Addressing the Concerns of the Lacks* Family: Quantification of Kin Genomic Privacy

Mathias Humbert Erman Ayday
Jean-Pierre Hubaux
Laboratory for Communications and Applications
EPFL, Lausanne, Switzerland
firstname.lastname@epfl.ch

Amalio Telenti
Institute of Microbiology
University Hospital of Lausanne
Lausanne, Switzerland
amalio.telenti@chuv.ch

ABSTRACT

The rapid progress in human-genome sequencing is leading to a high availability of genomic data. This data is notoriously very sensitive and stable in time. It is also highly correlated among relatives. A growing number of genomes are becoming accessible online (e.g., because of leakage, or after their posting on genome-sharing websites). What are then the implications for kin genomic privacy? We formalize the problem and detail an efficient reconstruction attack based on graphical models and belief propagation. With this approach, an attacker can infer the genomes of the relatives of an individual whose genome is observed, relying notably on Mendel's Laws and statistical relationships between the nucleotides (on the DNA sequence). Then, to quantify the level of genomic privacy as a result of the proposed inference attack, we discuss possible definitions of *genomic privacy* metrics. Genomic data reveals Mendelian diseases and the likelihood of developing degenerative diseases such as Alzheimer's. We also introduce the quantification of *health privacy*, specifically the measure of how well the predisposition to a disease is concealed from an attacker. We evaluate our approach on actual genomic data from a pedigree and show the threat extent by combining data gathered from a genome-sharing website and from an online social network.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection*; J.3 [**Life and Medical Sciences**]: *Biology and genetics*; K.4.1 [**Computer and Society**]: Public Policy Issues—*Privacy*

Keywords

Genomic Privacy; Inference Algorithms; Metrics; Kinship

*The family of Henrietta Lacks (August 1, 1920 - October 4, 1951), whose DNA was sequenced and published online without the consent of her family.

CCS'13, November 4–8, 2013, Berlin, Germany.

http://dx.doi.org/10.1145/2508859.2516707.

1. INTRODUCTION

With the help of rapidly developing technology, DNA sequencing is becoming less expensive. As a consequence, the research in genomics has gained speed in paving the way to personalized (genomic) medicine, and geneticists need large collections of human genomes to further increase this speed. Furthermore, individuals are using their genomes to learn about their (genetic) predispositions to diseases, their ancestries, and even their (genetic) compatibilities with potential partners. This trend has also caused the launch of health-related websites and online social networks (OSNs), in which individuals share their genomic data (e.g., OpenSNP [1] or 23andMe [2]). Thus, already today, thousands of genomes are available online.

Even though most of the genomes on the Internet are anonymized, it is possible to find genomes with the identifiers of their owners (e.g., OpenSNP [1]). Furthermore, it has been shown that anonymization is not sufficient for protecting the real identities of the genome donors [29,47]. Once the owner of a genome is identified, he is faced with the risk of discrimination (e.g., by employers or insurance companies) [9]. Some believe that they have nothing to hide about their genetic structure, hence they might decide to give full consent for the publication of their genomes on the Internet to help genomic research. However, our DNA sequences are highly correlated to our relatives' sequences. The DNA sequences between two random human beings are 99.9% similar, and this value is even higher for closely related people. Consequently, somebody revealing his genome does not only damage his own genomic privacy, but also puts his relatives' privacy at risk [46]. Moreover, currently, a person does not need consent from his relatives to share his genome online. This is precisely where the interesting part of the story begins: *kin genomic privacy.*

A recent New York Times' article [3] reports the controversy about sequencing and publishing, without the permission of her family, the genome of Henrietta Lacks (who died in 1951). On the one hand, the family members think that her genome is private family information and it should not be published without the consent of the family. On the other hand, some scientists argued that the genomes of current family members have changed so much over time (due to gene mixing during reproduction), that nothing accurate could be told about the genomes of current family members by using Henrietta Lacks' genome. As we will also show in this work, they are wrong. Minutes after Henrietta Lacks' genome was uploaded to a public website called SNPedia, researchers produced a report full of personal information

about Henrietta Lacks. Later, the genome was taken offline, but it had already been downloaded by several people, hence both her and (partially) the Lacks family's genomic privacy was already lost.

Unfortunately, the Lacks, even though possibly the most publicized family facing this problem, are not the only family facing this threat. As we mentioned before, the genomes of thousands of individuals are available online. Once the identity of a genome donor is known, an attacker can learn about his relatives (or his family tree) by using an auxiliary side channel, such as an OSN, and infer significant information about the DNA sequences of the donor's relatives. We will show the feasibility of such an attack and evaluate the privacy risks by using publicly available data on the Web.

Although the researchers took Henrietta Lacks' genome offline from SNPedia, other databases continue to publish portions of her genomic data. Publishing only portions of a genome does not, however, completely hide the unpublished portions; even if a person reveals only a part of his genome, other parts can be inferred using the statistical relationships between the nucleotides in his DNA. For example, James Watson, co-discoverer of DNA, made his whole DNA sequence publicly available, with the exception of one gene known as Apolipoprotein E (ApoE), one of the strongest predictors for the development of Alzheimer's disease. However, later it was shown that the correlation (called *linkage disequilibrium* by geneticists) between one or multiple polymorphisms and ApoE can be used to predict the ApoE status [40]. Thus, an attacker can also use these statistical relationships (which are publicly available) to infer the DNA sequences of a donor's family members, even if the donor shares only part of his genome. It is important to note that these privacy threats not only jeopardize kin genomic privacy, but, if not properly addressed, these issues could also hamper genomic research due to untimely fear of potential misuse of genomic information.

In this work, we evaluate the genomic privacy of an individual threatened by his relatives revealing their genomes. Focusing on the most common genetic variant in human population, single nucleotide polymorphism (SNP), and considering the statistical relationships between the SNPs on the DNA sequence, we quantify the loss in genomic privacy of individuals when one or more of their family members' genomes are (either partially or fully) revealed. To achieve this goal, first, we design a reconstruction attack based on a well-known statistical inference technique. The computational complexity of the traditional ways of realizing such inference grows exponentially with the number of SNPs (which is on the order of tens of millions) and relatives. Therefore, in order to infer the values of the unknown SNPs in linear complexity, we represent the SNPs, family relationships and the statistical relationships between SNPs on a factor graph and use the belief propagation algorithm [37, 41] for inference. Then, using various metrics, we quantify the genomic privacy of individuals and show the decrease in their level of genomic privacy caused by the published genomes of their family members. We also quantify the health privacy of the individuals by considering their (genetic) predisposition to certain serious diseases. We evaluate the proposed inference attack and show its efficiency and accuracy by using real genomic data of a pedigree. More importantly, by using genomic data and pedigree information we collected from a public genome-sharing website and an OSN, we show that the proposed inference attack threatens not only the Lacks family, but also many other families.

The rest of the paper is organized as follows. In Section 2, we give a brief background on genomics and belief propagation. In Section 3, we present the proposed framework in detail. In Section 4, we evaluate the performance of the proposed inference attack using different metrics. In Section 5, we show how the proposed inference attack threatens the genomic and health privacy of several families gathered from OSNs. In Section 6, we summarize the related work on genetic inference and genomic-privacy protection. Finally, we conclude the paper in Section 7.

2. BACKGROUND

In this section, we briefly introduce the relevant genetic principles, as well as the concept of belief propagation.

2.1 Genomics 101

DNA is a double-helix structure that consists of two complementary polymer chains. Genetic information is encoded on the DNA as a sequence of nucleotides (A,T,G,C) and a human DNA includes around 3 billion nucleotide pairs. With the decreasing cost of DNA sequencing, genomic data is currently being used mainly in the following two areas: (i) clinical diagnostics, for personalized genomic medicine and genetic research (e.g., genome-wide association studies[1]), and (ii) direct-to-consumer genomics, for genetic risk estimation of various diseases or for recreational activities such as ancestry search. In the following, we briefly introduce some concepts, which we use throughout this paper, about the human genome and reproduction.

2.1.1 Single Nucleotide Polymorphism

As already mentioned, human beings have 99.9% of their DNA in common. Thus, there is no need to focus on the whole DNA but rather on the most important variants. Single nucleotide polymorphism (SNP) is the most common DNA variation in human population. A SNP occurs when a nucleotide (at a specific position on the DNA) varies between individuals of a given population (as illustrated in Fig. 1). There are approximately 50 million SNP positions in human population [4]. Recent discoveries show that the susceptibility of an individual to several diseases can be computed from his SNPs [5, 33]. For example, it has been reported that two particular SNPs (rs7412 and rs429358) on the Apolipoprotein E (ApoE) gene indicate an (increased) risk for Alzheimer's disease. SNPs carry privacy-sensitive information about individuals' health, hence we will quantify health privacy focusing on individuals' published (or inferred) SNPs and the diseases they reveal.

In general, two different nucleotides (called alleles) are observed at a given SNP position: (i) the major allele is the most frequently observed nucleotide, and (ii) the minor allele is the rare nucleotide.[2] From here on, we represent the major allele as B for a SNP position, and the minor allele as b (where both B and b are in $\{A, T, G, C\}$).

Furthermore, each SNP position contains two nucleotides (one inherited from the mother and one from the father, as we will discuss next). Thus, the content of a SNP position

[1]Examination of many genetic variants in different individuals to determine if any variant is associated with a trait.
[2]The two alleles for the SNP position in Fig. 1 are C and T.

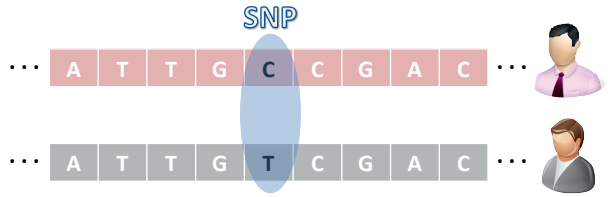


Figure 1: Single nucleotide polymorphism (SNP) with alleles C and T illustrated on a single string of two different individuals' DNAs.

		Father (*F*)		
		BB	Bb	bb
Mother (*M*)	BB	(1,0,0)	(0.5,0.5,0)	(0,1,0)
	Bb	(0.5,0.5,0)	(0.25,0.5,0.25)	(0,0.5,0.5)
	bb	(0,1,0)	(0,0.5,0.5)	(0,0,1)

Table 1: Mendelian inheritance probabilities for a SNP j, given different genotypes for the parents. The probabilities of the child's genotype is represented in parentheses. Each table entry represents $\big(\Pr(x_j^C = BB|x_j^M, x_j^F), \Pr(x_j^C = Bb|x_j^M, x_j^F), \Pr(x_j^C = bb|x_j^M, x_j^F)\big)$.

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. We represent the content of a SNP position as x_j^i for SNP j at individual i, where $x_j^i \in \{BB, Bb, bb\}$. For simplicity of presentation, in the rest of the paper, we denote BB as 0, Bb as 1, and bb as 2 (i.e., $x_j^i \in \{0, 1, 2\}$). Finally, each SNP i is assigned a minor allele frequency (MAF), p_i^b, which represents the frequency at which the minor allele (b) of the corresponding SNP occurs in a given population (typically, $0 < p_i^b < 0.5$).

2.1.2 Reproduction

Mendel's First Law states that alleles are passed independently from parents to children for different meioses (the process of cell division necessary for reproduction). For each SNP position, a child inherits one allele from his mother and one from his father. Each allele of a parent is inherited by a child with equal probability of 0.5. Let $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$ be the function modeling the Mendelian inheritance for a SNP j, where (M, F, C) represent mother, father, and child, respectively. We illustrate the Mendelian inheritance probabilities for a SNP j in Table 1.

Based on $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$, we can say that, given both parents' genomes, a child's genome is conditionally independent of all other ancestors' genomes.

2.1.3 Linkage Disequilibrium

As we discussed before, DNA sequences are highly correlated, leading to interdependent privacy risks. Linkage disequilibrium (LD) [24] is a correlation that appears between any pair of SNP positions in the whole genome due to the population's genetic history. Because of LD, the content of a SNP position can be inferred from the contents of other SNP positions. The strength of the LD between two SNP positions is usually represented by r^2 (or D'), where $r^2 = 1$ represents the strongest LD relationship.

2.2 Belief Propagation

Belief propagation [37, 41] is a message-passing algorithm for performing inference on graphical models (Bayesian networks, Markov random fields). It is typically used to compute marginal distributions of unobserved variables conditioned on 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) based on the dependences between variables. 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. Belief propagation is commonly used in artificial intelligence and information theory. It has demonstrated empirical success in numerous applications including LDPC codes [42], reputation management [11, 12], and recommender systems [10].

3. THE PROPOSED FRAMEWORK

In this section, we formalize our approach and present the different components that will allow us to quantify kin genomic privacy. Fig. 2 gives an overview of the framework.

In a nutshell, the goal of the adversary is to infer some *targeted SNPs* of a member (or multiple members) of a *targeted family*. We define $\mathbf{F}$ to be the set of family members in the targeted family (whose family tree, showing the familial connections between the members, is denoted as $\mathcal{G}_\mathbf{F}$) and $\mathbf{S}$ to be the set of SNP IDs (i.e., positions on the DNA sequence), where $|\mathbf{F}| = n$ and $|\mathbf{S}| = m$. Note that the SNP IDs are the same for all the members of the family. We also let 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 introduced in Section 2.1). Furthermore, $\mathbf{X}^i = \{x_j^i : j \in \mathbf{S}, i \in \mathbf{F}\}$ represents the set of SNPs for individual i. We let $\mathbb{X}$ be the $n \times m$ matrix that stores the values of the SNPs of all family members. Some entries of $\mathbb{X}$ might be known by the adversary (the observed genomic data of one or more family members) and others might be unknown. We denote the set of SNPs from $\mathbb{X}$ whose values are unknown as $\mathbb{X}_\mathrm{U}$, and the set of SNPs from $\mathbb{X}$ whose values are known (by the adversary) as $\mathbb{X}_\mathrm{K}$.

$\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$ is the function representing the Mendelian inheritance probabilities (in Table 1), where (M, F, C) represent mother, father, and child, respectively. The $m \times m$ matrix $\mathbb{L}$ represents the pairwise linkage disequilibrium (LD) between the SNPs in $\mathbf{S}$, that can be expressed by r^2 and

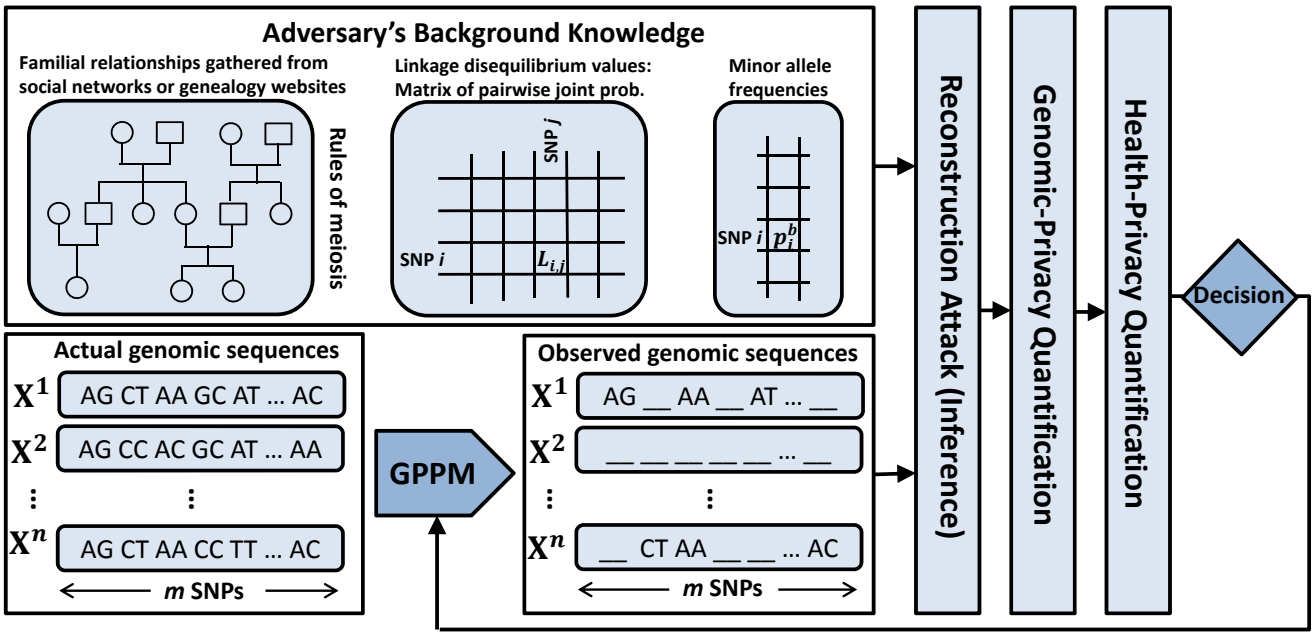


Figure 2: Overview of the proposed framework to quantify kin genomic privacy. Each vector $\mathbf{X}^i$ ($i \in \{1, \ldots, n\}$) includes the set of SNPs for an individual in the targeted family. Furthermore, each letter pair in $\mathbf{X}^i$ represents a SNP x_j^i; and for simplicity, each SNP x_j^i can be represented using $\{BB, Bb, bb\}$ (or $\{0, 1, 2\}$), as discussed in Section 2.1.1. Once the health privacy is quantified, the family should ideally decide whether to reveal less or more of their genomic information through the genomic-privacy preserving mechanism (GPPM).

D'; $\mathbb{L}_{i,j}$ refers to the matrix entry at row i and column j. $\mathbb{L}_{i,j} > 0$ if i and j are in LD, and $\mathbb{L}_{i,j} = 0$ if these two SNPs are independent (i.e., there is no LD between them). $\mathbf{P} = \{p_i^b : i \in \mathbf{S}\}$ represents the set of minor allele probabilities (or MAF) of the SNPs in $\mathbf{S}$. Finally, note that a joint probability $p(x_i, x_j)$ can be derived from $\mathbb{L}_{i,j}$, p_i^b, and p_j^b.

The adversary carries out a reconstruction attack to infer $\mathbb{X}_\mathrm{U}$ by relying on his background knowledge, $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$, $\mathbb{L}$, $\mathbf{P}$, and on his observation $\mathbb{X}_\mathrm{K}$. Once the targeted SNPs are inferred by the adversary, we evaluate genomic and health privacy of the family members based on the adversary's success and his certainty about the targeted SNPs and the diseases they reveal. Finally, we discuss some ideas to preserve the individuals' genomic and health privacy.

3.1 Adversary Model

An adversary is defined by his objective(s), attack(s), and knowledge. The objective of the adversary is to compute the values of the targeted SNPs for one or more members of a targeted family by using (i) the available genomic data of one or more family members, (ii) the familial relationships between the family members, (iii) the rules of reproduction (in Section 2.1.2), (iv) the minor allele frequencies (MAFs) of the nucleotides, and (v) the population LD values between the SNPs. We note that (i) and (ii) can be gathered online from genome-sharing websites and OSNs, and (iii), (iv), and (v) are publicly known information. Note that, in the future, the increasing possibility to accurately sequence, and to impute the actual haplotypes carried by an individual in each of the copies of the diploid genome will allow a more accurate inference of relatives' genotype than relying on population LD patterns only.

Various attacks can be launched, depending on the adversary's interest. The adversary might want to infer one particular SNP of a specific individual (targeted-SNP-targeted-relative attack) or one particular SNP of multiple relatives in the targeted family (targeted-SNP-multiple-relatives attack) by observing one or more other relatives' SNP at the same position. Furthermore, the adversary might also want to infer multiple SNPs of the same individual (multiple-SNP-targeted-relative attack) or multiple SNPs of multiple family members (multiple-SNP-multiple-relatives attack) by observing SNPs at various positions of different relatives. In this paper, we propose an algorithm that implements the latter attack, from which any other attacks can be carried out. We formulate this attack as a statistical inference problem.

3.2 Inference Attack

We formulate the reconstruction attack (on determining the values of the targeted SNPs) as finding the marginal probability distributions of unknown variables $\mathbb{X}_\mathrm{U}$, given the known values in $\mathbb{X}_\mathrm{K}$, familial relationships, and the publicly available statistical information. We represent the marginal distribution of a SNP j for an individual i as $p(x_j^i|\mathbb{X}_\mathrm{K})$.

These marginal probability distributions could traditionally be extracted from $p(\mathbb{X}_\mathrm{U}|\mathbb{X}_\mathrm{K}, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbb{L}, \mathcal{G}_\mathbf{F}, \mathbf{P})$, which is the joint probability distribution function of the variables in $\mathbb{X}_\mathrm{U}$, given the available side information and the observed SNPs. Then, clearly, each marginal probability distribution could be obtained as follows:

$$p(x_j^i|\mathbb{X}_\mathrm{K}) = \sum_{\mathbb{X}_\mathrm{U} \setminus \{x_j^i\}} p(\mathbb{X}_\mathrm{U}|\mathbb{X}_\mathrm{K}, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbb{L}, \mathcal{G}_\mathbf{F}, \mathbf{P}), \quad (1)$$

where the notation $\mathbb{X}_{\mathrm{U}} \backslash \{x_j^i\}$ implies all variables in $\mathbb{X}_{\mathrm{U}}$ except x_j^i. However, the number of terms in (1) grows exponentially with the number of variables, making the computation infeasible considering the scale of the human genome (which includes tens of million of SNPs). In the worst case, the computation of the marginal probabilities has a complexity of $\mathrm{O}(3^{nm})$. Thus, we propose to factorize the joint probability distribution function into products of simpler local functions, each of which depends on a subset of variables. These local functions represent the conditional dependences (due to LD and reproduction) between the different variables in $\mathbb{X}$. Then, by running the belief propagation algorithm on a factor graph, we can compute the marginal probability distributions in linear complexity (with respect to nm).

A factor graph is a bipartite graph containing two sets of nodes (corresponding to variables and factors) and edges connecting these two sets. Following [37], 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 two types of factor nodes: (i) *familial factor node*, representing the familial relationships and reproduction, and (ii) *LD factor node*, representing the LD relationships between the SNPs. We summarize the connections between the variable and factor nodes below (Fig. 3):

- Each variable node x_j^i has its familial factor node f_j^i and they are connected. Furthermore, x_j^k ($k \neq i$) is also connected to f_j^i if k is the mother or father of i (in $\mathcal{G}_{\mathbf{F}}$). Thus, the maximum degree of a familial factor node is 3.

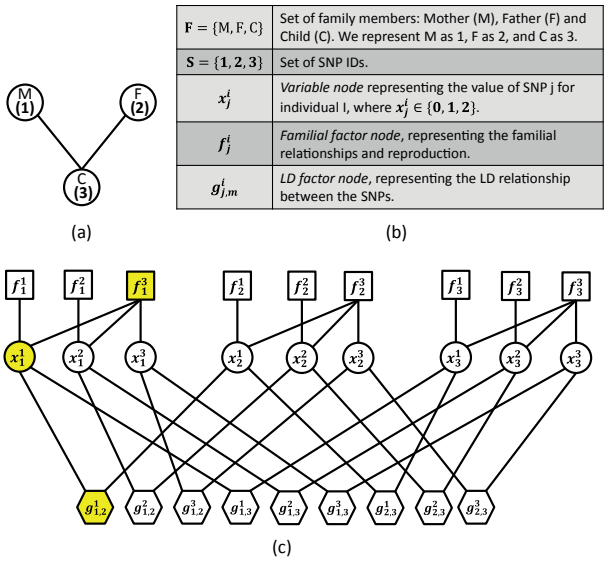
- Variable nodes x_j^i and x_m^i are connected to a LD factor node $g_{j,m}^i$ if SNP j is in LD with SNP m. Since the LD relationships are pairwise between the SNPs, the degree of a LD factor node is always 2.

Given the conditional dependences given by reproduction and LD, the global distribution $p(\mathbb{X}_{\mathrm{U}}|\mathbb{X}_{\mathrm{K}}, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbb{L}, \mathcal{G}_{\mathbf{F}}, \mathbf{P})$ can be factorized into products of several local functions, each having a subset of variables from $\mathbb{X}$ as arguments:

$$p(\mathbb{X}_{\mathrm{U}}|\mathbb{X}_{\mathrm{K}}, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbb{L}, \mathcal{G}_{\mathbf{F}}, \mathbf{P}) = \frac{1}{Z}\Big[\prod_{i\in\mathbf{F}}\prod_{j\in\mathbf{S}} f_j^i(x_j^i, \Theta(x_j^i), \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P})\Big] \times \Big[\prod_{i\in\mathbf{F}}\prod_{\substack{(j,m)\ \text{s.t.}\\ \mathbb{L}_{j,m}\neq 0}} g_{j,m}^i(x_j^i, x_m^i, \mathbb{L}_{j,m})\Big], \quad (2)$$

where Z is the normalization constant, and $\Theta(x_j^i)$ is the set of values of SNP j for the mother and father of i (in $\mathcal{G}_{\mathbf{F}}$).

Next, we introduce the messages between the factor and the variable nodes to compute the marginal probability distributions using belief propagation. We denote the messages from the variable nodes to the factor nodes as μ. We also denote the messages from familial factor nodes to variable nodes as λ, and from LD factor nodes to variable nodes as β. Let $\mathbb{X}^{(\nu)} = \{x_j^{i\,(\nu)} : j \in \mathbf{S}, i \in \mathbf{F}\}$ be the collection of variables representing the values of the variable nodes at the iteration ν of the algorithm. The message $\mu_{i\to k}^{(\nu)}(x_j^{i\,(\nu)})$ denotes the probability of $x_j^{i\,(\nu)} = \ell$ ($\ell \in \{0,1,2\}$), at the ν^{th} iteration. Furthermore, $\lambda_{k\to i}^{(\nu)}(x_j^{i\,(\nu)})$ denotes the probability that $x_j^{i\,(\nu)} = \ell$, for $\ell \in \{0,1,2\}$, at the ν^{th} iteration given $\Theta(x_j^i)$, $\mathcal{F}_R(x_j^M, x_j^F, x_j^C)$, and $\mathbf{P}$. Finally, $\beta_{k\to i}^{(\nu)}(x_j^{i\,(\nu)})$ denotes the probability that $x_j^{i\,(\nu)} = \ell$, for $\ell \in \{0,1,2\}$, at the ν^{th} iteration given the LD relationships between the SNPs.



$\mathbf{F} = \{\mathrm{M}, \mathrm{F}, \mathrm{C}\}$	Set of family members: Mother (M), Father (F) and Child (C). We represent M as 1, F as 2, and C as 3.
$\mathbf{S} = \{\mathbf{1}, \mathbf{2}, \mathbf{3}\}$	Set of SNP IDs.
x_j^i	*Variable node* representing the value of SNP j for individual I, where $x_j^i \in \{\mathbf{0}, \mathbf{1}, \mathbf{2}\}$.
f_j^i	*Familial factor node*, representing the familial relationships and reproduction.
$g_{j,m}^i$	*LD factor node*, representing the LD relationship between the SNPs.

Figure 3: The factor graph representation of a trio (mother, father, child) using 3 SNPs. (a) $\mathcal{G}_{\mathbf{F}}$, showing the familial connections among the trio. (b) descriptions of the notations in the factor graph. (c) factor graph representation of the trio using SNPs in $\mathbf{S} = \{1, 2, 3\}$. The message passing is described on the nodes (x_1^1, f_1^3, and $g_{1,2}^1$) highlighted in the graph.

For the clarity of presentation, we choose a simple family tree consisting of a trio (i.e., mother, father, and child) in Fig 3(a), and 3 SNPs (i.e., $|\mathbf{F}| = 3$ and $|\mathbf{S}| = 3$). In Fig. 3(c), we show how the trio and the SNPs are represented on a factor graph, where $i = 1$ represents the mother, $i = 2$ represents the father, and $i = 3$ 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_1^1), the familial factor node of the child (f_1^3), and the LD factor node $g_{1,2}^1$. The belief propagation algorithm iteratively exchanges messages between the factor and the variable nodes in Fig. 3(c), updating the beliefs on the values of the targeted SNPs (in $\mathbb{X}_{\mathrm{U}}$) at each iteration, until convergence. We denote the variable and factor nodes x_1^1, f_1^3, and $g_{1,2}^1$ with the letters i, k, and z, respectively.

The variable nodes generate their messages (μ) and send to their neighbors. Variable node i forms $\mu_{i\to k}^{(\nu)}(x_1^{1\,(\nu)})$ 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 ν^{th} iteration is given by

$$\mu_{i\to k}^{(\nu)}(x_1^{1\,(\nu)}) = \frac{1}{Z} \times \prod_{w\in(\sim k)} \lambda_{w\to i}^{(\nu-1)}(x_1^{1\,(\nu-1)}) \times \prod_{y\in\{z, g_{1,3}^1\}} \beta_{y\to i}^{(\nu-1)}(x_1^{1\,(\nu-1)}), \quad (3)$$

[3]The message $\mu_{i\to z}^{(\nu)}(x_1^{1\,(\nu)})$ from the variable node i LD factor node z is constructed similarly.

where Z is a normalization constant, and the notation ($\sim k$) means all familial factor node neighbors of the variable node i, except k. This computation is repeated for every neighbor of each variable node. It is important to note that the message in (3) is valid if the value of x_1^1 is unknown to the adversary (i.e., $x_1^1 \in \mathbb{X}_\mathrm{U}$). However, the value of x_1^1 can also be observed by the adversary (i.e., $x_1^1 \in \mathbb{X}_\mathrm{K}$). Thus, if $x_1^1 \in \mathbb{X}_\mathrm{K}$ and $x_1^1 = \rho$ ($\rho \in \{0,1,2\}$), then $\mu_{i\to k}^{(\nu)}(x_1^{1(\nu)} = \rho) = 1$ and $\mu_{i\to k}^{(\nu)}(x_1^{1(\nu)}) = 0$ for other potential values of x_1^1 (regardless of the values of the messages received by the variable node i from its neighbors).

Next, the factor nodes generate their messages. The message from the familial factor node k to the variable node i at the ν^{th} iteration is formed using the principles of belief propagation as

$$\lambda_{k\to i}^{(\nu)}(x_1^{1(\nu)}) = \sum_{\{x_1^2, x_1^3\}} f_1^3(x_1^1, \Theta(x_1^1), \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P}) \prod_{y\in\{x_1^2,x_1^3\}} \mu_{y\to k}^{(\nu)}(x_1^{1(\nu)}). \quad (4)$$

Note that $f_1^3(x_1^1, \Theta(x_1^1), \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P}) \propto p(x_1^1|\Theta(x_1^1), \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P})$, and this probability is computed using Table 1. Furthermore, if the degree of the familial factor node is 1 for a particular SNP, then the local function corresponding to the familial factor node only depends on the MAF of the corresponding SNP. For example, the degree of f_1^1 (in Fig. 3(c)) is 1, hence $f_1^1(x_1^1, \Theta(x_1^1), \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbf{P}) \propto p(x_1^1|p_1^b)$. The above computation must be performed for every neighbor of each familial factor node.

Similarly, the message from the LD factor node z to the variable node i at the ν^{th} iteration is formed as

$$\beta_{z\to i}^{(\nu)}(x_1^{1(\nu)}) = \sum_{x_2^1} g_{1,2}^1(x_1^1, x_2^1, \mathbb{L}_{1,2}) \prod_{y\in\{x_2^1\}} \mu_{y\to k}^{(\nu)}(x_1^{1(\nu)}). \quad (5)$$

As before, this computation is performed for every neighbor of each LD factor node. We further note that $g_{1,2}^1(x_1^1, x_2^1, \mathbb{L}_{1,2}) \propto p(x_1^1, x_2^1)$, which is derived from $\mathbb{L}_{1,2}$, p_1^b, and p_2^b. The algorithm proceeds to the next iteration in the same way as the ν^{th} iteration.

The algorithm starts at the variable nodes. Thus, at the first iteration of the algorithm (i.e., $\nu = 1$), the variable node i sends messages to its neighboring factor nodes based on the following rules: (i) If the value of x_1^1 is unknown to the adversary ($x_1^1 \in \mathbb{X}_\mathrm{U}$), $\mu_{i\to k}^{(1)}(x_1^{1(1)}) = 1$ for all potential values of x_1^1 and, (ii) if the value of x_1^1 is known to the adversary ($x_1^1 \in \mathbb{X}_\mathrm{K}$) and $x_1^1 = \rho$ ($\rho \in \{0,1,2\}$), $\mu_{i\to k}^{(1)}(x_1^{1(1)} = \rho) = 1$ and $\mu_{i\to k}^{(1)}(x_1^{1(1)}) = 0$ for other potential values of x_1^1. The iterations stop when all variables in $\mathbb{X}_\mathrm{U}$ have converged. The marginal probability of each variable in $\mathbb{X}_\mathrm{U}$ is given by multiplying all the incoming messages at each variable node.

3.3 Computational Complexity

The computational complexity of the proposed inference attack is proportional to the number of factor nodes. In our setting, there are nm familial factor nodes and a maximum of $nm(m-1)/2$ LD factor nodes. Hence, the worst-case computational complexity per iteration is $\mathrm{O}(nm^2)$. However, as each SNP is in LD with a limited number of other SNPs, the matrix $\mathbb{L}$ is sparse and the number of LD factor nodes grows with m rather than with $m(m-1)/2$, especially if we focus on SNPs in strong LD only. Thus, the average computational complexity per iteration is $\mathrm{O}(nm)$. Based on our experiments, we can state that the number of iterations before convergence is a small constant, between 10 and 15. Note finally that this complexity can be further reduced by using similar techniques developed for message-passing decoding of LDPC codes (e.g., working in log-domain [20]).

3.4 Privacy Metrics

A crucial step towards protecting kin genomic privacy is to quantify the privacy loss induced by the release of genomic information. Through the inference attack, the adversary infers the targeted SNPs (in $\mathbb{X}_\mathrm{U}$) belonging to the members of a targeted family by using his background knowledge and observed genomic data (of the family members). The inferred information can be expressed as the posterior distribution $p(\mathbb{X}_\mathrm{U}|\mathbb{X}_\mathrm{K}, \mathcal{F}_R(x_j^M, x_j^F, x_j^C), \mathbb{L}, \mathcal{G}_\mathbf{F}, \mathbf{P})$. Moreover, each posterior marginal probability distribution is represented as $p(x_j^i|\mathbb{X}_\mathrm{K})$, for all $i \in \mathbf{F}, j \in \mathbf{S}$. We propose to quantify kin genomic privacy using the following metrics: expected estimation error (incorrectness) and uncertainty.[4]

Correctness was already proposed in the context of location privacy [45]. In our scenario, correctness quantifies the adversary's success in inferring the targeted SNPs. That is, it quantifies the expected distance between the adversary's estimate on the value of a SNP, x_j^i ($x_j^i \in \mathbb{X}_\mathrm{U}$) and the true value of the corresponding SNP, $\hat{x_j^i}$. This distance can be expressed as the expected estimation error as follows:

$$E_j^i = \sum_{x_j^i \in \{0,1,2\}} p(x_j^i|\mathbb{X}_\mathrm{K})||x_j^i - \hat{x_j^i}||. \quad (6)$$

Privacy can also be represented as the adversary's *uncertainty* [22, 43], that is the ambiguity of $p(x_j^i|\mathbb{X}_\mathrm{K})$. This uncertainty is generally considered to be maximum if the posterior distribution is uniform. This definition of uncertainty can be quantified as the (normalized) entropy of $p(x_j^i|\mathbb{X}_\mathrm{K})$ as follows:

$$H_j^i = \frac{-\sum_{x_j^i\in\{0,1,2\}} p(x_j^i|\mathbb{X}_\mathrm{K}) \log p(x_j^i|\mathbb{X}_\mathrm{K})}{\log(3)}. \quad (7)$$

The higher the entropy is, the higher is the uncertainty.

Finally, we propose another entropy-based metrics that quantifies the mutual dependence between the hidden genomic data that the adversary is trying to reconstruct, and the observed data. This is quantified by mutual information $I(x_j^i; \mathbb{X}_\mathrm{K}) = H(x_j^i) - H(x_j^i|\mathbb{X}_\mathrm{K})$ [8]. As privacy decreases with mutual information, we propose the following (normalized) privacy metrics:

$$I_j^i = 1 - \frac{H(x_j^i) - H(x_j^i|\mathbb{X}_\mathrm{K})}{H(x_j^i)} = \frac{H(x_j^i|\mathbb{X}_\mathrm{K})}{H(x_j^i)}. \quad (8)$$

The aforementioned metrics are useful for quantifying the genomic privacy of individuals. In order to quantify a more tangible privacy, we must convert these genomic-privacy metrics into health-privacy metrics. To quantify an individual's health privacy, we focus on his predisposition to different diseases. Let $\mathbf{S_d}$ be the set of IDs of the SNPs that are associated with a disease d. Then, a metrics quantifying the

[4]These metrics are not specific to the proposed inference attack; they can be used to quantify genomic privacy in general.

health privacy for an individual i regarding the disease d can be defined as follows:

$$D_d^i = \frac{1}{\sum_{k \in \mathbf{S_d}} c_k} \sum_{k \in \mathbf{S_d}} c_k G_k^i, \quad (9)$$

where G_k^i is the genomic privacy of a SNP k for individual i, computed using (6), (7), or (8), and c_k is the contribution of SNP k to disease d.[5] Other health-privacy metrics based on non-linear combinations of genotypes or combinations of alleles will be defined in future work. Note that health-privacy metrics are valid at a given time, and cannot be used to evaluate future privacy provision, as genome research can change knowledge on the contribution of SNPs to diseases.

3.5 Genomic-Privacy Preserving Mechanism

Individuals willing to share genomic data for research or recreational purposes might be unwilling to share all their DNA sequence, and thus need to properly obfuscate the sensitive part(s) before releasing their genomic data. To do so, their DNA will go through an obfuscation process, that we call *genomic-privacy preserving mechanism* (GPPM). GPPM can be implemented using one of the following techniques: (i) hiding the SNPs, or (ii) reducing the precision or the quantity of the revealed SNPs.

Hiding all or specific SNPs can be achieved either by not releasing them or by encrypting them. Obviously, not releasing any of the SNPs would hinder genetic research, thus it is not a preferred way to protect the genomic privacy of individuals. Instead of not releasing the SNPs, the use of cryptographic algorithms to encrypt the genome is proposed. For example, Kantarcioglu *et al.* propose using homomorphic encryption on the SNPs of the individuals to perform genetic research on the encrypted SNPs [35]. However, the security of an individual's genome should be guaranteed for at least 70-100 years (i.e., during the typical lifetime of a human). As we show in this paper, even lifelong protection is not enough, considering kin privacy implications (e.g., for offsprings). It is known that even the best of the cryptographic algorithms we use today could be broken in around 30 years. Therefore, the appropriateness of cryptographic techniques for storing and processing the genomic data has been questioned due to long-term security requirements of the genomic data.

As an alternative to the cryptographic techniques, utility (i.e., precision and quantity of the revealed SNPs) can be traded for privacy. The precision of the revealed SNPs can be reduced, for example, by revealing only one of the two alleles of a SNP. Similarly, family members' SNPs can be selectively revealed by also considering the previously revealed SNPs from the corresponding family (to keep the genomic privacy of other family members above a desired threshold): we evaluate the privacy provided by this technique in Section 4 by assessing the inference power of the adversary for different fractions of observed data from a targeted family.

Eventually, using one of the above techniques, the GPPM will take $\mathbb{X}$ as input and output $\mathbb{X}_\mathrm{K}$ as the set of revealed SNPs. We note that a detailed implementation of the GPPM by using one of the aforementioned techniques is out of the scope of this work. We plan to study it in the future.

[5]These contributions are determined as a result of medical studies. Some SNPs might increase (or decrease) the risk for a disease more than others.

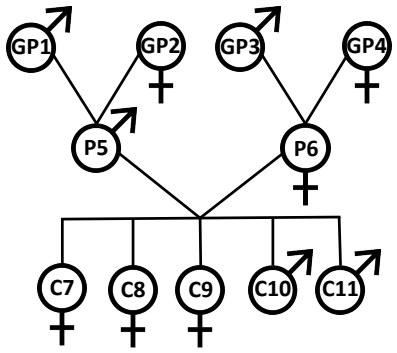


Figure 4: Family tree of *CEPH/Utah Pedigree 1463* consisting of the 11 family members that were considered. The symbols ♂ and ♀ represent the male and female family members, respectively.

4. EVALUATION

In this section, we first evaluate the performance of the proposed inference attack, then compare the performance of the inference with and without considering the linkage disequilibrium (LD) between SNPs, and finally evaluate the entropy-based metrics with respect to the expected estimation error in quantifying the genomic privacy.

For this evaluation, we use the *CEPH/Utah Pedigree 1463* that contains the partial DNA sequences of 17 family members (4 grandparents, 2 parents, and 11 children) [23]. We note in Fig. 4 that we only use 5 (out of 11) children for our evaluation because (i) 11 is much above the average number of children per family, (ii) we observe that the strength of adversary's inference does not increase further (due to the children's revealed genomes) when more that 5 children's genomes are revealed, and (iii) the belief propagation algorithm (in Section 3.2) might have convergence issues due to the number of loops in the factor graph, and this number increases with the number of children. As the SNPs related to important diseases, like Alzheimer's, are not included in this dataset, we quantify health privacy in Section 5 by using the data collected from a genome-sharing website.

To quantify the genomic privacy of the individuals in the CEPH family, we focus on their SNPs on chromosome 1 (which is the largest chromosome). We rely on the three metrics introduced in Section 3.4. That is, we compute the genomic privacy of each family member using the expected estimation error in (6), the (normalized) entropy in (7), and the (normalized) mutual information in (8) on the targeted SNPs, and we average the result based on the number of targeted SNPs for each individual. We rely on the L_1 norm to measure the distance between two SNP values in (6).

First, we assume that the adversary targets one family member and tries to infer his/her SNPs by using the published SNPs of other family members without considering the LD between the SNPs. We select an individual from the CEPH family and denote him as the target individual. We construct $\mathbf{S}$, the set of SNP IDs that we consider for evaluation, from 80k SNPs on chromosome 1. Thus, the set of targeted SNPs ($\mathbb{X}_\mathrm{U}$) includes 80k SNPs of the target individual. Furthermore, we gradually fill the set of observed SNPs ($\mathbb{X}_\mathrm{K}$) with the set of 80k SNPs of other family members. That is, we sequentially reveal 80k SNPs (whose IDs are in $\mathbf{S}$) of all family members (excluding the target in-

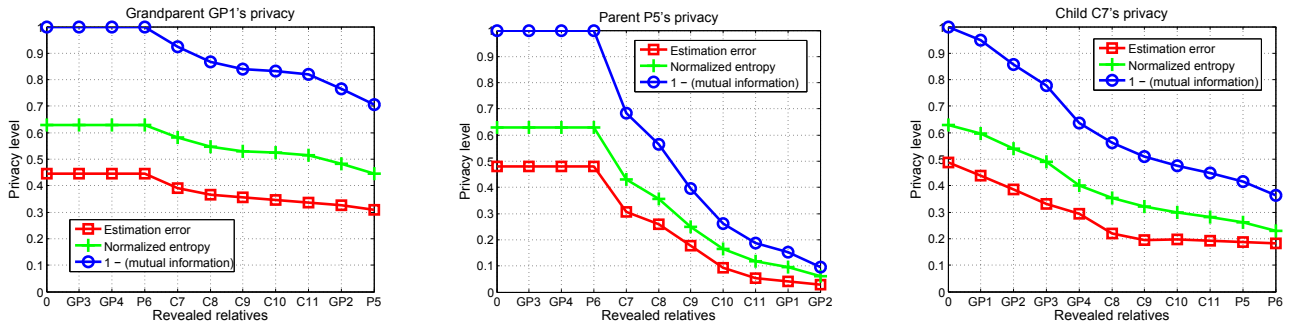


Figure 5: Evolution of the genomic privacy of the (a) grandparent (GP1), (b) parent (P5), and (c) child (C7). We reveal all the 80k SNPs on chromosome 1 of other family members starting from the most distant family members of the target individual (in terms of number of hops to the target individual in Fig. 4); the x-axis represents the disclosure sequence. We note that $x = 0$ represents the prior distribution, when no genomic data is observed by the adversary.

dividual) beginning with the most distant family members from the target individual (in terms of number of hops in Fig. 4) and we keep revealing relatives until we reach his/her closest family members.[6]

In Fig. 5 we show the evolution of the genomic privacy of three target individuals from the CEPH family (in Fig. 4): (i) grandparent (GP1), (ii) parent (P5), and (iii) child (C7). We note that all entropy-based metrics for each target individual start from the same values. We also observe that the parent's and the child's genomic privacy decreases considerably more than the grandparent's (the adversary's error for the grandparent's genome does not go below 0.3). Moreover, the observation of GP3, GP4 and P6's genomes has no effect on GP1 and P5's privacy as their genomes are independent (if no other relatives' genomes are observed). We observe in Fig. 5(a) that the grandparent's genomic privacy is mostly affected by the SNPs of the first revealed children (C7, C8), and also by those of his spouse and his child (P5). We also observe (in Fig. 5(b)) that, by revealing all family members' SNPs (expect P5), the adversary can almost reach an estimation error of 0. The target parent's genomic privacy significantly decreases only with the observation of his children's and spouse's SNPs. Finally, we observe in Fig. 5(c) that C7's genomic privacy decreases smoothly with the observation of his grandparents' SNPs, and then of his siblings'. We also observe a slight decrease of privacy once the parents' SNPs (P5 and P6) are also revealed, but the observation of parents (after the other children) does not have a significant effect on the adversary's error. It is important to note that the importance of a family member for the inference power of the adversary also depends on the sequence at which his/her SNPs are revealed in Fig. 5. For example, in Fig. 5(c), if the SNPs of the parents (P5 and P6) of the target child (C7) were revealed before her siblings (C8-C11), then the observation of her parents would reduce the genomic privacy of the target child more than her siblings (but the final genomic privacy would not change).

Next, we include the LD relationships and observe the change in the inference power of the adversary using the LD values. We construct $\mathbf{S}$ from 100 SNPs on chromosome 1. Among these 100 SNPs, each SNP is in LD with 5 other SNPs on average. Furthermore, the strength of the LD (r^2 value in Section 2.1.3) uniformly varies between 0.5 and 1 (where $r^2 = 1$ represents the strongest LD relationship, as discussed before). We note that we only use 100 SNPs for this study as the LD values are not yet completely defined over all SNPs, and the definition of such values is still an ongoing research. As before, we define a target individual from the CEPH family, construct the set $\mathbb{X}_\mathrm{U}$ from his/her SNPs, and sequentially reveal other family members' SNPs to observe the decrease in the genomic privacy of the target individual. We observe that individuals sometimes reveal different parts of their genomes (e.g., different sets of SNPs) on the Internet. Thus, we assume that for each family member (except for the target individual), the adversary observes 50 random SNPs from $\mathbf{S}$ only (instead of all the SNPs in $\mathbf{S}$), and these sets of observed SNPs are different for each family member. In Fig. 6, we show the evolution of genomic privacy of three target individuals when the adversary also uses the LD values. We observe that LD decreases genomic privacy, especially when few individuals' genomes are revealed. As more family member's genomes are observed, LD has less impact on the genomic privacy.

We also evaluate the inference power of the adversary to infer multiple SNPs among all family members, given a subset of SNPs belonging to some family members, and also considering the LD between SNPs. That is, we evaluate the inference power of the adversary for different fractions of observed data for the family members. Using the same set of 100 SNPs, we construct $\mathbb{X}_\mathrm{U}$ from $(\kappa \times 100 \times n)$ SNPs, randomly selected from all family members, where n is the number of family members in the family tree ($n = 11$ for this scenario), and $0 \leq \kappa \leq 1$. We assume that the SNPs that are not in $\mathbb{X}_\mathrm{U}$ are observed by the adversary (i.e., in $\mathbb{X}_\mathrm{K}$), and we observe the inference power of the adversary for the SNPs in $\mathbb{X}_\mathrm{U}$, for different values of κ. In Fig. 7, we observe an exponential decrease in the global genomic privacy (privacy of all family members), showing that the observation of a small portion of the family's SNPs can have a huge impact on genomic privacy. The estimation error is decreased by around 3 by observing only the first 10% of the SNPs.

[6]The exact sequence of the family members (whose SNPs are revealed) is indicated for each evaluation.

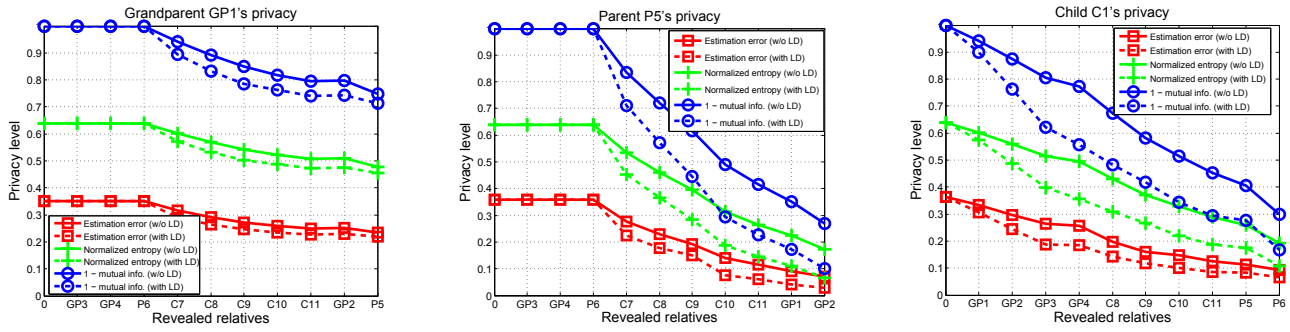


Figure 6: Evolution of the genomic privacy of the (a) grandparent (GP1), (b) parent (P5), and (c) child (C7), with and without considering LD. For each family member, we reveal 50 randomly picked SNPs (among 100 SNPs in S), starting from the most distant family members, and the x-axis represents the exact sequence of this disclosure. Note that $x = 0$ represents the prior distribution, when no genomic data is revealed.

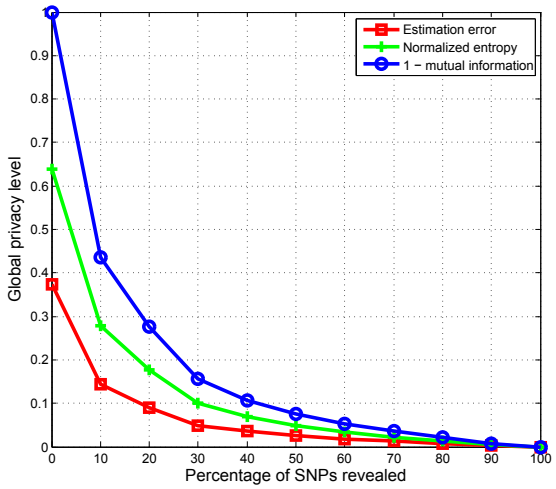


Figure 7: Evolution of the global privacy for the whole family by gradually revealing 10% of SNPs.

5. EXPLOITING GENOME-SHARING WEBSITES AND ONLINE SOCIAL NETWORKS

In order to show that the proposed inference attack threatens not only the Lacks family, but potentially *all* families, we collected publicly available data from a genome-sharing website and familial relationships from an OSN, and evaluated the decrease in genomic and health privacy of people due to the observation of their relatives' genomic data.

We gathered individuals' genomic data from OpenSNP [1], a website on which people can publicly share sets of SNPs. Then, we identified the owners of some gathered genomic profiles by using their names and sometimes profile pictures. Among these identified individuals, we managed to find family relationships of 6 of them (who publicly reveal the names of some of their relatives) on Facebook.[7] We expect this number to increase in the future, as more health-related OSNs (which let people share their genomic profiles, such as 23andMe [2]) emerge. Furthermore, we anticipate that the current widely used health-related OSNs (e.g., PatientsLikeMe [6]) will let users upload and share their genomic data. We identified 29 target individuals from 6 different families, whose genomic data can be inferred using the observed SNPs of the identified individuals.

We focus on 2 individuals I_1 and I_2 out of these 6 identified individuals and evaluate the genomic and health privacy for their family members. We observed that both I_1 and I_2 publicly disclosed around 1 million of their SNPs. Furthermore, we identified the names of (i) 1 mother, 2 sons, 2 daughters, 1 grandchild, 1 aunt, 2 nieces, and 1 nephew of I_1, and (ii) 1 sibling, 1 aunt, 1 uncle, and 6 cousins of I_2 on Facebook. We compute the genomic and health privacy of these target individuals using the (normalized) entropy in (7) on the targeted SNPs, and normalize the result based on the number of targeted SNPs for each individual. We do not use the expected estimation error in (6), as we do not have the ground truth for the genomes of the target individuals. Thus, privacy is quantified as the uncertainty of the adversary in this section.

To quantify the genomic privacy of the target individuals (i.e., family members of I_1 and I_2), we first construct $\mathbf{S}$ from all SNPs on chromosome 1 (from the observed genomes of I_1 and I_2). The set of observed SNPs ($\mathbb{X}_\mathrm{K}$) includes the observed SNPs of I_1 (respectively I_2) for the inference of family members of I_1 (respectively I_2). The set of targeted SNPs ($\mathbb{X}_\mathrm{U}$) includes 77k SNPs for I_1's family and 79k for I_2's family (from $\mathbf{S}$) for each evaluation. In Fig. 8, we show the decrease in the genomic privacy for different family members of I_1 (aunt, niece/nephew, grandchild, mother, child) and I_2 (cousin, aunt/uncle, sibling) as a result of our proposed inference attack, first without considering the LD dependencies (similarly to previous section). We observe that as expected, the decrease in the genomic privacy of close family members is significantly higher than that of more distant family members. However, as we have seen in Section 4, the observation of one (or more) additional family member(s) has often much more impact on the target's privacy than the observation of only one relative.

[7]According to [28], around 12% of Facebook users publicly share at least one family member on their profiles.

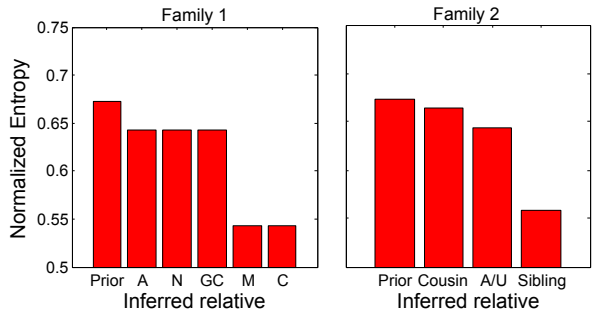


Figure 8: Attacker's uncertainty about all SNP values on chromosome 1 for two different families, without using LD. A stands for aunt, N for niece/nephew, GC for grandchild, M for mother, C for child, U for uncle. Same notations are used in Fig. 9 and 10.

.

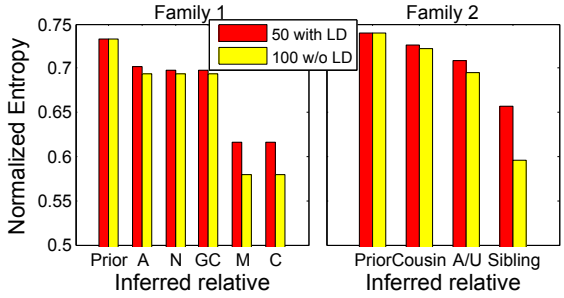


Figure 9: Attacker's uncertainty about values of 100 SNPs on chromosome 1 for two families, by observing (i) all 100 SNPs of the relative that reveals his/her genome, and (ii) only 50 SNPs but using LD.

In Fig. 9, we display the decrease of genomic privacy with respect to 100 SNPs of chromosome 1.[8] We first show the different privacy levels by using all 100 SNPs of the observed relative (i.e., I_1 or I_2), and then show the same by using only 50 SNPs of the observed relative and LD values. We note that the use of LD decreases privacy slightly more for the first familyt than for the second family. This is because we randomly picked 50 different SNPs for both families, and those picked in the second family had weaker LD relationships with other SNPs. We finally observe that the difference between the two observation cases (50 SNPs with LD and 100 SNPs without LD) is higher for close relatives (mother, child, or sibling) than for others.

We also evaluate the health privacy of the family members of I_1 and I_2 considering their predispositions to various diseases. We first noticed that almost all important SNPs for privacy-sensitive diseases affected by genomic factors, like Alzheimer's, ischemic heart disease, or macular degeneration, were revealed by I_1 and I_2. Due to lack of space, we focus on Alzheimer's as it is one of the most important diseases that are mainly attributable to genetic factors. Having two ApoE4 alleles (in SNPs rs7412 and rs429358 located on chromosome 19) dramatically increases an individual's probability of having Alzheimer's by the age of 80. Thus, the contents of these two SNPs carry privacy-sensitive information for individuals. We use the metrics in (9) to quantify the health privacy of family members for Alzheimer's disease. We assign equal weights to both associated SNPs (as their combination determines the predisposition to Alzheimer's disease). In Fig. 10, we show the attacker's uncertainty about the predisposition to Alzheimer's disease for the family members of I_1 and I_2. We notice a decrease of around 0.2 (from 0.5 to 0.3) in uncertainty between close relatives. Clearly, the knowledge of the SNPs of more relatives would further worsen the situation.

[8]We consider only 100 SNPs here for the same reason as in Section 4.

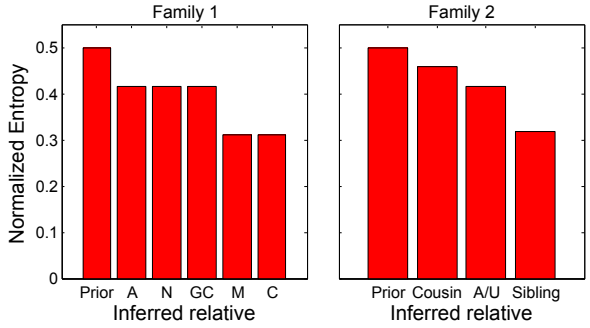


Figure 10: Adversary's uncertainty about Alzheimer's disease predisposition for 2 families.

6. RELATED WORK

Stajano *et al.* [46] were among the first to raise the issue of kin privacy in genomics. Cassa *et al.* [19] provide a framework for measuring the risks to siblings of someone who reveals his SNPs. They show that the inference error is substantially reduced when the sibling's SNPs are known, compared to when only the population frequencies are used. We push this work further, by considering any kind of family members, and LD relationship between SNPs, by proposing and evaluating different privacy metrics, and by presenting a real attack scenario using publicly available data. Our generic framework considers any observation of a family's genomic data, and the adversary's background knowledge.

Several algorithms for inference on graphical models have been proposed in the context of pedigree analysis. Exact inference techniques on Bayesian networks are used in order to map disease genes and construct genetic maps [26,34,38]. Monte Carlo methods (Gibbs sampling) were also proved to be efficient for genetic analyses in the case of complex pedigrees [31,44,48]. All these methods aim to infer specific genotypes given phenotypes (like diseases). Another paper relies on Gibbs sampling in order to infer haplotypes (used in association studies) from genotype data [36]. Genotype imputation [39] is another technique used by geneticists to complete missing SNPs based upon given genotyped data. A similar approach has recently been used to infer high-density genotypes in pedigrees, by relying notably on low-resolution genotypes and identity-by-descent regions of the genome [18]. None of these contributions addresses privacy.

We also briefly summarize the research on the privacy of genomic data in the following. Homer *et al.* [30] prove that de-identification is an ineffective way to protect the

privacy of genomic data, which is also supported by other works [27, 50, 52]. Most recently, Gymrek *et al.* [29] show how they identified DNAs of several individuals who participated in scientific studies. Fienberg *et al.* [25] propose using differential privacy to protect the identities of scientific study participants, however this approach reduces the accuracy of the research results. Some pieces of work also focus on protecting the privacy of genomic data and on preserving utility in medical tests such as (i) search of a particular pattern in the DNA sequence [16,49], (ii) comparing the similarity of DNA sequences [15,17,32], and (iii) performing statistical analysis on several DNA sequences [35]. Furthermore, Ayday *et al.* propose privacy-preserving schemes for medical tests and personalized medicine methods that use patients' genomic data [7,14]. For privacy-preserving clinical genomics, a group of researchers proposes to outsource some costly computations to a public cloud or semi-trusted service provider [21, 51]. Finally, Ayday *et al.* propose techniques for privacy-preserving management of raw genomes [13].

In contrast with these contributions, in this paper, we propose a novel and efficient inference attack in order to reconstruct genomic data of individuals given observed genomic data of their family members and special characteristics of genomic data. Furthermore, we quantify the genomic privacy of individuals as a result of this attack using different metrics, and show the real threat by using the data collected from different websites and OSNs.

7. CONCLUSION AND FUTURE WORK

We have proposed a novel reconstruction attack for inferring the genomic data of individuals from the observed genomes of their relatives, and we have compared several metrics to quantify genomic and health privacy.

As pointed out by Rebecca Skloot, the author of "The Immortal Life of Henrietta Lacks", the view we have today of genomes is like a world map, but Google Street View is coming very soon. This growing precision can be highly beneficial in terms of personalized medicine, but it can have devastating consequences on a family's peace of mind. As we already mentioned, the Lacks family is just one (albeit famous) example. In the future (and already today), people of the same family might have very different opinions on whether to reveal genomic data, and this can lead to dissent: relatives might have divergent perceptions of possible consequences. It is high time for the security research community to prepare itself for this formidable challenge. The genetic community is highly concerned about the fact that the proliferation of negative stories could potentially lead to a negative perception by the population and to tighter laws, thus hampering scientific progress in this field.

In future work, we plan to apply the proposed framework to more pedigrees, in order to fine tune our numerical results. We will also study the trade-off between utility and privacy of genomic data in order to design an optimized GPPM.

Acknowledgments

We would like to thank Jacques Fellay, Paul J. McLaren, Vincent Mooser and Jacques Rougemont for their insights and suggestions, Kévin Huguenin and Reza Shokri for their valuable comments on the submitted manuscript, and the anonymous reviewers for their helpful feedback.

8. REFERENCES

[1] `http://opensnp.org/`. Visited on 9-Aug-2013.

[2] `https://www.23andme.com/welcome/`. Visited on 9-Aug-2013.

[3] `http://www.nytimes.com/2013/03/24/opinion/sunday/the-immortal-life-of-henrietta-lacks-the-sequel.html?pagewanted=all`. Visited on 9-Aug-2013.

[4] `http://www.ncbi.nlm.nih.gov/projects/SNP/`. Visited on 9-Aug-2013.

[5] `http://www.eupedia.com/genetics/medical_dna_test.shtml`. Visited on 9-Aug-2013.

[6] `http://www.patientslikeme.com/`. Visited on 9-Aug-2013.

[7] `http://lca.epfl.ch/projects/genomic-privacy/`.

[8] D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 247–255. ACM, 2001.

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

[10] E. Ayday, A. Einolghozati, and F. Fekri. BPRS: Belief propagation based iterative recommender system. *IEEE ISIT*, 2012.

[11] E. Ayday and F. Fekri. Belief propagation based iterative trust and reputation management. *IEEE Transactions on Dependable and Secure Computing*, 9(3), 2012.

[12] E. Ayday and F. Fekri. BP-P2P: A belief propagation-based trust and reputation management for P2P networks. *SECON*, 2012.

[13] E. Ayday, J. L. Raisaro, U. Hengartner, A. Molyneaux, and J. P. Hubaux. Privacy-preserving processing of raw genomic data. *DPM 2013*, 2013.

[14] E. Ayday, J. L. Raisaro, P. J. McLaren, J. Fellay, and J. P. Hubaux. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. *HealthTech*, 2013.

[15] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik. Countering GATTACA: Efficient and secure testing of fully-sequenced human genomes. *CCS*, 2011.

[16] M. Blanton and M. Aliasgari. Secure outsourcing of DNA searching via finite automata. *DBSec*, 2010.

[17] F. Bruekers, S. Katzenbeisser, K. Kursawe, and P. Tuyls. Privacy-preserving matching of DNA profiles. Technical report, 2008.

[18] J. T. Burdick, W.-M. Chen, G. R. Abecasis, and V. G. Cheung. In silico method for inferring genotypes in pedigrees. *Nature genetics*, 38(9):1002–1004, 2006.

[19] C. A. Cassa, B. Schmidt, I. S. Kohane, and K. D. Mandl. My sister's keeper?: genomic research and the identifiability of siblings. *BMC Medical Genomics*, 1(1):32, 2008.

[20] J. Chen, A. Dholakia, E.Eleflhetiou, M. Fossotier, and X.-Y. Hu. Near optimum reduced-complexity decoding algorithm for LDPC codes. *IEEE ISIT*, 2002.

[21] Y. Chen, B. Peng, X. Wang, and H. Tang. Large-scale privacy-preserving mapping of human genomic sequences on hybrid clouds. *NDSS*, 2012.

[22] C. Diaz, S. Seys, J. Claessens, and B. Preneel. Towards measuring anonymity. In *Privacy Enhancing Technologies*, pages 54–68. Springer, 2003.

[23] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, N. L. Burns, B. G. Kermani, P. Carnevali, I. Nazarenko, G. B. Nilsen, G. Yeung, et al. Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. *Science*, 327(5961):78–81, 2010.

[24] D. S. Falconer and T. F. Mackay. *Introduction to Quantitative Genetics (4th Edition)*. Addison Wesley Longman, Harlow, Essex, UK, 1996.

[25] S. E. Fienberg, A. Slavkovic, and C. Uhler. Privacy preserving GWAS data sharing. *Proceedings of the IEEE 11th International Conference on Data Mining Workshops (ICDMW)*, Dec. 2011.

[26] M. Fishelson and D. Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18(suppl 1):S189–S198, 2002.

[27] J. Gitschier. Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project. *Am. J. Hum. Genet.*, 84:251–258, 2009.

[28] P. Gundecha, G. Barbier, and H. Liu. Exploiting vulnerability to secure user privacy on a social networking site. In *KDD*. ACM, 2011.

[29] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, , and Y. Erlich. Identifying personal genomes by surname inference. *Science: 339 (6117)*, Jan. 2013.

[30] 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*, 4, Aug. 2008.

[31] C. S. Jensen, A. Kong, and U. Kjærulff. Blocking gibbs sampling in very large probabilistic expert systems. *International Journal of Human Computer Studies*, 42(6):647–666, 1995.

[32] S. Jha, L. Kruger, and V. Shmatikov. Towards practical privacy for genomic computation. *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, pages 216–230, 2008.

[33] A. D. Johnson and C. J. O'Donnell. An open access database of genome-wide association results. *BMC Medical Genetics 10:6*, 2009.

[34] M. I. Jordan. Graphical models. *Statistical Science*, pages 140–155, 2004.

[35] M. Kantarcioglu, W. Jiang, Y. Liu, and B. Malin. A cryptographic approach to securely share and query genomic sequences. *IEEE Transactions on Information Technology in Biomedicine*, 12(5):606–617, 2008.

[36] B. Kirkpatrick, E. Halperin, and R. M. Karp. Haplotype inference in complex pedigrees. *Journal of Computational Biology*, 17(3):269–280, 2010.

[37] F. Kschischang, B. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47, 2001.

[38] S. L. Lauritzen and N. A. Sheehan. Graphical models for genetic analyses. *Statistical Science*, pages 489–514, 2003.

[39] Y. Li, C. Willer, S. Sanna, and G. Abecasis. Genotype imputation. *Annual review of genomics and human genetics*, 10:387, 2009.

[40] D. Nyholt, C. Yu, and P. Visscher. On Jim Watson's APOE status: Genetic information is hard to hide. *European Journal of Human Genetics*, 17:147–149, 2009.

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

[42] H. Pishro-Nik and F. Fekri. On decoding of low-density parity-check codes on the binary erasure channel. *IEEE Transactions on Information Theory*, 50:439–454, March 2004.

[43] A. Serjantov and G. Danezis. Towards an information theoretic metric for anonymity. In *Privacy Enhancing Technologies*, pages 41–53. Springer, 2003.

[44] N. Sheehan. On the application of markov chain monte carlo methods to genetic analyses on complex pedigrees. *International Statistical Review*, 68(1):83–110, 2000.

[45] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. Quantifying location privacy. In *IEEE Symposium on Security and Privacy*, 2011.

[46] F. Stajano, L. Bianchi, P. Liò, and D. Korff. Forensic genomics: kin privacy, driftnets and other open questions. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*, 2008.

[47] L. Sweeney, A. Abu, and J. Winn. Identifying participants in the personal genome project by name. *Available at SSRN 2257732*, 2013.

[48] A. Thomas, A. Gutin, V. Abkevich, and A. Bansal. Multilocus linkage analysis by blocked Gibbs sampling. *Statistics and Computing*, 10(3):259–269, 2000.

[49] J. R. Troncoso-Pastoriza, S. Katzenbeisser, and M. Celik. Privacy preserving error resilient DNA searching through oblivious automata. *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, 2007.

[50] R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou. Learning your identity and disease from research papers: Information leaks in genome wide association study. *Proceedings of the 16th ACM CCS*, pages 534–544, 2009.

[51] R. Wang, X. Wang, Z. Li, H. Tang, M. K. Reiter, and Z. Dong. Privacy-preserving genomic computation through program specialization. *Proceedings of the 16th ACM CCS*, pages 338–347, 2009.

[52] X. Zhou, B. Peng, Y. F. Li, Y. Chen, H. Tang, and X. Wang. To release or not to release: Evaluating information leaks in aggregate human-genome data. *ESORICS*, 2011.