Information Sciences 512 (2020) 207–218

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Inference attacks on genomic privacy with an improved HMM and an RCNN model for unrelated individuals☆

Hongfa Ding[a,b], Youliang Tian[c,*], Changgen Peng[a,d,*], Youshan Zhang[e], Shuwen Xiang[a]

[a] *State Key Laboratory of Public Big Data, College of Mathematics and Statistics, Guizhou University, Guiyang 550025, China*
[b] *College of Information, Guizhou University of Finances and Economics, Guiyang 550025, China*
[c] *College of Computer Science and Technology, Guizhou University, Guiyang 550025, China*
[d] *CETC Big Data Research Institute Co.,Ltd., Guiyang 550025, China*
[e] *Department of Computer Science and Engineering, Lehigh University, Bethlehem 18015, USA*

ARTICLE INFO

Article history:
Received 28 September 2018
Revised 3 September 2019
Accepted 21 September 2019
Available online 1 October 2019

MSC:
00-01
99-00

Keywords:
Genomic privacy
Inference attack
Privacy quantification
Hidden Markov model
Regressive convolutional neural network

ABSTRACT

In recent years, the collection of large-scale genomic data for individuals has become feasible and affordable. Concurrently, several practical attacks targeting genome reidentification and genotype inference have emerged to threaten the confidentiality of genomic data sharing, leading to security and privacy concerns regarding genomic data. The authors have shown that this problem can be even worse in this paper. Specifically, two possible large-scale genotype inference attack stretegies for nonrelatives have exposed. One is based on an improved hidden Markov model (iHMM), and the other is based on a regressive convolutional neural network (RCNN). By using a genomic privacy metric combining the attacker's incorrectness, the attacker's uncertainty, and the genomic privacy loss of the victims, it is shown that with these atrategies, the attack can be significantly more severe than those reported previously. It is also shown that machine learning can be applied to empower large-scale inference attacks against genomic privacy.

1. Introduction

With advances in sequencing technology, people are now able to sequence their DNAs more easily and cheaply. Human genomic data have become increasingly affordable and available. For example, in the 1000 Genomes Project [36], thousands of anonymous participants have donated their DNA to biomedical and precision medicine research. The governments of the United States, the United Kingdom, Canada, France, and China have also started genomic data collection programs for

☆ Funding: This work was supported by the National Natural Science Foundation of China [U1836205, 61662009 and 61772008], the Open Foundation of Guizhou Provincial Key Laboratory of Public Big Data [2017BDKFJJ023 and 2017BDKFJJ026], the Science and Technology Foundation of Guizhou [Guizhou-Science-Contract-Major-Program [2018]3001, Guizhou-Science-Contract-Major-Program [2018]3007, Guizhou-Science-Contract-Major-Program [2017]3002, Guizhou-Science-Contract-Support [2019]2004, Guizhou-Science-Contract-Support [2018]2162, Guizhou-Science-Contract-Support [2018]2159, Guizhou-Science-Contract[2017]1045 and Guizhou-Science-Contract[2019]1049], the Youth Science and Technology Talents Growth Project of the Guizhou Provincial Department of Education [Guizhou-Education-Contract-KY-Word [2016]171], and the Innovation Fund for Graduate Students of Guizhou University [Yan-Ligong2017068].

* Corresponding authors at: State Key Laboratory of Public Big Data, College of Mathematics and Statistics, Guizhou University, Guiyang 550025, China.
E-mail addresses: hongfa.ding@foxmail.com (H. Ding), youliangtian@163.com (Y. Tian), peng_stud@163.com (C. Peng), yoz217@lehigh.edu (Y. Zhang), shwxiang@vip.163.com (S. Xiang).

https://doi.org/10.1016/j.ins.2019.09.036
0020-0255/© 2019 Elsevier Inc. All rights reserved.

medicine and other reasons. Furthermore, an increasing number of people are sharing their genomic data online through sites such as 23andMe.com, PatientsLikeMe.com and Ancestry.com, either for fun or to find relatives. On the other hand, an individual can be identified by his or her DNA. Genomic data can also be used to identify specific traits and diseases. However, the increasing availability of genomic data introduces more prominent security and privacy challenges. Once these data is disclosed or misused, a person may face the risk of discrimination in many aspects including employment, insurance, and education [41].

Indeed, many research results and real-world cases have already raised concerns about the confidentiality and privacy of genomic data. In some cases, ananymousely collected genomic data can still introduce the exposure of sensitive information of individuals in various ways. For example, individuals were reidentified by Sweeney et al. [35] via linking names and contact information in publicly available profiles in the Personal Genome Project and by Gymrek et al. [9] via the profiling of short tandem repeats on the Y chromosome. The results of a genome-wide association study (GWAS) can be used to identify individuals [3]. Predisposition to certain diseases [33] and appearance characteristics [43] can also be inferred from genomic data. The disclosure of an individual's genomic data can cause threats to not only his or her own privacy but also the privacy of his or her relatives in the form of familial identification [27] or information on the genotypes of relatives [17]. Recently, it has been confirmed that geneticists can recover a specific individual's face from his or her genomic data [12], It has also possible that shared genomic data can be misused by malicious institutions [30].

The situation can be even worse. To protect an individual's own genomic privacy, typically he or she may choose to delete or hide certain parts of his or her genotype [31] and only share partial genomic data to third parties (e.g., hospitals or genome research institutions). It seems safe that many people without relative relationship can share their genomic data in this way. However, that is not effective. In this work, we will show that an adversary can robustly reconstruct an individual's genomic data from the shared partial genomic data of that individual and other publicly available genomic data.

In this paper, we will reveal two inference attack strategies for reconstructing an individual's genotype sequence: one is based on an improved discrete hidden Markov model (iHMM) and the other is based on a regressive convolutional neural network (RCNN) model. These inference attack models take both the observed genomic data of the victims and publicly available genomic data into consideration. We will also propose metrics to quantify the genomic privacy of victims and the severity level of an attack with regard to incorrectness, uncertainty and privacy loss. Compared with previous work by Samani et al. [28], our contributions are as follows:

- We will propose a unified adversary model for inference attacks on genomic privacy with the aim to reconstruct the genotype sequences of unrelated individuals from partially observed genomic data of the victims.
- We will present a genomic privacy inference attack strategy on unrelated individuals that employs the correlations of single nucleotide polymorphisms (SNPs) and the sampling and recombination model method in IMPUTE2 [15].
- We will reveal a genomic privacy inference attack strategy on unrelated individuals that employs an RCNN and investigate the large-scale capabilities of machine learning (e.g., RCNNs) in the context of genomic privacy attacks.
- We will evaluate the inference attack capability and quantify genomic privacy in terms of mutual information, which represents the decrement in the information uncertainty of the attacker and the increment in the privacy loss of the victim to the attacker.
- Compared with previous work, our results will be obtained with much higher accuracy, lower uncertainty on the inferred genomic data, and more loss of private information to the attacker.

The remainder of this paper is organized as follows. First, a brief summary on related previous works will be inluded in Section 2. Following that, some basic background to our work will be introduced in Section 3. An adversary model and evaluation metrics will be discussed in Section 4. The framework and details of the inference attack atrategies will be give in Section 5. The proposed inference attack strategies will be evaluated together with an introduction of a metric to quantify the privacy of genomic data in Section 6. Finally, Section 7 is used to conclude this paper.

2. Related work

2.1. Inference attacks on genomic privacy

Inference attacks, which employ available data to infer latent private information through data analysis [6], are very powerful atrategies of privacy and security attacks. Inference attacks have been widely studied in the contexts of location tracking [22], attribute privacy on social networks [8], membership and property privacy in machine learning [7,32], the vulnerability of advanced cryptography (e.g., encrypted databases and searchable encryption) [25] and genomic privacy (e.g., membership genomic privacy [44], genotype privacy [11,28] and kinship privacy [17]). As discussed in [1], inference attacks pose an enormous privacy threat to genomic data in the areas of social networking, genome sharing, GWAS research and clinical medicine.

In this work, we focus on how genotype privacy can be compromised in inference attacks based on shared SNP data of the victims (in which the sensitive SNPs are hidden) and publicly available genomic data.

2.2. Privacy breaches of genomic data

Although many works in the literature have addressed statistical genomic privacy breaches, most of them have been concerned with identification privacy and relied on pairwise linkage disequilibrium (LD). Following the genetic privacy studies on GWAS statistics conducted by Homer et al. [13], which showed that a GWAS participant's disease status can be inferred from his or her genotype, people start to consider donating genomic data for GWAS research and medical testing. Subsequently, deidentification are considered as insufficient to protect genetic privacy and confidentiality. For many public-domain databases, such as the database of Genotypes and Phenotypes (dbGaP) of the National Institutes of Health (NIH) [20,39], access rules have been changed to be controlled based on their genomic data. Wang et al. [44] have suggested that individuals' identities and diseases could be inferred based on GWAS results. Even though the data in public GWAS catalogs are differentially private, it can still contain the personal traits and identities of GWAS participants, and regular individual's privacy can be attacked through mining based on background information [45]. By using public trait loci and phenotype datasets, the genetic privacy of individuals can also be compromised by linking phenotypes to genotypes [10].

This paper focuses on the privacy of genotypes rather than that of identity [44,45] or disease status [13,44] based on genomic data. Although our work is also reply on publicly available genomic data, we do not require trait loci and phenotype data as in [10,44,45]. We only need the observed SNP sequences donated by individuals through gene sharing websites (e.g., PatientsLikeMe.com and 23andMe.com) and public genomic data from genetic research projects (e.g., the HapMap Project and the 1000 Genomes Project).

In [29], the authors have proposed a Bayesian method to predict individuals' genotypes at particular loci from gene expression data. Humbert et al. [17] have proposed a genotype inference attack by making use of both familial relationships and pairwise LD. Samani et al. [28] have explored genotype inference attacks on unrelated individuals with various higher-order single nucleotide variant (SNV) correlation models. An approach combining and extending the work of Humbert et al. [17] and Samani et al. [28] to infer family members' genotypes has been proposed in [4]. In this paper, we aim to infer the genotypes of large-scale SNP sequences for unrelated individuals rather than probing genotypes at particular loci [29] or kin genomic privacy, as in [17] and [4]. The inference attacks presented in this paper are designed for the same scenario considered in [28], namely, to affirm the privacy problems that arise when people publish their genomic data online. Comparing to the work reported in [28] make use of published genetic sequences with partially hidden information, publicly available reference panels and other genomic information, the attack model introduced in this paper is improved in both performance and methodology. The iHMM-based inference attack model to be discussed in this paper is an improvment of the recombination model-based inference attack presented in [28], by separating the inference of the genotypes of hidden SNPs into steps rather than inferring the genotypes directly. In such attack model, we combine the Markov chain Monte Carlo sampling strategy with an HMM to compute conditional distributions, which will significantly improve the attack power. Furthermore, the RCNN-based inference attack model proposed is based on a novel model for genotype reconstruction. Although machine learning has been widely applied in genomics study [18], very few works of this type have been reported to address genomic privacy issue. We take the initiative to apply RCNNs to the large-scale inference of the genotypes of hidden SNPs and the quantification of genomic privacy.

3. Background

In this section, we briefly present some information about genomics, HMMs, and RCNNs.

3.1. Genomics

A brief overview of the human genome is presented in Fig. 1 [28]. Human beings have 23 chromosome pairs. A human genome is encoded as DNA and includes approximately 3 billion nucleotide pairs. Each chromosome has a double-helix structure and consists of two complementary polymer chains of nucleotides (A, T, G, and C). Human beings can be identified by their DNA; 99% of human DNA is shared in common among all individuals, with only 0.5% differing among the genomes of different individuals. Human genomes are coded by different alleles (A, T, C, and G). The group of alleles on one chromosome is called the haploid genotype, and the group of allele pairs on a pair of chromosomes is called the diploid genotype [5].

A single nucleotide polymorphism (SNP) is a variation in a single nucleotide that occurs at a specific position in the genome. Each such variation is present to some appreciable degree within a population. By contrast, a single nucleotide variant (SNV) is a variation in a single nucleotide without any limitations on frequency. The SNP sequences of a given individual are very different from those of others. As such, an individual can be identified by his or her SNPs. SNPs are associated with certain traits and diseases, and a genome-wide association study (GWAS) is an observational study of SNPs of different individuals conducted to determine whether a given SNP is associated with a particular trait or disease.

For convenience, let the three possible states (i.e., AA, Aa and aa, respectively) of each SNP to be presented by 0, 1, or 2, depending on the number of minor alleles at each gene locus.

Linkage disequilibrium (LD) is defined as a correspondence or nonrandom association of alleles at two or more loci. Such an association is the consequence of the inheritance mechanism: given sufficient time for evolution, the occurrences

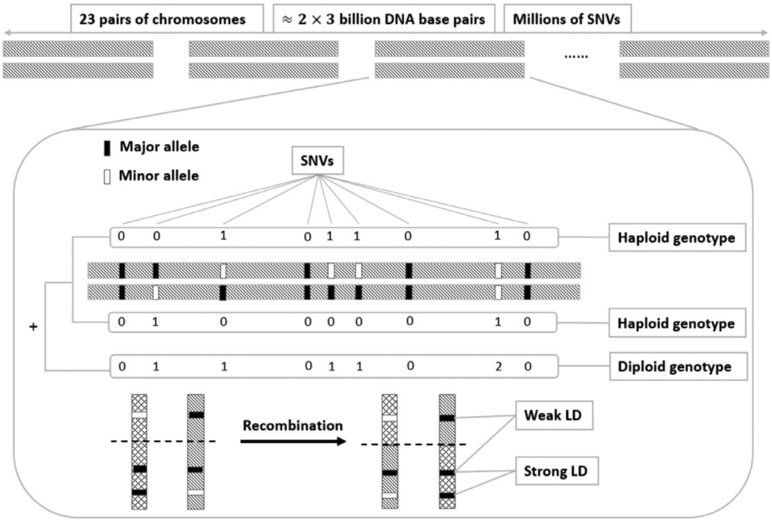

Fig. 1. Brief overview of the human genome [28].

of random recombination will produce an equilibrium distribution of alleles at all loci. There are several methods of LD modeling. We will focus on a mixed method that considers both a reference genotype dataset and a recommendation rate.

In the inheritance process, recombination is a subprocess in which some pieces of DNA are separated apart and recombined to form new combinations of alleles. The recombination process results in genetic diversity for all creatures. Recombination is intuitively related to LD.

3.2. HMMs

A hidden Markov model (HMM) [26,34] is a statistical Markov model with unobserved states, which can be represented by a simple dynamic Bayesian network. Specifically, three assumptions are adopted in our discussion: (1) the state at time t is generated by some process whose state S_t is hidden, (2) the process exhibits the Markov property, and (3) the hidden state variable is discrete.

HMMs can be used to characterize fundamental problems regarding likelihood, decoding and learning. At present, HMMs are widely applied in many areas, including speech recognition [26], handwriting recognition [16], and gene prediction [5].

The problem considered in this paper is similar to a parameter learning problem to some extent. Since the posterior marginals of all hidden state variables can be obtained through the computation of inference process given a sequence of observations/emissions, a forward-backward algorithm falls into our consideration.

3.3. RCNNs

Convolutional neural networks (CNNs) [19,30] have recently become a popular method of solving image classification, segmentation, and regression problems. However, regressive CNN (RCNN) architecture, a CNN in which the final layer is a regression layer, has yet to be developed for predicting genotype sequences. Different from traditional classification and segmentation problems that their output are discrete values [30], that of an RCNN is continuous.

In this work, we design an RCNN architecture for haplotype sequence forecasting, similar to that for missing value prediction. First, a public haplotype dataset is used to train and test our RCNN model. After the model is established, it is applied to attack the genotype sequences of individuals by phasing the observed genotypes into haplotypes and inferring the genotypes of the hidden SNPs.

4. Adversary model and quantitative evaluation metrics

4.1. Adversary model

We consider scenarios involving genomic data sharing in the real world. In such a scenario, a victim donates his or her SNP sequence for research, medical testing or finding relatives. Because of privacy concerns, the victim wants to hide certain sensitive SNPs that may be related to genetic diseases or private traits. Therefore, the victim shares a variation of his or her original SNP sequence $\hat{X} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n)$ (where $\hat{x}_i = \{0, 1, 2\}$) in which some of the SNPs are hidden. Let the hidden SNPs be denoted by X_H, the observable SNPs by X_O, and the published SNPs by $X = (x_1, x_2, \ldots, x_n) = X_H \cup X_O$ (where $x_i = \{-1, 0, 1, 2\}$, with a value of $x_i = -1$ indicating that $x_i \in X_H$ is a hidden SNP). Suppose that an adversary who can observe

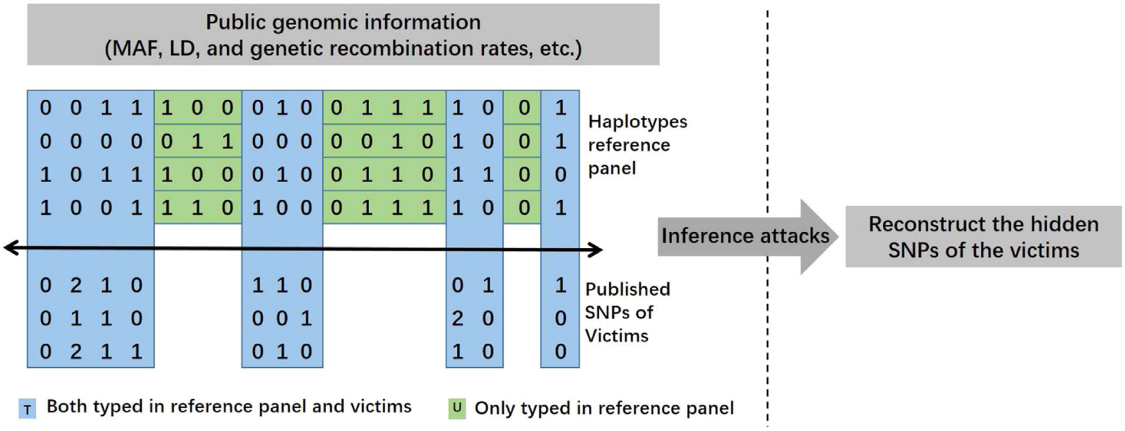
Public genomic information
(MAF, LD, and genetic recombination rates, etc.)
0 0 1 1 1 0 0 0 1 0 0 1 1 1 1 0 0 1
0 0 0 0 0 1 1 0 0 0 0 0 1 0 1 0 0 1
1 0 1 1 1 0 0 0 1 0 0 1 1 0 1 1 0 0
1 0 0 1 1 1 0 1 0 0 0 1 1 1 1 0 0 1
Haplotypes reference panel
0 2 1 0 1 1 0 0 1 1
0 1 1 0 0 0 1 2 0 0
0 2 1 1 0 1 0 1 0 0
Published SNPs of Victims
Inference attacks
Reconstruct the hidden SNPs of the victims
T Both typed in reference panel and victims
U Only typed in reference panel

Fig. 2. Overview of the adversary model. The adversary infers the hidden genotypes of the victims from their partially observed genotypes in combination with publicly available genomic data (a haplotype reference panel, MAFs, LD values, recombination rates, etc.).

the published SNPs X of the victim wants to reconstruct the original SNPs $\hat{X}$. For this purpose, the adversary can intrude on the genomic privacy of the victim (e.g., obtain his or her APOE gene status [23]) by means of an inference attack. To run such an inference attack, the adversary will collect some publicly available genomic information [14,37], such as minor allele frequencies (MAFs), LD values, genetic recombination rates and a haploid genotype reference panel for the population to which the victim belongs; see Fig. 2 for a visual summary.

Let the accessible public genomic information be denoted by $INFOR_{Pub}$ and the inferred SNPs by $\bar{X} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$. The adversary model of the inference attack *infer* on genotype privacy can be formally represented as

$$\begin{aligned}\bar{X} =& infer(X, INFOR_{Pub})\\ =& infer(X_H, H_O, INFOR_{Pub}).\end{aligned} \tag{1}$$

More specifically, the inference attack can be regarded as the process of computing the conditional marginal probability distribution of each hidden SNP given the published SNPs and the public genomic information, namely,

$$Prob(X = \{0, 1, 2\}) = Prob(X|(X_O, INFOR_{Pub})). \tag{2}$$

For each hidden SNP, the predicted value is the one with the highest conditional probability.

4.2. Quantitative evaluation metrics

To measure the capability of the adversary with regard to genomic privacy inference, it is necessary to employ a *genomic privacy metric*, as introduced by Ayday et al. [2], to evaluate the extent to which the adversary can compromise a victim's genomic privacy through an inference attack. As discussed by Wagner [42], several types of genomic privacy metrics are suitable to our discussion. In this paper, we assume that the adversary aims to infer the values of SNPs, and we consider only the SNPs that the individuals realy have. We apply the *normalized incorrectness*(i.e., the attacker's incorrectness), the *normalized entropy* (i.e., the attacker's uncertainty) and the *normalized mutual information* (i.e., the privacy loss of the victim) to quantify the capabilities of our inference attack models.

As a genomic privacy metric, the *normalized incorrectness* can be expressed as

$$E = 1 - \frac{\sum_1^n \mid \bar{x}_j - \hat{x}_j \mid}{\mid X_H \mid}, \tag{3}$$

where n is the number of SNPs that belong to the victim, $\bar{x}_j$ is the inferred genotype value of the SNP at locus j, $\hat{x}_j$ is the original genotype value of the SNP at locus j, and $|X_H|$ is the number of hidden SNPs that belong to the victim.

Although the incorrectness is a powerful metric for privacy quantification, it is not suitable to many scenarios because the original values of the victim's SNPs are not available. In these cases, we need other metrics. Here, we adopt the *normalized entropy* metric to represent the attacker's uncertainty. This metric can be evaluated based on the normalized entropy of the inferred SNPs. Specifically,

$$H = \frac{\sum_{j=1}^n \frac{H(X_j)}{log(3)}}{\mid X_H \mid}, \tag{4}$$

where $H(X_j) = -\sum_{\bar{x}_j=\in\{0,1,2\}} p(\bar{x}_j) log(p(\bar{x}_j))$ represents the entropy of the inferred SNP at locus j, $log(3)$ is the maximum entropy of the SNP at locus j, and $|X_H|$ is the number of hidden SNPs belonging to the victim.

This metric quantifies the confidence of the adversary in his or her inference attack in terms of the capability of the adversary rather than the victim's privacy loss. To this end, we use the decrement in the uncertainty to represent the change in the attacker's uncertainty regarding the hidden SNPs before and after the inference attack. The concept of mutual information [24] can serve as the basis for such a metric. Therefore, we use the normalized mutual information to quantify the average privacy loss of the victim to the attacker. i.e.,

$$I = \frac{\sum_{j=1}^{n} \frac{H_{MAF}(X_j)}{log(3)}}{| X_H |} - H, \tag{5}$$

where $H_{MAF}(X_j) = -\sum_{x_j=\in\{0,1,2\}} p_{MAF}(x_j) log(p_{MAF}(x_j))$ represents the natural entropy of the SNP at locus j and $p_{MAF}(x_j)$ is the probability of the SNP according to the MAF dataset. The metric defined in Eq. (5) represents the entropy change due to the inference attack, and thus, can measure the capability of the inference attack. It can also evaluate the genomic privacy loss of the victim in the face of the inference attack.

5. Proposed inference attack stratgies

In this section, we propose two inference attack strategies for use in the suggested adversary model. One is based on an improved HMM (iHMM) and the other is based on an RCNN model.

5.1. IHMM-based attack

To improve the performance of genomic privacy inference, we choose not to infer the hidden SNP genotypes of the victims directly as in [28]. Instead, inspired by the genotype imputation method of IMPUTE2 [15], we divide the attack process into three steps: (1) phasing the observed SNPs of the victims into haplotypes using the Markov chain Monte Carlo sampling strategy, (2) separately inferring the hidden haplotype pairs of each victim using the HMM model, and (3) combining the inferred haplotype pairs of each victim to form the inferred genotype sequences.

In the detailed construction of our model, we divide the SNPs of the reference panel and the victims into T (the SNPs that appear in both the reference panel and the victims) and U (the SNPs that do not appear in the victims but do appear in the reference panel). Let H_R^T denote the set of reference haplotypes for the SNPs in T, H_V^T denote the set of the victims' observed haplotypes for the SNPs in T, and H_V^U denote the set of the victims' hidden haplotypes corresponding to the SNPs in U. We assume that there are n victims. Then, $H_V^T = \{H_{V,1}^T, H_{V,2}^T, \dots, H_{V,n}^T\}$ represents the victims' haplotypes corresponding to the SNPs in T, where $H_{V,i}^T$ denotes the haplotype pairs of the ith victim and ρ is the population-scaled recombination map.

More specifically, the iHMM-based inference attack can proceed in three steps, detailed as follows:

(1) The attacker randomly generates haplotypes in H_V^T in consistency with the observed genotypes of the victims. Then, the attacker updates the haplotypes in H_V^T through multiple Markov chain Monte Carlo iterations. In each iteration, the attacker updates the phased haplotype pairs $H_{V,i}^T$ for the ith victim by sampling from $P(H_{V,i}^T | G_{V,i}^T, H_{V,-i}^T, H_R^T, \rho)$.
(2) The attacker infers the haplotypes in H_V^U using the HMM via the gene recombination model. In each iteration, the attacker infers the hidden haplotype pairs $H_{V,i}^U$ of the ith victim corresponding to the SNPs in U from the conditional distribution $P(H_{V,i}^U | H_{V,i}^T, H_R^{T \cup U}, \rho)$.
(3) The attacker combines the inferred haplotype pairs of each victim to obtain the inferred genotypes of the hidden SNPs of the victims.

In the phasing step of each iteration in step (1), the sampling is conditional to the k closest haplotypes, and the results are determined by its Hamming distance to the ith victim. The conditional distribution is computed using the HMM model based on the recombination process, and the phase space is reconstructed via Monte Carlo method. Because the state space includes all states of the haplotypes in H_R^T and the current-guess haplotypes in $H_{V,-i}^T$, more information can be available.

In step (2), the HMM state space can include all reference haplotypes $H_R^{T \cup U}$. This step is similar to the process of the recombination-based model of [28], which was inspired by [21]. However, we infer the haplotype pairs of each victim instead of inferring the genotypes directly.

This attack strategy described in steps (1) to (3) is different from that presented in [28], in which the genotype values of the hidden SNPs are inferred directly. Here, the adversary combines both Markov chain Monte Carlo sampling and HMM inference techniques to improve the results obtained for the conditional distributions of the target SNPs.

5.2. RCNN-based attack

The RCNN-based attack is also separated into three steps, and steps (1) and (3) are the same as those of the iHMM-based attack. Only step (2) is different. Similarly, the attacker observes the public genomic information and the victims' SNPs, phases the genotypes into haplotypes, infers the hidden haplotype pairs separately, and then combines the inferred haplotype pairs into genotypes. Here, we illustrate step (2) of the RCNN-based attack as follows:

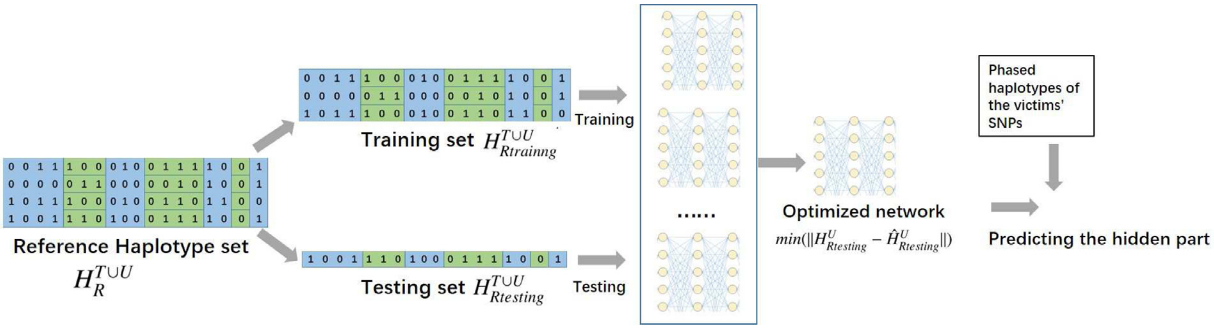

Fig. 3. Outline of the RCNN-based inference attack.

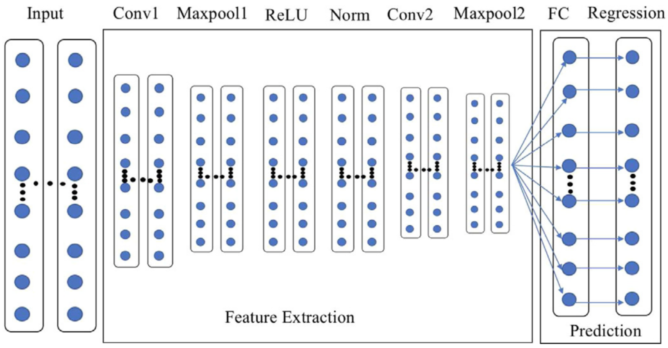

Fig. 4. The RCNN structure contains only eight layers, where the input layer is the haplotypes of the observed SNPs and the regression layer generates the haplotypes of the hidden SNPs. In the training stage, the RCNN will extract the features of the influential factors and check whether the MSE has converged. By using the trained RCNN classifier, we can predict the haplotypes of the hidden SNPs of the test data. (Conv: convolution layer, NA: normalization layer, FC: fully connected layer.) .

We model the objective of the RCNN as

$$H^{U}_{V,i} \leftarrow RCNN(H^{T}_{V,i}, H^{T\cup U}_{R}), \tag{6}$$

where given a reference haplotype set and a phased haplotype set based on the observed SNPs, the objective of Eq. (6) is to infer the values (i.e., 0 or 1) of the hidden parts.

Because both the public reference haplotypes and the observed SNPs of the victims belong to the same population (e.g., CEU or CHS [38]), these data have the same features for analysis by using a neural network.

We develop an RCNN model for the reference data $H^{T\cup U}_{R}$ by dividing these data into two sets: one is the training set $H^{T\cup U}_{Rtrain}$, and the other is the test set $H^{T\cup U}_{Rtest}$. We then choose the best trained network for the objective of $min(\|H^{U}_{Rtest} - \hat{H}^{U}_{Rtest}\|)$, where $\hat{H}^{U}_{Rtest}$ denotes the predicted values of the test set. The attacker can use this optimized network to infer the hidden values of the victims' haplotypes. Fig. 3 illustrates the outline of this process.

The resultant RCNN architecture can be shown in Fig. 4. The input consists of the haplotypes of the observed SNPs, and the last layer is a regression layer that represents the haplotypes of the hidden SNPs. This network can achieve two main tasks: feature extraction and prediction. It includes eight layers. Feature extraction is faciliatated by two convolution layers (Conv1 and Conv2), two max-pooling layers (Maxpool1 and Maxpool2), one rectified linear unit (ReLU) layer, and one normalization (Norm) layer. The ReLU layer reduces the number of epochs needed for training but consequently achieves a higher error rate comparing to traditional tanh units. The Norm layer increases the generalizability and reduces the error rate. Noticeably, the ReLU and Norm layers do not change the size of the feature map. The pooling layers summarize the outputs of adjacent pooling units. The prediction step is performed by a fully connected (FC) layer and a regression layer. The input layer comprises 8×1 influential factors (one month). Conv1 and Conv2 each has a filter size (F) of 1×1 and a number of filters (N) equal to 25, with a padding size (P) of 0. Maxpool1 and Maxpool2 each has a stride (S) of 2×2. Therefore, after each max-pooling layer, the dimensions of the feature map are divided by 2.

We minimize a loss function to train the RCNN model. We use the mean square error (MSE) as the loss function, which is defined as

$$\text{Loss} = \frac{1}{N}\sum_{i=1}^{N} |d^{i}_{t} - d^{i}_{o}|^{2}, \tag{7}$$

where N is the number of entries in the dataset and the subscript i represents the i-th entry in the dataset.

Table 1
The performance of inference attacks based on different models when 10% of the SNPs are hidden. M1-LD, M2 and RM denote the inference attacks based on a 1st-order Markov chain with public pairwise LD, a 2nd-order Markov chain and a recombination model, respectively, from Samani et al. [28], whereas iHMM and RCNN denote our proposed inference attacks.

	Error rate	Normalized entropy	Normalized privacy loss
M1-LD (Samani et al.)	0.3356	0.4872	0.1864
M2 (Samani et al.)	0.2400	0.3419	0.3316
RM (Samani et al.)	0.0578	0.069	0.6046
iHMM (Ours)	0.0085	0.0295	0.6520
RCNN (Ours)	0.0753	0.0973	0.5143

As shown in Fig. 1, once additional features have been extracted in the Maxpool2 layer, we can connect it to the FC layer and flatten all features into one dimension. During the training process, if the desired MSE is not reached in the current epoch, training will continue until either the maximal number of epochs or the desired MSE is reached. If the maximal number of epochs is reached, then the training process will stop regardless of the MSE value. To demonstrate the feasibility and practicability of the proposed method, the aggregate performance is evaluated by inputting the test dataset into the trained RCNN model and using it to predict the haplotypes of the hidden SNPs.

6. Evaluations and results

In this section, we will evaluate the performance of our proposed attack methods in terms of various metrics and compare our results with previous work based on the outcomes of a set of carefully designed experiments.

6.1. Dataset

For these experiments, we used a dataset from phase III of the HapMap Project [40], which is publicly available on the Internet. In this project anonymous genomic data is collected from 11 different populations worldwide for genetic research. Without loss of generalaty, we adopt the dataset on chromosome 22 of the Northern and Western European Ancestry (CEU) population, released in May 2010. This dataset contains haplotype sequences of individuals, and the MAFs, pairwise LD values, and recombination rates of these populations are also included. We consider these data as public background data. Additionally, the genotype sequences of 165 individuals are also included in the HapMap Project dataset. We will use these data as the genomic data of the chosen unrelated victims. This dataset has also been used in [28].

6.2. Results

In our experiments, we randomly hide different percentages (ranging from 5% to 60%) of the SNPs of the victims, infer the hidden SNPs using our proposed attack models, and quantify the genomic privacy results in terms of the three metrics described in Section 4.

First, we randomly hide 10% of the SNPs of the victims and evaluate the inference power of the adversary with the different attack models. Then, we run the experiment for 20 times and take the average value of each metric over all victims. We assess the attacks based on both the iHMM and the RCNN model. The results of the attacker's incorrectness, the attacker's uncertainty and the victim's privacy loss are shown in Table 1. In this table, M1-LD, M2 and RM denote the inference attacks presented in [28] based on a 1st-order Markov chain with public pairwise LD, a 2nd-order Markov chain, and a recombination model, respectively, whereas iHMM and RCNN denote our inference attacks based on the iHMM and the RCNN model, respectively. We compare the incorrectness of the different inference attacks in the *Error rate* column. The outcomes of both of our methods show significantly decreased incorrectness overall, with iHMM performing better and RCNN performing slightly worse than the RM method. Because the authors of [28] have not considered the metrics of uncertainty and privacy loss in their paper, we ammend their experiments as necessary to calculate these two metrics. The obtained results indicate that these two metrics are also suitable for measuring genomic privacy. The performance results in terms of uncertainty and privacy loss are shown in the *Normalized entropy* column and the *Normalized privacy loss* column of Table 1, respectively. The results show that with the iHMM-based inference attack, the attacker can achieve a lower uncertainty and obtain richer private information of the victims.

To further support our comparison results and for consistency with the experiment reported in [28], another experiment with 40% hidden SNPs was conducted. The performance results are shown in Table 2 and are consistent with those reported in Table 1.

Next, to observe the effects of the number of hidden SNPs on the different inference attacks, another set of experiments have been conducted by using the pairwise LD, 2-order Markov chain, recombination model, iHMM and RCNN attacks with different percentages (ranging from 5% to 60%) of hidden SNPs. The results in terms of the attacker's incorrectness, the attacker's uncertainty and the victim's genomic privacy loss to the attacker are shown in Figs. 5–7, respectively.

Table 2
The performance of inference attacks based on different models when 40% of the SNPs are hidden. M1-LD, M2 and RM denote the inference attacks based on a 1st-order Markov chain with public pairwise LD, a 2nd-order Markov chain and a recombination model, respectively, from Samani et al. [28], whereas iHMM and RCNN denote our proposed inference attacks.

	Error rate	Normalized entropy	Normalized privacy loss
M1-LD (Samani et al.)	0.3623	0.4867	0.1873
M2 (Samani et al.)	0.2873	0.3489	0.3251
RM (Samani et al.)	0.0923	0.0902	0.5838
iHMM (Ours)	0.0136	0.0430	0.6342
RCNN (Ours)	0.1028	0.1345	0.5347

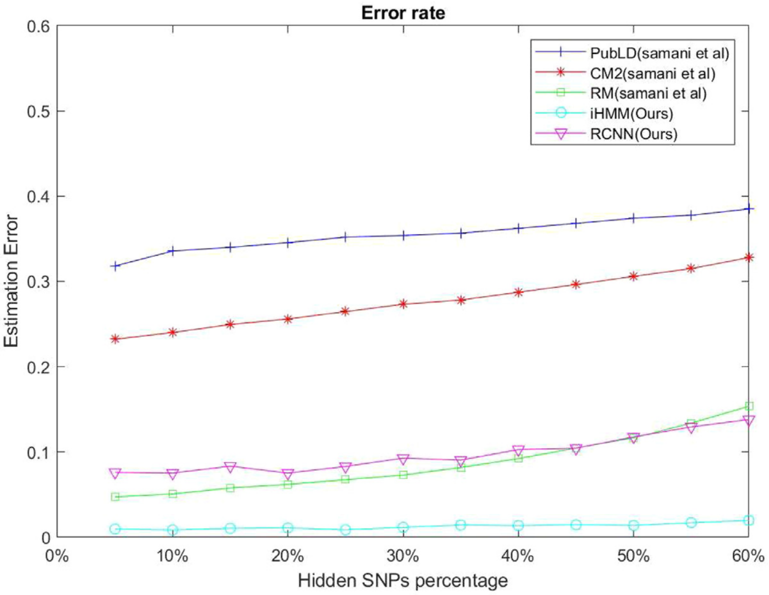

Fig. 5. Genomic privacy in terms of the attacker's incorrectness for inference attacks based on different models, with different numbers (from 5% to 60%) of the SNPs from the victims being randomly hidden.

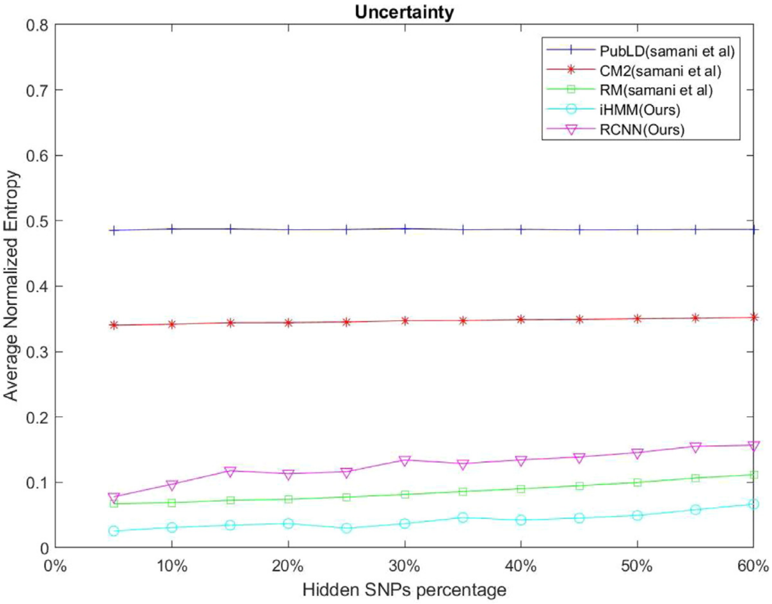

Fig. 6. Genomic privacy in terms of the attacker's uncertainty for inference attacks based on different models, with different numbers (from 5% to 60%) of the SNPs from the victims being randomly hidden.

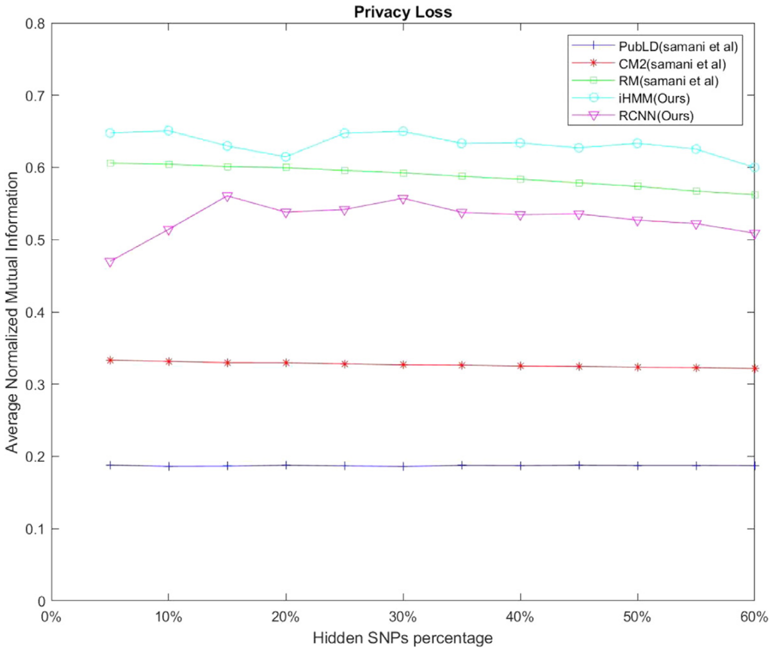

Fig. 7. Genomic privacy in terms of the victim's privacy loss to the attacker for inference attacks based on different models, with different numbers (from 5% to 60%) of the SNPs from the victims being randomly hidden.

In Fig. 5, we see the results of the inference attacks based on the different models in terms of the attacker's incorrectness. The inference power of these attacks can be observed increasing when fewer SNPs of the victims are hidden (i.e., the more SNPs of the victims that are revealed to the attacker, the lower the incorrectness). Both of our proposed attack models show improved inference power compared to the previous work in terms of incorrectness. The RCNN-based attack performs better than the recombination model-based attack when more SNPs are hidden (above 50%) and slightly worse than the recombination model-based attack when fewer SNPs are hidden (below 45%).

In Fig. 6, we show the results of inference attacks based on the different models in terms of the attacker's uncertainty. It can be observed that the inference power of these attacks increases when fewer SNPs of the victims are hidden (i.e., the more SNPs of the victims that are revealed to the attacker, the lower the uncertainty), and the results are consistent with the results in Fig. 5. The iHMM-based attack always performs better than the other attacks, whereas the RCNN-based attack does not perform well at all times.

Similarly, we see the results in terms of the victim's privacy loss in Fig. 7. Once again, the inference power of these attacks can be observed increasing when fewer SNPs of the victims are hidden (i.e., the more SNPs of the victims that are revealed to the attacker, the greater the privacy loss).

7. Conclusion

In this work, we have proposed several attack atrategies for inferring the genotypes of individuals from their own partial SNP sequences that are available online combined with public genomic information via either an improved hidden Markov model or a regressive convolutional neural network model. We have shown that the attacker can infer the private hidden SNPs of individuals with high accuracy, low uncertainty and high privacy loss. Experiments have shown that the proposed attacks extend and significantly improve upon existing work. By quantifying the genomic privacy of individuals based on publicly available genomic data, our work can help people to understand current risks to genomic privacy much better.

In future, we will further explore the potential of machine learning, extend this approach to attacks on kin genomic privacy, and identify suitable methods of resisting genomic privacy attack.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

Acknowlgedgments

This work was performed when Hongfa Ding was a visitor at Lehigh University with the support of the China Scholarship Council. The authors gratefully acknowledge Prof. Danchi Jiang his assistance in this study, and to reviewers for helpful comments.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ins.2019.09.036.

References

[1] E. Ayday, M. Humbert, Inference attacks against kin genomic privacy, IEEE Secur. Privacy 15 (5) (2017) 29–37, doi:10.1109/MSP.2017.3681052.

[2] E. Ayday, J.L. Raisaro, J. Hubaux, Personal use of the genomic data: privacy vs. storage cost, in: 2013 IEEE Global Communications Conference, GLOBECOM 2013, Atlanta, GA, USA, December 9–13, 2013, 2013, pp. 2723–2729, doi:10.1109/GLOCOM.2013.6831486.

[3] R. Cai, Z. Hao, M. Winslett, X. Xiao, Y. Yang, Z. Zhang, S. Zhou, Deterministic identification of specific individuals from GWAS results, Bioinformatics 31 (11) (2015) 1701–1707, doi:10.1093/bioinformatics/btv018.

[4] I. Deznabi, M. Mobayen, N. Jafari, O. Tastan, E. Ayday, An inference attack on genomic data using kinship, complex correlations, and phenotype information, IEEE/ACM Trans. Comput. Biol.Bioinf. 15 (4) (2018) 1333–1343, doi:10.1109/TCBB.2017.2709740.

[5] R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

[6] En.wikipedia.org, 2019, Inference attack, Accessed April 22. (https://en.wikipedia.org/wiki/Inference_attack).

[7] K. Ganju, Q. Wang, W. Yang, C.A. Gunter, N. Borisov, Property inference attacks on fully connected neural networks using permutation invariant representations, in: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15–19, 2018, 2018, pp. 619–633, doi:10.1145/3243734.3243834.

[8] N.Z. Gong, B. Liu, You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors, in: 25th *USENIX* Security Symposium (*USENIX* Security 16), 2016, pp. 979–995.

[9] M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying personal genomes by surname inference, Science 339 (6117) (2013) 321–324, doi:10.1126/science.1229566.

[10] A. Harmanci, M. Gerstein, Quantification of private information leakage from phenotype-genotype data: linking attacks, Nat. Methods 13 (3) (2016) 251–256, doi:10.1038/nmeth.3746.

[11] Z. He, Y. Li, J. Li, J. Yu, H. Gao, J. Wang, Addressing the threats of inference attacks on traits and genotypes from individual genomic data, in: Bioinformatics Research and Applications - 13th International Symposium, ISBRA 2017, Honolulu, HI, USA, May 29, - June 2, 2017, Proceedings, 2017, pp. 223–233, doi:10.1007/978-3-319-59575-7_20.

[12] P. Hess, Controversial geneticist warns: we can read your face in your dna., 2017, Accessed June 2, 2018. (https://www.inverse.com/article/36145-genetic-privacy-venter-23andme).

[13] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J.V. Pearson, D.A. Stephan, S.F. Nelson, D.W. Craig, Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays, PLOS Genet. 4 (8) (2008) 1–9, doi:10.1371/journal.pgen.1000167.

[14] B. Howie, J. Marchini, 2019, IMPUTE2, Accessed April 22. (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference).

[15] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLOS Genet. 5 (6) (2009) 1–15, doi:10.1371/journal.pgen.1000529.

[16] J. Hu, M.K. Brown, W. Turin, HMM based online handwriting recognition, IEEE Trans. Pattern Anal. Mach.Intell. 18 (10) (1996) 1039–1045.

[17] M. Humbert, E. Ayday, J.-P. Hubaux, 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, in: CCS '13, ACM, New York, NY, USA, 2013, pp. 1141–1152, doi:10.1145/2508859.2516707.

[18] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (6) (2015) 321–332, doi:10.1038/nrg3920.

[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach.Intell. 39 (4) (2017) 640–651, doi:10.1109/TPAMI.2016.2572683.

[20] M.D. Mailman, M. Feolo, Y. Jin, M. Kimura, K. Tryka, R. Bagoutdinov, L. Hao, A. Kiang, J. Paschall, L. Phan, et al., The NCBI dbGaP database of genotypes and phenotypes, Nat. Genet. 39 (10) (2007) 1181.

[21] J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly, A new multipoint method for genome-wide association studies by imputation of genotypes, Nat. Genet. 39 (7) (2007) 906–913, doi:10.1038/ng2088.

[22] S. Narain, T.D. Vo-Huu, K. Block, G. Noubir, Inferring user routes and locations using zero-permission mobile sensors, in: IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22–26, 2016, 2016, pp. 397–413, doi:10.1109/SP.2016.31.

[23] D.R. Nyholt, C.-E. Yu, P.M. Visscher, On Jim Watson's APOE status: genetic information is hard to hide, Eur. J. Hum. Genet. 17 (2) (2009) 147–149, doi:10.1038/ejhg.2008.198.

[24] C. Peng, H. Ding, Y. Zhu, Y. Tian, Z. Fu, Information entropy models and privacy metrics methods for privacy protection, J. Softw. 27 (8) (2016) 1891–1903, doi:10.13328/j.cnki.jos.005096.

[25] D. Pouliot, C.V. Wright, The shadow nemesis: Inference attacks on efficiently deployable, efficiently searchable encryption, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24–28, 2016, 2016, pp. 1341–1352, doi:10.1145/2976749.2978401.

[26] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.

[27] R.V. Rohlfs, S.M. Fullerton, B.S. Weir, Familial identification: population structure and relationship distinguishability, PLOS Genet. 8 (2) (2012) e1002469, doi:10.1371/journal.pgen.1002469.

[28] S.S. Samani, Z. Huang, E. Ayday, M. Elliot, J. Fellay, J.-P. Hubaux, Z. Kutalik, Quantifying genomic privacy via inference attack with high-order SNV correlations, in: Proceedings of the 2015 IEEE Security and Privacy Workshops, in: SPW '15, IEEE Computer Society, Washington, DC, USA, 2015, pp. 32–40, doi:10.1109/SPW.2015.21.

[29] E.E. Schadt, S. Woo, K. Hao, Bayesian method to predict individual SNP genotypes from gene expression data, Nat. Genet. 44 (5) (2012) 603–608, doi:10.1038/ng.2248.

[30] S. Scutti, What the golden state killer case means for your genetic privacy, 2018, Accessed May 28, 2018. (https://www.cnn.com/2018/04/27/health/golden-state-killer-genetic-privacy/index.html).

[31] X. Shi, X. Wu, An overview of human genetic privacy, Ann. New York Acad. Sci. 1387 (1) (2017) 61–72, doi:10.1111/nyas.13211.

[32] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22–26, 2017, 2017, pp. 3–18, doi:10.1109/SP.2017.41.

[33] S. Shringarpure, C. Bustamante, Privacy risks from genomic data-sharing beacons, Am. J. Hum. Genet. 97 (5) (2015) 631–646, doi:10.1016/j.ajhg.2015.09.010.

[34] M. Stamp, A revealing introduction to hidden Markov models, in: Department of Computer Science San Jose State University, 2004, pp. 26–56.

[35] L. Sweeney, A. Abu, J. Winn, Identifying participants in the personal genome project by name, 2013.

[36] The Genomes Project Consortium, A global reference for human genetic variation, Nature 526 (2015) 68, doi:10.1038/nature15393.

[37] IGSR: the international genome sample resource, 2019, Accessed April 22. (http://www.internationalgenome.org/),

[38] The International Genome Sample Resource (IGSR), 2019Which populations are part of your study?, Accessed April 22. (http://www.internationalgenome.org/category/population/).

[39] The National Human Genome Research Institute, 2019, Privacy in genomics, Accessed April 22. (https://www.genome.gov/27561246/privacy-in-genomics).

[40] G.A. Thorisson, A.V. Smith, L. Krishnan, L.D. Stein, The international HapMap project web site, Genome Res. 15 (11) (2005) 1592–1593.

[41] U.S. Equal Employment Opportunity Commission, Genetic information nondiscrimination act of 2008, 2008, = from Accessed 1 June 2018. https://www.eeoc.gov/laws/statutes/gina.cfm).

[42] I. Wagner, Evaluating the strength of genomic privacy metrics, ACM Trans. Priv. Secur. 20 (1) (2017) 2:1–2:34, doi:10.1145/3020003.

[43] S. Walsh, F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, 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. 5 (3) (2011) 170–180, doi:10.1016/j.fsigen.2010.02.004.

[44] R. Wang, Y.F. Li, X. Wang, H. Tang, X. Zhou, 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, in: CCS '09, ACM, New York, NY, USA, 2009, pp. 534–544, doi:10.1145/1653662.1653726.

[45] Y. Wang, J. Wen, X. Wu, X. Shi, Infringement of individual privacy via mining differentially private GWAS statistics, in: Y. Wang, G. Yu, Y. Zhang, Z. Han, G. Wang (Eds.), Big Data Computing and Communications, Springer International Publishing, Cham, 2016, pp. 355–366, doi:10.1007/978-3-319-42553-5_30.