BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 06/06 pp225–233
Volume 3, Number 3, September 2020
DOI: 10.26599/BDMA.2020.9020008

Inference Attacks on Genomic Data Based on Probabilistic Graphical Models

Zaobo He* and Junxiu Zhou

Abstract: The rapid progress and plummeting costs of human-genome sequencing enable the availability of large amount of personal biomedical information, leading to one of the most important concerns — genomic data privacy. Since personal biomedical data are highly correlated with relatives, with the increasing availability of genomes and personal traits online (i.e., leakage unwittingly, or after their releasing intentionally to genetic service platforms), kin-genomic data privacy is threatened. We propose new inference attacks to predict unknown Single Nucleotide Polymorphisms (SNPs) and human traits of individuals in a familial genomic dataset based on probabilistic graphical models and belief propagation. With this method, the adversary can predict the unobserved genomes or traits of targeted individuals in a family genomic dataset where some individuals' genomes and traits are observed, relying on SNP-trait association from Genome-Wide Association Study (GWAS), Mendel's Laws, and statistical relations between SNPs. Existing genome inferences have relatively high computational complexity with the input of tens of millions of SNPs and human traits. Then, we propose an approach to publish genomic data with differential privacy guarantee. After finding an approximate distribution of the input genomic dataset relying on Bayesian networks, a noisy distribution is obtained after injecting noise into the approximate distribution. Finally, synthetic genomic dataset is sampled and it is proved that any query on synthetic dataset satisfies differential privacy guarantee.

Key words: Single Nucleotide Polymorphism (SNP)-trait association; belief propagation; factor graph; data sanitization

1 Introduction

With the flourishing and technique advancement of whole-genome sequencing, there are large amount of personal genomic data available online. For instance, increasing amount of online social networks and health-related service providers provide Deoxyribonucleic Acid (DNA)-sequencing services, in which users share their DNA sequence to 23andMe[1] and OpenSNP[2], and share their diseases to PatientsLikeMe[3]. The increasing availability of genomic data has positively impacted the research in developing new medications for genetic diseases, through supporting the development of new research areas that are impossible previously. Moreover, through releasing their genomic data to data platforms or service providers, individuals are interested in learning about their predispositions to some genetic diseases. Therefore, there are large amount of genomic data available online for research or marketing purposes.

Although attractive, the growing availability of genomic data online brings serious privacy issues. Directly releasing genomic data may reveal private information of an individual even though the dataset is anonymized, since it is possible to de-anonymize it with background knowledge[4,5]. For example, membership attacks are general methods to identify participants in the genomic dataset by name. Hence anonymization is

• Zaobo He is with the Department of Computer Science and Software Engineering, Miami University, Oxford, OH 45011, USA. E-mail: hez26@miamioh.edu.
• Junxiu Zhou is with the Department of Computer Science, Northern Kentucky University, Highland Heights, KY 41099, USA. E-mail: zhouj2@nku.edu.

∗ To whom correspondence should be addressed.

Manuscript received: 2020-05-10; accepted: 2020-06-24

incapable for protecting membership privacy. Once a genome sequence is de-anonymized, the identified owner might bear the discrimination risk (for example, from insurance companies or employers)[6]. For example, a Genome-Wide Association Study (GWAS) releases a group of Single Nucleotide Polymorphism (SNP)-trait association, which reports there are 3 SNPs (rs7626795, rs2808630, and rs8034191 on chromosomes 3, 1, and 15, respectively) having a high correlation with an increasing susceptibility for lung cancer.

Due to the incapability of anonymization, on one hand, some believe that they can protect their genomic data privacy by hiding the key part of their genomes, and only releasing the remaining part. One the other hand, some believe that there is nothing to protect about their genomic data, therefore, they might determine to release their genomes to data platforms or service providers without any protection to support genomic research. With the growing amount of scandals regarding to genomic data leaking, the two thoughts above are certainly not desirable.

Firstly, our DNA is highly correlated to our relatives' DNA. Nowadays one can release her genomic data online instantly without his relatives' consent. Even the released data are anonymized, once the owner is identified by powerful adversaries with extensive background knowledge, the released data might put the relatives' privacy at risk. A typical example is Henrietta Lacks (who died in 1951), her DNA is published and sequenced without her and her relatives' consent[7]. One one hand, Lacks' family members think her DNA encodes the genetic information of whole family members and the data should not be released without the permission of all family members. On the other hand, some researchers argue that the DNA of current relatives has been diluted so much after many years with the process of reproduction for gene mixing with different people, such that nothing accurate genetic information can be learned about current family members.

Secondly, even though non-sensitive genomes (i.e., those genomes do not have close correlation with sensitive human traits, such as cancer) are released, the unreleased part might be predicted through inference attacks launched by powerful attackers with extensive background knowledge[8]. Several research organizations also make some research results publicly available which might be leveraged by adversaries. For example, GWAS catalog publishes the SNP-trait correlation (the statistical relationship between SNPs (Genotypes) and human traits (Phenotypes)) for research purpose, including at least 100 000 SNPs and corresponding associations with possible human traits[9]. SNP-trait associations indicate some SNPs (Genotypes) are associated with some human traits (Phenotypes).

A typical example is Nyholt et al.[10], the discoverer of DNA, and Watson released his whole genomes for genomic research, except one gene, i.e., Apolipoprotein E (ApoE). The status of ApoE has close correlation with the development of Alzheimer's disease. However, later studies show that the SNP-trait association, Mendel's Laws and the linkage disequilibrium[11] (the correlation between one or multiple SNPs and ApoE) can be leveraged to predict the value of ApoE, with the help of advanced machine learning algorithms[12].

In this paper, to mitigate the gap, we show that kin-genomic privacy can actually be jeopardized, unless we perturb genomic data prior to releasing to protect against genome inference. The proposed approach simulating inference attacks from powerful adversaries has linear complexity. Furthermore, a differentially private genomic data publishing approach is proposed, such that any query on released genomic data satisfies differential privacy guarantee. The adopted differential privacy[13, 14], as gold standard, allows unlimited background knowledge and reasoning power of adversaries. An algorithm that provides differential privacy guarantee such that the probability distribution of the algorithm output does not have significant change if the input dataset is changed to a neighboring dataset. Two datasets are neighboring if they are same except one entry.

Specifically, given SNP-trait associations collected from GWAS catalog[9], linkage disequilibrium[11], and Mendel's Laws, we propose an inference attack approach based on probabilistic graphical model to predict targeted unknown SNPs and traits. We model the process of inference attacks as computing the marginal distribution of unknown SNPs and traits. However, the computational complexity might be very high, considering there are tens of millions of SNPs and possible human traits. To solve this problem, we make use of belief propagation over probabilistic graphical model[15], which can compute the marginal probability distribution of unknown SNPs and traits with linear complexity.

We then investigate the problem of releasing differentially private genomic data. Given an input genomic dataset $\mathcal{D}$, we aim to release a sanitized $\mathcal{D}'$ that approximates the features of $\mathcal{D}$ as accurate as possible while any query on $\mathcal{D}'$ satisfies differential privacy guarantee. In this paper, we aim to seek a synthetic $\mathcal{D}'$ by sampling $\mathcal{D}$ based on the differentially private joint distribution of $\mathcal{D}$. However, the process above brings a key challenge. Previous studies[16, 17] show that it is difficult to directly apply differential privacy into high-dimensional data releasing. As shown in Ref. [18], if we directly inject differential privacy noise into the full dimensional distribution of genomic data, two problems are raised, namely, output scalability and signal-to-noise ratio. Output scalability means although the input genomic dataset is unwieldy, the output is very unwieldy and slow to utilize. Signal-to-noise ratio means the case that injected noises dominate the original signals of input datasets. To mitigate the gap, we propose to make use of Bayesian networks[19] to model the correlations among SNPs and traits, then approximate the distribution of genomic data with a set of local functions and each local function takes a small subset of variables as its arguments. Then differential privacy noise is injected into those local functions, hence we get a noisy distribution of dataset. Finally, the synthetic genomic dataset is sampled from the noisy distribution. The contributions of this work are summarized as follows:

- An inference attack algorithm based on belief propagation on factor graph is proposed with linear computational complexity which outperforms traditional algorithm with exponential complexity.
- A differentially private genomic data releasing algorithm is proposed for generating synthetic data.

2 Preliminary

This section introduces preliminary knowledge on belief propagation, SNP, and differential privacy. Then we introduce the problem formulation of inference attacks and differentially private genomic data publishing.

2.1 Single nucleotide polymorphism

Human beings have more than 99.9% of their genomes in common. Thus, the genomic data analysis does not need to concentrate on the whole DNA sequence, but rather concentrate on the most variant part. For human beings, the most important variations in our DNA are SNP[20]. The variation means a nucleotide (with value A, T, G, or C) on a certain location on the DNA sequence varies between different people of a specific population. For instance, given two DNA sequences, CAGGTCA and CAAGTCA, just one nucleotide: G and A are different. A pair of nucleotides, C and T, or G and A, is called alleles. A GWAS shows that the occurrence of a specific disease has a close correlation with one or multiple SNPs, and such data correlation is called SNP-trait association. As aforementioned example, the SNP-trait association released by GWAS catalog shows that one individual with special values of three SNPs (rs2808630, rs8034191, and rs7626795) gets a growing susceptibility for lung cancer.

In general, two types of nucleotides are known at a specific SNP position, namely, a major allele and a minor allele. The major allele is defined as the the most frequent nucleotide observed at a specific SNP position. The minor allele is defined as the rare nucleotide at a specific SNP position. We represent a minor allele and major allele as b and B, respectively. The two nucleotides at each SNP are inherited from parents (one from mother and one from father). Then, the content of a specific SNP can be represented as *BB* (there are two major alleles), *Bb* (one nucleotide is a major allele and one nucleotide is a minor allele), or *bb* (there are two minor alleles).

SNP-trait association. GWAS catalog periodically releases the SNP-trait association, which provides a significant reference for genomic research. In the process of identifying the SNPs correlated with human traits, the study splits individuals' genomic data in a given population into two sets: case set (with targeted traits) and control set (without targeted traits). Until now, GWAS catalog has released the association between SNPs and so many human traits, including height, type-2 diabetes, lung cancer, cervical cancer, Chronic kidney disease, etc.

Linkage Disequilibrium (LD). DNA sequences are correlated with each other and this interdependent associations lead to genetic risk. LD measures the degree that any two SNPs are dependent to each other. Due to LD, the content of SNPs locus could be inferred with the content of dependent SNPs. One of the commonly used LD metrics is Pearson correlation coefficient r^2. $r^2 = 1$ means the strongest LD correlation.

2.2 Belief propagation on factor graph

In general, belief propagation is used to compute marginal distribution of variables, given probability dependency among variables. A natural implementation of belief propagation is to execute belief propagation on probabilistic graphical model (i.e., Bayesian networks, factor graphs, Markov random fields, etc.). The operation of belief propagation is generally depicted as the process of message-passing between factor nodes and variable nodes on factor graph, to compute marginal distribution of unknown variables. A factor graph is an undirected, cyclic, or acyclic graph with two different types of nodes: variable nodes and factor nodes. An edge connects a variable node and a factor node if and only if the variable is an argument of the connected factor node. As a message-passing algorithm, belief propagation passes messages between two neighboring factor nodes and variable nodes. The passed message is the conditional probability of a variable node taking a specific value. Given the initial condition, the procedure of message passing on acyclic factor graph will converge to a stable situation.

2.3 Differential privacy

Let D represent a private dataset to be published. Differential privacy is an algorithm $\mathcal{A}$ that requests that prior to D's releasing, D should be added special noise employing the randomized algorithm $\mathcal{A}$, such that any query results on D have no significant difference if we change the input dataset to D', where D and D' are two neighboring datasets, i.e., they differ in only one entry. In other words, adversaries cannot learn significant information about any entry in D from the query results returned from the output of $\mathcal{A}$. Differential privacy offers provable privacy guarantee without any assumption of attackers' background knowledge.

Differential privacy can be formally defined as follows:

Difinition ε-Differential Privacy (ε-DP). A randomized algorithm $\mathcal{A}$ satisfies ε-DP, if for any two neighboring datasets, D and D' that differ in only one entry and for any possible output O of $\mathcal{A}$, the following condition holds:

$$\Pr[\mathcal{A}(D) = O] \leqslant \mathrm{e}^{\varepsilon} \cdot \Pr[\mathcal{A}(D') = O].$$

To satisfy differential privacy, there are two mechanisms used extensively, i.e., the Laplace mechanism[14] and the exponential mechanism[21].

Laplace mechanism. Given a query function F on D, the Laplace mechanism works on the output of F that takes D as input and outputs numeric values. To enable F to return differentially private results, Laplace mechanism add i.i.d. noise to every output of F; then F satisfies differential privacy:

$$\mathcal{A}(D) = F(D) + \mathrm{Lap}\left(\frac{\Delta F}{\varepsilon}\right)^d,$$

where ΔF denotes the sensitivity of F:

$$\Delta F = \max_{D \simeq D'} \|F(D) - F(D')\|_1,$$

and the injected noise η is sampled from a Laplace distribution $\mathrm{Lap}(\eta)$ with the probability distribution function:

$$\Pr[\eta = x] = \frac{1}{2\eta}\mathrm{e}^{-|x|/\eta}.$$

Reference [14] shows that the Laplace mechanism enables F to be an ε-differentially private function if $\eta \geqslant \Delta(F)/\varepsilon$.

Exponential mechanism. When F takes D as input and outputs categorical values, rather than numeric values, then the Laplace mechanism cannot be applied in noise injection. The exponential mechanism returns the output of F that takes D as input and outputs a categorical value. To enable the output of F to be differentially private, the exponential mechanism samples each output from the output domain Ω of $F(D)$. The sampling probability of every ω, $\omega \in \Omega$ can be computed in accordance to the quality of each output. Therefore, given a publisher-specified quality function $f_s(\omega, D)$, which measures the quality of each ω as output of $F(D)$. An ω with larger score indicates it is better to choose this ω as the output of $F(D)$. Therefore, given a dataset D, the exponential mechanism releases $F(D)$ by sampling ω, $\omega \in \Omega$, with a probability proportional to $\exp\left(\frac{\varepsilon f_s(\omega, D)}{2\Delta q}\right)$, where Δq=$\max_{\forall \omega, D, D'} |f_s(\omega, D)$– $f_s(\omega, D')|$ denotes the sensitivity of the publisher-specified score function. Reference [21] shows that injecting noise with exponential mechanism satisfies ε-differential privacy.

3 Problem Formulation

3.1 Genomic data model

In this paper, we consider a genomic dataset $\mathcal{D}(\mathcal{V}, \mathcal{X}, \mathcal{Y})$, where $\mathcal{V}$ represents $|\mathcal{V}|$ individuals in $\mathcal{D}$, $\mathcal{X}$ represents the set of SNPs $\mathcal{X}_i$ for each individual $i \in \mathcal{V}$, and $\mathcal{Y}$ represents the set of traits $\mathcal{Y}_i$ for each individual $i \in \mathcal{V}$. Given an SNP x, $x \in \mathcal{X}_i$ on a certain location on the DNA sequence, then the genotype of x takes value from $x \in \{BB, Bb, bb\}$ (as introduced in Section 2.1). We denote x_j^i as the value of SNP j $(j \in \mathcal{X}_i)$ of individual i $(i \in \mathcal{V})$. Likewise, we denote y_k^i as the value of trait

k ($k \in \mathcal{Y}_i$) of individual i ($i \in \mathcal{V}$), where $\mathcal{Y}_i$ is the set of traits of individual i.

Some privacy-unconcerned family members release partial or the whole DNA sequence, or traits for genomic research. For privacy concern, the most sensitive genomes or traits are kept secret. We denote the set of unknown SNPs and traits as $\mathcal{X}_U$ and $\mathcal{Y}_U$, respectively, and known SNPs and traits as $\mathcal{X}_K$ and $\mathcal{Y}_K$, respectively. According to the SNP-trait association released by GWAS, given a trait $y \in \mathcal{Y}_i$, we define the set of correlated SNPs is $\mathcal{X}(y)$.

3.2 Attacker model

The adversary aims to predict targeted SNPs and traits of targeted individuals in the input dataset, namely, $\mathcal{X}_U \cup \mathcal{Y}_U$. A powerful adversary is considered in this work with extensive background knowledge. In addition to the publicly available SNPs and traits, the adversary also collects publicly available SNP-trait associations, Mendel's Laws, and the linkage disequilibrium. For example, $\mathcal{F}(x_j^F, x_j^M, x_j^C)$ is the function denoting the probability dependency among three family members, i.e., father, mother, and child, which means the probability distribution of x_j^C is highly determined by the value of x_j^F and x_j^M, for any SNP j. $\mathcal{L}(x_j^i, x_k^i)$ is the function denoting the linkage disequilibrium among two SNPs x_j^i and x_k^i, which means there exists a statistical relationship between any two SNPs j and k for individual i because of population's genetic history. $\mathcal{G}(x_j^i, x_k^i, y_l^i)$ is the function denoting the SNP-trait associations, namely, the probability distribution of y_l^i in terms of x_j^i and x_k^i, for any individual i.

The adversary launches an inference attack to predict the value of targeted SNPs and traits, leveraging his background knowledge $\mathcal{X}_K, \mathcal{Y}_K$, $\mathcal{F}(x_j^F, x_j^M, x_j^C)$, $\mathcal{L}(x_j^i, x_k^i)$, and $\mathcal{G}(x_j^i, x_k^i, y_l^i)$.

3.3 Problem definition

The problem of inference attacks on genomic data is defined as follows:

Input: (1) Publicly available SNPs X_K and traits Y_K released by family members, and SNP-trait association $\mathcal{G}$; (2) differential privacy budget ε.

Output: (1) Inference attacks algorithm for predicting the unknown SNPs X_U and traits Y_U. (2) ε-differentially private genomic data publishing algorithms.

4 Inference Attack on Unknown SNPs and Traits

The problem of predicting targeted unknown SNPs and traits can be modeled as computing the marginal distribution of unknown variables, given $\mathcal{X}_K, \mathcal{Y}_K$, $\mathcal{F}(x_j^F, x_j^M, x_j^C)$, $\mathcal{L}(x_j^i, x_k^i)$, and $\mathcal{G}(x_j^i, x_k^i, y_l^i)$. The marginal distribution of an arbitrary SNP of individual i, x_j^i, $x_j^i \in \mathcal{X}_U$ can be computed as

$$\Pr(x_j^i|\mathcal{X}_K, \mathcal{Y}_K, \mathcal{F}(x_j^F, x_j^M, x_j^C), \mathcal{L}(x_j^i, x_k^i), \mathcal{G}(\cdot)) = \sum_{\mathcal{X}_U \backslash x_j^i} \Pr(X_U|\mathcal{X}_K, \mathcal{Y}_K, \mathcal{F}(x_j^F, x_j^M, x_j^C), \mathcal{L}(x_j^i, x_k^i), \mathcal{G}(\cdot)) \tag{1}$$

where $X_U \backslash x_j^i$ is all the unknown attributes in X_U except x_j^i, and $\Pr(X_U|\mathcal{X}_K, \mathcal{Y}_K, \mathcal{F}(x_j^F, x_j^M, x_j^C)$, $\mathcal{L}(x_j^i, x_k^i), \mathcal{G}(\cdot))$ is the joint distribution of all unknown SNPs. The computation of the marginal distribution of unknown traits can be formulated with the same way.

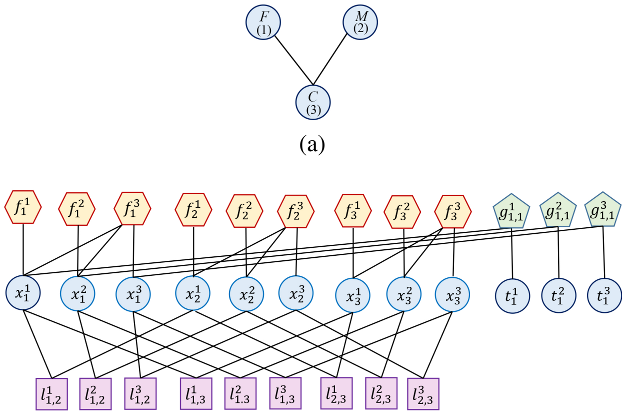
However, the computational complexity of computing the joint distribution of all unknown SNPs is very high, considering there are tens of millions of SNPs in human DNA sequence. Actually, the computation of Eq. (1) has exponential complexity, due to the requirement to sum over the items with a exponential scale. Therefore, it is undesirable to compute joint distribution with the above way, instead of leveraging the data correlations among SNPs and traits. We propose to develop a probabilistic graphical model, specifically, a factor graph, to encode the probability dependency among different variables. Then we execute belief propagation on the factor graph, such that the joint distribution of all unknown variables can be factorized into a set of local functions and each function takes a small subset of variables as their arguments. Since each local function just holds a small size of SNPs and traits as variables, the computation of $\Pr(X_U|\mathcal{X}_K, \mathcal{Y}_K, \mathcal{F}(x_j^F, x_j^M, x_j^C), \mathcal{L}(x_j^i, x_k^i), \mathcal{G}(\cdot))$ now has linear computational complexity. Compared with Eq. (1) adopted by most of existing works, the proposed approach is efficient. Constructing a factor graph and executing belief propagation on it are challenging since identifying the relationship between a factor node and a variable node is hard, given tens of millions of variables. To mitigate the gap, we construct a factor graph by integrating known and unknown SNPs and traits, family relationship $\mathcal{F}(x_j^F, x_j^M, x_j^C)$, linkage disequilibrium $\mathcal{L}(x_j^i, x_k^i)$, and SNP-trait association $\mathcal{G}(\cdot)$ into a graph. Since $\mathcal{F}(x_j^F, x_j^M, x_j^C)$, $\mathcal{L}(x_j^i, x_k^i)$, and $\mathcal{G}(\cdot)$ inherently incorporate a set of variables as arguments, they can act as set of factor nodes, and known and unknown variable nodes can act as variable nodes.

The factor graph can be constructed with 5 types of nodes: *SNP variable node* — a known or unknown SNP; *trait variable node* — a known or unknown trait; *familial factor node* — the probability distribution of a set of SNPs derived from Mendel's Laws; *LD factor node* — the probability distribution of a pair of SNPs derived from linkage disequilibrium; *association factor node* — the probability distribution of a set of SNPs and traits derived from the SNP-trait association released by GWAS. Then, we can link a factor node and a variable node in the following manners:

- Each SNP variable node x_j^i connects to its own familial factor nodes f_j^i. Moreover, SNP variable nodes x_j^i and x_j^k $(k \neq i)$ are connect to familial factor node $\mathcal{F}_j^i$ if k is the father or mother of i.
- SNP variable nodes x_j^i and x_k^i are connected to LD factor node $\mathcal{L}_{j,k}^i$ if SNP j is in linkage disequilibrium with SNP k.
- Both SNP variable node x_j^i and trait variable node y_k^i are connected to association factor node $\mathcal{G}_{j,k}^i$, if the value of trait k is highly determined by the value of SNP j.

For instance, a factor graph for a family with three members (father, mother, and child) is shown in Fig. 1. As a simple case, Fig. 1 shows that a factor graph with 1 trait and 3 SNP variables. From Fig. 1, we observe that trait $t_1^i (i = 1, 2, 3)$ is associated with x_1^i.

After the message-passing procedure by executing belief propagation on a factor graph, the joint distribution $\Pr(X_U | \mathcal{X}_K, \mathcal{Y}_K, \mathcal{F}(x_j^F, x_j^M, x_j^C), \mathcal{L}(x_j^i, x_k^i), \mathcal{G}(\cdot))$ can be factorized into the product of a set of local functions and each local function takes a small subset of variables as arguments:



(a)

(b)

Fig. 1 Factor graph representation of a family with three members *M*, *F*, and *C* with 3 SNPs and 1 trait each individual. (a) Familial relationship. (b) A factor graph integrates all variables and data correlations.

$$\Pr(X_U | \mathcal{X}_K, \mathcal{Y}_K, \mathcal{F}(x_j^F, x_j^M, x_j^C), \mathcal{L}(x_j^i, x_k^i), \mathcal{G}(\cdot)) = \frac{1}{Z}\left[\prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{X}} f_j^i(x_j^i, \Theta(x_j^i), \mathcal{F}(x_j^F, x_j^M, x_j^C))\right] \times \left[\prod_{i \in \mathcal{V}} \prod_{(j,k)\ \text{s.t.}\ \mathcal{L}_{j,k} \neq 0} l_{j,k}^i(x_j^i, x_k^i, \mathcal{L}(x_j^i, x_k^i))\right] \times \left[\prod_{i \in \mathcal{V}} \prod_{j \in \mathcal{X}} g_{j,k}^i(x_j^i, \Lambda(x_j^i), \mathcal{G}(\cdot))\right] \tag{2}$$

where Z ia a normalization factor, $\Theta(x_j^i)$ represents the set of SNP j of all family members of i, and $\Lambda(x_j^i)$ represents the set of traits of individual i which are correlated with j, according to SNP-trait association.

As a message-passing algorithm, belief propagation iteratively passes messages between factor nodes and variable nodes, where the message is the "belief" of the probability distribution of a variable node, in terms of correlated variables. We denote the messages passed from SNP variable nodes to factor nodes as μ and the message passed from trait variable nodes to factor nodes as τ. The message passed from familial factor nodes to variable nodes is defined as ν, while the message passed from LD factor nodes to variable nodes is denoted by λ, and the message passed from association factor nodes to variable nodes is defined as β.

The message passed from SNP variable nodes to neighboring factor nodes $\mu_{v_s \to s}^T ({x_j^i}^T = j_l)$ represents the probability of $x_j^i = j_l$ at the T-th iteration, where j_l is the l-th possible value for SNP j. Likewise, the message passed from trait variable node to neighboring node $\tau_{v_t \to s}^T ({y_j^i}^T = k_l)$ represents the probability of $y_j^i = k_l$ at the T-th iteration, where k_l is the l-th possible value for trait j. Furthermore, $\nu_{f \to v}^T ({x_j^i}^T = j_l)$ represents the probability of $x_j^i = j_l$ at the T-th iteration, given Mendel's Laws $\mathcal{F}(\cdot)$. Likewise, $\lambda_{ld \to v}^T (x_j^i = j_l)$ is the probability of ${x_j^i}^T = j_l$ at the T-th iteration, given linkage disequilibrium $\mathcal{L}$ between any pair of SNPs. Finally, $\beta_{a \to v}^T ({x_j^i}^T = j_l)$ is the probability of $x_j^i = j_l$ at the T-th iteration, given the probability distribution of correlated traits of SNP j, when the message sent from SNP factor nodes to trait variable nodes. Likewise, $\beta_{a \to v}^T ({x_j^i}^T = j_l)$ represents the probability of ${y_j^i}^T = j_l$ at the T-th iteration, given the probability distribution of correlated SNPs of trait j, when the messages sent from trait factor nodes to SNP variable nodes.

In the procedure of message passing, a variable node v passes messages to one of its neighbor factor nodes s. The passed message is $\mu^T_{v\to s}({z^i_j}^T)$ by computing the product of all messages it receives from its neighbor factor nodes except s, where s can be $\mathcal{F}(\cdot)$, $\mathcal{L}(\cdot)$, and $\mathcal{G}(\cdot)$, and $z \in \{x, y\}$. The above computation is valid when z^i_j is an unknown variable. However, when $z^i_j \in \mathcal{X}_K$ or $z^i_j \in \mathcal{Y}_K$, for example, we observe $x^i_j = Bb$, then $\mu^T_{v\to s}({x^i_j}^T = Bb) = 1$, $\mu^T_{v\to f}({x^i_j}^T = BB) = 0$, and $\mu^T_{v\to f}({x^i_j}^T = bb) = 0$. Furthermore, familial factor node f sends messages to its neighbor variable node v by computing the product of all assages from f' neighbor variable nodes except v, and then multiplying the factor f with the obtained results; finally sum the messages from all the neighbors of f excluding v.

Initially, since factor nodes do not hold any information about their neighbors, each variable node passes messages to its neighbor factor nodes. At the first iteration, each known variable node z^i_j passes the message $\mu^1_{v\to s}({z^i_j}^1) = 1$ for every possible SNP or trait values, since currently there is no available information for $z^i_j \in \mathcal{X}_U$ or $\mathcal{Y}_U$. However, for a known variable with $z^i_j = \rho$, send message $\mu^1_{v\to s}({z^i_j}^T = \rho) = 1$ and $\mu^1_{v\to s}({z^i_j}^T = \rho') = 0$ for each possible SNP or trait value ρ' except ρ. Until the value of all unknown SNPs and traits is converged after several rounds of iteration (or the content of passed massages are converged), the iteration procedure can be stopped.

After the procedure of message passing, the marginal distribution of unknown SNPs and traits is computed by multiplying all passed messages to the unknown variable.

Since the messages encode the conditional probability distribution of SNPs and traits, we need to figure out the content of each message. Firstly, the prevalence rate of a trait can be collected from CDC[22] as a prior knowledge. Next we need to compute the conditional probability of an SNP s_i on its correlated traits. Therefore, we turn to compute the probability of the nucleotide on an SNP location, conditional on one of its correlated traits. For a specific SNP location on the human DNA sequence, there are two nucleotides: non-risk allele and risk allele. For a specific human trait k, GWAS studies its properties by dividing the volunteers into two groups: one is case group (all volunteers carry k) and one is control group (all volunteers do not carry k). Through comparing the DNA sequence between case and control groups, GWAS can identify the most common nucleotide in case group that indicates if one individual carries this nucleotide, there exists high probability to carry k for this individual. Such nucleotide in DNA sequence is called risk allele of trait k, while the other nucleotide in a specific SNP location is called non-risk allele. Table 1 shows the conditional probability of the risk allele and non-risk allele of SNP j on the probability of a correlated trait k.

Based on the conditional probability in Table 1, it is trivial to compute the conditional probability of an SNP on one of the correlated traits.

Table 2 shows the probability dependency between an SNP j and one of its correlated traits. Likewise, the probability distribution of a trait conditioned on one of its correlated SNP can be trivially computed from Table 2 based on the Bayesian posterior probability.

5 Differentially Private Genomic Data Publishing

In this section, we propose an approach to publish a genomic dataset $\mathcal{D}(\mathcal{V}, \mathcal{X}, \mathcal{Y})$ with differential privacy guarantee. As discussed in Ref. [18], publishing high-dimensional dataset generally arises two key challenges: scalability and signal-to-noise ratio. Given a genomic dataset with tens of millions of genomes, how to achieve a tradeoff between privacy and utility becomes a serious problem. We study the problem above by adopting differential privacy to sanitize the huge-dimensional dataset. To mitigate the gaps, we make use of a Bayesian network[23] that is a probabilistic graphical model to describe the conditional independence among a set of variables. A Bayesian network is a directed acyclic graph in which each node represents a variable and

Table 1 Probability of risk allele r^k_j and non-risk allele ρ^k_j on the SNP j' position conditioned on one of the associated trait k of SNP j. Here $p^{k^o}_j$ and $p^{k^a}_j$ represent the risk allele frequency in control and case group, respectively.

Allele	k (with trait k)	$\bar{k}$ (without trait k)
r^k_j	${p^k_j}^a$	${p^k_j}^o$
ρ^k_j	$1-{p^k_j}^a$	$1-{p^k_j}^o$

Table 2 Probability of genotype of SNP j ($r^k_j r^k_j$, $r^k_j \rho^k_j$, and $\rho^k_j \rho^k_j$) conditioned on one of its correlated traits k.

Genotype of SNP j	k (with trait k)	$\bar{t_k}$ (without trait k)
$r^k_j r^k_j$	$\sqrt{{p^k_j}^a}$	$\sqrt{{p^k_j}^o}$
$r^k_j \rho^k_j$	${p^k_j}^a(1-{p^k_j}^a)$	${p^k_j}^a(1-{p^k_j}^o)$
$\rho^k_j \rho^k_j$	$\sqrt{1-{p^k_j}^a}$	$\sqrt{1-{p^k_j}^o}$

the edge between two nodes represents the conditional probability of one variable on the other variables. Actually, we can transfer a factor graph to a Bayesian network easily. Compared with private data releasing algorithms in Ref. [24, 25], generating synthetic data with differential privacy guarantee can provide provable privacy guarantee.

Therefore, our proposed approach works in three steps:

(1) Construct a factor graph $\mathcal{N}$ over the SNPs and traits in $\mathcal{D}(\mathcal{V}, \mathcal{X}, \mathcal{Y})$, using an $(\varepsilon/2)$-differentially private approach.

(2) Develop an $(\varepsilon/2)$-differentially private approach to generate the conditional distributions of (a) each pair of SNPs (from linkage disequilibrium $\mathcal{L}$); (b) set of SNPs and their correlated traits (from SNP-trait association $\mathcal{G}$), and (c) set of SNPs (from familial relationship $\mathcal{F}$).

(3) Use the factor graph $\mathcal{N}$ constructed and the set of noise conditional distributions to compute an approximate joint distribution to sample a synthetic genomic dataset $\mathcal{D}^*$.

According to Eq. (2), we define the number of familial factor nodes $f_j^i(x_j^i, \Theta(x_j^i), \mathcal{F}(x_j^F, x_j^M, x_j^C))$, LD factor nodes $l_{j,k}^i(x_j^i, x_k^i, \mathcal{L}(x_j^i, x_k^i))$, and association factor node $g_{j,k}^i(x_j^i, \Lambda(x_j^i), \mathcal{G}(\cdot))$ as m, n, and k, respectively. For each familial factor node $f_j^i(\cdot)$, the scale of Laplace noise injected to it is $\frac{9}{m\varepsilon}$, in order for guaranteeing $(\varepsilon/3)$-differential privacy, since the sensitivity of $f_j^i(\cdot)$ is $\frac{3}{m}$. For each LD factor node $l_{j,k}^i(\cdot)$, we inject Laplace noise to it with scale $\frac{6}{n\varepsilon}$, in order for guaranteeing $(\varepsilon/3)$-differential privacy, since the sensitivity of $g_{j.m}^i(\cdot)$ is $\frac{2}{n}$. Likewise, for each connection factor node $g_{j,k}^i(x_j^i, \Lambda(x_j^i), \mathcal{G}(\cdot))$, we inject Laplace noise to it with scale $\frac{6}{k\varepsilon}$, in order for guaranteeing $(\varepsilon/3)$-differential privacy, since the sensitivity of $g_{j.m}^i(\cdot)$ is $\frac{2}{k}$. According to the compensability property of differential privacy, our method satisfies ε-differential privacy.

6 Conclusion

In this paper, we propose an SNP and trait inference model to predict the value of unknown SNPs and traits based on probabilistic graphical models and belief propagation. Our approach has linear computation complexity by integrating SNP-trait associations, linkage disequilibrium, and Mendel's Laws into a factor graph then executing belief propagation on it to compute the marginal probability distribution of unknown SNPs or traits. We also propose an approach to publish high-dimensional genomic data with differential privacy guarantee. We make use of Bayesian networks to find an approximate distribution of the input genomic dataset, then inject differential privacy noise into the approximate distribution. Finally, synthetic genomic dataset is sampled based on the noisy distribution of input genomic dataset.

References

[1] 23andMe, Find out what your DNA says about you and your family, https://www.23andme.com/, 2020.

[2] openSNP, https://opensnp.org/, 2020.

[3] Patientslikeme, https://www.patientslikeme.com/, 2020.

[4] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, Identifying personal genomes by surname inference, *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[5] L. Sweeney, A. Abu, and J. Winn, Identifying participants in the personal genome project by name (A re-identification experiment), arXiv preprint arXiv: 1304.7605, 2013.

[6] E. Ayday, E. De Cristofaro, J. P. Hubaux, and G. Tsudik, The chills and thrills of whole genome sequencing, arXiv preprint arXiv: 1306.1264, 2013.

[7] New York Times, The immortal life of Henrietta Lacks, the sequel by Rebecca Skloot, https://www.nytimes.com/2013/03/24/opinion/sunday/the-immortal-life-of-henrietta-lacks-the-sequel.html, 2013.

[8] Z. B. He, Z. P. Cai, Y. C. Sun, Y. S. Li, and X. Z. Cheng, Customized privacy preserving for inherent data and latent data, *Personal and Ubiquit. Comput.*, vol. 21, no. 1, pp. 43–54, 2017.

[9] GWAS Catalog, The NHGRI-EBI catalog of human genome-wide association studies, https://www.ebi.ac.uk/gwas/docs/about, 2020.

[10] D. R. Nyholt, C. E. Yu, and P. M. Visscher, On Jim Watson's APOE status: Genetic information is hard to hide, *European Journal of Human Genetics*, vol. 17, no. 2, pp. 147–149, 2009.

[11] D. S. Falconer and T. F. C. Mackay, *Introduction to Quantitative Genetics*, *4th ed.* Harlow, UK: Longmans, 1996.

[12] Y. Yu, M. Li, L. L. Liu, Y. H. Li, and J. X. Wang, Clinical big data and deep learning: Applications, challenges, and future outlooks, *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 288–305, 2019.

[13] S. Kumar and M. Singh, Big data analytics for healthcare industry: Impact, applications, and tools, *Big Data Mining and Analytics*, vol. 2, no. 1, pp. 48–57, 2019.

[14] C. Dwork, Differential privacy, in *Proc. 33rd Int. Colloquium on Automata*, *Languages and Programming*, Venice, Italy, 2006, pp. 1–12.

[15] F. R. Kschischang, B. J. Frey, H. A. Loeliger, Factor

graphs and the sum-product algorithm, *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[16] F. R. Kschischang, B. J. Frey, H. A. Loeliger, Factor graphs and the sum-product algorithm, *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, 2001.

[17] B. Liu, S. Feng, X. Guo, and J. Zhang, Bayesian analysis of complex mutations in HBV, HCV, and HIV studies, *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 145–158, 2019.

[18] X. Ding and X. Guo, A survey of SNP data analysis, *Big Data Mining and Analytics*, vol. 1, no. 3, pp. 173–190, 2018.

[19] Z. B. He, J. G. Yu, J. Li, Q. L. Han, G. C. Luo, and Y. S. Li, Inference attacks and controls on genotypes and phenotypes for individual genomic data, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 17, no. 3, pp. 930–937, 2020.

[20] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. K. Xiao, PrivBayes: Private data release via Bayesian networks, in *Proc. 2014 ACM SIGMOD Int. Conf. on Management of Data*, Snowbird, UT, USA, 2014, pp. 1423–1434.

[21] F. McSherry and K. Talwar, Mechanism design via differential privacy, in *Proc. 48th Ann. IEEE Symp. on Foundations of Computer Science*, Providence, RI, USA, 2007, pp. 94–103.

[22] Centers for Disease Control and Prevention, Hypertension, https://www.cdc.gov/nchs/fastats/hypertension.htm, 2018.

[23] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[24] Z. B. He, Z. P. Cai, Q. L. Han, W. T. Tong, L. M. Sun, and Y. S. Li, An energy efficient privacy-preserving content sharing scheme in mobile social networks, *Personal Ubiquit. Comput.*, vol. 20, no. 5, pp. 833–846, 2016.

[25] Z. B. He, Z. P. Cai, and J. G. Yu, Latent-data privacy preserving with customized data utility for social network data, *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 665–673, 2018.

Zaobo He received the PhD degree from Georgia State University in 2018. He received the MS and BS degrees from Shaanxi Normal University and Yanan University in 2014 and 2011, respectively. He is currently an assistant professor at Department of Computer Science and Software Engineering, Miami University. His research interests include data privacy, big data analytics, and cybersecurity.

Junxiu Zhou received the PhD degree from University of Arkansas at Little Rock in 2018. She is currently an assistant professor at Department of Computer Science, Northern Kentucky University. Her research interests include machine learning and deep learning, computer vision, image processing, and network optimization.