Big Data Privacy in Biomedical Research

Shuang Wang, *IEEE Senior Member*, Luca Bonomi, Wenrui Dai, IEEE Member, Feng Chen, Cynthia Cheung, Cinnamon S. Bloss, Samuel Cheng, *IEEE Senior Member,* and Xiaoqian Jiang, *IEEE Member*

Abstract—Biomedical research often involves studying patient data that contain personal information. Inappropriate use of these data might lead to leakage of sensitive information, which can put patient privacy at risk. The problem of preserving patient privacy has received increasing attentions in the era of big data. Many privacy methods have been developed to protect against various attack models. This paper reviews relevant topics in the context of biomedical research. We discuss privacy preserving technologies related to (1) record linkage, (2) synthetic data generation, and (3) genomic data privacy. We also discuss the ethical implications of big data privacy in biomedicine and present challenges in future research directions for improving data privacy in biomedical research.

Index Terms—Biomedical research, data privacy, data security, genome analysis, bioethics.

1 Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act [1] in the U.S. has mandated the adoption of electronic health records (EHRs) to improve the quality of health care, and by January 2015, 83% of office-based physicians had adopted EHRs. The massive adoption of EHR systems allows healthcare providers and researchers to create and collect large-scale phenotypic data from patients with various diseases (e.g., cancers, cardiovascular diseases, etc.). Besides EHR data, advancements in sequencing technology have made human genomic data increasingly affordable and available. The Precision Medicine Initiative, recently announced by the President Obama, will build a national cohort to cover one million Americans with genomic data sequenced. To reach this goal, it will integrate genomic data and EHR data from existing networks and recruit new volunteers. These recent progresses open the door to big data science and have a great potential to speedup biomedical discoveries. On the other hand, the increasing biomedical data, which include a lot of sensitive information about patients, also make the privacy challenge more prominent than ever. These data need to be carefully protected, otherwise, could lead to information disclosure and privacy breach and will negatively impact patients and may have serious implications (e.g., discrimination for employment, insurance, or education [2]).

In the U.S., the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) safeguards the security and privacy of health records. HIPAA provides two different approaches to achieve de-identification: the first, which is seldom exercised, is Expert Determination, where the re-identification risk inherent in the data should be assessed by an expert to be sufficiently low; the second is Safe Harbor, where a list of 18 identifiers need to be removed [3]. In reality, Safe Harbor seems to be preferred to Expert Determination because Safe Harbor is more operational, requiring data disclosure to follow a predefined list [4] to eliminate certain sensitive information to generate de-identified data. While still the dominant approach in practice, there are numerous debates on these HIPAA privacy rules [5]–[8]. Some think that protections from data de-idientification are not sufficient [5]. Current privacy rules do not deal with longitudinal data and transactional data, which can be used to re-identify an individual. Personal genome data are also not covered despite that HIPAA clearly states the protection of biometrics like finger and voice prints. Others contend that privacy safeguards will hamper biomedical research, and that realizing them will impede meaningful biomedical studies that depend on suppressed attributes, e.g., fine-grained geriatric studies with people over 89 in areas which have less than 20,000 residents [9]. Someone also concerns that the efficiencies using computerized health records in studies may be eroded due to the privacy rule [5]. Essentially, any data access policy implicitly involves a trade-off between privacy risks and data usefulness: at one end of the spectrum one can gain unrestrained access to the data as collected (highest risk and highest usability); at the other end of the spectrum we do not disseminate any data (lowest risk and lowest usability) [10]. In reality, most clinical data owners make a compromise by choosing

- *S.W., L. B. W.D., F.C. and X. J. are with Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093. E-mail: {shw070, lbonomi, wed004, f4chen, x1jiang}@ucsd.edu*
- *C.C. and C.S.B. are with Department of Psychiatry, University of California San Diego, La Jolla, CA, 92093. Email: {cyc077, cbloss} @eng.ucsd.edu*
- *S.C is with School of Electrical and Computer Engineering, University of Oklahoma, Tulsa, OK, 74135, Email: samuel.cheng@ou.edu*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2608848, IEEE Transactions on Big Data

one of the following two strategies: (1) altering data to make it difficult to trace information to a particular individual, or (2) restricting the amount of information that is released. An appropriate solution should take into consideration both the context of the application and the possible background knowledge of attackers to strike the right balance.

To set the scope of this paper, we selected a few important and practical topics in biomedical research to discuss related privacy preserving technologies. These topics include: (1) record linkage, (2) distributed data analysis, (3) synthetic data generation, and (4) secure genome analyses. We will focus on both privacy protection technologies for both electronic health records (EHR) and genomic data. The rest of this paper is organized as follows: First, Section 2 introduces the backgrounds of the data privacy problem in terms of ethics and privacy criteria. Second, we review the data privacy preserving methods in Section 3. Section 4 discusses the challenges and future directions. Finally, we draw the conclusion in Section 5.

2 BACKGROUNDS

2.1 Ethics and privacy

The definition of privacy is an individual's right to control access to her/his personal data, such as identifying information, bio-speciman and so on. The term confidentiality usually refers to the protection process of individual privacy. The use of individual's information must fulfill the requirements in the informed consent and the data use agreement. The laws and government policies can be used to ensure obedience and provide guidelines in privacy protection. The Ethical, Legal, and Social Implications (ELSI) studies are essential for biomedical data privacy, which provides a new methodology to biomedical research by recognizing, investigating and tackling the ELSI of studies involving human subjects. The impacts of ELSI studies include, but are not limited to, the risks and benefits associated with a scientific study, the attitudes of research participants, legal and normative aspects, etc. For example, the benefits of maintaining confidentiality include (1) it establishs trust between research participants and researchers; (2) it provides respects to participants; (3) it increases participants' willingness in sharing data for research. The basic principle of research ethics is to ensure that no individual should risk harm due to the participation in a research study. It is also important to assess participant's understanding of the benefit, risk and the research to be conducted. Furthermore, all participants have to be voluntarily involved in research. For instance, the ethical concerns about informed consent in genomic research include, but are not limited to, participants' attitudes toward sharing data for primary research [11]–[14] or secondary use [15], [16], participants' understanding of privacy risk [17], [18] or study details [19], voluntary participation [20], and disclosure of results [21]. Section 4 will discuss some ethical implications of big data privacy.

2.2 Privacy foundation and terminology

2.2.1 Protection of sensitive data dissemination

Differential privacy (DP) has emerged as the de facto standard of privacy, which provides robust privacy guarantee, regardless of the adversary's prior knowledge. DP provides a general solution for privacy protection that could cover the full spectrum of potential information that an attacker may possess.

The foundation of DP is based on the premise that the query result of a "private" dataset should not change drastically with an addition, deletion or modification of a single record. Otherwise, there can be a chance for malicious users to infer the identity or sensitive information of a private record. Note that DP is a rather strong condition since it does not assume the methods a malicious user could use in extracting that information. DP does not try to limit what records can or cannot exist in a dataset. Actually, the definition is blind to (independent of) the actual data inside the dataset. To ensure that the statistics of the query results do not change drastically with the change of a single record, the query outcome is typically distorted by the addition of noise.

ϵ-Differential Privacy [22]: An algorithm $K(\cdot)$ is ε-differentially private if and only if for any output S,

$$\Pr(K(D_1) \in S) \leq e^{\varepsilon}\Pr(K(D_2) \in S),$$

where D_1 and D_2 are any two datasets where they differ by at most one record. The parameter ε is called the privacy budget, where a smaller value corresponds to a stronger protection. To achieve ε-DP, the two most common approaches are Laplace mechanism [22] and exponential mechanism [23].

2.2.2 Protection of sensitive data computation

In this subsection, we will introduce several technologies that can be used to safeguard the computation process of sensitive data.

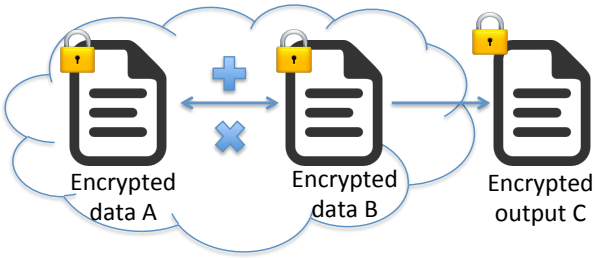


Figure 1: Illustrative diagram of Homomorpic encryption, which allows computation over encrypted data with encrypted output.

Homomorphic Encryption (HME) [24]. As illustrated in Figure 1, HME allows the direct computation over encrypted data using certain arithmetic operations (i.e., multiplication and addition), where the returned output is also encrypted under the same encryption key. Three are different types of HME cryptosystems: (1) partially HME [25] allows a single type of HME operation (e.g., either addition or multiplication), (2) somewhat HME

[26] enables both HME operations with a limited number of iterations and additional computational costs, (3) fully HME [27] supports unlimited number of both operations with considerable computational costs. As HME is often computationally demanding, section 3.3.1 discussed how to leverage task-oriented optimizations to maximize the efficiency [28] for secure genome analyses.

Secure multiparty computation (SMC): SMC is a set of cryptographic protocols that enable two or more parties to jointly compute functions over their private inputs without leaking their sensitive information. *Garbled Circuit* [29] is widely used to achieve secure two-party computation [30], [31]. In a garbled circuit, the inputs of each gate will be mapped to garbled values, and the truth table of each gate will be encrypted by these garbled inputs. In other words, the inputs of the garbled circuit are the garbled values of the binary presentation of the original inputs. The garbled circuit and garbled value mapping tables are created by one party (i.e., circuit generator (CG)), and is evaluated by the other party (i.e., circuit evaluator (CE)). To evaluate the garbled circuit, the CE also needs to receive the garbled inputs from CG. Due to the encrypted nature, CE cannot infer the originals of the garbled inputs from the CG and therefore the data privacy is protected. The mathematical proofs of the security can be found in the article of the Yao's protocol [32] and it is widely adopted. CE securely converts its original inputs into garbled values based on CG's garbled value mapping tables by using oblivious transfer (OT) protocol, by which the generator does not know which garbled value the evaluator gets. By using the garbled inputs from both CG and CE, the CE will evaluate the gate one by one until obtaining the garbled output. Since the CE has no knowledge about the corresponding values of the garbled output, it will send the garbled output to the CG to decrypt the results. CG cannot learn CE's inputs due to the adoption of OT protocol. Garbled circuit is used to secure the computation over two parties which are "semi-honest" (i.e., each party exactly follows the common procedures, but may be curious to gain the inputs from the other party. *Secret sharing (SS)* [33] is another SMC framework, which supports the secure computation over more than two parties. The traditional definition of secret sharing scheme includes a secret owner and n parties. Then, an SS scheme enables the secret owner to distribute shares of the secret among n parties subject to: (1) shares from a given number of parties can be used to recover the secret, and (2) share in each party alone cannot reveal any partial information on the secret. Many cryptographic protocols (secure multi-party computation of general functions) are developed based on SS scheme (see [34] for a survey of results). Typically, secret sharing primitives provide secure protocols for computing simple arithmetic functions across multiple parties. These protocols combined yield to efficient and secure solution for computing a wide range of functions. Specifically, for SMC protocols, secret sharing schemes are used to enable the multiple parties to jointly compute a common function without leaking each individual's input to other parties. Due to their strong secure guarantees and distributed nature, these techniques are cornerstones for many biomedical applications. For example, in privacy-preserving record linkage, SMC protocols are applied to securely compute the similarity of the medical records without disclosing any sensitive information about the patients' input.

3 PRIVACY PRESERVING METHODS

In this section, we will review privacy-preserving methods under the following four categories, including EHR patient linkage, synthetic data generation and genomic data privacy.

3.1 EHR Patient Linkage

In the biomedical domain, information systems have been undergone significant improvement making it possible to collect, store, and process huge amount of data. Despite efforts in managing these information systems, the healthcare data often result to be fragmented, redundant, prone to errors, and heterogeneous, making the task of finding useful information among such data very challenging. A crucial step in biomedical research is the *record linkage process*, also referred to as duplicate detection [35] or entity resolution [36], which consists in identifying records that refer to the same real-world entity across two or more data sources.

3.1.1 The privacy-preserving record linkage process

In the traditional record linkage setting, two data owners with datasets $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively, transform their records into vectors, where each component corresponds to a record's attribute or field. Then, in this representation, the matching records can be identified by looking at the components in their vectors. In general, the final goal is assign each of all the possible pair (r_A, r_B), with $r_A \in \boldsymbol{A}$ and $r_B \in \boldsymbol{B}$, to one of these three disjoint sets: $\boldsymbol{M}$, $\boldsymbol{U}$, and $\boldsymbol{C}$ representing respectively the set of *matches*, the *non-matches*, and those records that require human intervention in order to classify. As an example, consider two hospitals that are interested in finding their common patients. Each patient's record can be represented as a vector, where each component represents a patient's record attribute, such as: name, zipcode and date of birth. Therefore, by looking at similarity between each component the original records can be classified as a matching records, non-matching or if they require further investigation.

A major challenge in record linkage is often the lack of unique identifiers for the entities, which may compromise the *linkage quality*. In fact, even though attributes such as: Social Security Number (SSN) and/or demographics information are often used for administrative purposes [37] within the database hospital. Their use as common identifiers with external parties in the linkage process is prohibited or limited by federal regulations that protect patient's sensitive information. As a consequence, the lack of such a universal patient identifier forces record linkage solutions to rely on other attrib-

utes which are not so sensitive. Examples of attributes commonly used are names, dates of birth, zipcode, etc. In such a setting, the integration process could be extremely challenging due to ambiguity of the attribute values (e.g. same name or address), errors (e.g. typos in the attribute values) or missing values. To overcome this challenge, many existing techniques use approximate and probabilistic algorithms to perform the linkage and minimize the number of records in the set $\boldsymbol{C}$ that requires human intervention to be further matched.

Given the sensitivity of medical data, the privacy is a major concern in data integration process. Since the fundamental goal of record linkage is exactly identifying the entity represented by the records being linked, researches proposed privacy-preserving record linkage (PPRL) solutions that aim to privately compute the set of matching records between the parties without revealing any information about the non-linked records. Modern PPRL solutions provide privacy guarantees such as: k-anonymity [38], encryption/hashing [39], and DP [40]. In addition to the privacy model adopted by PPRL solutions, we can identify two typical protocol settings, in which the linkage is performed: *two-party* and *three-party*. In the former, the data owners directly determine the matching pairs through a series of encrypted messages. Solutions that adopt this setting typically employ SMC to privately compute the set of matching records, an example of which is the techniques in [41], [42]. On the other hand, in the three-party setting, the linkage is performed through a third-party, which is typically in charge of matching the records between the two original data owners. While in some cases, the third-party is a trusted institution (e.g. National Center for Health Statistics in USA [43]), modern PPRL solutions tend to consider the third-party untrusted. In both these settings, privacy threats may come from one of the parties in the protocol. A popular threat model considered in PPRL assumes a "semi-honest" model.

3.1.2 Secure transformation using bloom filters

Secure transformations typically proceed by decomposing the original string attributes of the records into short subsequences called *n-grams*. These in combination with *hash functions* are used to map the original records into a *Bloom filter* [44]. The bloom filter (BF) is a probabilistic data structure in which the original string is represented with a bit array. Between each gram in the original string and each bit in the array there exists a probabilistic map determined by a set of hash functions. Initially in this transformation, all the bits in the array are set to 0, then only the bits selected by hashing the grams of the original string are turned to 1. Traditionally, this method yields to a field-level BF (FBF), where the final similarity between records BF is measured adopting set-to-set distance metrics; for example, the Dice coefficient is commonly used. While PPRL solutions based on this encoding provide high utility results, the FBF representation of the original string attribute could still disclose some information to adversary [45] (e.g. frequency based attacks). To improve the security of PPRL techniques based on BF, Durham et al. [46] proposed a new approach, which combines multiple FBFs of each record attribute into one composite BF per record. The intuition behind such a strategy relies on the fact that the composite nature of the structure introduces additional dependency between the bit values making difficult for the adversary to exploit the frequency information of the grams to identify the original records from the single bit values.

3.1.3 Secure computation using scalar product

A common framework for privately computing the similarity between the vector representations of the records is based on secure scalar product protocols. An example of PPRL approach using this framework is the work proposed by Yakout et al. [41] where a two-party PPRL approach is proposed. The overall solution combines secure transformation and SMC. Initially, the protocol follows the steps in [47], where the original string records are embedded into vectors. In this phase, a collection of reference sets, called *base*, each containing random strings is used to embed the original records into a vector space. Specifically, each data owner creates a vector of distances for each record, where the *i*-th component stores the minimum Edit distance between the record and the strings in the *i*-th set of the base. Then, each vector is represented by the first coefficient of its Discrete Fourier Transform (DFT). Using this map, it can be shown that if two vectors are close prior to mapping on the complex plane, then their mapped values will be closed in the complex plane as well. Thus, the similarity between the original records is computed evaluating the scalar product in the complex plane. Authors in [41] demonstrated that this technique provides high linkage utility (no false negative and minimal false positive) and has small execution time compare to the secure transformation in [47].

3.1.4 Hybrid solution with privacy-preserving blocking

In the overall record linkage process, the matching step may become a major performance bottleneck when the pairwise record comparisons are performed naively. In general, it can be noticed that evaluating all records pairwise similarity is not necessary. In fact, by just looking at differences at attribute level, it is often sufficient to identify the small portion of the overall pairs that likely represent possible matching records and discard those that likely are non-matching. To take advantage of this, current PPRL solutions employ effective blocking/indexing algorithms to reduce the unnecessary pairwise comparisons, while preserving the real matching pairs. While the blocking step significantly reduces the overhead in the matching step it may compromise the final utility if the indexing/blocking approach is not carefully design. In PPRL solution the blocking step is generally followed by a SMC phase where the records within the same block are securely matched. For example, a recent approach based on differentially private blocking in a three-party setting has been proposed by Kuzu et al. [48]. This privacy-preserving blocking ap-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2608848, IEEE Transactions on Big Data

proach greatly reduces the cost of the SMC in evaluating the pairwise similarity and the overall solution provides accurate results in linking dataset containing personal identifiers.

3.1.5 PPRL Comparison

We provide a brief comparison of the PPRL methods presented in the previous sections with respect to five dimensions grouped into three categories: *privacy guarantee*, *scalability*, and *linkage quality*, as illustrated in Table 1. First of all, we notice that except for the method proposed by Yakout et al. [41], all the other PPRL solutions require the presence of an additional party (i.e. third party) to perform the linkage. Second, from the scalability perspective, even though all these techniques have quadratic computational complexity in the number of records, the solution proposed by Yakout et al. [41] and Kuzu et al. [48] implement indexing techniques that reduce the overall computational cost. Finally, regarding the linkage quality, all the described solutions with the exception of the approach proposed by Durham et al. [46] allow to match each record's attribute separately. Overall, each PPRL technique described in the previous sections is suitable for different biomedical applications depending on the specific setting. Specifically, for linkage tasks in a Big Data scenario, it is preferable to use PPRL solutions that take advantage of indexing techniques for better scalability. Another important consideration relates to the presence of a third party in the protocol. While the third party may be beneficial in reducing the complexity in matching the records, it may introduce additional privacy risks; for example, exposing the overall protocol to attacks based on collusion.

Table 1: PPRL methods comparison

PPRL Methods	Privacy		Scalability		Linkage
	Protocols	# Parties	Indexing	Comp.	Matching
Durham [46]	Hashing Bloom Filter	3	None	Quadratic	Record
Yakout [41]	Embedding SMC	2	Sorted	Quadratic	Attribute
Scannapieco [47]	Embedding	3	None	Quadratic	Attribute
Kuzu [48]	DP & SMC	3	Spatial	Quadratic	Attribute

3.2 EHR Data Anonymization

EHR data anonymization is important to protect sensitive private information while supporting population-level data analysis and research. For example, users can build generalized linear models on anonymized data that preserve the first and second order statistics (i.e., arithmetic mean, and variance) of the original data. The challenge is to release an anonymous dataset that maximumly preserves the data utility. In this section, we will review state-of-the-art methods for sanitizing individual records satisfying DP assuming attackers have arbitrary background knowledge. Interested readers can refer to an early survey on privacy-preserving data publishing models satisfying other privacy criteria [49] (covering k-anonymity [50], l-diversity [51], and t-closeness [52]). Because most DP algorithms are developed for statistical databases, we will focus on their applications to structured relational data, which are very common in biomedical data analysis. There are other interesting studies related to unstructured clinical notes, but we are not discussing them here due to space limitation.

3.2.1 Partition-based methods

The main idea of partition-based methods is to divide data into partitions and perturb them in each partition.

Dwork et al. [22] proposed a straightforward method by introducing independent Laplacian perturbation to each cell count of the original histogram. This method is often used as a baseline by later studies and it does not scale well to high dimensionality or large domain data (often leading to large perturbation error to small counts in each histogram bin). Xiao et al. [53] proposed DPCube algorithm to partition data based on domains and to generate a fine-grained and equal-width cell histogram, which approximates the original distribution. Then, they generate noisy synthetic data like [22], followed by the KD-tree-based multidiemsnional partition scheme to obtain close-to-uniform partitions. Noises are added to the original counts of these partitions to generate the final synthetic data. This approach saves privacy budgets as it perturbs elements in a close-to-uniform partition together rather than separately. Cormode et al. [54] developed a similar technique for differentially private spatial decompositions (PSDs), which partitions the space into smaller regions and reports statistics of observations that reside within each region. They proposed two partition mechanisms based on data independent tree structure (Quadtree) and data dependent tree structure (KD-tree). The former structure does not need protection and data owner only needs to add noises to the partition. The latter structure needs to be constructed privately and the paper suggested four different methods to compute a private median (i.e., smooth sensitivity, Exponential mechanism, cell-based method, and noisy mean).

Ace et al. [55] proposed a P-HPartition framework that is based on a divisible hierarchical clustering method. The idea is that there are similar counts for histogram bins within the same cluster and therefore we can use the mean values or cluster centers to obtain a good approximation. Therefore, the release of the noisy cluster centers can achieve a better utility, as it has a smaller sensitivity comparing to individual samples. It leverages a binary tree structure to partition the data in two steps: (1) pick a partition that has not been split more than certain times and (2) use Exponential mechanism to determine how to split the selected partition. Mohammed et al. [56] proposed a synthetic data generation method (DiffGen) based on the decision tree. This method is designed for data with several predictor attributes and a class attribute. DiffGen first generalizes the predictor attributes and uses a decision tree to iteratively divide the original data into equivalent groups, in which all records share the same attribute values. The division process leverages a taxonomy tree (generated from the private data) to specify different attributes using Exponential mechanism.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2608848, IEEE Transactions on Big Data

Finally, DiffGen adds noise to the counts of each leaf node and releases the synthetic data. Xu et al. [57] developed two partition-based models called NoiseFirst and StructFirst, which differ in the order of histogram structure computation and noise injection. The first approach uses [22] to obtain noisy cells followed by a post-optimization step to merge adjacent cells in terms of noisy counts in partitions. The second approach is to first construct the optimal histogram by choosing boundaries using Exponential mechanism followed by adding Laplacian noise on the average of these partitions. These methods are mostly applicable to low dimensional data for their high computational complexity.

3.2.2 Transformation-based methods

Another category of approaches is to transform the data into some compact representation (e.g., bases), on which perturbation is applied and synthetic data are generated accordingly.

Barak et al. [58] improved a differentially private Fourial perturbation algorithm (FPA) using a two-step approach. They first calculate a DP frequency matrix and transformed it to the frequency domain using Fourier transformation, where Laplacian noises are added to the Fourier coefficients. Then, a linear programming technique is employed to reconstruct a non-negative frequency matrix, from which synthetic data can be sampled. The computation is very challenging in this model, as it needs to solve a linear program with the number of variables equal to the items in the frequency matrix. Ace et al. [55] developed an enhanced discrete Fourier perturbation algorithm (EFPA) for generating differentially private histograms. They improved the performance of FPA by developing a more precise score function for the Exponential mechanism to remove high frequency components and utilizing the intrinsic correlation among the Fourier coefficients of real-valued histograms.

Jiang et al. [59] developed another technique based on principal component analysis (PCA) and linear discriminant analysis (LDA). The idea is to perturb the first and second order statistics (mean and co-variance) before eigendecomposition. Then, they used noisy eigenvectors to reconstruct a synthetic matrix. The problem with this approach is that it does not scale well to high dimension when too much noise is added to the co-variance matrix. Xiao et al. [60] developed a Privelet method by applying a wavelet transformation (an invertible linear function) on the original histogram and adding polylogarithmic noises to the transformed data. Similar to the PCA approach [59], Privelet has three steps: transformation, perturbation, and reconstruction to generate synthetic data.

3.2.3 Statistical model-based methods

These methods are to build statistical models from private data and to sample points from the model to be released.

Machanavajjhala et al. [61] proposed a differentially private data synthesizer by fitting the private data to a multinomial dirichlet model and sampled data from this distribution. This is a typical parametric model, which has a finite number of parameters. Cormode et al. [62] developed several sampling and filtering methods to create a compact histogram summary of sparse data under DP. Their simplest model is a high-pass filter that attenuates signals with frequencies lower than the cutoff threshold. In this method, every non-zero contingency table cell is independently perturbed and released while only k cells with zero counts are selected uniformly at random using a binomial distribution. For these selected cells, the values are drawn from a special distribution to ensure that the final outputs have the same distribution as the baseline approach [22]. Two additional advanced methods with different priority sampling algorithms were also introduced.

Li et al. [63] proposed a DPCopula model that generates high dimensional and large domain synthetic data. The model generates differentially private copula, which has a multivariate probability distribution function with uniform marginal probability distribution for each of its variables, to model the empirical histogram of all dimensions. This is considered as a semi-parametric model because it estimates marginal distributions using non-parametric technique and models the joint dependence of each dimension by the correlation matrix in a parametric manner. Ji et al. [64] developed another synthetic data generation method based on importance weighting. The method is based on computing weights that make an existing dataset (reference data), for which there are no confidentiality issues, analogous to the dataset that must be kept private. In a non-parametric manner, the method computes differentially private weights between samples in the private data to samples in the reference data. Synethic data can be easily generated by sampling the reference data using the noisy weights.

The methods above are applicable to anonymize structured tabular data (such as demographics) but EHR also involve set-valued data (multiple diagnosis at the same encounter) and sequence data (a set of consequtive lab results). Directly applying previous methods to these data will be infeasible either due to the computation complexity or due to the introduction of too much noise. Recent studies shed lights to these problems and we just review some important techniques. Chen et al. developed a novel method for publishing set-valued data via differential privacy [65]. This approach is similar to [56], which uses context-free taxonomy trees to conduct probabilistic top-down partitioning for data release. Zeng et al. proposed differentially private frequent itemset mining by striking the right balance between approximation error and perturbation error (due to differential privacy) [66]. For sequence data, Cheng et al. developed a differentially private maximal frequent sequence mining algorithm [67], which relies on a multi-stage approach to estimate expected frequency followed by candidate extraction/validation. Xu et al. developed a similar model for frequent sequence mining [68] where noisy local support of candidate sequences in the sample databases is used to evaluate these potentially frequent sequences. They also

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2608848, IEEE Transactions on Big Data

Table 2: Comparison of different protection methods for secure genomic data analysis

Methods	Analytic tasks	Performance	Notes	Secure protocol
Kim et al. [95]	Minor allele frequence	26.31 s	Tested with 610 SNPs and 400 subjects based on the report in iDASH 2015 genome privacy protection competition	HME
Lu et al. [98]		112.32 s		
Kim et al. [95]	Chi-sequare	27.11 s		
Lu et al. [98]		112.32 s		
Zhang et al. [103]		13.00 s		SMC
Kim et al. [95]	Hamming distance	80.03 s	Two sequences with length of 10,000	HME
Zhang et al. [103]		604 s		SMC
Cheon et al. [94]	Exact Edit distance	25.43 s	Two sequences with length of 8	HME
Zhang et al. [92]		0.30 s		
Kim et al. [95]	Approximate Edit distance	181.92 s	Two sequences with length of 10,000	
Wang et al. [87]	Exact logistic regerssion	46.49 s	Average computing time per SNP	
Lauter et al. [93]	Pearson Goodness-of-Fit	1.36 s	1000 genotype and phenotype data	
	EM algorithm	6.85 s		
	CATT	3.63 s		
	LD computation	0.74 s		

developed novel shrinkage algorithm to enforce the length constraint.

3.3 Genomic data Privacy

With the dramatic reductions in sequencing costs [69], it is becoming more affordable to obtain high-throughput human genomic data for healthcare and biomedical research. Massive collection of genomic data [70] enables the developments of effective diagnosis methods and the discovery of new treatments. Owing to these potentials, many initiatives have been established. For example, the Precision Medicine Initiative aims to expand the health records and linked genomic data to one million participants in the United States [71]. However, the collection of large-scale genomic data also raises several concerns about data confidentiality and privacy. It is because genomic data even with the removal of explicit identifiers (e.g., name) can still leak sensitive information about individuals, such as identity [72]–[74], predisposition to diseases [75]–[77], appearance [78], [79], etc. Furthermore, genomic data may disclose information about people beyond the individual from whom the data were sampled. Therefore the privacy risk may propagate to blood relatives of the individual [80]. As genomic data are irrevocable once they are made public, the privacy risks may increase over time with the accumulation of knowledge about human genetics and the development of advanced attacking methods. Because of these potential privacy risks [75], the NIH has taken most aggregated results off the public domain [81]. Many efforts have also been made in protecting genome privacy, but with different focuses on genome privacy, including legal [2], ethics [82]–[84] and technologies [85]–[87]. In this section, we will discuss the genome privacy protection methods in the following two categories including, secure genomic data computation and privacy-presrving genomic data dissemination.

3.3.1 Secure genomic data computation

The recent NIH data sharing policy change allows users to store and analyze human genomic data using cloud-computing services, which address some of the concerns about efficiently handling large scale genomic data. But on the other hand, the privacy challenge becomes more prominent with cloud computing as owners lose the full control of the data any more. It becomes more complicated as copies of data can be stored in a distributed file system or automatically backed up by the cloud service provider. Without necessary protection, it is risky to use the cloud for handling human genomic data, of which information leakage can lead to re-identification [74], [76], [79], [88] and might negatively impact patients. The NIH Security Best Practices for Controlled-Access Data Subject to the NIH Genomic data Sharing (GDS) Policy [89] also states that researchers and their institutions are accountable for ensuring the confidentiality of human genomic data, instead of the cloud service provider.

Existing studies of secure outsourcing on genomic data analysis are mainly based on the HME technology. The first fully HME scheme that supports both addition and multiplication operations over encrypted data was proposed by Gentry in [90]. Brakerski et al. [26], [27] improved performance of HME by using learning with errors (LWE). Recently, HME has been extended to support different operations and analytic tasks, such as division [91], comparison [92] and sampling [87] operations, as well as regression model evaluation [93] and learning [87], sequence comparison [94], and genome wide association studies [95]. Lauter et al. [93] studied several secure statistical algorithms for genetic association studies based on HME, which include pearson Goodness-of-Fit and Chi-squared tests, Expectation Maximization (EM) algorithm for estimating haplotype frequencies, Cochran-Armitage Test for Trend (CATT) for allele-disease association as well as computation of linkage disequilibrium (LD). Through task-oriented optimizations (e.g., customized data encoding scheme) to achieve 80 bits of security, Lauter et

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TBDATA.2016.2608848, IEEE Transactions on Big Data

al. [93] showed that all above analysis can be finished between 0.19s and 6.85s to handle 1000 genotype and phenotype data. The problem of HME-based the integer comparison is studied by Togan et al. [96]. Cheon et al. [94] extended the HME-based integer comparison algorithm and developed a framework for secure outsourcing of Edit distance calculation, which employed a greedy algorithm to compute the upper bound of exact Edit distance. Zhang et al. [92] improved Cheon's method by combining path-finding algorithm and integer comparison, which enables exact homomorphic Edit distance computation. However, HME can only handle small-scale exact Edit distance computation for sequence length (< 10). For the concerns of scalability using HEM, Kim et al. [95] showed that HME can support efficient secure approximate Edit distance computation between two sequences with length of 10,000. Furthermore, Graepel et al. [97] and Naehrig et al. [28] also demonstrated that certain machine learning algorithms can be implemented using HME. Recently, Wang et al. [87] proposed a framework for homomorphic computation of exact logistic regression for rare disease study in GWAS. Zhang et al. [91], Kim et al. [95] and Lu et al. [98] also developed HME-based chi-squared statistic computation.

Although HME-based secure outsourcing schemes are promising for computing in public clouds, HME is usually computational intensive and requires huge storage overhead. For certain task-specific applications, may efficient solutions have been studied. For example, Ayday et al. [86] developed a privacy-preserving mechanism, which enables a medical unit to privately retrieve of patients' short reads. Based on honey encryption, the GenoGuard framework [99] provides strong protection for genomic data against brute-force attacks. The problem of privacy-preserving disease susceptibility testing was studied in [100]. Verle et al. [101] developed a secure exact logistic regression algorithm, which allows multiple parties to collaborate via SMC. Kantarcioglu et al. [102] proposed a secure framework to query and share genomic sequences, where different parties can compute a function jointly over their private inputs without leaking sensitive information. Recently, many SMC-based secure genomic data analysis frameworks have been proposed in [103], [104]. Table 2 provided a comparison of different protection methods for secure genomic data analysis.

3.3.2 Privacy-preserving genomic data dissemination

For the protection of genomic data dissemination, Malin et al. [105] proposed a generalized lattice method based on k-anonymity [38] to anonymize genomic sequences, which ensure that an attacker cannot differentiate one genetic sequence from $k-1$ other entries in a dataset. Loukides et al. [106] developed an anonymization model to protect the association between individual's phenotypic and genotypic data in GWAS. By adopting DP, Yu et al. [107] developed a privacy-presreving logistic regression model for detecting disease association in GWAS databases. Recently, Wang et al. [108] proposed a top-down specialization-based DP framework for privacy-preserving genomic data dissemination. In [109], Johnathon et al. proposed a mechanism to conduct differentially private chi-square test statistics on genomic data. Uhler et al. [110] studied another solution for chi-squared test under differential privacy protection, which allows to release the top M most significant SNPs in GWAS. In addition, Yu et al. [111], [112] improved the Uhler's methodology with better utility and privacy tradeoffs as well as formal proofs. Zhao et al. [113] also developed a synthetic genomic data generation technique leveraging linkage disequilibrium, which is a feature reduction technique that

Table 3: Summary of different protection methods using parallelization techniques

Methods	Application scenarios	Framework	Protection mechanism
McSherry [115]	Data analysis	PING	Outcome after computation (DP)
Tran et. al.[119]	Distributed computation	MapReduce	Access control Outcome after computation (DP)
Xiao et. al. [120]	Accountability test		Accountability
Roy *et al.* [121]	MapReduce computation		Access control Outcome after computation (DP)
Han *et al.* [122]	Top-k query		Outcome after computation (DP)
Chen *et al.* [123]	Feature selection		Computation process (Encryption)
Zhang *et al.* [124] [125]	Data anonymization		
Zhang *et al.* [126]	Hybrid cloud computation		
Santos *et al.* [127]	Cloud computation		
Chen et al. [128]	Distributed computation		Computation process (HME)
Kamara et al. [129]	Function evaluation		
Xu *et al.* [130]	DNA read-mapping		Computation process (FPGA)
Chen *et al.* [131]			Computation process (Encryption)
Raisaro *et al.* [132]	Replication/Fine-mapping association studies		Access control Computation process (HME)
Zhao *et al.* [133]	Genomic signature search		Computation process (Encryption)

takes advantage the fact that the strong correlation exists between SNPs, a unique feature of the genome, to save privacy budget in generating high dimensional genomic data.

3.4 Privacy-Preserving Parallelization Techniques

In order to handle big biomedical data, parallelization techniques have been considered in privacy and security models for scaling up computation and storage [114]. As shown in Table 3, McSherry [115] developed a Privacy Integrated Queries (PING) platform to satisfy differential privacy in privacy-preserving data analysis. The PING platform assembled privacy theory, language design and implementation to provide tools for both analysts and providers. MapReduce [116] is a programming model that efficiently support parallel computation and processing for large scale datasets with a large cluster of machines (nodes). For real-world tasks, it commonly consists of two procedures: *Map* procedure to arrange data by mapping input key/value pairs to intermediate ones and *Reduce* procedure to summarize these intermediate pairs. However, the MapReduce framework does not consider security issues in its computation model [117]. The security and privacy challenges include providing accountability and access control for users and protecting privacy in computing model [118]. Tran and Sato [119] utilized role-based access control (RBAC) and type enforcement (TE) to prevent malicious MapReduce framework from leaking sensitive data. Accountable MapReduce [120] performed Accountability Test (A-Test) to detect malicious machines (nodes) with optimized utilization resource based on assigned auditors. In the sense of protection mechanism, differential privacy (DP) and homomorphic encryption (HME) are considered for MapReduce to protect the output and process of aggregated computation, respectively. Airavat [121] integrated mandatory access control and differential privacy with the MapReduce framework to guarantee security and privacy of the output of aggregated computation. Han *et al.* [122] developed the *DiffMR* algorithm that satisfied differential privacy by using exponential mechanism to process top-k query. Iterative selection with reject rates based on score function is performed to guarantee accuracy of the query on large-scale datasets. Chen *et al.* [123] proposed a privacy-preserving distributed algorithm based on differential privacy for feature selection, where MapReduce framework was adopted for Gini index-based methods over large scale datasets. To efficiently preserve privacy, Zhang *et al.* [124] proposed a heuristic algorithm for data anonymization to identify partially encrypted intermediate datasets with their generation relationship. Constrained optimization is made to constrain privacy disclosure based on an upper bound of quantified joint privacy leakage of multiple datasets. Using the MapReduce framework, a scalable two-phase top-down specialization (TDS) [125] approach was developed to anonymize large-scale datasets. Zhang *et al.* [126] proposed a privacy-aware data-intensive computation method for hybrid cloud computing. The proposed *Sedic* system modified MapReduce to partition the computation tasks by moving sanitized data to the public cloud and maintaining sensitive data in the private cloud. Santos *et al.* [127] presented an Excalibur system to introduce policy-sealed data for intermediate data, where a policy is defined by the customer for encryption and decryption. To enable computation over encrypted intermediate data, Chen and Huang [128] presented a modified framework that employed fully homomorphic encryption (FHE) in the MapReduce framework. In [129], Parallel homomorphic encryption (PHE) was developed to securely outsource computation on massive dataset to a cluster of machines. PHE enabled MapReduce operations like element testing and keyword search on encrypted dataset to support general evaluation for parallelizable functions.

In genomic research, privacy-preserving mechanisms have been developed for applications over large datasets, including DNS read-mapping, association study and genomic signature search. For DNA read-mapping, FPGA was incorporated to exploit its inherent tamper resistant properties to preserve the privacy of intermediate data in the MapReduce framework [130]. Chen *et al.* [131] developed a secure and scalable read mapping technique for the hybrid clouds. Raisaro *et al.* [132] proposed a parallelizable and flexible privacy-preserving architecture for replication and fine-mapping genetic association studies over encrypted genotypes and phenotypes. A MapReduce implementation was developed to support parallelization for encrypted large-scale dataset including proteomic, transcriptomic and metabolomics data. Zhao *et al.* [133] presented a site-wise encryption approach implemented with Hadoop framework to support secure searching of genomic signatures from whole human genome sequences.

4 Challenges and Future Directions of Big Data Privacy

Concers of ethical implecations in healtchcare data: Specifically, one issue central to the discussion of privacy ethics and policy is that of accountability. While biomedical research for academic purposes is subject to institutional review, uses of information gleaned from secondary data mining processes often exists outside the jurisdiction of conventional ethics boards [134]. For example, there is a lack of clarity as to who is ultimately responsible for ensuring adequate protections from the consequences of re-identification, as well as empirical research delineating the nature and impact of such consequences, if any. Another issue to consider is that of informed consent. The protections traditionally afforded to participants and researchers by virtue of informed consent from a legal standpoint are quickly becoming untenable given this new data ecosystem. This suggests that a critical component of informed consent going forward will be increased transparency regarding the limits of data anonymity and unknowable future risks in the context of

ever-changing policies of data storage and information sharing.

Concerns of temporal information in EHR linkage: In EHR linkage, the same patient records may change over time; therefore, the temporal information has to be taken into account for designing effective linkage solutions. While in the biomedical community record linkage mostly relies on a static model (i.e. no use of temporal information), in the database community recently few works have been proposed to linkage temporal records [135], [136]. For example, in the pioneering work of Li et al. [135], the temporal record linkage is performed in three major phases: similarity measure, temporal model construction, and temporal clustering. Typically, the similarity between two records are measured using attribute value similarities. Nevertheless, if attribute values evolve over time, they may become less reliable indicators for record matching. Thus a temporal model is typically constructed to learn how entities evolve over time, which determines the weight of each attribute in the matching step. Temporal clustering is used to process records in increasing temporal order aiming to identify the entity temporal evolution. The effectiveness of these approaches has been demonstrated on several temporal databases. However, future work is needed to extend these solutions in the biomedical setting.

Concerns of computational and storage efficiency: One may resort to the parallel computation in terms of task-level parallelization (TLP) or data-level parallelization (DLP). TLP can be used in the scenario with multiple study tasks, where different computing nodes can handle different tasks simultaneously. TLP is easy to implement, but it is a challenge to properly schedule different tasks to achieve the maximum throughput. DLP provides a fine-grained parallelism, which could achieve better performance than TLP with careful instruction-level optimizations. In DLP, multiple processors can concurrently perform the same instruction on multiple data points. The design challenges in DLP are to efficiently organize data and instructions to minimize the latency during context switch (e.g., switching from one condition to another in an algorithm with branches). For storage efficiency optimization, the ciphertext in HME or SMC mainly consists of numbers represented by '0'-'9'. Thus, each digit only requires 4 bits to represent. As a result, the size of encrypted data can be reduced by half with substitution-based compression schemes [137], [138]. The authors in [87] demonstrated an entropy-based compression scheme to achieve more than 55% reduction.

Community efforts on genome privacy: Despite of the exciting progress, many emerging problems also need to be addressed for genome privacy, which motivates several community efforts. For example, genome privacy workshops originated from the computer security community, such as GenoPri'14 [139], GenoPri'15 [140] and PrivaGen'15 [141]. Additionally, the iDASH Genome Privacy and Security Challenges [142], [143] have attracted leading researchers from data privacy, security, genetics, bioethics, and law. The competitions were reported by GenomeWeb [144], [145] and Nature News [146]. The outcomes of the competitions identified several limitations in the current genome privacy protection studies. First, genomic data protection using perturbation-based protection methods [108], [112] often present too much noise, where practical genomic applications may not be able to trustworthly rely on these noise outputs. Second, cryptography-based protection methods [91], [104] for secure collaboration or outsourcing currently only support limited genomic computations due to their complexity. Third, the ethical implications of these protection methods are not yet clear. Therefore, further investigations of genome privacy are important and necessary, which motivates researchers to develop advanced genome privacy protection technologies and to investigate their ethical implications.

5 Conclusion

We discussed multiple aspects of big data privacy in the context of biomedical research. The "big" part of data privacy is because healthcare data often contain large scale clinical and genomic data, which are big in size and large in dimension. There are some unique challenges and off-the-shelf tools have difficulties in handling them. For example, the scalability concerns about fully homomorphic encryption and secure multiparty computing algorithms to deal with whole genome sequencing (WGS) data. There are also challenges in safeguarding the outcomes of computation on high dimensional genomic data, which can easily exhaust the budget if not allocated carefully. We reviewed state-of-the-art privacy-preserving technologies for record linkage, synthetic data generation, and genomic data analysis. Despite of exciting progresses, there are many problems and emerging challenges need to be addressed and we believe good solutions to mitigate privacy risks in biomedical research require a joint effort from different communities (e.g, computer security, ELSI, biomedicine, genomics, etc.).

Acknowledgment

Please address all correspondence concerning this manuscript to S. W. (Email: shw070@ucsd.edu). This work was supported in part by grants from, NHGRI (R00HG008175, R01HG008753), NLM (R00LM011392, R21LM012060,), and NHLBI (U54HL108460).

References

[1] Health Information Technology for Economic and Clinical Health. 2010.
[2] L. Slaughter, Genetic Information Nondiscrimination Act of 2008, vol. 50. HeinOnline, 2008, p. 41.
[3] "Health Insurance Portability and Accountability Act (HIPAA)." [Online]. Available: http://www.hhs.gov/ocr/hipaa.
[4] D. Lafky, "The Safe Harbor method of de-identification: An empirical test," Fourth Natl. HIPAA Summit West, 2010.

[5] D. McGraw, "Why the HIPAA privacy rules would not adequately protect personal health records: Center for Democracy and Technology (CDT) brief," 2008. [Online]. Available: http://www.cdt.org/brief/why-hipaa-privacy-rules-would-not-adequately-protect-personal-health-records. [Accessed: 20-Sep-2015].

[6] K. Benitez and B. Malin, "Evaluating re-identification risks with respect to the HIPAA privacy rule," J. Am. Med. Informatics Assoc., vol. 17, no. 2, pp. 169–177, 2010.

[7] P. Kwok, M. Davern, E. Hair, and D. Lafky, "Harder than you think: a case study of re-identification risk of HIPAA-compliant records," Chicago NORC Univ. Chicago. Abstr., vol. 302255, 2011.

[8] L. Sweeney, "Data sharing under HIPAA: 12 years later," in Workshop on the HIPAA Privacy Rule's De-Identification Standard, 2010.

[9] S. J. Nass, L. A. Levit, and L. O. Gostin, Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. The National Academies Press, 2009.

[10] X. Jiang, A. D. Sarwate, and L. Ohno-Machado, "Privacy technology to support data sharing for comparative effectiveness research: a systematic review.," Med. Care, vol. 51, no. 8 Suppl 3, pp. S58–65, Aug. 2013.

[11] B. A. Bernhardt, E. S. Tambor, G. Fraser, L. S. Wissow, and G. Geller, "Parents' and children's attitudes toward the enrollment of minors in genetic susceptibility research: implications for informed consent," Am. J. Med. Genet. Part A, vol. 116, no. 4, pp. 315–323, 2003.

[12] A. L. McGuire, J. M. Oliver, M. J. Slashinski, J. L. Graves, T. Wang, P. A. Kelly, W. Fisher, C. C. Lau, J. Goss, M. Okcu, and others, "To share or not to share: a randomized trial of consent for data sharing in genome research," Genet. Med., vol. 13, no. 11, pp. 948–955, 2011.

[13] J. M. Oliver, M. J. Slashinski, T. Wang, P. A. Kelly, S. G. Hilsenbeck, and A. L. McGuire, "Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives," Public Health Genomics, vol. 15, no. 2, pp. 106–114, 2012.

[14] L. Jamal, J. C. Sapp, K. Lewis, T. Yanes, F. M. Facio, L. G. Biesecker, and B. B. Biesecker, "Research participants' attitudes towards the confidentiality of genomic sequence information," Eur. J. Hum. Genet., vol. 22, no. 8, pp. 964–968, 2014.

[15] D. Levy, G. L. Splansky, N. K. Strand, L. D. Atwood, E. J. Benjamin, S. Blease, L. A. Cupples, R. B. D'Agostino, C. S. Fox, M. Kelly-Hayes, and others, "Consent for genetic research in the Framingham heart study," Am. J. Med. Genet. Part A, vol. 152, no. 5, pp. 1250–1256, 2010.

[16] P. Boddington, L. Curren, J. Kaye, N. Kanellopoulou, K. Melham, H. Gowans, and N. Hawkins, "Consent forms in genomics: the difference between law and practice," Eur. J. Health Law, vol. 18, no. 5, pp. 491–519, 2011.

[17] J. Peppercorn, I. Shapira, T. Deshields, D. Kroetz, P. Friedman, P. Spears, D. E. Collyar, L. N. Shulman, L. Dressler, and M. M. Bertagnolli, "Ethical aspects of participation in the Database of Genotypes and Phenotypes of the National Center for Biotechnology Information," Cancer, vol. 118, no. 20, pp. 5060–5068, 2012.

[18] F. D'Abramo, J. Schildmann, and J. Vollmann, "Research participants' perceptions and views on consent for biobank research: a review of empirical data and ethical analysis," BMC Med. Ethics, vol. 16, no. 1, p. 60, 2015.

[19] A. AL-RIYAMI, D. Jaju, S. Jaju, and H. J. Silverman, "The adequacy of informed consent forms in genetic research in Oman: a pilot study," Dev. World Bioeth., vol. 11, no. 2, pp. 57–62, 2011.

[20] P. A. Marshall, C. A. Adebamowo, A. A. Adeyemo, T. O. Ogundiran, M. Vekich, T. Strenski, J. Zhou, T. E. Prewitt, R. S. Cooper, and C. N. Rotimi, "Voluntary participation and informed consent to international genetic research," Am. J. Public Health, vol. 96, no. 11, pp. 1989–1995, 2006.

[21] J. K. Williams, S. Daack-Hirsch, M. Driessnack, N. Downing, L. Shinkunas, D. Brandt, and C. Simon, "Researcher and institutional review board chair perspectives on incidental findings in genomic research," Genet. Test. Mol. Biomarkers, vol. 16, no. 6, pp. 508–513, 2012.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," Theory Cryptogr., vol. 3876, no. 1, pp. 265–284, 2006.

[23] F. McSherry and K. Talwar, "Mechanism Design via Differential Privacy," in 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), 2007, pp. 94–103.

[24] R. Frederick, "Core Concept: Homomorphic encryption," Proc. Natl. Acad. Sci., vol. 112, no. 28, pp. 8515–8516, 2015.

[25] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in Advances in cryptology—EUROCRYPT'99, 1999, pp. 223–238.

[26] Z. Brakerski, C. Gentry, and V. Vaikuntanathan, "(Leveled) fully homomorphic encryption without bootstrapping," in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12, 2012, pp. 309–325.

[27] Z. Brakerski and V. Vaikuntanathan, "Efficient fully homomorphic encryption from (standard) LWE," SIAM J. Comput., vol. 43, no. 2, pp. 831–871, 2011.

[28] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?," in Proceedings of the 3rd ACM workshop on Cloud computing security workshop - CCSW '11, 2011, p. 113.

[29] Y. Huang, D. Evans, J. Katz, and L. Malka, "Faster Secure Two-Party Computation Using Garbled Circuits.," USENIX Secur. Symp., no. August, pp. 8–12, 2011.

[30] F. Chen, S. Wang, N. Mohammed, S. Cheng, and X. Jiang, "PRECISE:PRivacy-prEserving Cloud-assisted quality Improvement Service in hEalthcare.," IEEE Int. Conf. Syst. Biol. [proceedings]. IEEE Int. Conf. Syst. Biol., vol. 2014, pp. 176–183, Oct. 2014.

[31] F. Chen, N. Mohammed, S. Wang, W. He, S. Cheng, and X. Jiang, "Cloud-Assisted Distributed Private Data Sharing," in The ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 2015.

[32] A. C. Yao, "Protocols for secure computations," in 23rd Annual Symposium on Foundations of Computer Science (sfcs 1982), 1982, pp. 160–164.

[33] S. G. Choi, K.-W. Hwang, J. Katz, T. Malkin, and D. Rubenstein, "Secure multi-party computation of boolean circuits with applications to privacy in on-line marketplaces," in Topics in Cryptology--CT-RSA 2012, Springer, 2012, pp. 416–432.

[34] A. Beimel, "Secret-sharing schemes: a survey," in Coding and cryptology, Springer, 2011, pp. 11–46.

[35] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE Trans. Knowl. Data Eng., vol. 19, no. 1, pp. 1–16, 2007.

[36] L. Getoor and A. Machanavajjhala, "Entity Resolution: Theory, Practice & Open Challenges," Proc. VLDB Endow., vol. 5, no. 12, pp. 2018–2019, Aug. 2012.

[37] M. Kantarcioglu, W. Jiang, and B. Malin, "A Privacy-Preserving Framework for Integrating Person-Specific Databases," in Privacy in Statistical Databases, vol. 5262, J. Domingo-Ferrer and Y. Saygın, Eds. Springer Berlin Heidelberg, 2008, pp. 298–314.

[38] L. Sweeney, "k-anonymity: A model for protecting privacy," Int. J. Uncertainty, Fuzziness Knowledge-Based Syst., vol. 10, no. 05, pp. 557–570, 2002.

[39] B. Schneier, Applied Cryptography. John Wiley & Sons, 1995.

[40] C. Dwork, "Differential privacy," Int. Colloq. Autom. Lang. Program., vol. 4052, no. d, pp. 1–12, 2006.

[41] M. Yakout, M. J. Atallah, and A. Elmagarmid, "Efficient and Practical Approach for Private Record Linkage," J. Data Inf. Qual., vol. 3, no. 3, pp. 5:1–5:28, Aug. 2012.

[42] A. Al-Lawati, D. Lee, and P. McDaniel, "Blocking-aware private record linkage," in Proceedings of the 2nd international workshop on Information quality in information systems, 2005, pp. 59–68.

[43] H.-C. Kum, A. Krishnamurthy, A. Machanavajjhala, M. K. Reiter, and S. C. Ahalt, "Privacy preserving interactive record linkage (PPIRL)," JAMIA, vol. 21, no. 2, pp. 212–220, 2014.

[44] B. H. Bloom, "Space/Time Trade-offs in Hash Coding with Allowable Errors," Commun. ACM, vol. 13, no. 7, pp. 422–426, Jul. 1970.

[45] M. Kuzu, M. Kantarcioglu, E. Durham, and B. Malin, "A

Constraint Satisfaction Cryptanalysis of Bloom Filters in Private Record Linkage," in Privacy Enhancing Technologies, vol. 6794, S. Fischer-Hübner and N. Hopper, Eds. Springer Berlin Heidelberg, 2011, pp. 226–245.

[46] E. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, B. Malin, and others, "Composite Bloom filters for secure record linkage," Knowl. Data Eng. IEEE Trans., vol. 26, no. 12, pp. 2956–2968, 2014.

[47] M. Scannapieco, I. Figotin, E. Bertino, and A. K. Elmagarmid, "Privacy Preserving Schema and Data Matching," in Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, 2007, pp. 653–664.

[48] M. Kuzu, M. Kantarcioglu, A. Inan, E. Bertino, E. Durham, and B. Malin, "Efficient Privacy-aware Record Integration," in Proceedings of the 16th International Conference on Extending Database Technology, 2013, pp. 167–178.

[49] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput. Surv., vol. 42, no. 4, p. 14, 2010.

[50] L. Sweeney and others, "k anonymity: A model for protecting privacy," Int. J. Uncertain. Fuzziness Knowl. Based Syst., vol. 10, no. 5, pp. 557–570, 2002.

[51] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: privacy beyond k-anonymity," in Proceedings of the 22nd International Conference on Data Engineering, 2006, pp. 1–12.

[52] N. Li, T. Li, and S. Venkatasubramanian, "t Closeness : Privacy Beyond k-Anonymity and l-Diversity," in Data Engineering, IEEE 23rd International Conference on, 2007, no. 2, pp. 106–115.

[53] Y. Xiao, L. Xiong, L. Fan, S. Goryczka, and H. Li, "DPCube: Differentially Private Histogram Release through Multidimensional Partitioning," Trans. Data Priv., vol. 7, no. 3, pp. 195–222, 2014.

[54] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, "Differentially Private Spatial Decompositions," in 2012 IEEE 28th International Conference on Data Engineering, 2012, pp. 20–31.

[55] G. Acs, C. Castelluccia, and R. Chen, "Differentially Private Histogram Publishing through Lossy Compression," in 2012 IEEE 12th International Conference on Data Mining, 2012, pp. 1–10.

[56] N. Mohammed, R. Chen, B. C. M. M. B. C. M. Fung, P. S. Yu, and S. Y. Philip, "Differentially private data release for data mining," Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min., vol. 18, no. 40, pp. 493–501, 2011.

[57] J. Xu, Z. Zhang, X. Xiao, Y. Yang, and G. Yu, "Differentially Private Histogram Publication," in 2012 IEEE 28th International Conference on Data Engineering, 2012, pp. 32–43.

[58] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar, "Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release," in Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS), 2007, pp. 273–282.

[59] X. Jiang, Z. Ji, S. Wang, N. Mohammed, S. Cheng, and L. Ohno-Machado, "Differential-Private Data Publishing Through Component Analysis," Trans. Data Priv., vol. 6, no. 1, pp. 19–34, 2013.

[60] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," in Proceedings of the 26th IEEE International Conference on Data Engineering (ICDE), 2010, pp. 225–236.

[61] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: Theory meets Practice on the Map," in 2008 IEEE 24th International Conference on Data Engineering, 2008, pp. 277–286.

[62] G. Cormode, C. Procopiuc, D. Srivastava, and T. T. L. Tran, "Differentially private summaries for sparse data," in Proceedings of the 15th International Conference on Database Theory - ICDT '12, 2012, p. 299.

[63] H. Li, L. Xiong, and X. Jiang, "Differentially Private Synthesization of Multi-Dimensional Data using Copula Functions," in 17th International Conference on Extending Database Technology (EDBT 2014), 2014.

[64] Z. Ji and C. Elkan, "Differential privacy based on importance weighting.," Mach. Learn., vol. 93, no. 1, pp. 163–183, Oct. 2013.

[65] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong, "Publishing set-valued data via differential privacy," Proc. VLDB Endow., vol. 4, no. 2, pp. 1087–1098, 2011.

[66] C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," Proc. VLDB Endow., vol. 6, no. 1, pp. 25–36, Nov. 2012.

[67] X. Cheng, S. Su, S. Xu, P. Tang, and Z. Li, "Differentially private maximal frequent sequence mining," Comput. Secur., vol. 55, no. C, pp. 175–192, Nov. 2015.

[68] S. Xu, S. Su, X. Cheng, Z. Li, and L. Xiong, "Differentially Private Frequent Sequence Mining via Sampling-based Candidate Pruning.," Proc. Int. Conf. Data Eng., vol. 2015, pp. 1035–1046, Apr. 2015.

[69] "DNA sequencing costs." [Online]. Available: http://www.genome.gov/sequencingcosts/. [Accessed: 04-Oct-2015].

[70] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big Data: Astronomical or Genomical?," PLoS Biol, vol. 13, no. 7, p. e1002195, 2015.

[71] "Obama Precision Medicine Initiative Aims to Collect One Million Genomes," MIT Technol. Rev., 2015.

[72] Z. Lin, A. B. Owen, R. B. Altman, M. R. Anderlik, M. A. Rothstein, P. Sankar, W. H. Li, L. A. Sadler, L. Carey, L. Mitnik, H. D. Cash, L. C. R. J. Willenborg, T. D. Waal, T. E. Klein, R. Chadwick, L. Frank, M. A. Austin, and V. Barbour, "Genetics. Genomic research and human subject privacy.," Science, vol. 305, no. 5681, p. 183, Jul. 2004.

[73] L. Sweeney, A. Abu, and J. Winn, "Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment)," Computers and Society, arXiv, Apr. 2013.

[74] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, Y. Erlich, B. Sykes, C. Irven, T. E. King, S. J. Ballereau, K. E. Schürer, M. A. Jobling, B. McEvoy, D. G. Bradley, T. E. King, M. A. Jobling, T. E. King, M. A. Jobling, A. Motluk, J. E. Lunshof, R. Chadwick, D. B. Vorhaus, G. M. Church, J. Gitschier, L. L. Rodriguez, L. D. Brooks, J. H. Greenberg, E. D. Green, M. Gymrek, D. Golan, S. Rosset, Y. Erlich, N. Leat, L. Ehrenreich, M. Benjeddou, K. Cloete, S. Davison, S. K. Lim, Y. Xue, E. J. Parkin, C. Tyler-Smith, S. Levy, S. M. Prescott, J. M. Lalouel, M. Leppert, Z. Lin, A. B. Owen, R. B. Altman, F. R. Bieber, C. H. Brenner, D. Lazer, N. Homer, K. B. Jacobs, H. K. Im, E. R. Gamazon, D. L. Nicolae, N. J. Cox, D. W. Craig, E. E. Schadt, S. Woo, K. Hao, A. L. McGuire, and R. A. Gibbs, "Identifying personal genomes by surname inference.," Science, vol. 339, no. 6117, pp. 321–4, Jan. 2013.

[75] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays.," PLoS Genet., vol. 4, no. 8, p. e1000167, Aug. 2008.

[76] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, "Learning your identity and disease from research papers," in Proceedings of the 16th ACM conference on Computer and communications security - CCS '09, 2009, pp. 534–44.

[77] S. S. Shringarpure and C. D. Bustamante, "Privacy leaks from genomic data-sharing beacons," Am. J. Hum. Genet., vol. 97, pp. 631–646, 2015.

[78] S. Walsh, F. Liu, K. N. Ballantyne, M. van Oven, O. Lao, and M. Kayser, "IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information," Forensic Sci. Int. Genet., vol. 5, no. 3, pp. 170–180, 2011.

[79] P. Claes, D. K. Liberton, K. Daniels, K. M. Rosana, E. E. Quillen, L. N. Pearson, B. McEvoy, M. Bauchet, A. A. Zaidi, W. Yao, and others, "Modeling 3D facial shape from DNA," PLoS Genet, vol. 10, no. 3, p. e1004224, 2014.

[80] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the lacks family: Quantification of kin genomic privacy," in Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, 2013, pp. 1141–1152.

[81] NIH, "Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)," 2007. [Online]. Available: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html. [Accessed: 01-Jul-2015].

[82] A. L. McGuire, T. Caulfield, and M. K. Cho, "Research ethics and the challenge of whole-genome sequencing.," Nat. Rev. Genet., vol. 9, no. 2, pp. 152–6, Feb. 2008.

[83] T. Caulfield, A. L. McGuire, M. Cho, J. A. Buchanan, M. M. Burgess, U. Danilczyk, C. M. Diaz, K. Fryer-Edwards, S. K. Green, M. A. Hodosh, E. T. Juengst, J. Kaye, L. Kedes, B. M. Knoppers, T. Lemmens, E. M. Meslin, J. Murphy, R. L. Nussbaum, M. Otlowski, D. Pullman, P. N. Ray, J. Sugarman, and M. Timmons, "Research ethics recommendations for whole-genome research: consensus statement.," PLoS Biol., vol. 6, no. 3, p. e73, Mar. 2008.

[84] W. J. Dondorp and G. M. W. R. de Wert, "The 'thousand-dollar genome': an ethical exploration," Eur. J. Hum. Genet., vol. 21, pp. S6–S26, 2013.

[85] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool.," Nat. Genet., vol. 41, no. 9, pp. 965–7, Sep. 2009.

[86] E. Ayday, J. L. Raisaro, U. Hengartner, A. Molyneaux, and J.-P. Hubaux, "Privacy-Preserving Processing of Raw Genomic Data," Data Priv. Manag. Auton. Spontaneous Secur., vol. 8247, pp. 133–147, 2014.

[87] S. Wang, Y. Zhang, W. Dai, K. Lauter, M. Kim, Y. Tang, H. Xiong, and X. Jiang, "HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS.," Bioinformatics, vol. 32, no. 2, pp. 211–8, Jan. 2016.

[88] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy.," Nat. Rev. Genet., vol. 15, no. 6, pp. 409–21, Jun. 2014.

[89] "NOT-OD-15-086: Notice for Use of Cloud Computing Services for Storage and Analysis of Controlled-Access Data Subject to the NIH Genomic Data Sharing (GDS) Policy." [Online]. Available: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-086.html. [Accessed: 10-Apr-2015].

[90] C. Gentry, "Fully homomorphic encryption using ideal lattices," in Proceedings of the 41st annual ACM symposium on Symposium on theory of computing - STOC '09, 2009, pp. 169–178.

[91] Y. Zhang, W. Dai, X. Jiang, H. Xiong, and S. Wang, "FORESEE: Fully Outsourced secuRe gEnome Study basEd on homomorphic Encryption.," BMC Med. Inform. Decis. Mak., vol. 15 Suppl 5, no. Suppl 5, p. S5, Dec. 2015.

[92] Y. Zhang, W. Dai, S. Wang, M. Kim, K. Lauter, J. Sakuma, H. Xiong, and X. Jiang, "SECRET: Secure Edit-distance Computation over homomoRphic Encrypted daTa," in 5th Annual Translational Bioinformatics Conference (TBC), 2015.

[93] K. Lauter, A. López-Alt, and M. Naehrig, "Private computation on encrypted genomic data," in 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy. ht tp://seclab. soic. indiana. edu/GenomePrivacy/papers/Genome% 20Privacy-paper9. pdf.(29 July 2014, date last accessed), 2014.

[94] J. H. Cheon, M. Kim, and K. Lauter, "Homomorphic Computation of Edit Distance," in WAHC'15 - 3rd Workshop on Encrypted Computing and Applied Homomorphic Cryptography, 2015.

[95] M. Kim and K. Lauter, "Private genome analysis through homomorphic encryption.," BMC Med. Inform. Decis. Mak., vol. 15 Suppl 5, no. Suppl 5, p. S3, Dec. 2015.

[96] M. Togan and C. Plesca, "Comparison-based computations over fully homomorphic encrypted data," in Communications (COMM), 2014 10th International Conference on, 2014, pp. 1–6.

[97] T. Graepel, K. Lauter, and M. Naehrig, "ML confidential: Machine learning on encrypted data," in Information Security and Cryptology--ICISC 2012, Springer, 2013, pp. 1–21.

[98] W.-J. Lu, Y. Yamada, and J. Sakuma, "Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption.," BMC Med. Inform. Decis. Mak., vol. 15 Suppl 5, no. Suppl 5, p. S1, Dec. 2015.

[99] Z. Huang, E. Ayday, J. Fellay, J.-P. Hubaux, and A. Juels, "GenoGuard: Protecting Genomic Data against Brute-Force Attacks," in 36th IEEE Symposium on Security and Privacy, 2015.

[100] G. Danezis, "Simpler Protocols for Privacy-Preserving Disease Susceptibility Testing," in 14th Privacy Enhancing Technologies Symposium, Workshop on Genome Privacy (GenoPri'14), 2014.

[101] D. Du Verle, S. Kawasaki, Y. Yamada, J. Sakuma, and K. Tsuda, "Privacy-Preserving Statistical Analysis by Exact Logistic Regression," in 2nd International Workshop on Genome Privacy and Security (GenoPri'15), 2015.

[102] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin, "A cryptographic approach to securely share and query genomic sequences," IEEE Trans. Inf. Technol. Biomed., vol. 12, no. 5, pp. 606–617, 2008.

[103] Y. Zhang, M. Blanton, and G. Almashaqbeh, "Secure distributed genome analysis for GWAS and sequence comparison computation.," BMC Med. Inform. Decis. Mak., vol. 15 Suppl 5, no. Suppl 5, p. S4, Dec. 2015.

[104] S. D. Constable, Y. Tang, S. Wang, X. Jiang, and S. Chapin, "Privacy-preserving GWAS analysis on federated genomic datasets," BMC Med. Inform. Decis. Mak., vol. 15, no. Suppl 5, p. S2, Dec. 2015.

[105] B. A. Malin, "Protecting genomic sequence anonymity with generalization lattices.," Methods Inf. Med., vol. 44, no. 5, pp. 687–92, Jan. 2005.

[106] G. Loukides, A. Gkoulalas-Divanis, and B. Malin, "Anonymization of electronic medical records for validating genome-wide association studies," Proc. Natl. Acad. Sci. U. S. A., vol. 107, no. 17, pp. 7898–7903, 2010.

[107] F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg, "Differentially-Private Logistic Regression for Detecting Multiple-SNP Association in GWAS Databases," in Privacy in Statistical Databases, vol. 8744, J. Domingo-Ferrer, Ed. Cham: Springer International Publishing, 2010, pp. 170–184.

[108] S. Wang, N. Mohammed, and R. Chen, "Differentially private genome data dissemination through top-down specialization.," BMC Med. Inform. Decis. Mak., vol. 14, no. Suppl 1, p. S2, Dec. 2014.

[109] A. Johnson and V. Shmatikov, "Privacy-preserving data exploration in genome-wide association studies," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13, 2013, p. 1079.

[110] C. Uhler, A. B. Slavkovic, and S. E. Fienberg, "Privacy-preserving data sharing for genome-wide association studies," J. Priv. Confidentiality, vol. 5, no. 1, pp. 137–166, 2013.

[111] F. Yu, S. E. Fienberg, A. B. Slavković, and C. Uhler, "Scalable privacy-preserving data sharing methodology for genome-wide association studies.," J. Biomed. Inform., vol. 50, no. 50C, pp. 133–141, Feb. 2014.

[112] F. Yu and Z. Ji, "Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies: An Application to iDASH Healthcare Privacy Protection Challenge," BMC Med. Informatics Decis. Mak. [submitted], 2014.

[113] Y. Zhao, X. Wang, X. Jiang, L. Ohno-Machado, and H. Tang, "Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery.," J. Am. Med. Inform. Assoc., vol. 22, no. 1, pp. 100–8, Jan. 2015.

[114] D. Chen and H. Zhao, "Data Security and Privacy Protection Issues in Cloud Computing," 2012 Int. Conf. Comput. Sci. Electron. Eng., vol. 1, no. 973, pp. 647–651, 2012.

[115] F. McSherry, "Privacy integrated queries," in Proceedings of the 35th SIGMOD International Conference on Management of Data (SIGMOD), 2009, pp. 19–30.

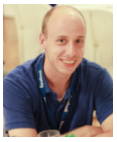
[116] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Commun. ACM, vol. 51, no. 1, p. 107, 2008.

[117] D. A. B. Fernandes, L. F. B. Soares, J. V. Gomes, M. M. Freire, and P. R. M. In??cio, "Security issues in cloud environments: A survey," Int. J. Inf. Secur., vol. 13, no. 2, pp. 113–170, 2014.

[118] K. Grolinger, M. Hayes, W. a. Higashino, A. L'Heureux, D. S. Allison, and M. a. M. Capretz, "Challenges for MapReduce in

Big Data," in Proc. of the SERVICES - IEEE World Congress on Services, 2014, pp. 182–189.
[119] Q. Tran and H. Sato, "A solution for privacy protection in mapreduce," in Proceedings - International Computer Software and Applications Conference, 2012, pp. 515–520.
[120] Z. Xiao and Y. Xiao, "Achieving Accountable MapReduce in cloud computing," Futur. Gener. Comput. Syst., vol. 30, no. 1, pp. 1–13, 2014.
[121] I. Roy, S. T. V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and Privacy for MapReduce," in 7th USENIX Conference on Networked Systems Design and Implementation (NSDI'10), 2010, pp. 297–312.
[122] X. Han, M. Wang, X. Zhang, and X. Meng, "Differentially Private Top-k Query over Map-Reduce," in Proceedings of the fourth international workshop on Cloud data management, 2012, pp. 25–32.
[123] K. CHEN, W. WAN, and Y. LI, "Differentially private feature selection under MapReduce framework," J. China Univ. Posts Telecommun., vol. 20, no. 5, pp. 85–103, 2013.
[124] X. Zhang, C. Liu, S. Nepal, S. Pandey, and J. Chen, "A privacy leakage upper bound constraint-based approach for cost-effective privacy preserving of intermediate data sets in cloud," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1192–1202, 2013.
[125] X. Zhang, L. T. Yang, C. Liu, and J. Chen, "A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 2, pp. 363–373, 2014.
[126] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan, "Sedic: Privacy-Aware Data Intensive Computing on Hybrid Clouds," in Proceedings of the 18th ACM conference on Computer and communications security - CCS '11, 2011, pp. 515–526.
[127] N. Santos, R. Rodrigues, K. P. Gummadi, and S. Saroiu, "Policy-sealed data: A new abstraction for building trusted cloud services," in Security'12: Proceedings of the 21st USENIX conference on Security symposium, 2012, pp. 175–188.
[128] X. Chen and Q. Huang, "The data protection of mapreduce using homomorphic encryption," in Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 2013, pp. 419–421.
[129] S. Kamara and M. Raykova, "Parallel Homomorphic Encryption," in Financial Cryptography and Data Security, no. volume 7862 of the series Lecture Notes in Computer Sciences, A. A. Adams, M. Brenner, and M. Smith, Eds. Springer Berlin Heidelberg, 2013, pp. 213–225.
[130] L. Xu, H. Kim, X. Wang, W. Shi, and T. Suh, "Privacy preserving large scale DNA read-mapping in MapReduce framework using FPGAs," in 2014 24th International Conference on Field Programmable Logic and Applications (FPL), 2014, no. 1, pp. 1–4.
[131] Y. Chen, B. Peng, X. Wang, and H. Tang, "Large-Scale Privacy-Preserving Mapping of Human Genomic Sequences on Hybrid Clouds.," in Proceeding of NDSS Symposium 2012, 2012.
[132] J. L. Raisaro, E. Ayday, P. McLaren, A. Telenti, and J.-P. Hubaux, "On a novel privacy-preserving framework for both personalized medicine and genetic association studies," in PRIVAGEN, 2015.
[133] Y. Zhao, X. Wang, and H. Tang, "Secure Genomic Computation through Site-Wise Encryption," in AMIA Summits on Translational Science Proceedings, 2015, vol. 2015, pp. 227–231.
[134] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," Information, Commun. Soc., vol. 15, no. 5, pp. 662–679, 2012.
[135] P. Li, X. L. Dong, A. Maurino, and D. Srivastava, "Linking Temporal Records," PVLDB, vol. 4, no. 11, pp. 956–967, 2011.
[136] Y.-H. Chiang, A. Doan, and J. F. Naughton, "Modeling Entity Evolution for Temporal Record Matching," in Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, 2014, pp. 1175–1186.
[137] L. P. Deutsch, "GZIP file format specification version 4.3," 1996. [Online]. Available: https://tools.ietf.org/html/rfc1952. [Accessed: 08-Aug-2015].
[138] I. Pavlov, "7zip file archive application," 2007. [Online]. Available: https://ford.ischool.utexas.edu/handle/2081/8999. [Accessed: 01-Sep-2015].
[139] "GenoPri 2014." [Online]. Available: http://seclab.soic.indiana.edu/GenomePrivacy/. [Accessed: 27-Jul-2015].
[140] "Genopri 2015 - Home." [Online]. Available: http://www.genopri.org/. [Accessed: 30-Mar-2015].
[141] "Privacy-aware computational genomics 2015 (PRIVAGEN 2015)." [Online]. Available: http://aistcrypt.github.io/Privacy-Aware-Computational-Genomics/. [Accessed: 25-Jul-2015].
[142] "2014 iDASH Genome Privacy Protection Challenge Workshop," 2014. [Online]. Available: http://www.humangenomeprivacy.org/2014. [Accessed: 24-Mar-2015].
[143] "2015 iDASH Genome Privacy Protection Challenge Workshop." [Online]. Available: http://www.humangenomeprivacy.org/2015/. [Accessed: 24-Mar-2015].
[144] "New Community Challenge Seeks to Evaluate Methods of Computing on Encrypted Genomic Data | GenomeWeb." [Online]. Available: https://www.genomeweb.com/informatics/new-community-challenge-seeks-evaluate-methods-computing-encrypted-genomic-data. [Accessed: 13-Apr-2015].
[145] "To Keep It Safe and Sound | GenomeWeb." [Online]. Available: https://www.genomeweb.com/scan/keep-it-safe-and-sound. [Accessed: 13-Apr-2015].
[146] E. Check Hayden, "Cloud cover protects gene data.," Nature, vol. 519, no. 7544, pp. 400–1, Mar. 2015.

Shuang Wang (S'08–M'12) received the B.S. degree in applied physics and the M.S. degree in biomedical engineering from the Dalian University of Technology, China, and the Ph.D. degree in electrical and computer engineering from the University of Oklahoma, OK, USA, in 2012. He was worked as a postdoc researcher with the Department of Biomedical Informatics (DBMI), University of California, San Diego (UCSD), CA, USA, 2012 - 2015. Currently, he is an assistant professor at the DBMI, UCSD. His research interests include machine learning, and healthcare data privacy/security. He has published more than 60 journal/conference papers, 1 book and 2 book chapters. He was awarded a NGHRI K99/R00 career grant. Dr. Wang is a senior member of IEEE.



Luca Bonomi (M'15) received B.S. and M.S. in computer engineering, University of Padova, and Ph.D. degree in computer science, Emory University in 2006, 2008, and 2015, respectively. He is currently a postdoctoral scholar with the DBMI, UCSD. His research interests include data privacy & security and Data Mining Algorithms.

Wenrui Dai (M'15) received B.S., M.S., and Ph.D. degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China in 2005, 2008, and 2014. He is currently a postdoc at DBMI, UCSD. His research interests include learning-based image/video coding, image/signal processing and predictive modeling.

Feng Chen (M'15) received his B.S. degree, M.S. degree and Ph.D. degree from China University of Geosciences-Beijing, Beihang University and the University of Oklahoma-Tulsa, respectively. He is now a postdoctoral researcher at DBMI, UCSD. His research interests include

secure computation over EHR/ genome data and pure mathematical problems.

Cynthia Cheung is a research associate at UCSD. Cynthia obtained her bachelor's degrees in Human Development and Spanish Literature from UCSD and is currently working toward a Master of Public Health and a Master of Arts in Latin American Studies from San Diego State University.

Cinnamon S. Bloss is an Assistant Professor in the Departments of Psychiatry and Family Medicine and Public Health, Division of Health Policy at the UCSD. Dr. Bloss' current research focuses on the individual and societal impacts of emerging biomedical technologies.

Samuel Cheng (S'01–M'04) received the B.S. degree in electrical and electronic engineering from the University of Hong Kong, the M.Phil. degree in physics from Hong Kong University of Science and Technology, the M.S. degree in electrical engineering from and the University of Hawaii, Honolulu, HI, USA, and the Ph.D. degree in electrical engineering from Texas A&M University in 2004. In 2006, he joined the School of Electrical and Computer Engineering at the University of Oklahoma and is currently an Associate Professor.

Xiaoqian Jiang (S'06–M'10) is an assistant professor in the Department of Biomedical Informatics, UCSD. He received his PhD in computer science from Carnegie Mellon University. He is an associate editor of BMC Medical Informatics and Decision Making and serves as an editorial board member of Journal of American Medical Informatics Association. He works primarily in health data privacy and predictive models in biomedicine. Dr. Jiang is a recipient of NIH K99/R00 award and he won the distinguished paper award from AMIA Clinical Research Informatics (CRI) Summit in 2012 and 2013.