functional genomics data.

Genomic data-sharing beacons

The Global Alliance for Genomics and Health (GA4GH) launched the Beacon Project to enable genomic and clinical data sharing across federated networks in a privacy-preserving manner[139]. A genomic data-sharing beacon is a framework for public web servers that respond to queries about genomic data collection. With this framework, a query ('Does any individual in this project have a SNP in this location of the genome?') receives a binary response ('yes' or 'no') to protect the privacy of the research participants of the project. Although beacons were originally designed to query specific alleles from a genomic data collection, an extension called MBeacon for functional genomics data, specifically for DNA methylation data, has been implemented[140]. Since the original beacons for genomics data have been shown extensively to be vulnerable to privacy attacks[36,141–143], MBeacon incorporates a novel differential privacy mechanism called SVT[2] that can successfully mitigate privacy attacks without harming the utility of the data.

Conclusions

Advances in sequencing technologies and laboratory techniques have enabled the development of many assays to probe the epigenetic and transcriptomic states of the cell by measuring gene expression or DNA-binding protein levels. Although functional genomics data are generated to elucidate the activities of nucleic acids and proteins and are not obtained for the purpose of genotyping, they can potentially be intrusive because, by virtue of the data generation technology, the data include snippets of DNA sequences containing genetic variants in addition to other metadata such as pre-determinable conditions, traits, sex and race. With increases in different omics techniques, genomic privacy studies have shifted in focus and have shown that new privacy breaches are possible for mining obvious and cryptic information about individuals. Herein, we outline the types of information that can be gleaned from functional genomics data and discuss mitigation strategies grounded in privacy and utility.

What will the privacy of functional genomics data look like in the future? On one hand, the size of genomic databases will increase and this will make it difficult to re-identify individuals with the noisy genotypes inferred from functional genomics datasets (FIG. 6). On the other hand, there are multiple trends that are in contrast to this statement. As more epigenetics data, even in a summarized form such as gene expression values, are collected and shared, more associations between diseases and genomic activities will likely be found, which could lead to the inference of patient phenotypes. That is, with the increase in technologies that characterize the activities of the genome, the risk of inferring characteristic and potentially private information of a single individual will increase. In addition, more functional genomics data from a single individual will lead to better direct genotyping and also greater inference of more cryptic genotypes through eQTLs and chromatin QTLs (FIG. 6). Moreover, longitudinal assaying of functional genomic features may become part of clinical routine in precision medicine or even part of daily life, similar to monitoring simple physiological markers such as heart rate. It has indeed been shown that personal omics profiling can elucidate important molecular and medical phenotypes[144]. This will increase the amount of characterizing data about an individual. However, it also underscores the fact that functional genomics data, especially in summarized form, are transient compared with the genome. Whereas the genome is characteristic and identifying for a person's entire life cycle, functional genomics can characterize a moment in time, and the risk to privacy can deteriorate as time passes from the time biosamples were collected.

Given these trends, we reiterate the importance of proactively developing appropriate data-sharing modes for functional genomics data instead of relying on traditional DNA sequencing data-sharing modes.

Gamze Gürsoy[1,2], *Tianxiao Li*[1], *Susanna Liu*[3,4], *Eric Ni*[1], *Charlotte M. Brannon*[1,2,6] *and Mark B. Gerstein*[1,2,4,5]✉

[1]*Computational Biology and Bioinformatics Program, Yale University, New Haven, CT, USA.*

[2]*Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA.*

[3]*Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT, USA.*

[4]*Statistics and Data Science, Yale University, New Haven, CT, USA.*

[5]*Computer Science, Yale University, New Haven, CT, USA.*

[6]*Present address: Department of Biology, Stanford University, Stanford, CA, USA.*

✉*e-mail: mark@gersteinlab.org*

https://doi.org/10.1038/s41576-021-00428-7
Published online 10 November 2021

1. Hirst, M. & Marra, M. A. Next generation sequencing based approaches to epigenomics. *Brief. Funct. Genomics* **9**, 455–465 (2010).
2. Werner, T. Next generation sequencing in functional genomics. *Brief. Bioinform.* **11**, 499–511 (2010).
3. Bonifer, C. & Cockerill, P. N. Chromatin mechanisms regulating gene expression in health and disease. *Adv. Exp. Med. Biol.* **711**, 12–25 (2011).
4. Byron, S. et al. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

6. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
7. Boyle, A. P. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
8. Buenrostro, J. et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods.* **10**, 1213–1218 (2013).
9. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
10. Gasperskaja, E. & Kučinskas, V. The most common technologies and tools for functional genome analysis. *Acta Med. Litu.* **24**, 1–11 (2017).
11. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
12. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
13. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
14. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
15. Rodriguez-Esteban, R. & Jiang, X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med. Genomics* **10**, 59 (2017).
16. Gürsoy, G. et al. Data sanitization to reduce private information leakage from functional genomics. *Cell* **183**, 905–917.e16 (2020).
17. Harmanci, A. & Gerstein, M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat. Methods* **13**, 251–256 (2016).
18. Gürsoy, G., Lu, N., Wagner, S. & Gerstein, M. Recovering genotypes and phenotypes using allele-specific genes. *Genome Biol.* **22**, 263 (2021).
19. Schadt, E. E., Woo, S. & Hao, K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat. Genet.* **44**, 603–608 (2012).
20. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
21. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
22. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
23. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
24. Rozowsky, J. et al. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.* **27**, 66–75 (2009).
25. Harmanci, A., Rozowsky, J. & Gerstein, M. MUSIC: identification of enriched regions in ChIP-seq experiments using a mappability-corrected multiscale signal processing framework. *Genome Biol.* **15**, 474 (2014).
26. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
27. Zhao, Y. et al. A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics* **20**, 160 (2019).
28. Harmanci, A. & Gerstein, M. Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions. *Nat. Commun.* **9**, 2453 (2018).
29. Mangul, S. et al. ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol.* **19**, 36 (2018).
30. Tierney, B. T. et al. The predictive power of the microbiome exceeds that of genome-wide association studies in the discrimination of complex human disease. Preprint at https://doi.org/10.1101/2019.12.31.891978 (2020).
31. Danko, D. et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* **184**, 3376–3393.e17 (2021).
32. Tovino, S. A. HIPAA compliance. in *The Cambridge Handbook of Compliance* 895–908 (Cambridge University Press, 2021).
33. Rothstein, M. A. Putting the Genetic Information Nondiscrimination Act in context. *Genet. Med.* **10**, 655–656 (2008).
34. Yordanov, A. Nature and ideal steps of the data protection impact assessment under the general data protection regulation. *Eur. Data Prot. Law Rev.* **3**, 486–495 (2017).
35. Greenbaum, D., Harmanci, A. & Gerstein, M. Proposed social and technological solutions to issues of data privacy in personal genomics. In *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering* (IEEE, 2014).
36. Ayoz, K., Ayday, E. & Cicek, A. E. Genome reconstruction attacks against genomic data-sharing beacons. *Proc. Priv. Enh. Technol.* **2021**, 28–48 (2021).
37. Berger, B. & Cho, H. Emerging technologies towards enhancing privacy in genomic data sharing. *Genome Biol.* **20**, 128 (2019).
38. Mittos, A., Malin, B. & De Cristofaro, E. Systematizing genome privacy research: a privacy-enhancing technologies perspective. *Proc. Priv. Enh. Technol.* **2019**, 87–107 (2019).
39. Huang, Z. et al. A privacy-preserving solution for compressed storage and selective retrieval of genomic data. *Genome Res.* **26**, 1687–1696 (2016).
40. Dyke, S. O. M. et al. Epigenome data release: a participant-centered approach to privacy protection. *Genome Biol.* **16**, 142 (2015).
41. He, D. et al. Identifying genetic relatives without compromising privacy. *Genome Res.* **24**, 664–672 (2014).
42. Uhlerop, C., Slavković, A. & Fienberg, S. E. Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confid.* **5**, 137–166 (2013).
43. Romeo Casabona, C. M. Genetic privacy and non-discrimination. *Rev. Derecho Genoma Hum.* **34**, 141–151 (2011).
44. Ducato, R., Perra, S. & Zuddas, C. The legal fate of biobanks between privacy, IPRs and crisis of a firm: a preliminary study on the case of "bio-bankruptcy". *Rev. Derecho Genoma Hum.* **41**, 89–102 (2014).
45. Moniz, H. Privacy and intra-familiy communication of genetic information. *Rev. Derecho Genoma Hum.* **21**, 103–124 (2004).
46. Andrews, L. B. Genetic privacy: from the laboratory to the legislature. *Genome Res.* **5**, 209–213 (1995).
47. Erlich, Y. & Narayanan, A. Routes for breaching and protecting genetic privacy. *Nat. Rev. Genet.* **15**, 409–421 (2014).
48. Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12**, 125 (2011).
49. Rodriguez, L. L., Brooks, L. D., Greenberg, J. H. & Green, E. D. Research ethics. The complexities of genomic identifiability. *Science* **339**, 275–276 (2013).
50. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
51. Lockhart, N. C. et al. Development of a consensus approach for return of pathology incidental findings in the Genotype-Tissue Expression (GTEx) project. *J. Med. Ethics* **44**, 643–645 (2018).
52. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
53. Flynn, M. The culprit's name remains unknown. But he licked a stamp, and now his DNA stands indicted. *Washington Post*, 17 October 2018.
54. Claw, K. G. et al. A framework for enhancing ethical genomic research with Indigenous communities. *Nat. Commun.* **9**, 2957 (2018).
55. Garrison, N. A. et al. Genomic research through an Indigenous lens: understanding the expectations. *Annu. Rev. Genomics Hum. Genet.* **20**, 495–517 (2019).
56. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. Identity inference of genomic data using long-range familial searches. *Science* **362**, 690–694 (2018).
57. Tsosie, K. S., Yracheta, J. M., Kolopenuk, J. A. & Geary, J. We have "gifted" enough: indigenous genomic data sovereignty in precision medicine. *Am. J. Bioeth.* **21**, 72–75 (2021).
58. Fox, K. The illusion of inclusion - the "all of us" research program and indigenous peoples' DNA. *N. Engl. J. Med.* **383**, 411–413 (2020).
59. Rozowsky, J. et al. ExceRpt: a comprehensive analytic platform for extracellular RNA profiling. *Cell Syst.* **8**, 352–357.e3 (2019).
60. All of Us Research Program Investigators. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
61. 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).
62. 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).
63. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321–324 (2013).
64. Sweeney, L. et al. Re-identification risks in HIPAA Safe Harbor Data: a study of data from one environmental health study. *Technol. Sci.* **2017**, 2017082801 (2017).
65. Narayanan, A. & Shmatikov, V. Robust DE-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (IEEE, 2008).
66. Knoppers, B. M. & Beauvais M. J. S. Three decades of genetic privacy: a metaphoric journey. *Hum. Mol. Genet.* **30**, R156–R160 (2021).
67. Bonomi, L., Huang, Y. & Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* **52**, 646–654 (2020).
68. Arellano, A. M., Dai, W., Wang, S., Jiang, X. & Ohno-Machado, L. Privacy policy and technology in biomedical data science. *Annu. Rev. Biomed. Data Sci.* **1**, 115–129 (2018).
69. Wang, S. et al. Big data privacy in biomedical research. *IEEE Trans. Big Data* **6**, 296–308 (2020).
70. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).
71. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
72. Davies, R. W. et al. Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* **53**, 1104–1111 (2021).
73. Chen, S.-F. et al. Genotype imputation and variability in polygenic risk score estimation. *Genome Med.* **12**, 100 (2020).
74. Gürsoy, G., Brannon, C. M., Navarro, F. C. P. & Gerstein, M. "FANCY: fast estimation of privacy risk in functional genomics data". *Bioinformatics* **36**, 5145–5150 (2020).
75. Backes, M. et al. Identifying personal DNA methylation profiles by genotype inference. In *2017 IEEE Symposium on Security and Privacy (SP)* (IEEE, 2017).
76. Philibert, R. A. et al. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. *Clin. Epigenetics* **6**, 28 (2014).
77. Liang, P. & Pardee, A. B. Analysing differential gene expression in cancer. *Nat. Rev. Cancer* **3**, 869–876 (2003).
78. Balgobind, B. V. et al. Evaluation of gene expression signatures predictive of cytogenetic and molecular subtypes of pediatric acute myeloid leukemia. *Haematologica* **96**, 221–230 (2011).
79. Bergman, Y. & Cedar, H. DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* **20**, 274–281 (2013).
80. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
81. Liu, Z. et al. Underlying features of epigenetic aging clocks in vivo and in vitro. *Aging Cell* **19**, e13229 (2020).
82. Kuo, C.-L., Pilling, L. C., Liu, Z., Atkins, J. L. & Levine, M. E. Genetic associations for two biological age measures point to distinct aging phenotypes. *Aging Cell* **20**, e13376 (2021).
83. Leung, D. & Levine, M. Epigenetic signatures of cell states in aging. *Innov. Aging* **4**, 132–132 (2020).
84. Office for Human Research Protections. Genetic Information Nondiscrimination Act (GINA): OHRP Guidance. *U.S. Department of Health & Human Services* (2009).
85. Manor, O. et al. Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11**, 5206 (2020).
86. Franzosa, E. A. et al. Identifying personal microbiomes using metagenomic codes. *Proc. Natl Acad. Sci. USA* **112**, E2930–E2938 (2015).
87. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
88. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
89. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).

90. Tryka, K. A. et al. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res.* **42**, D975–D979 (2014).
91. Fernandez-Orth, D., Lloret-Villas, A. & Rambla de Argila, J. European genome-phenome archive (EGA)-granular solutions for the next 10 years. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (IEEE, 2019).
92. Paltoo, D. N. et al. Data use under the NIH GWAS data sharing policy and future directions. *Nat. Genet.* **46**, 934–938 (2014).
93. Joly, Y., Dyke, S. O. M., Knoppers, B. M. & Pastinen, T. Are data sharing and privacy protection mutually exclusive? *Cell* **167**, 1150–1154 (2016).
94. Wang, X. et al. iDASH secure genome analysis competition 2017. *BMC Med. Genomics* **11**, 85 (2018).
95. Kuo, T.-T. et al. iDASH secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching. *BMC Med. Genomics* **13**, 98 (2020).
96. Rivest, R. L., Adleman, L. & Dertouzos, M. L. *On Data Banks and Privacy Homomorphisms* (Massachusetts Institute of Technology, 1978).
97. Gentry, C. Fully homomorphic encryption using ideal lattices. In *Proceedings of the 41st Annual ACM Symposium on Symposium on Theory of Computing- STOC '09* (ACM Press, 2009).
98. Zheng, W. et al. A survey of Intel SGX and its applications. *Front. Comput. Sci.* **15**, 153808 (2021).
99. Yao, A. C.-C. How to generate and exchange secrets. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)* (IEEE, 1986).
100. Kairouz, P. et al. Advances and open problems in federated learning. *Found. Trends Mach. Learn.* **14**, 1–210 (2021).
101. Chong, K. S., Yap, C. N. & Tew, Z. H. Multi-key homomorphic encryption create new multiple logic gates and arithmetic circuit. In *2020 8th International Symposium on Digital Forensics and Security (ISDFS)* (IEEE, 2020).
102. Xu, J., Cui, B., Shi, R. & Feng, Q. Outsourced privacy-aware task allocation with flexible expressions in crowdsourcing. *Future Gener. Comput. Syst.* **112**, 383–393 (2020).
103. Zolotareva, O. et al. Flimma: a federated and privacy-preserving tool for differential gene expression analysis. Preprint at https://arxiv.org/abs/2010.16403 (2020).
104. Subramanian, S. K. & Duraipandian. Artificial neural network based method for classification of gene expression data of human diseases along with privacy preserving. *Int. J. Comput. Technol.* **4**, 722–730 (2005).
105. Carpov, S. & Tortech, T. Secure top most significant genome variants search: iDASH 2017 competition. *BMC Med. Genomics* **11**, 82 (2018).
106. Yu, F. & Ji, Z. Scalable privacy-preserving data sharing methodology for genome-wide association studies: an application to iDASH healthcare privacy protection challenge. *BMC Med. Inform. Decis. Mak.* **14** (Suppl. 1), S3 (2014).
107. Chen, H. et al. Logistic regression over encrypted data from fully homomorphic encryption. *BMC Med. Genomics* **11**, 81 (2018).
108. Ohno-Machado, L. et al. iDASH: integrating data for analysis, anonymization, and sharing. *J. Am. Med. Inform. Assoc.* **19**, 196–201 (2012).
109. Warnat-Herresthal, S. et al. Swarm learning for decentralized and confidential clinical machine learning. *Nature* **594**, 265–270 (2021).
110. Cho, H., Wu, D. J. & Berger, B. Secure genome-wide association analysis using multiparty computation. *Nat. Biotechnol.* **36**, 547–551 (2018).
111. Kockan, C. et al. Sketching algorithms for genomic data analysis and querying in a secure enclave. *Nat. Methods* **17**, 295–301 (2020).
112. Kim, D. et al. Privacy-preserving approximate GWAS computation based on homomorphic encryption. *BMC Med. Genomics* **13**, 77 (2020).
113. Kim, M. & Lauter, K. Private genome analysis through homomorphic encryption. *BMC Med. Inform. Decis. Mak.* **15** (Suppl. 5), S3 (2015).
114. Sarkar, E. et al. Fast and scalable private genotype imputation using machine learning and partially homomorphic encryption. *IEEE Access* **9**, 93097–93110 (2021).
115. Kim, M. et al. Ultrafast homomorphic encryption models enable secure outsourcing of genotype imputation. *Cell Systems* **12**, 1–13 (2021).
116. Gürsoy, G., Chielle, E., Brannon, C. M., Maniatakos, M. & Gerstein, M. Privacy-preserving genotype imputation with fully homomorphic encryption. Preprint at https://doi.org/10.1101/2020.05.29.124412 (2020).
117. Froelicher, D. et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. Preprint at https://doi.org/10.1101/2021.02.24.432489 (2021).
118. Dokmai, N. et al. Privacy-preserving genotype imputation in a trusted execution environment. *Cell Systems* **12**, 983–993 (2021).
119. Hie, B., Cho, H. & Berger, B. Realizing private and practical pharmacological collaboration. *Science* **362**, 347–350 (2018).
120. Mandl, K. D. et al. The Genomics Research and Innovation Network: creating an interoperable, federated, genomics learning system. *Genet. Med.* **22**, 371–380 (2020).
121. Kim, M., Gunlu, O. & Schaefer, R. F. Federated learning with local differential privacy: Trade-offs between privacy, utility, and communication. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2021).
122. Li, N., Lyu, M., Su, D. & Yang, W. *Differential Privacy: from Theory to Practice* (Morgan & Claypool, 2016).
123. Pfitzner, B., Steckhan, N. & Arnrich, B. Federated learning in a medical context: a systematic literature review. *ACM Trans. Internet Technol.* **21**, 1–31 (2021).
124. Dwork, C. & Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9**, 211–407 (2013).
125. Ozdayi, M. S., Kantarcioglu, M. & Malin, B. Leveraging blockchain for immutable logging and querying across multiple sites. *BMC Med. Genomics* **13**, 82 (2020).
126. Pattengale, N. D. & Hudson, C. M. Decentralized genomics audit logging via permissioned blockchain ledgering. *BMC Med. Genomics* **13**, 102 (2020).
127. Ma, S., Cao, Y. & Xiong, L. Efficient logging and querying for blockchain-based cross-site genomic dataset access audit. *BMC Med. Genomics* **13**, 91 (2020).
128. Kuo, T.-T. The anatomy of a distributed predictive modeling framework: online learning, blockchain network, and consensus algorithm. *JAMIA Open.* **3**, 201–208 (2020).
129. Kuo, T.-T., Gabriel, R. A., Cidambi, K. R. & Ohno-Machado, L. EXpectation Propagation LOgistic REgRession on permissioned blockCHAIN (ExplorerChain): decentralized online healthcare/genomics predictive model learning. *J. Am. Med. Inform. Assoc.* **27**, 747–756 (2020).
130. Kuo, T.-T., Kim, J. & Gabriel, R. A. Privacy-preserving model learning on a blockchain network-of-networks. *J. Am. Med. Inform. Assoc.* **27**, 343–354 (2020).
131. Mackey, T. K. Fit-for-purpose?' — challenges and opportunities for applications of blockchain technology in the future of healthcare. *BMC Med.* **17**, 68 (2019).
132. Kuo, T.-T., Gabriel, R. A. & Ohno-Machado, L. Fair compute loads enabled by blockchain: sharing models by alternating client and server roles. *J. Am. Med. Inform. Assoc.* **26**, 392–403 (2019).
133. Kuo, T.-T., Kim, H.-E. & Ohno-Machado, L. Blockchain distributed ledger technologies for biomedical and health care applications. *J. Am. Med. Inform. Assoc.* **24**, 1211–1220 (2017).
134. Gürsoy, G., Brannon, C. M., Wagner, S. & Gerstein, M. Storing and analyzing a genome on a blockchain. Preprint at https://doi.org/10.1101/2020.03.03.975334 (2020).
135. Gürsoy, G., Bjornson, R., Green, M. E. & Gerstein, M. Using blockchain to log genome dataset access: efficient storage and query. *BMC Med. Genomics* **13**, 78 (2020).
136. Gürsoy, G., Brannon, C. M. & Gerstein, M. Using Ethereum blockchain to store and query pharmacogenomics data via smart contracts. *BMC Med. Genomics* **13**, 74 (2020).
137. Grishin, D. et al. Citizen-centered, auditable, and privacy-preserving population genomics. Preprint at https://doi.org/10.1101/799999 (2019).
138. Ozercan, H. I., Ileri, A. M., Ayday, E. & Alkan, C. Realizing the potential of blockchain technologies in genomics. *Genome Res.* **28**, 1255–1263 (2018).
139. Fiume, M. et al. Federated discovery and sharing of genomic data using beacons. *Nat. Biotechnol.* **37**, 220–224 (2019).
140. Hagestedt, I. et al. MBeacon: privacy-preserving beacons for DNA methylation data. In *Proceedings 2019 Network and Distributed System Security Symposium* (Internet Society, 2019).
141. Shringarpure, S. S. & Bustamante, C. D. Privacy risks from genomic data-sharing beacons. *Am. J. Hum. Genet.* **97**, 631–646 (2015).
142. Raisaro, J. L. et al. Addressing beacon re-identification attacks: quantification and mitigation of privacy risks. *J. Am. Med. Inform. Assoc.* **24**, 799–805 (2017).
143. Bu, D., Wang, X. & Tang, H. Haplotype-based membership inference from summary genomic data. *Bioinformatics* **37**, i161–i168 (2021).
144. Chen, R. et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293–1307 (2012).
145. PsychENCODE Consortium. Revealing the brain's molecular architecture. *Science* **362**, 1262–1263 (2018).
146. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nat. Rev. Genet.* **7**, 862–872 (2006).
147. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. Lond. B* **368**, 20120362 (2013).
148. Michaelson, J. J., Loguercio, S. & Beyer, A. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* **48**, 265–276 (2009).

Acknowledgements
This study was supported by grants from the US National Institutes of Health (R01 HG010749 to M.B.G. and K99 HG010909 to G.G.). This work is also supported by the A.L. Williams Professorship Fund.

Author contributions
All authors researched the literature. G.G., T.L., S.L., E.N. and M.B.G. provided substantial contributions to discussions of the content. G.G., T.L., E.N., C.M.B. and M.B.G. wrote the article. G.G., C.M.B. and M.B.G. reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Genetics thanks Yann Joly, A. Ercüment Çiçek and Xinghua Mindy Shi for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.