An Inference Attack on Genomic Data Using Kinship, Complex Correlations, and Phenotype Information

Iman Deznabi, Mohammad Mobayen, Nazanin Jafari, Oznur Tastan, and Erman Ayday

Abstract—Individuals (and their family members) share (partial) genomic data on public platforms. However, using special characteristics of genomic data, background knowledge that can be obtained from the Web, and family relationship between the individuals, it is possible to infer the hidden parts of shared (and unshared) genomes. Existing work in this field considers simple correlations in the genome (as well as Mendel's law and partial genomes of a victim and his family members). In this paper, we improve the existing work on inference attacks on genomic privacy. We mainly consider complex correlations in the genome by using an observable Markov model and recombination model between the haplotypes. We also utilize the phenotype information about the victims. We propose an efficient message passing algorithm to consider all aforementioned background information for the inference. We show that the proposed framework improves inference with significantly less information compared to existing work.

Index Terms—Privacy, security, genomic privacy, inference attacks

1 Introduction

Substantial progress has been achieved towards reducing the cost of DNA sequencing. As a consequence, research in genomics has gained speed towards paving the way to personalized (genomic) medicine, and geneticists now need large collections of human genomes to further increase this speed. Furthermore, individuals are using their genomes to learn about their health, their origins, and even their (genetic) compatibilities with potential partners. This trend has led to the launch of health-related websites and online social networks (OSNs), in which individuals can share their genomic data (e.g., OpenSNP or 23andMe). There are, however, significant risks in sharing this genomic data which carries a lot of sensitive information about its owner. By analyzing the DNA of an individual, it is now possible to learn about his disease predispositions (e.g., for Alzheimer's), ancestors, and physical attributes. This threat to genomic privacy is magnified by the fact that a person's genome is correlated to his family members' genomes, thus leading to interdependent privacy risks.

Individuals (either directly or indirectly) share vast amount of personal information on the Web, and some of this information can be used to infer their genomic data. Hence, there is a need to clearly understand the nature and extent of privacy risks on the genomic data of individuals considering publicly available information on the Web. In this paper, we propose to establish a unifying framework to quantify the genomic privacy of individuals using all publicly available resources.

Humbert et al. previously proposed a framework to quantify genomic privacy of individuals considering (i) partial genomic data that is publicly shared by the individual and his family members, (ii) simple pairwise correlations in the genome (i.e., linkage disequilibrium), and (iii) other public genomic knowledge (e.g., minor allele frequencies) [1]. In a recent study, Samani et al. showed that higher order correlations in the genome actually enables stronger inference power compared to the pairwise correlations [2]. However, in that work, authors did not study the implications of this result on kin genomic privacy.

Motivated by these recent studies, in this work, our two main contributions are showing the extend of privacy risk on the individuals and their family members due to (i) complex correlations (i.e., high order correlations) in the genome, and (ii) publicly available phenotype information (e.g., physical traits or disease information) about the individuals. The main objective of this work is to develop a new unifying framework for quantification of genomic privacy of individuals. Similar to the previous work, we use a graph-based, iterative algorithm to build this framework efficiently. Our results show that the attacker's inference power (on the genomic data of individuals) significantly improves by using complex correlations and phenotype information (along with information about their family bonds). We believe that this paper would be a significant step towards establishing a greater understanding of the privacy risks on the genomic data of individuals.

• *The authors are with the Computer Engineering Department, Bilkent University, Ankara 06800, Turkey. E-mail: {iman.deznabi, mohammadm, nazanin.jafari}@bilkent.edu.tr, {oznur.tastan, erman}@cs.bilkent.edu.tr.*

Manuscript received 7 June 2016; revised 10 Apr. 2017; accepted 4 May 2017. Date of publication 30 May 2017; date of current version 6 Aug. 2018. (Corresponding author: Erman Ayday.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2017.2709740

The rest of the paper is organized as follows. In the next section, we summarize the existing work on genomic privacy. In Section 3, we provide brief background information about genomics and the belief propagation algorithm (which is the core of the proposed framework). In Section 4, we briefly summarize the work by Humbert et al., as we build our framework on top of this previous work. In Section 5, we describe the proposed scheme in detail. In Section 6, we evaluate the proposed scheme using real genomic data. Finally, in Section 7, we conclude the paper.

2 Related Work

In the last few years, there have been several works addressing the security and privacy concerns on genomic data [3]. A large part of the related work on genomic privacy focuses on the problem of private pattern-matching and the comparison of genomic sequences. For example, Troncoso-Pastoriza et al. propose an algorithm for private string searching on the DNA sequence by using a finite state machine [4]. Their work is then revisited by Blanton et al., who develop an efficient method for sequence comparison using garbled circuits [5]. Furthermore, Baldi et al. make use of private-set intersection and present an effective algorithm for privacy-preserving substring matching on DNA sequences [6]. Chen et al. propose a privacy-preserving method to align short sequences to a reference genome by outsourcing the computation to the cloud [7]. Jha et al. propose a privacy-preserving implementation of fundamental genomic computations for sequence alignment [8]. Furthermore, Naveed et al. proposed a scheme based on functional encryption for privacy-preserving similarity test on genomic data [9]. To hide access patterns to genomic data that is stored at a cloud environment, Karvelas et al. proposed using the ORAM mechanism [10]. Recently, Wang et al. proposed an efficient privacy-preserving protocol to find genetically similar patients in a distributed environment [11].

Another line of investigation is represented by works focusing on private clinical genomics. De Cristofaro et al. propose a secure protocol between two parties for testing genomic sequences without the leaking of any private information about the genomic sequence or the nature of the test [12]. Wang et al. propose techniques for computing on genomic data by distributing the task between a data provider and consumer through program specialization [11]. Ayday et al. proposed a scheme to protect the privacy of users' genomic data while enabling medical units to access the genomic data in order to conduct medical tests or to develop personalized medicine methods [13].

A third area of interest addresses the problem of protecting genomic privacy, while still allowing for both basic and translational medical research on the data. It has been shown that deanonymization is a serious threat for genomic data [14], [15]. Thus, many solutions have been proposed for privacy-preserving genomic research either by using statistical techniques (such as differential privacy) [16] or cryptographic techniques [17].

Independent of these categories, Ayday et al. proposed a technique for privacy-preserving storage and retrieval of raw-genomic data [18] and Huang et al. proposed an information-theoretical technique for secure storage of genomic data [19]. In this paper, building on top of the previous work on kin genomic privacy [1], we develop a unifying framework for quantification of genomic privacy of individuals notably by using complex correlations in the genome, family bonds, and publicly available phototype information.

3 Background

In this section we give a brief background on genomics and the belief propagation algorithm.

3.1 Genomics

Single nucleotide polymorphism (SNP). Around 99.9 percent of an individual's genome is identical to the reference human genome and the rest is human genetic variation. The most common genetic variations in humans are the SNPs. SNP is a variation in the genome in which a single nucleotide (A, C, G, or T) differs between members of the same species or paired chromosomes of an individual. There are usually two different alleles (nucleotides) that are observed at a SNP position; one is called the *minor allele* and the other is the *major allele*. Furthermore, each SNP carries two alleles in total. Hence, the content of a SNP position can be in one of the following states: (i) BB (homozygous-major genotype), if an individual receives the same major allele from both parents; (ii) Bb (heterozygous genotype), if he receives a different allele from each parent (one minor and one major); or (iii) bb (homozygous-minor genotype), if he inherits the same minor allele from both parents (this is also shown in Fig. 1(a)). For simplicity, in the rest of the paper, we denote the value (content) of a SNP as the number of minor alleles it carries. Thus, we denote BB as 0, Bb as 1 and bb as 2.

Reproduction: The Mendel's first law, the Law of Segregation, states that a child's SNPs are independent from his ancestors', given the SNPs of his parents. Each child inherits one allele (nucleotide) of a SNP from his mother and the other one from his father, and each allele is inherited with a probability of 0.5. In [1] authors model this law by a function (introduced in Section 4) that simply considers the Mendelian inheritance probabilities as in Fig. 1b. We also use this inheritance information in this work.

Correlations in the genome: It is shown that SNPs on the DNA sequence are correlated. For example, pairwise correlations between the SNPs in the genome are referred to as linkage disequilibrium (LD) [20]. In [1], the authors use the LD values between the SNPs as an input to their inference algorithm. In this work, we show that more complex, higher order correlations in the genome threaten kin genomic privacy more than the pairwise correlations.

Phenotypes: Phenotypes are observable characteristics of individuals (e.g., physical traits or diseases) that may be related to both their genotype and the environment. For example, SNP *Rs12821256* on chromosome 12 is associated with having blonde hair. If an individual has (C,C) nucleotide pair for this SNP, he is 4 times more likely to have blonde hair compared to other individuals. We use phenotype information of individuals to improve the inference power of the proposed algorithm.

3.2 Belief Propagation

Belief propagation [21] is a message-passing algorithm for performing inference on graphical models (e.g., Bayesian networks or Markov random fields). It is typically used to

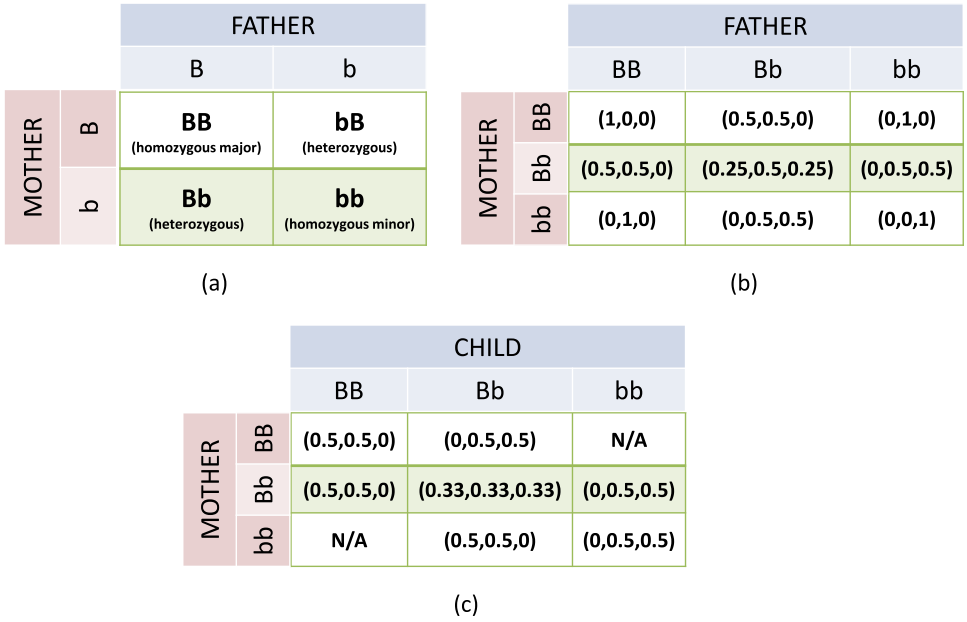


Fig. 1. (a) Mendelian inheritance for a child. (b) Inheritance probabilities for a SNP, given different genotypes for the parents. The probabilities of the child's genotype are represented in parentheses. (c) Inheritance probabilities for a SNP, given different genotypes for the child and the mother. The probabilities of the father's genotype are represented in parentheses (given the child and the father, the probabilities for the mother are also the same).

compute marginal distributions of unobserved variables conditioned on the observed ones. Computing marginal distributions is hard in general as it might require summing over an exponentially large number of terms. The belief propagation algorithm can be described in terms of operations on a factor graph, a graphical model that is represented as a bipartite graph. One of the two disjoint sets of the factor graph's vertices represents the (random) variables of interest, and the second set represents the functions that factor the joint probability distribution (or global function) of the variables based on the dependencies between them. An edge connects a variable node to a factor node if and only if the variable is an argument of the function corresponding to the factor node. The marginal distribution of an unobserved variable can be exactly computed by using the belief propagation algorithm if the factor graph has no cycles. However, the algorithm is still well defined and often gives good approximate results for factor graphs with cycles (as it has been observed in decoding of LDPC codes) [22]. Belief propagation is commonly used in artificial intelligence and information theory.

4 Quantifying Kin Genomic Privacy [1]

In [1], authors evaluate the genomic privacy of an individual threatened by his relatives revealing their genomes. Focusing on the SNPs in the genome, they quantify the loss in genomic privacy of individuals when one or more of their family members' genomes are (either partially or fully) revealed. They design a reconstruction attack, in which they formulate the SNPs, family relationships, and the pairwise correlations (LD) between SNPs on a factor graph and use the belief propagation algorithm for inference. Then, using various metrics, they quantify the genomic privacy of individuals and reveal the decrease in their level of genomic privacy caused by the published genomes of their family members. In the following, we briefly summarize the framework of [1] as we build the proposed scheme on top of this framework.

The goal of the adversary is to infer some *targeted SNPs* of a member (or multiple members) of a *targeted family*. Let $\mathbf{F}$ be the set of family members in the targeted family (whose family tree is $\mathcal{G}_{\mathbf{F}}$) and $\mathbf{S}$ be the set of SNP IDs (on the DNA sequence), where $|\mathbf{F}| = n$ and $|\mathbf{S}| = m$. Let also x_j^i be the value of SNP j ($j \in \mathbf{S}$) for individual i ($i \in \mathbf{F}$), where $x_j^i \in \{0, 1, 2\}$ (as discussed in Section 3.1). Also, $\mathbb{X}$ is an $n \times m$ matrix that stores the values of the SNPs of all family members. Among the SNPs in $\mathbb{X}$, the ones whose values are unknown are in set $\mathbb{X}_{\mathrm{U}}$, and the ones whose values are known (by the adversary) are in set $\mathbb{X}_{\mathrm{K}}$. $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$ is the function representing the Mendelian inheritance probabilities (as in Fig. 1b), where (M, F, C) represent mother, father, and child, respectively. Finally, $\mathbf{P} = \{p_i^b : i \in \mathbf{S}\}$ represents the set of minor allele probabilities (or MAF) of the SNPs in $\mathbf{S}$.

The adversary carries out a reconstruction attack to infer $\mathbb{X}_{\mathrm{U}}$ by relying on his background knowledge, $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$, $\mathbb{L}$[1], $\mathbf{P}$, and on his observation $\mathbb{X}_{\mathrm{K}}$. The authors formulate this reconstruction attack as finding the marginal probability distributions of unknown variables $\mathbb{X}_{\mathrm{U}}$, and to run this attack in an efficient way, they formulate the problem on a factor graph and use the belief propagation algorithm for inference. In this work, we formulate the attack by also considering complex correlations in the genome and publicly available phenotype information. We show that the inference attack is significantly stronger when these additional factors are also considered. In the following, we provide the details of the proposed framework emphasizing the differences from [1].

5 Proposed Framework

Our main objective is to develop a unifying framework for the quantification of the genomic privacy of individuals

1. $\mathbb{L}$ is a $m \times m$ matrix representing the pairwise linkage disequilibrium between each pair of SNPs. Instead of the LD values, we use higher order correlations in this work for inference.

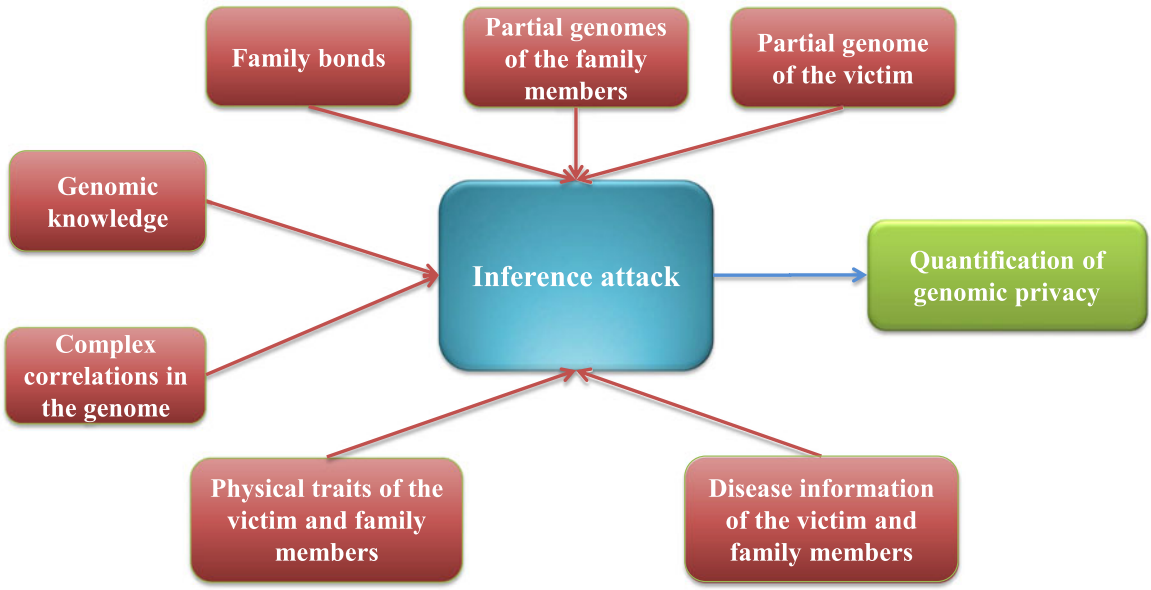


Fig. 2. Overview of the proposed framework for quantification of genomic privacy.

using all available public data on the Web and background knowledge on genomics. We assume that the attacker has access to the following resources about the target individuals: (i) the partial genomic data of individuals (from public genomic databases and genome sharing websites), (ii) phenotype information (physical characteristics) of individuals from OSNs, (iii) health related information of individuals from OSNs and health related social networks, and (iv) family bonds of individuals (e.g., their family trees) from OSNs or genealogy websites. Our proposed framework is also sketched in Fig. 2.

The objective is to infer the missing parts of the genomes of individuals in the target individuals set. For this, we use family bonds between the individuals in the target set, probabilistic relationship between the phenotype and genotype, similar relationship between diseases and the genotype, and some genomic tools for inference such as high order correlations in the genome and the recombination model. To run this inference attack efficiently, similar to the previous work, we rely on the belief propagation algorithm on a factor graph. Then, we quantify genomic privacy of individuals and show the risk for each individual.

Constructing the Factor Graph: A factor graph is a bipartite graph containing two sets of nodes (corresponding to variables and factors) and edges connecting these two sets. We form a factor graph by setting a variable node for each SNP x_j^i ($j \in \mathbf{S}$ and $i \in \mathbf{F}$). We use three types of factor nodes[2]: (i) *familial factor node*, representing the familial relationships and reproduction, (ii) *correlation factor node*, representing the higher order correlations between the SNPs either by using a Markov chain or hidden Markov model, and (iii) *phenotype factor node*, representing the correlation between the SNPs and the phenotypes (e.g., physical traits or diseases) of individuals. The factor graph representation of our proposed framework is shown in Fig. 3. We summarize the connections between the variable and factor nodes below:

- Each variable node x_j^i has its familial factor node f_j^i if at least one parent of individual i is in the target family. Furthermore, x_j^k ($k \neq i$) is also connected to the familial factor node of x_j^i if k is the mother or father of i. If an individual i's both parents are not present in the target family, we do not assign familial factor nodes corresponding to the variable nodes of that individual. For example, in Fig. 3, all familial factor nodes belong to the child as his parents are present in the toy example. However, father's and mother's variable nodes do not have separate familial factor nodes.
- Variable nodes in set $\mathbf{C}$ are connected to a correlation factor node $g_{\mathbf{C}}^i$ (of individual i) if SNPs in $\mathbf{C}$ have correlation among each other. In particular, we consider higher order correlations in the genome. We model these correlations either using a Markov chain or a hidden Markov model, HMM (i.e., recombination model). When we use a Markov chainwith order of k the correlation set of node i is $\mathbf{C}_i = \{node_{i-k}, node_{i-k+1}, node_{i-k+2}, \ldots, node_{i-1}\}$ if $i > k$, and $\mathbf{C}_i = \{node_1, node_2, node_3, \ldots, node_{i-1}\}$ if $i \leq k$, and when we use HMM, $\mathbf{C}$ includes all SNPs in a chromosome.
- Variable nodes of individual i in set $\mathbf{H}_\alpha^i$ are connected to a phenotype factor node ph_α^i if SNPs in $\mathbf{H}_\alpha^i$ are associated with the phenotype ph_α. Note that more than one SNP can be associated with a given phenotype. Similarly, a SNP may be associated with more than one phenotype.

Messages between the Nodes: As shown in [23], following the rules of belief propagation, the global probability distribution of the variable nodes can be factorized into products of local functions that are defined by the factor nodes following the rules of the belief propagation algorithm. The iterative belief propagation algorithm is based on exchanging messages between the variable and the factor nodes. We represent these messages as in the following:

2. There are two types of factor nodes in [1] representing the family relationships and the LD between the SNPs.

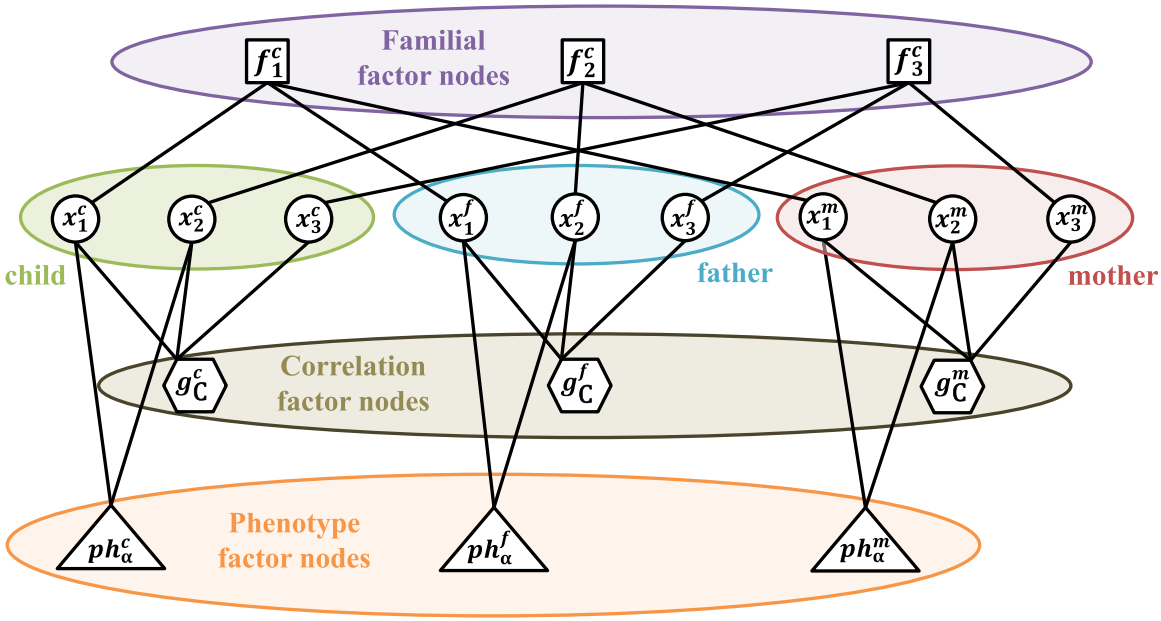


Fig. 3. Factor graph representation of the proposed framework.

- The message $\mu_{i\to k}^{(\nu)}(x_j^{i(\nu)})$ (from a variable node i to a factor node k) denotes the probability of $x_j^{i(\nu)} = \ell$ ($\ell \in \{0,1,2\}$), at the νth iteration.
- The message $\lambda_{k\to i}^{(\nu)}(x_j^{i(\nu)})$ (from a familial factor node to a variable node) denotes the probability that $x_j^{i(\nu)} = \ell$, for $\ell \in \{0,1,2\}$, at the νth iteration given $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$, $\mathbf{P}$, and the values of SNP j for the other two family members (other than individual i) that are connected to the corresponding familial factor node.
- The message $\beta_{k\to i}^{(\nu)}(\mathbf{C}, x_j^{i(\nu)})$ (from a correlation factor node to a variable node) denotes the probability that $x_j^{i(\nu)} = \ell$, for $\ell \in \{0,1,2\}$, at the νth iteration given the high order correlation between the SNPs in set $\mathbf{C}$.
- The message $\delta_{k\to i}^{(\nu)}(x_j^{i(\nu)})$ (from a phenotype factor node to a variable node) denotes the probability that $x_j^{i(\nu)} = \ell$, for $\ell \in \{0,1,2\}$, at the νth iteration given the phenotype ph_k for individual i and the association of the corresponding phenotype with SNP j.

Toy Example on a Trio: Following [1], we choose a simple family tree consisting of a trio (i.e., mother, father, and child) and 3 SNPs (i.e., $|\mathbf{F}| = 3$ and $|\mathbf{S}| = 3$). In Fig. 3, we show how the trio and the SNPs are represented on a factor graph, where $i = m$ represents the mother, $i = f$ represents the father, and $i = c$ represents the child. Furthermore, the 3 SNPs are represented as $j = 1$, $j = 2$, and $j = 3$, respectively. We describe the message exchange between the variable node representing the first SNP of the mother (x_1^m), the familial factor node of the child (f_1^c), the correlation factor node $g_{\mathbf{C}}^m$, and the phenotype factor node ph_α^m (representing the phenotype α for the mother). Here we assume that variable nodes in set $\mathbf{C}$ are SNPs 1, 2, and 3. We also assume that the phenotype α is associated with SNPs 1 and 2 (that are in set $\mathbf{H}_\alpha^m$). The belief propagation algorithm iteratively exchanges messages between the factor and the variable nodes, updating the beliefs on the values of the targeted SNPs (in $\mathbb{X}_\mathrm{U}$) at each iteration, until convergence. For simplicity, we denote the variable and factor nodes x_1^m, f_1^c, $g_{\mathbf{C}}^m$, and ph_α^m with the letters i, k, z, and s, respectively.

Messages from variable nodes: Variable node i forms $\mu_{i\to k}^{(\nu)}(x_1^{m(\nu)})$ by multiplying all information it receives from its neighbors excluding the familial factor node k.[3] Hence, the message from variable node i to the familial factor node k at the νth iteration is given by

$$\mu_{i\to k}^{(\nu)}(x_1^{m(\nu)}) = \frac{1}{Z} \times \beta_{z\to i}^{(\nu-1)}(\mathbf{C}, x_1^{m(\nu-1)}) \times \delta_{s\to i}^{(\nu-1)}(x_1^{m(\nu-1)}), \quad (1)$$

where Z is a normalization constant. This computation is repeated for every neighbor of each variable node. If $x_1^m \in \mathbb{X}_\mathrm{K}$ (i.e., it is one of the SNPs that is observed by the attacker), then the message $\mu_{i\to k}^{(\nu)}(x_1^{m(\nu)})$ is constructed as a constant, depending on the value of x_1^m. Note that following the rules of belief propagation, to prevent self-bias, the message $\lambda_{k\to i}^{(\nu-1)}(x_1^{m(\nu-1)})$ is not used while generating $\mu_{i\to k}^{(\nu)}(x_1^{m(\nu)})$. Also, if the parents of the mother (m) were also in the graph, x_1^m would have its corresponding familial factor node f_1^m, and hence the λ message generated from this factor node would have been also used when generating $\mu_{i\to k}^{(\nu)}(x_1^{m(\nu)})$. Similarly, if SNP x_1 is associated with other phenotypes, δ messages from those phenotype factor nodes are also used while generating the message.

Messages from familial factor nodes: The message from the familial factor node k to the variable node i at the νth iteration is formed using the principles of belief propagation as

$$\lambda_{k\to i}^{(\nu)}(x_1^{m(\nu)}) = \sum_{\{x_1^f, x_1^c\}} f_1^c(x_1^m, x_1^f, x_1^c, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P}) \times \prod_{y\in\{f,c\}} \mu_{x_1^y\to k}^{(\nu)}(x_1^{y(\nu)}), \quad (2)$$

3. Other messages from the variable node i to the other factor nodes (z and s) are also constructed similarly.

where, $f_1^c(x_1^m, x_1^f, x_1^c, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P})$ is proportional to $p(x_1^m | x_1^f, x_1^c, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P})$, and this probability is computed using the table in Fig. 1b. This computation is performed for every neighbor of each familial factor node.

Messages from correlation factor nodes: The message from the correlation factor node z to the variable node i at the νth iteration is formed as

$$\beta_{z\to i}^{(\nu)}(\mathbf{C}, x_1^{m(\nu)}) = \sum_{x_2^m, x_3^m} g_C^m(x_1^m, x_2^m, x_3^m) \times \prod_{y\in\{2,3\}} \mu_{x_y^m \to k}^{(\nu)}(x_y^{m(\nu)}). \quad (3)$$

β messages are generated for every neighbor of each correlation factor node. As mentioned, as opposed to [1], in this work, we consider higher order correlations in the genome to make the inference stronger, and hence the function $g_C^m(x_1^m, x_2^m, x_3^m)$ depends on the correlation model we use. We consider two different correlation models on the genome: (i) Markov chain, in which we consider the genome as a sequence of SNPs, where the value of each SNP depends on the values of neighboring k SNPs. In this scenario, $g_C^m(x_1^m, x_2^m, x_3^m) = p(x_1^m | x_2^m, x_3^m)$, for $k = 2$ (note that LD is a special case of this formalization when $k = 1$). And, (ii) hidden Markov model (HMM), in which the genome is modeled as a Markov process with unobserved (hidden) states. We realize the HMM model for the genome by using the recombination model [24].

Messages from phenotype factor nodes: Finally, the message from the phenotype factor node s to the variable node i at the νth iteration is formed as

$$\delta_{s\to i}^{(\nu)}(x_1^{m(\nu)}) = \sum_{x_2^m} ph_\alpha^m(x_1^m, x_2^m) \times \mu_{x_2^m \to s}^{(\nu)}(x_2^{m(\nu)}). \quad (4)$$

Note that in this toy example, the phenotype α is associated with SNPs x_1 and x_2 only. The function $ph_\alpha^m(x_1^m, x_2^m)$ is computed based on the association of both SNPs with the corresponding phenotype. In some cases, it is observed that the associations of the SNPs to a phenotype are independent from each other. On the other hand, in some cases, we observe that the association depends on the values of both SNPs. Similarly, in some cases, the association is probabilistic, while in some cases the association may be deterministic. For example, having blonde hair color is associated with SNP *Rs12821256* [25]. If an individual has blonde hair, the probability distribution of the corresponding SNP is shown to be (0.01,0.4,0.59),[4] while if he does not have blonde hair, this distribution is shown to be (0.7,0.28,0.02). Thus, the attacker can improve his inference power by obtaining phenotype information about the individuals in the target family.

At each iteration of the algorithm, all variable and factor nodes generate their messages and send to all of their neighbors as described above. At the end of each iteration, we compute the marginal probabilities of each variable nodes (by multiplying all incoming messages), and we stop the algorithm when the values of the marginal probabilities stop changing. Note that the computational complexity of this inference attack is linear with the number of variable or factor nodes in the factor graph.

4. Each entry represents the probability that the value of the SNP is 0, 1, and 2, respectively.

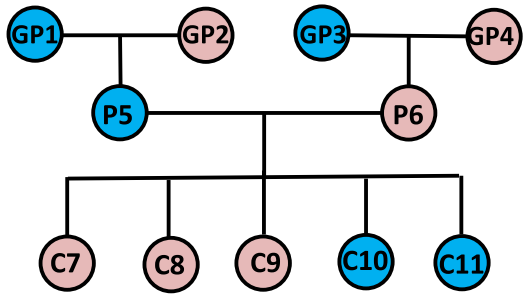


Fig. 4. Family tree of CEPH/Utah pedigree 1463 consisting of the 11 family members that were considered. The blue nodes (i.e., darker ones) represent the male and the pink ones (i.e., lighter ones) represent the female family members.

6 Evaluation

Here, we summarize our methodology to evaluate the proposed inference framework.

6.1 Datasets

In order to evaluate our method we used two datasets:

- CEPH/Utah Pedigree 1463
- Manuel Corpas Family Pedigree

6.1.1 CEPH/UTAH Pedigree 1463

To evaluate the proposed inference algorithm, we used the CEPH/Utah Pedigree 1463 dataset [26][5]. We obtained the SNP data both in the genome variant (GVF) and variant call (VCF) formats. Dataset contains partial DNA sequences of 17 family members and we used 11 of these 17 individuals (to be consistent with the previous work). The family bonds between these 11 individuals are illustrated in Fig. 4.

We focused on 100 neighboring SNPs (on the DNA sequence) of the target family on the 22nd chromosome. We also collected data for calculating MAF and to model the higher order correlations in the genome. For this purpose, we used data of the CEU population from the 1000 Genomes Project and HapMap.

6.1.2 Manuel Corpas Family Pedigree

Manuel Corpas is a scientist, who released his family DNA dataset in variant call format on his website [27]. The dataset consists DNA sequences of father, mother, son (Manuel Corpas), daughter, and aunt. The family tree of the individuals in this dataset is illustrated in Fig. 5. Similar to the CEPH/UTAH Pedigree dataset setup, for this dataset, we focused on the 22nd chromosome and selected 100 neighboring SNPs of each family member.

6.2 Evaluation Metrics

Similar to [1], we evaluated the proposed framework in terms of both attacker's *incorrectness* and *uncertainty*. Incorrectness quantifies the adversary's error in inferring the

5. The previous work by Humbert et al. also use the same dataset.

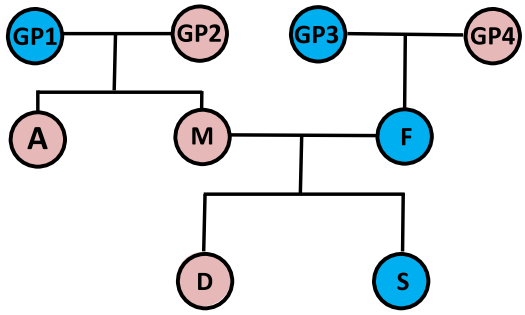


Fig. 5. Family tree of manuel corpas consisting of the nine family members that were considered. The blue nodes (i.e., darker ones) represent the male and the pink ones (i.e., lighter ones) represent the female family members. Genomic data for the grandparents (GP1, GP2, GP3, and GP4) is missing in the original dataset.

SNPs of the individuals in the target set. This metric can be expressed as follows:

$$E_j^i = \sum_{x_j^i \in \{0,1,2\}} p(x_j^i|\Psi)||x_j^i - \hat{x}_j^i||. \quad (5)$$

where, $\hat{x}_j^i$ is the true value of the inferred SNP, and Ψ includes all the information that is available to the attacker (as in Fig. 2). The incorrectness metric quantifies how far the adversary is away from the actual value of a SNP in his inference. We also evaluated the proposed scheme based on the attacker's uncertainty. For this purpose, we used the following normalized entropy metric from [1],

$$H_j^i = \frac{-\sum_{x_j^i \in \{0,1,2\}} p(x_j^i|\Psi)\log(x_j^i|\Psi)}{\log(3)}. \quad (6)$$

This can be described as the entropy of the adversary for an unobserved SNP. This metric quantifies the confidence of the adversary about his inference. Note that one needs the ground truth data in order to evaluate the incorrectness of the attacker. Here, by using both incorrectness and uncertainty metrics, we show the correlation between two, as in practice, it is not trivial to possess the ground truth data in order to evaluate the incorrectness of the attacker. That is, we show that one can also use the normalized entropy to quantify an individual's genomic privacy (and hence the strength of an inference attack). In fact, a recent work about genomic privacy metrics also reports that both incorrectness and uncertainty (normalized entropy) are suitable metrics to quantify genomic privacy (and hence the inference attack power) [28]. We compute the metrics in Equations (5) and (6) for each SNP and then take the average for all the SNPs in the unknown set $\mathbb{X}_U$.

6.3 Results

Due to the nature of kinship and characteristics of genomic data, we cannot avoid having cycles in our factor graph. Although there is no theoretical proof that our solution (and belief propagation algorithm in general) will converge to an optimal result in the presence of cycles, according to several runs of the algorithm on different SNPs, we observed that belief propagation converges with a significantly low error.

6.3.1 CEPH/UTAH Pedigree 1463

We conducted experiments for both high order correlation models (Markov chain and HMM). In the first experiment,

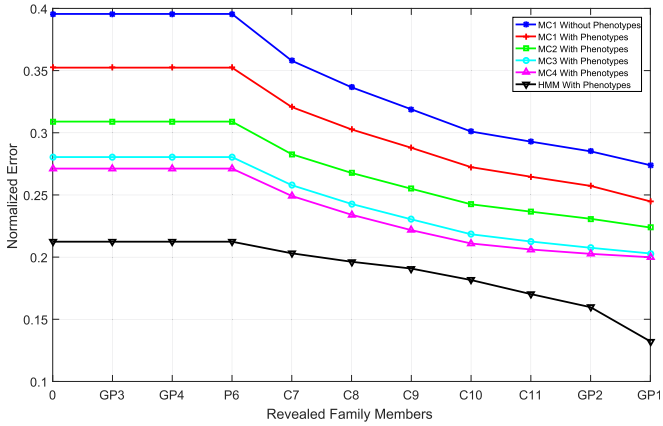


Fig. 6. Decrease in genomic privacy of P5 (in Fig. 4) in terms of the *incorrectness* of the attacker. We reveal partial genomes of other family members for different high order correlation models in the genome. MC stands for the Markov chain model (with different orders) and HMM stands for the hidden Markov model.

among the 100 SNPs we considered, we randomly hide 50 SNPs belonging to P5 in the CEPH/UTAH family (in Fig. 4) and tried to infer them by gradually increasing the background information of the attacker. We also assumed that the attacker knows the following 3 phenotypes of each family member (that are associated with the considered SNPs) [25].

- Verbal declarative memory-associated to *Rs5747035*
- Neurofibromatosis-associated to *Rs121434260*
- Crohn's disease-associated to *Rs4820425*

Because the information about these phenotypes in family members are not publicly available, we probabilistically simulated these phenotypes for the family members (using real probabilities obtained from [25]) and used these simulated phenotypes for the inference. Thus, the contribution of the phenotype information to the inference attack will remain the same if we use the real phenotype information about the individuals as well.

We started revealing 50 random SNPs (out of 100) of other family members (starting from the most distant one to the P5 in terms of number of hops in Fig. 4) and observe how the inference power of the attacker changes. We run each experiment 50 times and take the average of each privacy metric. We modelled the high order correlations via both the Markov chain model (for different orders-k) and HMM. We show our results for the attacker's incorrectness and uncertainty in Figs. 6 and 7, respectively. Note that the case when $k = 1$ (with no phenotype information) represents the previous work by Humbert et al. We observed that both the incorrectness and uncertainty of the attacker decreases by revealing more data. More importantly, our results show that high order correlations and phenotype information contributes significantly to the inference power of the attacker. In both figures, we see that for the Markov chain model, attacker's inference does not improve much for orders of Markov chain (k) that is larger than 3. We further discuss the relation between the amount of unobserved (hidden) SNPs and this bottleneck (about the order of the Markov chain) in Appendix B. We also observed that the HMM increases the attacker's inference power compared to the Markov chain model. In all experiments, the accuracy of the HMM is better than the Markov chain's accuracy, which is also consistent with the previous work [2].

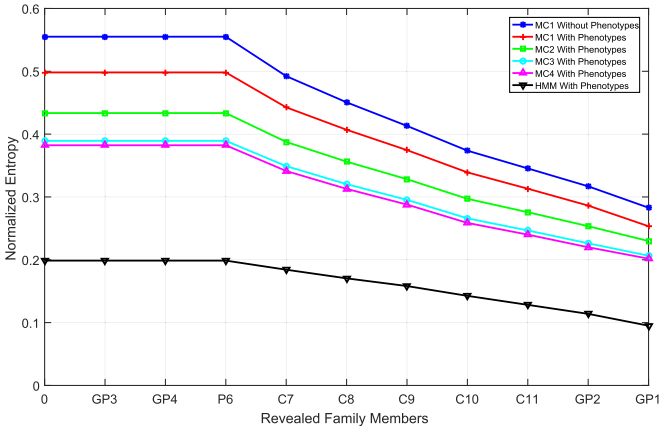


Fig. 7. Decrease in genomic privacy of P5 (in Fig. 4) in terms of the *uncertainty* of the attacker. We reveal partial genomes of other family members for different high order correlation models in the genome. MC stands for the Markov chain model (with different orders) and HMM stands for the hidden Markov model.

Next, to observe the effect of number of hidden SNPs to the high order correlation model, we run the same experiment for the Markov chain model and HMM by hiding different number of SNPs from the victim (P5) and the other family members. This time, we started revealing varying number of random SNPs (out of 100) of other family members (starting from the most distant one to the P5 as before) and observe the inference power of the attacker. In Figs. 8 and 9, we show our results for the Markov chain model when the order of the Markov chain (k) is 3. We observed that the inference power of the Markov chain model increases as more SNPs of the family members are observed. We obtained similar results for the HMM model (as before, we observed that HMM gives better accuracy compared to Markov chain for varying number of hidden SNPs). In order to show the standard deviations of the experiments, we also show the results with error bars in Appendix A.

6.3.2 Manuel Corpas Family Pedigree

We also evaluated our proposed attack on the Manuel Corpas Family Pedigree dataset. Here, we set our target as the mother (M in Fig. 5) and try to infer her unobserved SNPs.

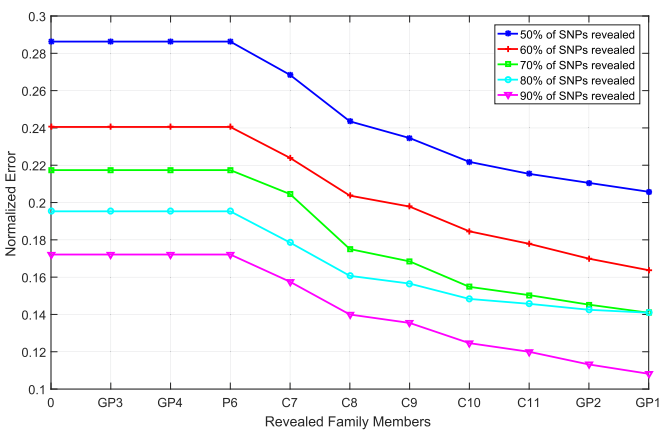


Fig. 8. Decrease in genomic privacy of P5 (in Fig. 4) in terms of the *incorrectness* of the attacker. We reveal different number of random SNPs from other family members and use the Markov chain model (with $k = 3$) to model the high order correlation in the genome.

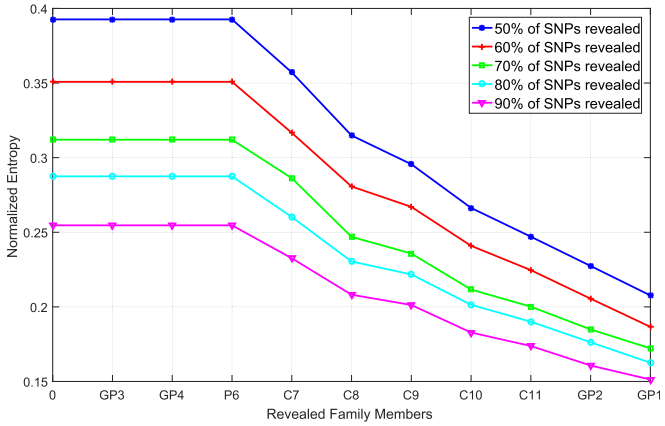


Fig. 9. Decrease in genomic privacy of P5 (in Fig. 4) in terms of the *uncertainty* of the attacker. We reveal different number of random SNPs from other family members and use the Markov chain model (with $k = 3$) to model the high order correlation in the genome.

Unlike the previous experiment, here, we started revealing from the closest family members to the farthest member to show that the strength of the proposed inference attack is independent of the dataset and evaluation methodology. Similar to the previous experiment, we assumed that the attacker knows the same set of three phenotypes about each member of this family and we revealed 50 random SNPs (out of 100) of other family members. We run each experiment 50 times and take the average of each privacy metric.

The results for this experiment (in terms of normalized error and normalized entropy) are given in Figs. 10 and 11. Obtained results are consistent with our expectations (error and entropy decrease with each revealed family member). Similar to the previous results, it can be seen that high order correlation and phenotype information contributes significantly to inference power of the attacker. In general, we observed that the results are consistent with CEPH/UTAH pedigree experiments. However, since we changed the order of revealing family members, unlike the previous results, here we observed a continuous decrease in error and entropy for the genomic privacy of the victim. This is because each family member has a direct effect on our inference power.

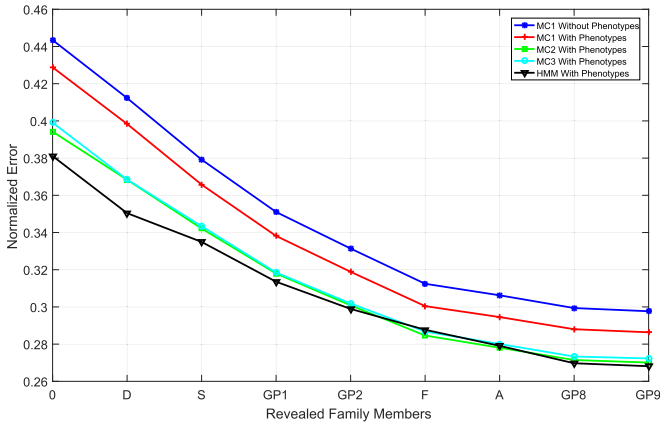


Fig. 10. Decrease in genomic privacy of M (in Fig. 5) in terms of the *incorrectness* of the attacker. We reveal partial genomes of other family members for different high order correlation models in the genome. MC stands for the Markov chain model (with different orders) and HMM stands for the hidden Markov model.

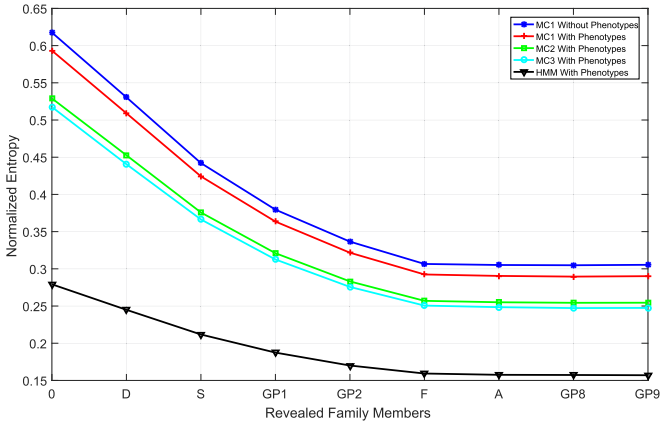


Fig. 11. Decrease in genomic privacy of M (in Fig. 5) in terms of the *uncertainty* of the attacker. We reveal partial genomes of other family members for different high order correlation models in the genome. MC stands for the Markov chain model (with different orders) and HMM stands for the hidden Markov model.

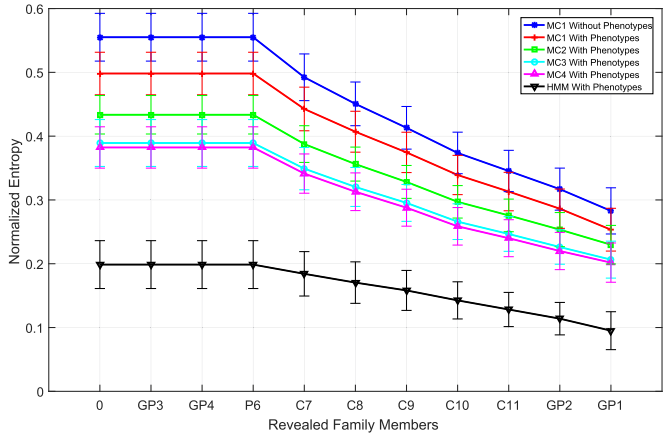


Fig. 13. Decrease in genomic privacy of P5 (in Fig. 4) in terms of the *uncertainty* of the attacker. We reveal partial genomes of other family members for different high order correlation models in the genome. MC stands for the Markov chain model (with different orders) and HMM stands for the hidden Markov model.

7 Conclusion and Future Work

In this work, we proposed an attack for inferring genomic data of individuals from publicly available data about themselves, their family members, and about genomics. We showed that the attacker can efficiently infer privacy-sensitive point mutations of an individual with high accuracy. We also showed that the proposed framework extends and significantly improves the existing work in this area. Establishing a unifying framework to quantify the genomic privacy of individuals using all publicly available resources, we believe that this work would be a significant step towards establishing a greater understanding of the privacy risks on the genomic data of individuals. As future work, we will extend this work and study the balance between privacy and utility. Once the genomic privacy of an individual is quantified, the proposed framework will provide recommendations to the individual (about sharing his genomic-related data) to reduce the risk on his genomic privacy. Furthermore, we will study different models for high order correlations, such as recurrent neural networks (which is shown to be a powerful technique for classifying time series data) to capture potential nonlinear relationships between the SNPs. We will also extend our evaluation on different chromosomes and other phenotypes.

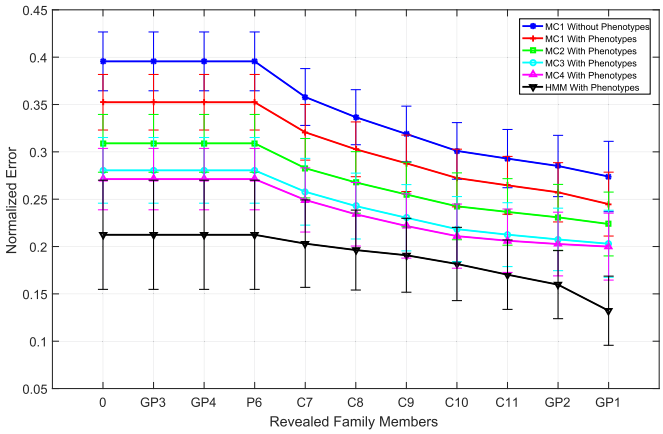


Fig. 12. Decrease in genomic privacy of P5 (in Fig. 4) in terms of the *incorrectness* of the attacker. We reveal partial genomes of other family members for different high order correlation models in the genome. MC stands for the Markov chain model (with different orders) and HMM stands for the hidden Markov model.

Appendix A
Standard Deviation of the Conducted Experiments

We computed and plotted the standard deviations of the experiments. In Figs. 12 and 13 we show CEPH/UTAH pedigree results with error bars which represents the standard deviation of 50 runs over uncertainty and incorrectness. As shown, the results from the experiments do not have significant deviations from the average.

Appendix B
Bottleneck of the Markov Chain Order

We have conducted two experiments on the UTAH family in order to see the relation between bottleneck of the Markov chain order and number of hidden SNPs. We hide 10 and 90 percent of SNPs of each family member and then start to infer the missing SNPs. In Figs. 14 and 15 we show the effect of

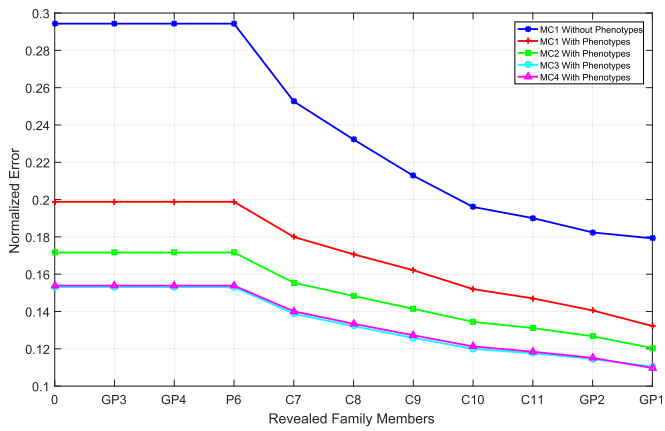


Fig. 14. Decrease in genomic privacy of P5 in terms of the *incorrectness* of the attacker, when we reveal 90 percent of random SNPs from other family members.

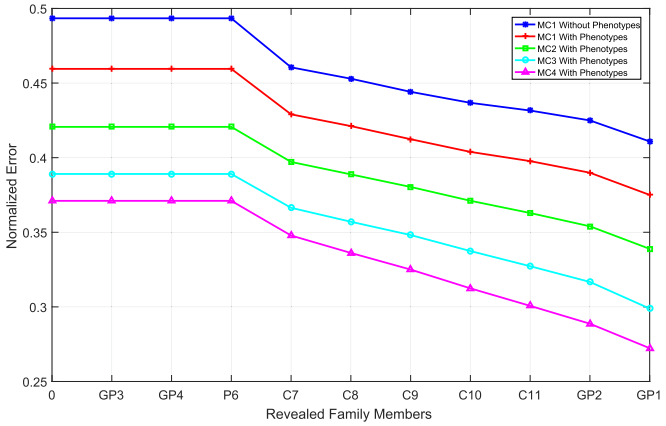


Fig. 15. Decrease in genomic privacy of P5 in terms of the *incorrectness* of the attacker, when we reveal 10 percent of random SNPs from other family members.

number of hidden SNPs on error (uncertainty) while inferring SNPs of P5. We conclude that the Markov chain bottleneck is related to the number of SNPs we try to infer. When the number of observed SNPs (by the attacker) is a lot, Markov models have more data to work with, and hence they converge to a small error value even with low order models. Thus, higher order models would not make the error any smaller. On the other hand, when the attacker observes fewer SNPs, increasing the order of the Markov chain model also increases the chance of inferring an unobserved SNP. For instance, in Fig. 14, when we reveal 90 percent of each family member's SNPs (i.e., when the attacker already observes a significant amount of data), results obtained by Markov order 3 and 4 are totally overlapping. However, in Fig. 15, when we reveal only 10 percent of each family member's SNPs, Markov order 4 does a significantly better job than Markov order 3.

Acknowledgments

Iman Deznabi and Mohammad Mobayen contributed equally. Erman Ayday is supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 707135 and by the Scientific and Technological Research Council of Turkey, TUBITAK, under Grant No. 115C130.

References

[1] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the concerns of the Lacks family: Quantification of kin genomic privacy," *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2013, pp. 1141–1152.

[2] S. S. Samani, et al., "Quantifying genomic privacy via inference attack with high-order SNV correlations," in *Proc. Workshop. Genome Privacy. Secur.*, 2015, pp. 32–40.

[3] M. Naveed, et al., "Privacy in the genomic era," *ACM Comput. Surveys*, vol. 48, no. 1, Sep. 2015, Art. no 6.

[4] J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik, "Privacy preserving error resilient DNA searching through oblivious automata," *Proc. ACM ACM SIGSAC Conf. Comput. Commun. Secur.*, 2007, pp. 519–528.

[5] M. Blanton and M. Aliasgari, "Secure outsourcing of DNA searching via finite automata," in *Proc. 24th Annu. IFIP WG 11.3 Work. Conf. Data Appl. Security Privacy*, 2010, pp. 49–64.

[6] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik, "Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes," *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2011, pp. 691–702.

[7] Y. Chen, B. Peng, X. Wang, and H. Tang, "Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds," in *Proc. 19th Netw. Distrib. Syst. Secur. Symp.*, 2012.

[8] S. Jha, L. Kruger, and V. Shmatikov, "Towards practical privacy for genomic computation," in *Proc. IEEE Symp. Secur. Privacy*, 2008, pp. 216–230.

[9] M. Naveed, et al., "Controlled functional encryption," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2014, pp. 1280–1291.

[10] N. Karvelas, A. Peter, S. Katzenbeisser, E. Tews, and K. Hamacher, "Privacy-preserving whole genome sequence processing through proxy-aided ORAM," in *Proc. 13th Workshop Privacy Electron. Soc.*, 2014, pp. 1–10.

[11] R. Wang, X. Wang, Z. Li, H. Tang, M. K. Reiter, and Z. Dong, "Privacy-preserving genomic computation through program specialization," *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2009, pp. 338–347.

[12] E. De Cristofaro, S. Faber, and G. Tsudik, "Secure genomic testing with size- and position-hiding private substring matching," in *Proc. 12th ACM Workshop Privacy Electron. Soc.*, 2013, pp. 107–118.

[13] E. Ayday, J. L. Raisaro, J. Rougemont, and J.-P. Hubaux, "Protecting and evaluating genomic privacy in medical tests and personalized medicine," in *Proc. 12th ACM Workshop Privacy Electron. Soc.*, 2013, pp. 95–106.

[14] N. Homer, S. Szelinger, M. Redman, D. Duggan, and W. Tembe, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, Aug. 2008, Art. no. e1000167.

[15] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Sci.*, vol. 339, Jan. 2013, Art. no. 6117.

[16] A. Johnson and V. Shmatikov, "Privacy-preserving data exploration in genome-wide association studies," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 1079–1087.

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

[18] E. Ayday, J. L. Raisaro, U. Hengartner, A. Molyneaux, and J.-P. Hubaux, "Privacy-preserving processing of raw genomic data," in *Data Privacy Management and Autonomous Spontaneous Security*. Berlin, Germany: Springer, 2013.

[19] Z. Huang, E. Ayday, J.-P. Hubaux, J. Fellay, and A. Juels, "GenoGuard: Protecting genomic data against brute-force attacks," in *Proc. IEEE Symp. Secur. Privacy*, 2015, pp. 447–462.

[20] D. S. Falconer and T. F. Mackay, *Introduction to Quantitative Genetics*, 4th Ed.. Harlow, Essex, U.K.: Addison Wesley Longman, 1996.

[21] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Burlington, MA, USA: Morgan Kaufmann Publishers, Inc., 1988.

[22] A. T. Ihler, W. F. John III, and A. S. Willsky, "Loopy belief propagation: Convergence and effects of message errors," *J. Mach. Learning Res.*, vol. 6, pp. 905–936, May 2005.

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

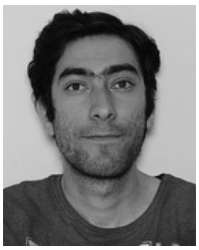
[24] N. Li and M. Stephens, "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data," *Genetics*, vol. 165, 2003, Art. no. 1039.

[25] (2016, May 1). [Online]. Available: http://www.snpedia.com/

[26] R. Drmanac, et al., "Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays," *Sci.*, vol. 327, no. 5961, pp. 78–81, 2010.

[27] (2016, Nov. 8). [Online]. Available: https://manuelcorpas.com/2016/05/24/my-personal-exome-analysis-part-i-first-findings-2/

[28] I. Wagner, "Evaluating the strength of genomic privacy metrics," *ACM Trans. Privacy Secur.*, vol. 20, no. 1, 2017, Art. no. 2.



Iman Deznabi is working toward the master's degree in the Department of Computer Engineering, Bilkent University, Ankara, Turkey. His research interests include bioinformatics and machine learning. He is a student member of the IEEE.

Mohammad Mobayen is working toward the master's degree in the Department of Computer Engineering, Bilkent University, Ankara, Turkey. His research interests include privacy-enhancing technologies and machine learning. He is a student member of the IEEE.

Nazanin Jafari is working toward the master's degree in the Department of Computer Engineering, Bilkent University, Ankara, Turkey. Her research interests include high performance computing and big data analysis. She is a student member of the IEEE.

Oznur Tastan received the BSc degree in biological sciences and bioengineering from Sabanci University and the PhD degree from Carnegie Mellon University, School of Computer Science, in 2011. Before joining Bilkent, she worked as a post-doctoral researcher at Microsoft Research New England Lab, Cambridge, Massachusetts. Since 2012, she has been affiliated with the Department of Computer Engineering at Bilkent University. She has worked on diverse problems in computational biology and machine learning. She is a member of the IEEE.

Erman Ayday received the MS and PhD degrees from Georgia Tech Information Processing, Communications and Security Research Lab (IPCAS) in the School of Electrical and Computer Engineering (ECE), Georgia Institute of Technology, Atlanta, Georgia, in 2007 and 2011, respectively, under the supervision of Dr. Faramarz Fekri. He is an assistant professor of computer science with Bilkent University, Ankara, Turkey. Before that, he was a post-doctoral researcher at EPFL, Switzerland, in the Laboratory for Communications and Applications 1 (LCA1) led by Prof. Jean-Pierre Hubaux. His research interests include privacy-enhancing technologies (including big data and genomic privacy), wireless network security, trust and reputation management, and applied cryptography. He is the recipient of the Distinguished Student Paper Award at IEEE S&P 2015, 2010 Outstanding Research Award from the Center of Signal and Image Processing (CSIP) at Georgia Tech, and 2011 ECE Graduate Research Assistant (GRA) Excellence Award from Georgia Tech. Other various accomplishments of his include several patents, research grants, and the H2020 Marie Curie individual fellowship. He is a member of the IEEE and the ACM.