Chapter 17
Private Genome Data Dissemination

Noman Mohammed, Shuang Wang, Rui Chen, and Xiaoqian Jiang

Abstract With the rapid advances in genome sequencing technology, the collection and analysis of genome data have been made easier than ever before. In this course, sharing genome data plays a key role in enabling and facilitating significant medical breakthroughs. However, substantial privacy concerns have been raised on genome data dissemination. Such concerns are further exacerbated by several recently discovered privacy attacks. In this chapter, we review some of these privacy attacks on genome data and the current practices for privacy protection. We discuss the existing work on privacy protection strategies for genome data. We also introduce a very recent effort to disseminating genome data while satisfying differential privacy, a rigorous privacy model that is widely adopted for privacy protection. The proposed algorithm splits raw genome sequences into blocks, subdivides the blocks in a top-down fashion, and finally adds noise to counts in order to preserve privacy. It has been empirically shown that it can retain essential data utility to support different genome data analysis tasks.

17.1 Introduction

In the past decades, genome sequencing technology has experienced unprecedented development. The Human Genome Project (HGP), initiated in 1990, took 13 years to complete at a total cost of $3 billion, while nowadays it takes only 2 or 3 days for a whole genome sequence at the cost of $6K [1]. The readily availability of genome data has spawned many new exciting research areas, ranging from understanding the mechanisms of cellular functions to identifying criminals. There has been no

N. Mohammed (✉)
Department of Computer Science, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
e-mail: noman@cs.umanitoba.ca

S. Wang • X. Jiang
Department of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA
e-mail: shw070@ucsd.edu; x1jiang@ucsd.edu

R. Chen
Samsung Research America, Mountain View, CA, USA
e-mail: rui.chen1@samsung.com

A. Gkoulalas-Divanis, G. Loukides (eds.), *Medical Data Privacy Handbook*,
DOI 10.1007/978-3-319-23633-9_17

doubt that genome data analysis generates interesting scientific discoveries and enables significant medical breakthroughs. However, substantial privacy concerns have been raised on the dissemination of genome data. The public has realized that a personally identifiable genomic segment already gives an adversary access to a wealth of information about the individual and his or her genetic relatives [2], exposing their privacy to considerable risks.

Genomic data leakage has serious implications for research participants such as discrimination for employment, insurance, or education [1]. Recent research results show that given some background information about an individual, an adversary can identify or learn sensitive information about the victim from the de-identified data. For example, Homer's attack [3] demonstrated that it is possible to identify a genome-wide association study (GWAS) participant from the allele frequencies of a large number of single-nucleotide polymorphisms (SNPs). As a consequence, the U.S. National Institutes of Health (NIH) has forbidden public access to most aggregate research results to protect privacy. Later, Wang et al. [4] showed an even higher risk that individuals could be actually identified from a relatively small set of statistics, such as those routinely published in GWAS papers. There are also many other attacks revealed recently [5–7], which could result in harm to the privacy of individuals. Therefore, there has been a growing demand to promote privacy protection for genome data dissemination.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) [8] establishes the Privacy Rule to protect health information. The Privacy Rule defines an operational approach, called Safe Harbor that removes 18 HIPAA-specified identifiers to achieve some degree of "de-identification". Since genome data are biometrics, it would be natural to remove these data from "de-identified" data sets. However, there is no explicit clarification of de-identified genomic data by the Institute of Medicine (IOM) or HIPAA regulations.

There have been long and vigorous debates [9, 10] about the current privacy rules for Human Genomic Studies (HGS). Some researchers contend that existing privacy rules are not adequate for the protection of genomic information [4, 11], as the technological evolution and the increasing accessibility of data cause the "de-identified" genome data to be re-identifiable. Others complain that privacy regulations impede effective data access and use for research [9, 12], as genomic data are most useful when presented in high quality, sufficient samples, and associated with an individual's medical history, etc. Recently, the Presidential Commission for the Study of Bioethical Issues published a report about privacy and progress in Whole Genome Sequencing (WGS) [12]. The report concludes that under current privacy rules, genome privacy is not adequately protected and that at the same time genomic researchers and data owners cannot effectively access and share data. To address these limitations, there have been some pioneering efforts on developing practical privacy-preserving technology solutions to genome data sharing.

In this chapter, we present an approach to disseminate genomic data in a privacy-preserving manner. The privacy guarantee is guarded by the rigorous *differential privacy* model [13]. Differential privacy is a rigorous privacy model that makes no

assumption about an adversary's background knowledge. A differentially-private mechanism ensures that the probability of any output (released data) is equally likely from all nearly identical input datasets and thus guarantees that all outputs are insensitive to any individual's data. In other words, an individual's privacy is not at risk because of his or her participation in the dataset. The proposed approach uses a top-down structure to split long sequences into segments before adding noise to mask record owners' identity, which demonstrates promising utility with a desirable computational complexity.

The rest of the chapter is organized as follows. An overview of the related work is presented in Sect. 17.2. Section 17.3 describes the problem statement and details the data privacy requirement and the utility criteria. Section 17.4 describes the proposed algorithm and analyzes the privacy guarantee and the computational cost of the proposed method. Experimental results are presented in Sect. 17.5, and finally, Sect. 17.6 concludes the chapter.

17.2 Literature Review

17.2.1 Privacy Attacks and Current Practices

In general, the traditional consent mechanism that allows data disclosure without de-identification is not suitable for genomic data [14]. It is often impossible to obtain consent for an unknown future research study without creating any bias in the sample. In addition, the consent mechanism also raises a number of ethical issues (e.g., withdrawal from future research, promise of proper de-identification, etc.) [15].

A common de-identification technique is to remove all explicit identifying attributes (e.g., name and SSN) and to assign each record a random number prior to data sharing. However, this technique is vulnerable to identity disclosure attacks as demonstrated by a recent study that successfully identified the participants of the Personal Genome Project (PGP) through public demographic data [16].

Given that a genome sequence is a strong personal identifier, several organizations including the U.S. National Institutes of Health (NIH) initially adopted a two-tier access model: controlled access and open access [17]. Individual-level biomedical data are only available to approved researchers (i.e., data requesting institutions) based on proper agreements, while summary statistics, which are useful for meta-GWAS analyses, can be disclosed publicly. Unfortunately, some recent studies have shown that from summary statistics adversaries can already learn sensitive information [3, 4]. This is known as an attribute disclosure attack. Currently, there are a number of techniques for breaching biomedical data privacy [18]. In response, the NIH and other data custodians have moved summary statistics from open access to controlled access. For example, the Database of Genotypes and Phenotypes (dbGaP) [19], which is a popular public database recommended by

Genome Canada for sharing biomedical data, no longer provides open access to summary statistics. Although such policy provides enhanced privacy protection, it largely limits researchers' ability to conduct timely research.

In the United States, the Health Insurance Portability and Accountability Act (HIPAA) provides two standards for health data sharing without requiring patients' consent. The Safe Harbor Standard, which is also used by health organizations in Canada [20], considers a dataset properly de-identified if the 18 specific attributes are removed. However, genomic data along with other types of data (e.g., diagnostic code) are not part of these specified attributes [21]. The alternative Statistical Standard requires an expert to certify that the risk of re-identification is very small from the disclosed data. However, there is no specific guideline to address how to achieve Statistical Standard for genomic data.

The current practice of genomic data sharing is based on the controlled access model, and privacy is ensured through a data-use certification (DUC) agreement. However, an agreement cannot prevent an insider from intentionally performing privacy attacks or even stealing data. A recent article reported several violations of such agreements (e.g., investigators sharing controlled-access data with unapproved parties) that occurred in the last 6-year period [22].

17.2.2 Privacy Preserving Techniques

Privacy-preserving data sharing techniques study how to transform raw data into a version that is immune to privacy attacks but that still preserves useful information for data analysis. Existing techniques are primarily based on two major privacy models: k-anonymity [23] and differential privacy [13]. Data sharing techniques adopting the k-anonymity model require that no individual should be identifiable from a group of size smaller than k based on the values of quasi-identifiers (e.g., age, gender, date of birth). In spite of its wide application in the healthcare domain [21, 24], recent research results indicate that k-anonymity based techniques are vulnerable to an adversary's background knowledge [25–28]. This has stimulated a discussion in the research community in favor of the differential privacy model, which provides provable privacy guarantees independent of an adversary's background knowledge.

To satisfy a specific privacy model, certain anonymization techniques have to be developed to achieve a reasonable trade-off between privacy and utility. While many anonymization techniques have been proposed for various types of data (i.e., relational [29, 30], set-valued [31], spatio-temporal data [32]), the problem of genomic data anonymization has been little studied. Recent methods [33–35] propose to generalize genomic data to achieve the k-anonymity privacy model. Malin [35] presents how to anonymize genome sequences. Loukides et al. [33] and Heatherly et al. [34] propose to anonymize only health data (i.e., no protection for genome sequences). However, as mentioned before, all methods adopting k-anonymity as the underlying privacy model are vulnerable to the recently discovered

privacy attacks [25–28]. More recently, differentially private mechanisms [36–38] have been proposed for genomic data. However, these techniques only release some aggregate information or target specific data analysis tasks (e.g., minor allele frequencies, top-k most relevant SNPs). They do not support individual-level data sharing, which could be of much greater interest to the research community.

As another possible direction, cryptography-based methods have also been suggested for distributed genomic data sharing [39, 40]. Data custodians outsource their data securely through homomorphic encryption to a third party that carries out computations on the encrypted data. However, these techniques have several shortcomings. First, they assume the existence of a trusted third party [39] or tamper-resistant hardware [40]. These assumptions may not be practical in most real-life applications. Second, these techniques only support a limited set of aggregate queries and do not enable to share individual-level data. Finally, these techniques suffer from one main drawback: the aggregate output has no privacy guarantee. In the rest of this chapter, we will introduce a novel differentially private data dissemination technique that supports individual-level genome data sharing.

17.3 Problem Statement

Suppose a data owner has a data table $D(A^i, A^{snp})$ and wants to release an anonymous data table $\hat{D}$ to the public for data analysis. The attributes in D are classified into two categories: (1) An *explicit identifier* attribute that explicitly identifies an individual, such as *SSN*, and *Name*. These attributes are removed before releasing the data as per the HIPAA Privacy Rule [41]; (2) A multi-set of SNPs (genomic data), which is denoted by A^{snp}, for each individual in the data table D. For example in Table 17.1, A^i and A^{snp} are *ID* and *Genomic data* attributes, respectively.

Given a data table D, our objective is to generate an anonymized data table $\hat{D}$ such that $\hat{D}$ satisfies ϵ-differential privacy, and preserves as much utility as possible for data analysis. Next, we introduce differential privacy and data utility models.

Table 17.1 Raw genome data

ID	Genomic data
1	AG CC CC GG CT GG AA CC
2	AG CC CC GG TT GG AA CC
3	AA CC CC GG TT GG AA CC
4	AG CT CT AG CT AG AG CT
5	GG CT CT AG CC GG AA CC
6	AA CC CC GG TT GG AA CC
7	AG CT CT AG CT AG AG CT
8	AA CC CC GG TT GG AA CC
9	GG CT TT AG CC AG AA CC
10	AG CT CT GG CT AG AA CC

17.3.1 Privacy Protection Model

Differential privacy is a recent privacy definition that provides a strong privacy guarantee. It guarantees that an adversary (even with arbitrary background knowledge) learns nothing more about an individual from the released data set, regardless of whether her record is present or absent in the original data. Informally, differential privacy requires that any computational output be insensitive to any particular record. Therefore, from an individual's point of view, the output is computed as if from a data set that does not contain his or her record. Formally, differential privacy is defined as follows.

Definition 17.1 (ϵ-Differential Privacy [13]). A randomized algorithm Ag is differentially private if for all data sets D and D' whose symmetric difference is at most one record (i.e., $|D \triangle D'| \leq 1$), and for all possible anonymized data sets $\hat{D}$,

$$\Pr[Ag(D) = \hat{D}] \leq e^{\epsilon} \times \Pr[Ag(D') = \hat{D}], \tag{17.1}$$

where the probability is taken over the randomness of Ag.

The privacy level is controlled by the parameter ϵ. A smaller value of ϵ provides a strong privacy guarantee. A standard mechanism to achieve differential privacy is the Laplace mechanism [13]. Its key idea is to add properly calibrated Laplace noise to the true output of a function in order to mask the impact of any single record. The maximal impact of a record to a function f's output is called its *sensitivity*.

Definition 17.2 (Sensitivity [13]). For a function $f : D \rightarrow \mathbb{R}^d$, the sensitivity of f is $\Delta f = \max_{D,D'} \|f(D) - f(D')\|_1$ for all databases such that $|D \triangle D'| \leq 1$.

Given a function's sensitivity and a privacy parameter, the Laplace mechanism is given as follows.

Theorem 17.1 ([13]). *For any function $f : D \rightarrow \mathbb{R}^d$, the mechanism Ag,*

$$Ag(D) = f(D) + \langle \mathtt{Lap}_1(\frac{\Delta f}{\epsilon}), \ldots, \mathtt{Lap}_d(\frac{\Delta f}{\epsilon}) \rangle \tag{17.2}$$

gives ϵ-differential privacy, where $\mathtt{Lap}_i(\frac{\Delta f}{\epsilon})$ are i.i.d Laplace variables with scale parameter $\frac{\Delta f}{\epsilon}$.

17.3.2 Privacy Attack Model

The likelihood ratio test [42] provides an upper bound on the power of any method for the detection of an individual in a cohort, using the following formula:

$$\overline{L} = \sum_{j}^{m} \left(x_j \log \frac{\hat{p}_j}{p_j} + (1 - x_j) \log \frac{1 - \hat{p}_j}{1 - p_j} \right),$$

where x_j is either 0 (i.e., major allele) or 1 (i.e., minor allele), m is the number of SNPs, p_j is the allele frequency of SNP j in the population and $\hat{p}_j$ is that in a pool.

A statistic $\overline{L}$ measure the probability that a subject in the case group will be re-identified. The re-identification risk is considered to be high if the LR test statistic of individual's SNVs is significantly greater than those of individuals who are not in the same group.

17.3.3 Utility Criteria

We use a case-control association χ^2 test to evaluate the utility of a differentially private data. The test has the following form: $\chi^2 = \sum_i^r \sum_j^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$, where r is the number of rows, c is the number of columns, $O_{i,j}$ is observed frequencies, and $E_{i,j}$ is expected frequencies. The χ^2 test statistic provide a measure of how close the observed frequencies are to the expected frequencies. Suppose that the observed allele counts (e.g., for allele "A" and "T") for the case group are a and b, respectively, for total of $r = a + b$. Similarly, we define the observed counts for the same allele (i.e., "A" and "T") in the control group as c and d, respectively, for total of $s = c + d$. Then, the expected allele frequencies in the case group can be expressed as $(a + c)r/(r + s)$ and $(b + d)r/(r + s)$. The expected allele frequencies in the control group can be expressed as $(a + c)s/(r + s)$ and $(b + d)s/(r + s)$, which measure the expected allele frequencies in both case and control populations.

17.4 Genomic Data Anonymization

In this section, we first present our genomic data anonymization algorithm as described in Algorithm 17.1 and prove that the algorithm is ϵ-differentially private. We then analyze the runtime complexity of the algorithm.

17.4.1 Anonymization Algorithm

The proposed algorithm first divides the genomic data into blocks and then generalizes each block. Thus, the algorithm divides the raw data into several equivalence groups, where all the records within an equivalence group have the same block values. Finally, the algorithm publishes the noisy counts of the groups. Next we elaborate each line of the algorithm.

Table 17.2 Genome data partitioned into blocks

	Genomic data			
ID	Block 1	Block 2	Block 3	Block 4
1	AG CC	CC GG	CT GG	AA CC
2	AG CC	CC GG	TT GG	AA CC
3	AA CC	CC GG	TT GG	AA CC
4	AG CT	CT AG	CT AG	AG CT
5	GG CT	CT AG	CC GG	AA CC
6	AA CC	CC GG	TT GG	AA CC
7	AG CT	CT AG	CT AG	AG CT
8	AA CC	CC GG	TT GG	AA CC
9	GG CT	TT AG	CC AG	AA CC
10	AG CT	CT GG	CT AG	AA CC

Dividing the Raw Data (Line 1) Algorithm 17.1 first divides the raw genomic data into multiple blocks. Each block consists of a number of SNPs. For example, the raw genomic data of Table 17.1 can be divided into four blocks as shown in Table 17.2, where each block consists of two SNPs. These blocks are treated like different attributes and thus enable the proposed algorithm to anonymize high-dimensional genomic data effectively. We denote each block by A_i^{snp} and thus $A^{snp} = \cup A_i^{snp}$.

Note that the sizes of all the blocks do not need to be equal. For example, if there were nine SNPs in Table 17.1 instead of 8, it would be impossible to have all blocks of size two. In such a case, the last block can be bigger than the other blocks. In principle, each block may have a different size, and the proposed algorithm can handle such a scenario.

We do not use any heuristic to determine the size of each block. Block size is always constant, and hence, this step does not use any privacy budget (See Sect. 17.4.2). Experimental results suggest that six SNPs per block yield good result. However, this number may vary depending on the data set in question. It is an interesting research problem to design a heuristic that can determine the optimal size of each block so as to maximize the data utility for a given data set.

Algorithm 17.1 Genomic data anonymization algorithm.

- **Input:** Raw data set D, privacy budget ϵ, and number of specializations h
- **Output:** Anonymized genomic data set $\hat{D}$

1: Divide the genome data into blocks;
2: Generate the taxonomy tree for each block;
3: Initialize every block in D to the topmost value;
4: Initialize Cut_i to include the topmost value;
5: **for** $i = 1$ to h **do**
6: Select $v \in \cup Cut_i$ randomly;
7: Specialize v on D and update $\cup Cut_i$;
8: **return** each leaf node with noisy count $(C + \texttt{Lap}(1/\epsilon))$

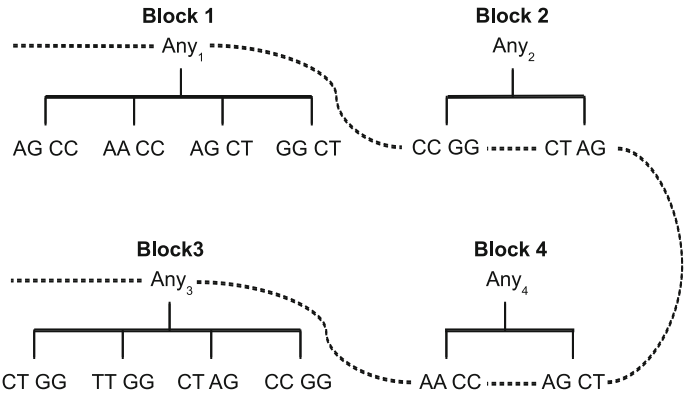

Fig. 17.1 Taxonomy tree of blocks

Generating the Taxonomy Tree (Line 2) A taxonomy tree of a block A_i^{snp} specifies the hierarchy among the values. Figure 17.1 presents the taxonomy trees of Blocks $1-4$ (ignore the dashed curve for now) in Table 17.2. A *cut* of the taxonomy tree for a block A_i^{snp}, denoted by Cut_i, contains exactly one value on each root-to-leaf path (more discussion follows).

Ideally, the data owner should provide a taxonomy tree for each block as the knowledge of the taxonomy tree is domain specific. However, if no taxonomy tree is provided, Algorithm 17.1 can generate it by scanning the data set once for each block. For each unique value that appears in the data set, a leaf node is created from the root node Any_1. For example, four unique values (i.e., AG CC, AA CC, AG CT, and GG CT) appear in Table 17.2 for Block 1; therefore, the corresponding taxonomy tree also has four leaves as shown in Fig. 17.1.

All the generated taxonomy trees have only two levels (i.e., the root and the leaf nodes). However, a data owner can define a multilevel taxonomy tree for each block [43]. A multilevel taxonomy tree provides more flexibility and may preserve more data utility; further investigation is needed in order to validate the benefit of multilevel taxonomy trees.

Data Anonymization (Lines 3–8) The data anonymization starts by creating a single root partition by generalizing all values in $\cup A_i^{snp}$ to the top-most value in their taxonomy trees (Line 3). The initial Cut_i contains the topmost value for each block A_i^{snp} (Line 4).

The specialization starts from the topmost cut and pushes down the cut iteratively by specializing some value in the current cut. The general idea is to anonymize the raw data by a sequence of specializations, starting from the topmost general state as shown in Fig. 17.2. A *specialization*, denoted by $v \rightarrow child(v)$, where $child(v)$ is the set of child values of v, replaces the parent value v with a child value. The specialization process can be viewed as pushing the "cut" of each taxonomy

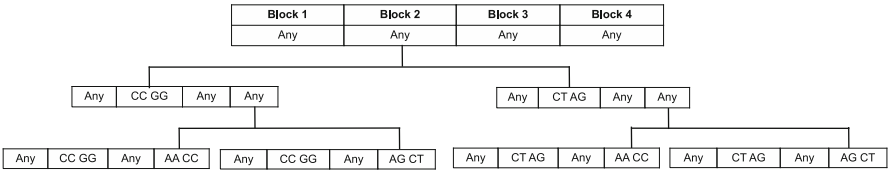

Fig. 17.2 Tree for partitioning records

Table 17.3 Anonymous data ($\epsilon = 1, h = 2$)

Genomic data	Noisy count
Any CC GG Any AA CC	3
Any CC GG Any AG CT	2
Any CT AG Any AA CC	1
Any CT AG Any AG CT	3

tree downwards. Figure 17.1 shows a solution cut indicated by the dashed curve corresponding to the anonymous Table 17.3.

At each iteration, Algorithm 17.1 randomly selects a candidate $v \in \cup Cut_i$ for specialization (Line 6). Candidates can be selected based on their score values, and different heuristics (e.g., information gain) can be used to determine candidates' scores. In future work, we will investigate how to design a scoring function tailored to a specific data utility requirement.

Then, the algorithm specializes v and updates $\cup Cut_i$ (Line 7). Algorithm 17.1 specializes v by recursively distributing the records from the parent partition into disjoint child partitions with more specific values based on the taxonomy tree. The algorithm terminates after a given number of specializations.

Example 17.1. Consider Table 17.1 with $\epsilon = 1$ and $h = 2$, where ϵ is the privacy budget and h is the number of specializations. Initially the algorithm creates one root partition containing all the records that are generalized to $\langle Any_1, Any_2, Any_3, Any_4 \rangle$. $\cup Cut_i$ includes $\{Any_1, Any_2, Any_3, Any_4\}$. Let the first specialization be $Any_2 \rightarrow \{CC\ GG, CT\ AG\}$. The algorithm creates two new partitions under the root, as shown in Fig. 17.2, and splits data records between them. $\cup Cut_i$ is updated to $\{Any_1, Any_3, Any_4\}$. Suppose that the next specialization is $Any_4 \rightarrow \{AA\ CC, AG\ CT\}$, which creates further specialized partitions, as illustrated in Fig. 17.2. ■

Returning the Noisy Counts (Line 9) Finally, Algorithm 17.1 computes the noisy count of each leaf partition to construct the anonymous data table $\hat{D}$ as shown in Table 17.3. The number of leaf partitions is at least 2^h and the exact number depends on the taxonomy tree of the blocks.

Publishing the true counts of each partition violates differential privacy; therefore, a random variable $\texttt{Lap}(\Delta f/\epsilon)$ is added to the true count of each leaf partition, where $\Delta f = 1$, and the noisy counts are published instead.

17.4.2 Privacy Analysis

We now analyze the privacy implication of each of the above steps and quantify the information leakage in terms of privacy budget.

Line 1 The algorithm divides the raw data into blocks, where the block size is a given constant irrespective of the given data set. Since the block generation process is data independent, this step does not require any privacy budget. However, if a heuristic were used to determine the block size, then a portion of privacy budget should be allocated to satisfy differential privacy.

Line 2 We assume that the data owner provides the taxonomy trees. In such a case, this step incurs no privacy leakage and no privacy budget is consumed as the taxonomy trees are generated from public knowledge that is independent of any particular data set.

On the other hand, the alternative approach that we outlined, for a scenario when the taxonomy trees are not provided, needs additional treatment to satisfy differential privacy. It is because, for a different data set $\hat{D}$, a taxonomy tree may have one more or less leaf node. We argue that taxonomy trees represent the domain knowledge, and therefore, should be part of public information.

Lines 3–8 The algorithm selects a candidate for specialization randomly (Line 7) and iteratively creates child partitions based on the given taxonomy trees (Line 8). Both operations are independent of the underlining data set (the selection process is random and the partitioning process is fixed due to the given taxonomy trees), and therefore no privacy budget is required for the h number of iterations.

Line 9 The algorithm adds Laplace noise $\texttt{Lap}(1/\epsilon)$ to the true count of each leaf partition and the requisite privacy budget is ϵ due to the *parallel composition property* [44]. The Parallel composition property guarantees that if a sequence of computations are conducted on *disjoint* data sets, then the privacy cost does not accumulate but depends only on the worst guarantee of all the computations. Since the leaf partitions are disjoint (i.e., a record can fall into exactly one leaf partition), the total privacy cost (i.e., the budget required) for this step is ϵ.

In conclusion, Line 1, Line 2, Lines 3–8, and Line 9 use 0, 0, 0, and ϵ privacy budgets, respectively. According to the *sequential composition property* of differential privacy [44], any sequence of computations that each provides differential privacy in isolation also provides differential privacy in sequence. Therefore, Algorithm 17.1 satisfies ϵ-differential privacy.

17.4.3 Computational Complexity

The proposed algorithm is scalable and the runtime is linear to the size of the data set. This is an important property to achieve in the age of big data. In this section, we present a brief analysis of the computational complexity of Algorithm 17.1.

Line 1 Algorithm 17.1 generates the blocks from the raw data. This can be done by scanning the data set once. Thus, the runtime of this step is $O(|D| \times m)$, where $|D|$ is the number of records and m is the number of SNPs.

Line 2 In case, Algorithm 17.1 can also generate the taxonomy trees (if not given) by scanning the data set once. This is can be achieved simultaneously with the previous step (Line 1); hence, there is no additional cost for generating taxonomy trees. Therefore, if there are d/n number of blocks, where n is the block size, then the runtime of this step is $O(|D| \times \frac{d}{n})$.

Lines 3–8 Candidates are selected randomly in each iteration, which requires constant $O(1)$ time (Line 6).

To perform a specialization $v \rightarrow child(v)$, we need to retrieve D_v, the set of data records generalized to v. To facilitate this operation we organize the records in a tree structure as shown in Fig. 17.2. Each leaf partition (node) stores the set of data records having the same generalized block values. This will allow us to calculate the noisy counts in Line 9.

Initially, the tree has only one leaf partition containing all data records, generalized to the topmost value on every block. In each iteration we perform a specialization by refining the leaf partitions and splitting the records among the new child partitions. This operation also requires scanning all the records once per iteration. Thus, the runtime of this step is $O(|D| \times h)$. The value of h is constant and usually very small (around 10), and therefore, can be ignored.

Line 9 The cost of adding Laplace noise is proportional to the number of leaf nodes, which is 2^h. For a small value of h, the number of leaf nodes is insignificant with respect to the size of the data set $|D|$. We therefore can ignore the cost of this step. Note that, we can easily determine the true count of a leaf partition as it keeps track of the set of data records it represents.

Hence, the total runtime of the algorithm is: $O(|D| \times m + |D|) = O(|D| \times m)$.

17.5 Experimental Results

The goal of the proposed framework is to generate differentially private data that can mitigate the attack of likelihood ratio tests, while preserving highly significant SNPs as much as possible. Two data sets (i.e., chr2 and chr10) with 200 participants in case, control and test groups were used in our experiments. The 200 cases are from Personal Genome Project (PGP: http://www.personalgenomes.org/), missing values filled by using fastPHASE. The 200 Controls are simulated based on the haplotypes of 174 individuals from CEU population of International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/). Besides, the chr2 and chr10 data sets contain 311 SNPs and 610 SNPs, respectively.

Table 17.4 Data utility of chr2 data set with privacy budget of 1.0 and power of 0.01

Cutoff p-value	Accuracy	Sensitivity	Precision	F1-score	# of significant SNPs
5E-02	0.178	1.000	0.079	0.147	22
1E-02	0.211	0.999	0.075	0.140	20
1E-03	0.250	0.948	0.072	0.134	19
1E-05	0.297	1.000	0.060	0.114	14

Table 17.5 Data utility of chr10 data set with privacy budget of 1.0 and power of 0.09

Cutoff p-value	Accuracy	Sensitivity	Precision	F1-score	# of significant SNPs
5E-02	0.301	0.956	0.092	0.168	45
1E-02	0.317	0.903	0.048	0.091	23
1E-03	0.431	1.000	0.041	0.080	15
1E-05	0.577	1.000	0.030	0.058	8

The number of specializations used in our experiment was 5. SNP data were split evenly into $N/6$ blocks, where N is the number of SNP. All the results are based on the average of 100 trials.

Tables 17.4 and 17.5 illustrate the results of the proposed method on chr2 and chr10 data sets with privacy budget of 1.0, where power indicates the ratio of identifiable individuals using the likelihood ratio test in the case group. The power serves as a measurement of the remaining privacy risk in the differentially private results. In Tables 17.4 and 17.5, cutoff p-value thresholds of 5E-2, 1E-2, 1E-3, 1E-5 were used in our experiment, for which four measurements (accuracy, sensitivity, precision and F1-score) were calculated under each method. The last column corresponds to the number of significant SNPs discovered in the original data without adding noise. We can see that the proposed results showed high sensitivities but low precisions on both data sets, which means our method can correctly preserve most true significant SNPs, but with a large amount of false positive reports. Because most SNPs of interest can pass the same filter (e.g., p-value) on both the original data and our differentially private outputs, the latter can serve as a proxy (for exploratory analysis) of the former without losing too much critical information.

Figure 17.3 show the box plots of the data utility in terms of sensitivity and precision for both testing data sets with privacy budget of 1.0 under different cutoff p-values. We can see that the proposed method achieved high sensitivity on both data sets for all cutoff p-values. Moreover, Fig. 17.3 also depict that the precision decreases as the cutoff p-value decreases. Comparing the experiments, results on chr2 are less sensitive in precision than results on chr10 when p-values changes.

Figure 17.4 present the test statistics calculated on case and test groups (i.e., individuals unrelated to both case and control) for both chr2 and chr10 data sets. An individual in the case group can be re-identified with a high confidence if the test statistic obtained from his/her SNP sequence is significantly higher than these of the test group using likelihood ratio test [42]. Figure 17.4 depict that 2 and 18 case

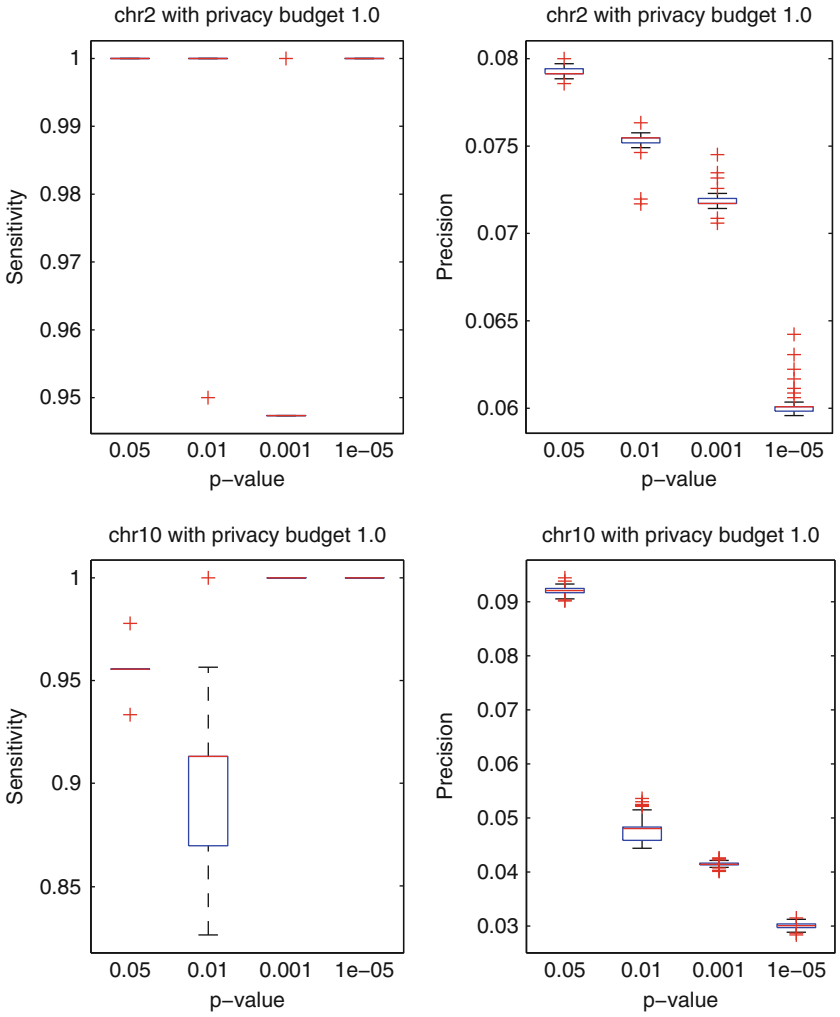

Fig. 17.3 Boxplots of data utility of chr2 and chr10 data with different p-values

individuals have higher test statistic values than 95 % test individuals (i.e., a 5 % false positive rate) in both data sets. The results suggest that the proposed method provides a better privacy protection on a small data set (i.e., chr2 data set) under the same privacy budget.

Finally, Fig. 17.5 show both utility and privacy risk for chr2 and chr10 data sets. To measure the utility, we set the cutoff threshold at 0.1 and measure the sensitivity. By changing privacy budget from 0.1 to 1, we observed no performance gain of sensitivity nor much privacy risk change on chr2 data set, as shown in Fig. 17.5. This seems to indicate that privacy budget 0.1 is sufficient to provide enough protection without destroying the utility.

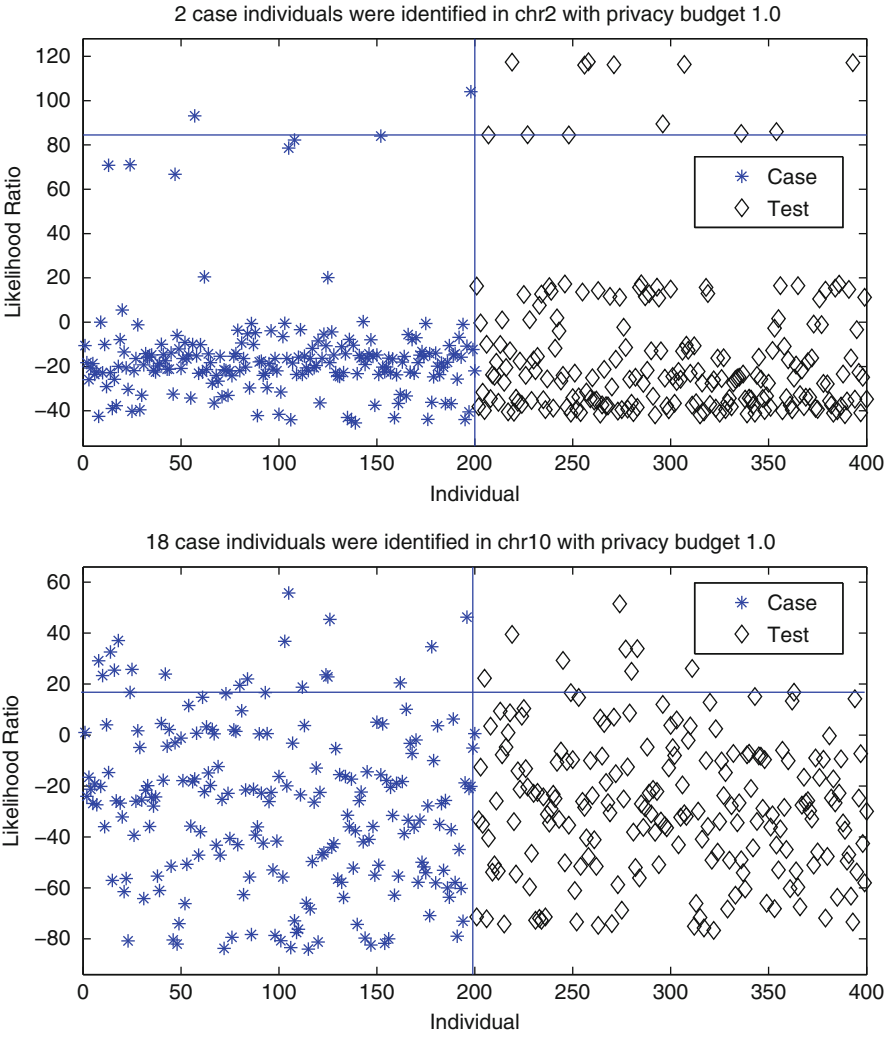

Fig. 17.4 Privacy risk of chr2 and chr10 data. The *star* and *diamond* markers represent the test value of a specific individual in the case (*left*) or test (*right*) group, respectively. The *horizontal line* indicates the 0.95 confidence level for identifying case individuals that are estimated based on the test statistic values of test individuals

We also tested the proposed algorithm on a larger data set (i.e., chr10). Figure 17.5 shows that the proposed algorithm achieves the best sensitivity and the highest number of re-identification risk with privacy budget of 1.0. There is a non-negligible difference in terms of re-identification risk when the privacy budget changes from 0.1 to 0.5 but the difference is not obvious between privacy budgets 0.5 and 1. This indicates that a larger dataset like chr10 needs more privacy budget to protect the privacy of its entities.

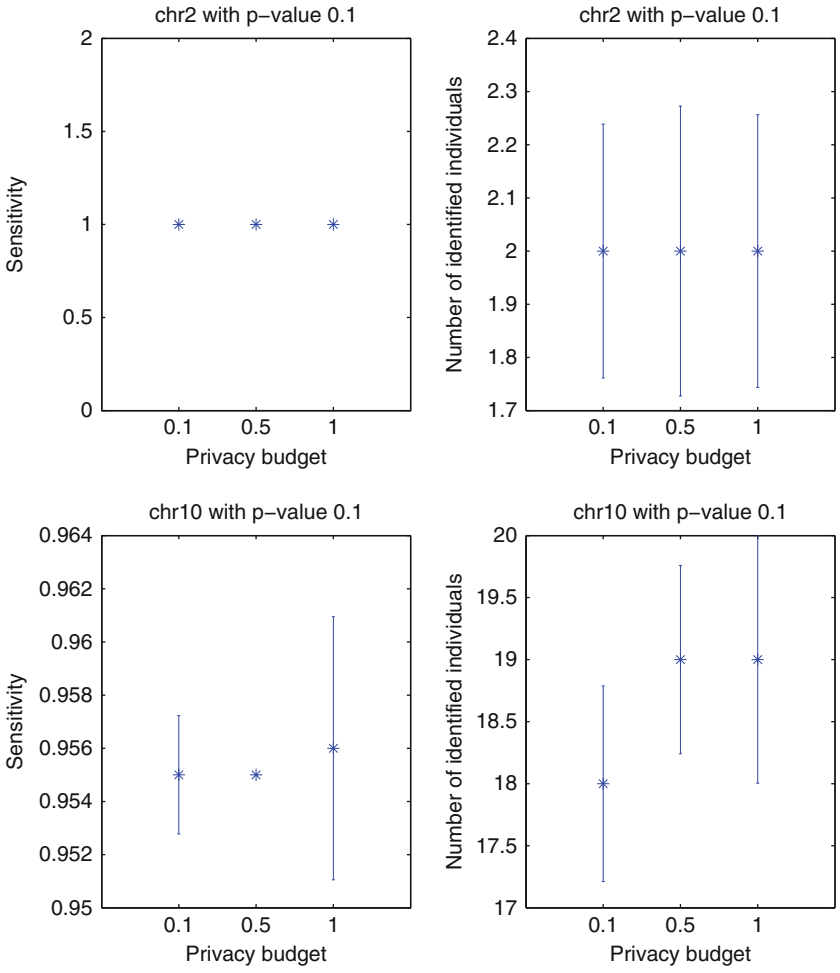

Fig. 17.5 Comparison of data utility and privacy risk for chr2 and chr10 data with different privacy budget

17.6 Conclusion

As an important type of modern medical data, genome data has been incorporated into diverse research disciplines and real-life applications. More and more patient data networks are incorporating or prepare to incorporate genome data along with clinical information to support precision medicine. However, its personal identifying nature has aroused much public concern about the privacy implications of its dissemination, which has been increasingly confirmed by several recently discovered privacy attacks. The current practice of genomic data sharing heavily relies on the controlled access model. Despite the restriction the model poses, it still

cannot provide guaranteed privacy protection. For example, side channel leakage remains to be a big problem to controlled access model.

Advanced technical efforts are indispensable to private genome data dissemination. In this chapter, we gave an overview of the recent developments toward accomplishing this goal. Our focus was on a new differentially private genome data dissemination algorithm. This algorithm supports individual-level data sharing, which is a desideratum for many research areas.

While the existing studies have demonstrated the promise of private genome data sharing, there are still some notable limits. For example, the precision performance of the proposed framework is relatively poor. To make this anonymization process effective, several challenges must be addressed. First, further study is needed to understand how to partition the genome sequences into blocks that are still meaningful. Genomic data is high dimensional and it includes hundreds of Single Nucleotide Polymorphisms (SNPs). Dividing the long sequences into blocks can reduce the high-dimensionality challenge. Second, it might be useful to construct a taxonomy tree for each genome block, which was not tried before. While multilevel taxonomy trees have been proposed for a SNP [35], it is not clear how to construct multilevel trees for a block. Finally, genomic data are shared for different data analysis tasks (e.g., association test, logistic regression); in which case introducing a task specific utility function in the data specialization process may preserve better data utility.

Acknowledgements This article was funded by iDASH (U54HL108460), NHGRI (K99HG 008175), NLM (R00LM011392, R21LM012060), NCBC-linked grant (R01HG007078) and NSERC Discovery Grants (RGPIN-2015-04147).

References

1. Naveed, M., Ayday, E., Clayton, E.W., Fellay, J., Gunter, C.A., Hubaux, J.-P., Malin, B.A., Wang, X.F.: Privacy in the Genomic Era. ACM Comput. Surv. to appear
2. Roche, P.A., Annas, G.J.: DNA testing, banking and genetic privacy. N. Engl. J. Med. **355**, 545–546 (2006)
3. Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J.V., Stephan, D.A., Nelson, S.F., Craig, D.W.:: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using highdensity SNP genotyping microarrays. PLoS Genet. **4**(8), e1000167 (2008)
4. Wang, R., Li, Y.F., Wang, X.F., Tang, H., Zhou, X.: Learning your identity and disease from research papers: information leaks in genome wide association study. In: Proceedings of the 16th ACM Conference on Computer and Communications Security (CCS), New York, pp. 534–544 (2009)
5. Goodrich, M.T.: The mastermind attack on genomic data. In: Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P), pp. 204–218 (2009)
6. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E., Erlich, Y.: Identifying personal genomes by surname inference. Science **339**(6117), 321–324 (2013)

7. Rodriguez, L.L., Brooks, L.D., Greenberg, J.H., Green, E.D.: The complexities of genomic identifiability. Science **339**(6117), 275–276 (2013)
8. Health Insurance Portability and Accountability Act of 1996. Public L. No. 104–191, 110 Stat. 1936, 1996. http://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf.
9. Zhou, X., Peng, B., Li, Y., Chen, Y.: To release or not to release: evaluating information leaks in aggregate human-genome data. In: Security ESORICS, Leuven, pp. 1–27 (2011)
10. Weaver, T., Maurer, J., Hayashizaki, Y.: Sharing genomes: an integrated approach to funding, managing and distributing genomic clone resources. Nat. Rev. Genet. **5**(11), 861–866 (2004)
11. Malin, B.A., Sweeney, L.A.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J. Biomed. Inform. **37**(3), 179–192 (2004)
12. Presidential Commission for the Study of Bioethical Issues: Privacy and Progress in Whole Genome Sequencing (October) (2012)
13. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Conference on Theory of Cryptography (TCC), pp. 265–284 (2006)
14. Caulfield, T., Knoppers, B.: Consent, privacy and research biobanks: policy brief No. 1. Genomics, Public Policy and Society, Genome Canada (2010)
15. Ogbogu, U., Burningham, S.: Privacy protection and genetic research: where does the public interest lie? Alberta Law Rev. 51(3), 471–496 (2014)
16. Sweeney, L., Abu, A., Winn, J.: Identifying participants in the personal genome project by name (a re-identification experiment) (2013) [arXiv:1304.7605]
17. National Institutes of Health, Modifications to Genome-Wide Association Studies (GWAS) Data Access, 28 August 2008
18. Erlich, Y., Narayanan, A.: Routes for breaching and protecting genetic privacy. Nat. Rev. Genet. **15**(6), 409–21 (2014)
19. Mailman, M., et al.: The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet. **39**(10), 1181–1186 (2007)
20. Emam, K.: Data anonymization practices in clinical research: a descriptive study. Health Canada, Access to Information and Privacy Division (2006).
21. Emam, K.: Methods for the de-identification of electronic health records for genomic research. Genome Med. **3**, 25 (2011). doi:10.1186/gm239
22. Paltoo, D., et al.: Data use under the NIH GWAS data sharing policy and future directions. Nat. Genet. **46**, 934–938 (2014)
23. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl. Based Syst. **10**, 557–570 (2002)
24. Emam, K.: A globally optimal k-anonymity method for the de-identification of health data. J. Am. Med. Inform. Assoc. **16**(5), 670–682 (2009)
25. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data **1**(1), Article No. 3 (2007)
26. Li, N., Li, T., Venkatasubramanian, S.: *t*-closeness: a new privacy measure for data publishing. IEEE Trans. Knowl. Data Eng. **22**(7), 943–956 (2010)
27. Zhang, L., Jajodia, S., Brodsky, A.: Information disclosure under realistic assumptions: privacy versus optimality. In Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS), pp. 573–583 (2007)
28. Ganta, S., Kasiviswanathan, S., Smith, A.: Composition attacks and auxiliary information in data privacy. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 265–273 (2008)
29. Fung, B., Wang, K., Yu, P.: Anonymizing classification data for privacy preservation. IEEE Trans. Knowl. Data Eng. **19**(5), 711–725 (2007)
30. Mohammed, N., Chen, R., Fung, B.C.M., Yu, P.S.: Differentially private data release for data mining. In Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 493–501, San Diego, CA (2011)

31. Terrovitis, M., Mamoulis, N., Kalnis, P.: Local and global recoding methods for anonymizing set-valued data. J. Very Large Data Bases **20**(1), 83–106 (2011)
32. Fan, L., Xiong, L., Sunderam, V.: Differentially private multi-dimensional time-series release for traffic monitoring. In Proceedings of the 27th IFIP WG 11.3 Conference on Data and Applications Security and Privacy (2013)
33. Loukides, G., Gkoulalas-Divanis, A., Malin, B.:. Anonymization of electronic medical records for validating genome-wide association studies. Proc. Natl. Acad. Sci. U. S. A. **107**(17), 7898–7903 (2010)
34. Heatherly, R., Loukides, G., Denny, J., Haines, J., Roden, D., Malin, B.: Enabling genomic–phenomic association discovery without sacrificing anonymity. PLoS ONE **8**(2), e53875 (2013)
35. Malin, B.A.: Protecting DNA sequences anonymity with generalization lattices. Methods Inf. Med. **12**(1), 687–692 (2005)
36. Johnson, A., Shmatikov, V.: Privacy-preserving data exploration in genome-wide association studies. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp. 1079–1087 (2013)
37. Chen, R., Peng, Y., Choi, B., Xu, J., Hu, H.: A private DNA motif finding algorithm. J. Biomed. Inform. **50**, 122–132 (2014)
38. Yu, F., Fienberg, S.E., Slavkovic, A.B., Uhler, C.: Scalable privacy-preserving data sharing methodology for genome-wide association studies. J. Biomed. Inform. **50**, 133–141 (2014)
39. Kantarcioglu, M., Jiang, W., Liu, Y., Malin, B.: A cryptographic approach to securely share and query genomic sequences. IEEE Trans. Inf. Technol. Biomed. 12(5), 606–617 (2008).
40. Canim, M., Kantarcioglu, M., Malin, B.: Secure management of biomedical data with cryptographic hardware. IEEE Trans. Inf. Technol. Biomed. **16**(1), 166–175 (2012)
41. Malin, B., Benitez, K., Masys, D.: Never too old for anonymity: a statistical standard for demographic data sharing via the hipaa privacy rule. J. Am. Med. Inform. Assoc. **18**(1), 3–10 (2011)
42. Sankararaman, S., Obozinski, G., Jordan, M.I., Halperin, E.: Genomic privacy and limits of individual detection in a pool. Nat. Genet. **41**(9), 965–967 (2009)
43. Malin, B.A.: An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. J. Am. Med. Inform. Assoc. **12**(1), 28–34 (2005)
44. McSherry, F.: Privacy integrated queries. In: Proceedings of the 35th SIGMOD International Conference on Management of Data (SIGMOD), pp. 19–30 (2009)