Dissecting Privacy Risks in Biomedical Data

Pascal Berrang*, Mathias Humbert†, Yang Zhang*, Irina Lehmann‡, Roland Eils§¶, Michael Backes‖
*CISPA, Saarland Informatics Campus
†Swiss Data Science Center, ETH Zurich and EPFL
‡Helmholtz Center for Environmental Research Leipzig, UFZ
§German Cancer Research Center (DKFZ)
¶University of Heidelberg
‖CISPA Helmholtz Center i.G., Saarland Informatics Campus

Abstract—**The decreasing costs of molecular profiling has fueled the biomedical research community with a plethora of new types of biomedical data, enabling a breakthrough towards a more precise and personalized medicine. However, the release of these intrinsically highly sensitive data poses a new severe privacy threat. While biomedical data is largely associated with our health, there also exist various correlations between different types of biomedical data, along the temporal dimension, and also in-between family members. However, so far, the security community has focused on privacy risks stemming from genomic data, largely overlooking the manifold interdependencies between other biomedical data.**

In this paper, we present a generic framework for quantifying the privacy risks in biomedical data taking into account the various interdependencies between data (i) of different types, (ii) from different individuals, and (iii) at different time. To this end, we rely on a Bayesian network model that allows us to take all aforementioned dependencies into account and run exact probabilistic inference attacks very efficiently. Furthermore, we introduce a generic algorithm for building the Bayesian network, which encompasses expert knowledge for known dependencies, such as genetic inheritance laws, and learns previously unknown dependencies from the data. Then, we conduct a thorough inference risk evaluation with a very rich dataset containing genomic and epigenomic data of mothers and children over multiple years. Besides effective probabilistic inference, we further demonstrate that our Bayesian network model can also serve as a building block for other attacks. We show that, with our framework, an adversary can efficiently identify the parent-child relationships based on methylation data with a success rate of 95%.

1. Introduction

Over the last decade, the plummeting costs of molecular profiling have dramatically transformed biomedical science and enabled new breakthroughs towards a more personalized and precise medicine. This radical transformation has been made possible by a deluge of new types of biomedical data, such as genomic, epigenomic, and transcriptomic data, avaibable for research. For example, millions of genotypes (the most important part of the human genome) are now available to scientists, medical practitioners, and private companies (such as 23andMe [1]), and this number will certainly keep increasing in the future.

The main negative aspect of this new data-driven medicine is its impact on privacy. Indeed, all sorts of biomedical data are intrinsically highly privacy sensitive, since they often closely reflect our health status and the diseases we carry. For example, DNA methylation is one of the most important epigenetic elements influencing human health and anomalous DNA methylation patterns have been associated with various types of cancer [2], [3], [4]. The privacy concerns are further exacerbated by the fact that different kinds of biomedical data are increasingly available through multiple public databases or third-party providers. Moreover, biomedical data other than genetic data might not be considered as genetic information in the legal meaning and thus not be protected by legal frameworks, such as the US Genetic Information Nondiscrimination Act (GINA) [5], [6]. Finally, the various correlations between different types of biomedical data, between family members, and along the temporal dimension must be taken into account to provide guarantees that biomedical data privacy is preserved. Although some types of biomedical data are influenced by external factors, and thus vary over the course of time, recent research indicates that even these data still contain enough information to jeopardize the privacy of their owners [7].

The security community has so far focused on privacy risks stemming from genomic data only, largely overlooking the major challenges brought by the increasing availability of data at other layers of the human biological stack. Very recently, attempts have been made towards better understanding and thwarting the privacy risks concerning epigenomic and transcriptomic data [8], [9], [6], [10], [11]. However, none of these works have tackled the biomedical-data privacy problem by jointly studying the different kinds of data, and their temporal and familial dimensions. This paper aims at filling this gap by proposing a generic framework for dissecting and quantifying privacy risks in biomedical data on a large scale.

Contributions. Specifically, we present a Bayesian network model that encompasses genomic data and epigenomic data from related individuals at different points in time. This probabilistic graphical model enables us to consider all probabilistic dependencies between these biomedical data, including temporal and familial correlations, and perform inference attacks very efficiently.

Among all kinds of data considered in our framework, some data dependencies are known from expert knowledge, such as genetic inherance laws, while others need to be learned from data, such as the correlations between methylation and genomic data or those between different time points. Therefore, we develop a general algorithm which considers both external knowledge and data-learned dependencies to automatically learn the structure of the Bayesian network. Then, we apply maximum likelihood estimation together with external knowledge to obtain the parameters, i.e., the conditional probabilities of the network. Finally, we perform probabilistic inference attacks using variable elimination to eventually get the exact posterior probabilities of targeted variables given observed data.

Based on the posterior probabilities output by our Bayesian network model, we evaluate how privacy evolves with respect to various scenarios of data disclosure. We quantify privacy levels with well-established privacy metrics such as entropy and estimation error, generalizing the estimation error to data other than the genome. Given the limited genomic and epigenomic data available together, we evaluate the privacy risks stemming from familial interdependencies and temporal correlations in separate settings.

Predicting the DNA methylation of a child given his/her genome and his/her mother's data (genome and DNA methylation) yields an estimation error as small as 0.1 for almost 60% of the DNA methylation positions. When considering the prior probability on the child's methylation data, the same estimation error is only achieved for 10% of the DNA methylation positions, demonstrating that the percentage of positions that are highly at risk is multiplied by around six for an informed adversary. Moreover, we found that observing more evidence reduces the average adversary uncertainty.

When predicting the DNA methylation of an individual given another DNA methylation sample observed one year before, the Bayesian network allows us to achieve an estimation error of less than 0.2 for approximately 82% of all methylation positions while the inference relying on the prior probability achieves the same estimation error at only 40% methylation regions. Further examining this strong performance, we found that, even for a longer time span of four years, the estimation error remains stable. This could typically enable an attacker to perform a temporal linkability attack against methylation profiles in the same vein as the one proposed against microRNA expression data [7].

Although we focus on a specific set of biomedical data due to the scarcity of rich datasets, the fundamental framework underlying our Bayesian network is still general enough to be easily extended to incorporate other types of data such as transcriptomic data (e.g., microRNA or gene expression [8], [7]). In particular, the structure learning algorithm we propose is not specific to our application, and thus can be used in any setting in which the Bayesian network can be constructed by learning some dependencies from data and embedding others from expert knowledge.

Finally, we demonstrate that our Bayesian network model can also serve as a fundamental building block to other applications: We study a linking attack that infers the mother-child relation. More precisely, we match children's methylation profiles to their mothers' (and vice versa) by comparing the posterior probabilities output by our Bayesian network given mother's methylation data with the real methylation profiles of the children. We also present a strong heuristic limiting the number of DNA methylation positions to consider, which significantly outperforms the approach with all positions taken into account. Our results show that using our framework for this kind of attack results in successfully linking 95% of mother-child pairs This corresponds to only a single incorrectly matched pair in our dataset.

Organization. In Section 2, we present the relevant biomedical background used in the paper. In Section 3, we introduce the adversarial model. In Section 4, we describe our quantification framework based on Bayesian networks. We detail the dataset in Section 5 and use it in Section 6 to evaluate various attack scenarios with our Bayesian framework. In Section 7, we apply our framework to link children to mothers based on their methylation data. We present related work in Section 8 before concluding in Section 9.

2. Preliminaries

In this section, we introduce the relevant background on the biomedical data used throughout the paper, i.e., genomic and DNA methylation data.

2.1. Genomics

The DNA is a double-helix structure consisting of complementary polymer chains. The genetic information is encoded on each of these chains as a sequence of nucleotides (A, T, G, C). Since 99.9% of the human DNA of two different individuals is exactly the same, the interesting parts are the remaining 0.1% of the positions. These positions that may vary throughout a population are referred to as *single nucleotide polymorphisms* (SNP).

Generally, two possible nucleotides can be observed at a given SNP. One is called the *major allele*, and is the most frequently occurring nucleotide at this SNP in the population. The other nucleotide is called the *minor allele*, thus is the least frequently occurring nucleotide. We usually denote the major allele using an uppercase letter $B \in \{A, T, G, C\}$ and the minor allele using a lowercase letter $b \in \{A, T, G, C\}$, with $b \neq B$.

Furthermore, each SNP position contains two alleles, one inherited from the father and one inherited from the mother. Thus, a SNP (also called genotype) can take three different values:

BB: if an individual inherits the same major allele from both parents (homozygous-major genotype),

Bb: if an individual inherits different alleles from the parents (heterozygous genotype),

bb: if an individual inherits the same minor allele from both parents (homozygous-minor genotype).

For simplicity, BB is often encoded as 0, Bb as 1 and bb as 2. We will follow the same encoding in the rest of our paper. Finally, we rely on Mendel's First Law which states that, for each SNP, a child inherits one allele from his mother and one allele from his father. Each allele of a parent is passed on to the child with uniform probability of 0.5.

2.2. DNA Methylation

One of the most important epigenetic elements in the DNA is called methylation. Its consequences affect the structure and the activity of the DNA molecule [12], [13].

In humans, DNA methylation so far has only been observed as the addition of a methyl group to the cytosines by specific enzymes called methyltransferases. This type of cytosine methylation in CpG-dinucleotides leads to the formation of 5-methylcytosine. That means that methylation can only occur at positions in the DNA where a C nucleotide is followed by a G nucleotide (called CpG-dinucleotide). Essentially, each such position can only have two possible states regarding DNA methylation: methylated or not.

However, since DNA methylation in principle can vary between copies of the DNA, e.g., in different cells, DNA methylation at a given CpG-dinucleotide is usually measured as a real value between 0 and 1. This value represents the fraction of methylated dinucleotides at this position.

Anomalous changes in the DNA methylation patterns, which are frequently observed in cancer, can lead to the hyper-activation of genes such as oncogenes, or the silencing of tumor suppressor genes [2]. However, while changes in the DNA methylation can have a dramatic effect on cancer, such changes in normal tissues can also be caused by, e.g., environmental influences. Recent studies showed that environmental cues such as pollution, exposure to stress or cigarette smoke can lead to changes in the methylation [14], [15], [16], [17].

Besides these external factors, the genotype of an individual can also affect the methylation of some regions [18], [19], [20]. Carrying particular alleles at certain SNPs can cause specific DNA methylation patterns at other positions or regions. Such SNPs having an influence on the DNA methylation are also called methylation quantitative trait loci (meQTLs).

3. Threat Model

The adversary's very general objective is to infer some hidden biomedical data, given observed ones. To do so, the adversary first needs to construct some (graphical) model that he will use during his attack. Therefore, we assume that the adversary has access to a set of training samples, which consist of DNA methylation profiles and genotypes. The adversary's training set may be further annotated with kinship relations between mothers and their children, or it may contain samples from the same individuals, taken at different points in time.

After this knowledge construction step, the adversary carries out his inference attack by observing part of the data (e.g., a DNA methylation profile or a genome) of a target or close relatives of the target (i.e., parents and children), potentially at a different time point. We thoroughly analyze the adversary's ability to predict information about his targets and their close relatives, varying the amount of additional information the adversary observes. Inferring genomic, epigenomic or transcriptomic information about targets may also reveal some sensitive information about those individuals, as shown later in the chapter. For example, both the genome and the DNA methylation contain information about phenotypic traits and the health status of a person [21], [2], [3], [4]. Moreover, this kind of information and also the kinship between individuals can be further matched to side channels such as surname-genome associations databases [22] or online social networks [23].

The adversary can further use the inference attack outcome to carry out a more tangible attack, such as linking DNA methylation profiles of a mother or a child to the corresponding DNA methylation profiles of the child or the mother, respectively. Our framework can in general cope with (i) any background knowledge from domain experts, (ii) any knowledge the adversary can construct based on auxiliary datasets, and (iii) any data the adversary observes during his inference attack.

4. The Framework

In this section, we formalize our approach and present the methodology that allows us to quantify the privacy of interdependent biomedical data.

We rely on a Bayesian network model to build a general privacy framework that we instantiate with genomic and epigenomic data. Bayesian networks allow us to perform a wide range of inferences. Moreover, in contrast to many other machine learning models, Bayesian networks can naturally handle missing data, i.e., they are able to perform inferences given any observed subset of evidence. Both of these advantages largely increase the generality of our framework. Besides, Bayesian networks allow us to take various biological layers (from genomic to transcriptomic via epigenomic layers) and their interrelation into account, while also providing ways to incorporate external domain knowledge easily. Lastly, there exist efficient algorithms for parameter learning and inference.

Our framework encompasses the three major steps in Bayesian network inference: (i) learning the structure of the Bayesian network, (ii) learning the necessary parameters of the network, and (iii) performing probabilistic inference on the network given observed evidence. We eventually rely on a set of privacy metrics which can be directly coupled with the Bayesian network in order to quantify the privacy of a given individual.

4.1. Bayesian Networks

Given a set of random variables, a Bayesian network is a probabilistic graphical model encoding a complex distribution over the random variables in a directed acyclic graph (DAG) $G = (V, E)$. Formally, each node $X_1, \dots, X_l \in V$ in the graph corresponds to a random variable. An edge $X_i \rightarrow X_j \in E$ between nodes $X_i, X_j \in V$ corresponds to a direct interaction between these nodes. Conversely, missing edges represent conditional independencies between nodes.

We now recall the basic definitions relevant to Bayesian networks to define the exact set of independencies induced by the graphical representation. These definitions will be used in Section 4.3 to describe our structure learning algorithm.

A structure $X \rightarrow Z \leftarrow Y$ in a graph is called a *v-structure*. A *trail* between X_1 and X_n is a sequence of nodes connected by edges $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$, where $X \rightleftharpoons Y$ denotes an edge of arbitrary direction between X and Y. Based on these notations, we next introduce the concept of an *active trail.*

Definition 1 (Active Trail [24])**.** *Let G be a DAG structure and $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$ a trail in G. Let $\mathbf{Z}$ be a subset of observed variables. The trail $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$ is active given $\mathbf{Z}$ if:*

- *Whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants are in $\mathbf{Z}$;*
- *no other node along the trail is in $\mathbf{Z}$.*

Intuitively, information can flow through the network along active trails. This notion then allows us to formally define the set of independencies induced by a graph based on a concept called *d-separation.*

Definition 2 (d-separation and Independencies [24])**.** *Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three sets of nodes in G. We state that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated given $\mathbf{Z}$, denoted by $\mathsf{d{-}sep}_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$, if there is no active trail between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$.*

We use $\mathcal{I}(G)$ to denote the set of independencies that correspond to d-separation:

$$\mathcal{I}(G) = \{(X \perp Y \mid \mathbf{Z}) \mid \mathsf{d{-}sep}_G(X, Y \mid \mathbf{Z})\}.$$

We state that a Bayesian network G is an I-map (independency map) for a probability distribution P over the same set of random variables if $\mathcal{I}(G) \subseteq \mathcal{I}(P)$ with $\mathcal{I}(P)$ being the set of all independencies holding in P.

Let $\mathsf{Parents}(X_i) \subseteq V$ denote the parent nodes of X_i in a Bayesian network G, and $\mathsf{NonDescendants}(X_i)$ denote the nodes in the graph that are not descendants of X_i. Given that G is an I-map for P, the graph structure can be translated into a factorization for the joint probability distribution as:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid Parents(X_i)).$$

Hence, we only need to know the distributions of these factors in order to obtain the whole distribution. These factors are also called the *parameters of the model.*

4.2. Notation and Networks

We now introduce the notations needed to construct the Bayesian networks for our particular scenario.

Let $\mathcal{A}$ be a set of individuals containing mothers and their children, $\mathcal{S}$ be a set of SNP IDs (i.e., positions on the DNA sequence), $\mathcal{R}$ be a set of methylation regions, and $\mathcal{T}$ be a set of points in time. Let t_i denote the time point at year i. We define $g_a^i \in \{0, 1, 2\}$ to be the value of SNP $i \in \mathcal{S}$ for an individual $a \in \mathcal{A}$. Similarly, $m_{a,t}^r \in [0, 1]$ denotes the average methylation within region $r \in \mathcal{R}$ for an individual $a \in \mathcal{A}$ at time point $t \in \mathcal{T}$. Let $\mathcal{M}$ denote the set of mothers, each member of which has a corresponding child in $\mathcal{A}$. Also, let $\mathcal{C}$ be the set to represent children who have their corresponding mothers in $\mathcal{A}$. For simplicity reasons, we assume that $\mathcal{M} \cap \mathcal{C} = \emptyset$. Also, note that $\mathcal{M} \cup \mathcal{C} \subseteq \mathcal{A}$.

Let G^i and M_t^r be random variables modeling the genome at position i and the average methylation in region r at time point t. Whenever we want to specify the set of individuals a certain random variable should capture, we will add the group of individuals the variable should refer to as a subscript. For example, $M_{\mathcal{C},t_0}^r$ denotes a random variable of a child's methylation in a region r at a given time point t_0.

Naively encoding these settings in *one* Bayesian network would yield a graph with $2 \cdot (|\mathcal{T}| \cdot |\mathcal{R}| + |\mathcal{S}|)$ vertices. In this paper, however, we take a different approach and separate the random variables as much as possible, designing independent Bayesian networks. To this end, we assume we have a set $\mathcal{Q} \subseteq \mathcal{S} \times \mathcal{R}$, containing pairs (i, r) of SNP IDs and methylation regions, such that there are no dependencies between any two such pairs. A similar assumption about SNPs independence has also been made in the genomic privacy context [25], [26]. This is a key element in simplifying the network structure as it allows us to build $|\mathcal{Q}|$ independent Bayesian networks. In Section 5, we show that such an independency assumption can be made if the SNP-methylation pairs are sufficiently far apart from each other.

Although our framework consisting of $|\mathcal{Q}|$ networks is general enough to consider all kinds of inference tasks, we focus on two particularly interesting settings in this work: analyzing mother-child interdependencies, and the temporal inference of methylation values. For the interdependencies of related individuals, we thus only consider data from a single time point t_0, while, for the temporal inference, we do not consider separate nodes for mothers and children. This also allows us to model adversaries having access to either data of related individuals or samples of the same individuals taken at multiple points in time.

Since we now consider separate networks, we will further simplify our notation when referring to exactly one *pair* $(i, r) \in \mathcal{Q}$ and only a single time point $t_0 \in \mathcal{T}$. By G, M we will denote the genome at a specific position and the methylation in a specific region (at time point t_0 if not stated otherwise), respectively. If we want to restrict the set of possible individuals, we will add a subscript containing the set of individuals. For example, $G_{\mathcal{M}}$ describes the mothers' genotypes at position i. Moreover, we will use P or $P_{(i,r)}$ to denote the probability distribution over the random variables of interest, given this specific pair in $\mathcal{Q}$.

4.3. Structure Learning

The first step of our approach is to construct the actual network and, contrary to previous work where the structure is already given [27], [28], we have to learn most of the edges between the nodes in the Bayesian networks.

In literature, there exist general algorithms (as listed in [24]) which learn the structure of a Bayesian network based on data. These algorithms can generally be divided into two categories: scoring-based algorithms and constraint-based algorithms. Scoring-based algorithms usually aim at finding a DAG structure, such that the probability model corresponding to the Bayesian network best fits the probability distribution of the data. Constraint-based algorithms learn the network structure by testing for independencies based on data and subsequently constructing an I-map for the learned independencies.

However, we cannot directly apply those algorithms, since they build the structure solely based on data. In our case, we additionally have external knowledge about certain parts of our model. We can classify our external knowledge into three categories: (1) existing edges, (2) directions of edges, and (3) known independencies.

Algorithm. Since most of our external knowledge can be translated into a set of known independency statements, we rely on an approach similar to constraint-based algorithms for learning the structure. In particular, we first limit the set of possible Bayesian networks by our external knowledge. Then, we use independency tests to decide which of the unknown edges should be part of the network. In particular, we test for statistical independence by applying the χ^2-test at a significance level of α.

In this paper, we introduce the novel notion of a minimal I-map given external knowledge.

Definition 3 (External Knowledge)**.** *We denote our external knowledge by the letter κ, and state that a graph is consistent with κ if the external knowledge holds in the graph. We denote this by writing $G \vDash \kappa$.*

A minimal I-map given external knowledge captures the idea that G should closely reflect the independencies of P. Ideally, both sets of independencies should be the same.

Definition 4 (Minimal I-map given External Knowledge)**.** *We state that a DAG G is a minimal I-map for a set of independencies $\mathcal{I}$ if*

1) *G is an I-map for $\mathcal{I}$, i.e., $\mathcal{I}(G) \subseteq \mathcal{I}$;*
2) *G is consistent with the external knowledge κ, i.e., $G \vDash \kappa$;*
3) *and the removal of any edge from G results in either $G \nvDash \kappa$ or it renders G not an I-map for $\mathcal{I}$.*

Algorithm 1 Build a minimal I-map given external knowledge.

```
Input: External knowledge $\kappa$, a set of independencies $\mathcal{I}$
    over the variables $V$.
Output: DAG $G = (V, E)$, which is a minimal I-map for
    $\mathcal{I}$ given $\kappa$.
 1: Let $\mathbf{G} = \{G \mid G \vDash \kappa, G = (V, E)\}$ be the set of all
    directed acyclic graphs with nodes $V$, for which the
    external knowledge holds.
 2: Let $G^* =$ None.
 3: for $G \in \mathbf{G}$ do
 4:     Calculate $\mathcal{I}(G) = \{(X \perp Y \mid \mathbf{Z}) \mid \mathsf{d-sep}_G(X, Y \mid \mathbf{Z})\}$.
 5:     if $\mathcal{I}(G) \subseteq \mathcal{I}$ then
 6:         if $G^*$ is None or $|\mathcal{I}(G)| > |\mathcal{I}(G^*)|$ then
 7:             $G^* = G$
 8:         end if
 9:     end if
10: end for
11: return $G^*$
```

We propose an algorithm that achieves this definition in Algorithm 1: We first enumerate the set $\mathbf{G}$ of all graphs that contain the necessary nodes and are consistent with our external knowledge (line 1). Then, we attempt to find a graph in $\mathbf{G}$ that is a minimal I-map for a given set of independencies $\mathcal{I}$. To this end, we return the graph $G^* \in \mathbf{G}$, which is an I-map for $\mathcal{I}$ (line 5) and encodes the highest number of independencies (i.e., the least number of edges) of all I-maps in $\mathbf{G}$ (line 6). If none of the graphs in $\mathbf{G}$ is an I-map for $\mathcal{I}$, the algorithm returns None. If none of the graphs in $\mathbf{G}$ is an I-map for $\mathcal{I}$, the algorithm returns False.

Theorem 1 (Correctness of Algorithm 1)**.** *Algorithm 1 returns either* None *if there is no minimal I-map given κ or a DAG G^*, which is a minimal I-map for $\mathcal{I}$ given κ.*

Theorem 1 states the correctness of our algorithm, and we prove its validity in Appendix A.

Scalability. The pseudocode given in Algorithm 1 scales with the number of possible graphs for which the external knowledge holds. However, as the structure learning only has to be done once, we do not consider it a time-critical step. Moreover, the algorithm's efficiency can be further improved, leveraging Proposition 2 from Appendix A. The proposition states that removing an edge $e \in E$ from a graph $G = (V, E)$ – yielding $G' = (V, E \setminus \{e\})$ – only introduces new independencies, i.e., $\mathcal{I}(G) \subsetneq \mathcal{I}(G')$.

Viewing our algorithm as a search problem starting with the *full graph and subsequently removing edges*, we can apply classical search algorithms, such as A^* to our problem and only need to add new independencies. During the search, we do not need to further follow branches for which $\mathcal{I}(G) \nsubseteq \mathcal{I}$, as this criterion cannot be reached anymore by removing edges. States that do not fulfill the external knowledge will have to be excluded from finding the minimum across the branches, however.

Conversely, depending on the concrete scenario and the number of constraints, we can also view our algorithm

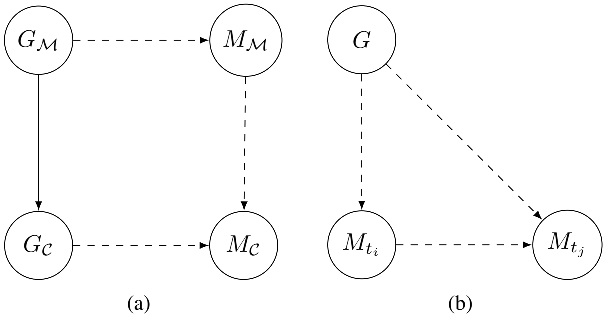


Figure 1. (a) Graphical model for mother-child dependencies. The full edge represents external expert knowledge that is given, and dashed edges represent dependencies that need to be learned: if they exist (structure learning) and, if so, what is the magnitude of the dependency (parameter learning); (b) graphical model for temporal inference of DNA methylation.

as a search problem starting with the *empty graph and subsequently add edges*. Similar pruning techniques as the ones mentioned above also apply in this case.

Mother-Child Networks. Next, we describe how the algorithm can be applied to the networks capturing the mother-child interdependencies. The set of random variables being considered are $V = \{G_{\mathcal{M}}, G_{\mathcal{C}}, M_{\mathcal{M}}, M_{\mathcal{C}}\}$, and we assume the following external knowledge κ:

- $G_{\mathcal{M}} \to G_{\mathcal{C}} \in E$, i.e., Mendelian inheritance laws state that the genotype of the mother influences the one of the child (i.e., it is an existing edge),
- $\forall X$: $M_X \to G_X \notin E$, i.e., there is never an edge from the methylation of a mother/child to her genome,
- $M_{\mathcal{C}} \to M_{\mathcal{M}} \notin E$, i.e., analogously to the genome, it is impossible for the mother to inherit methylation patterns from her child,
- $\forall X, Y$: $\{G_X \to M_Y, M_Y \to G_X\} \cap E \neq \emptyset \Rightarrow X = Y$, i.e., there is no direct connection between a genome and the methylation value of different individuals.

Incorporating this external knowledge leaves us with the potential DAG as shown in Figure 1(a). While the edge between the genomes is fixed, the dashed edges are subject to our analysis. In total, applying our external knowledge results in 24 possible independencies and eight possible graph structures.

Next, we iterate over all 24 possible conditional independencies and leverage the χ^2 test for each of these in order to obtain the set $\mathcal{I}(P)$ of independencies being justified by our data. We then run our algorithm with the given external knowledge and $\mathcal{I}(P)$ and obtain the graph structure that best represents $\mathcal{I}(P)$.

The algorithm hence provides us with the graph structure that best represents $\mathcal{I}(P)$.

Temporal Inference. Similarly, we can use our algorithm for finding the smallest I-map given the external knowledge for the temporal inference of DNA methylation. We consider two time points t_i and t_j and the set of random variables $V = \{G, M_{t_i}, M_{t_j}\}$.

Below, we list the external knowledge κ incorporated for the temporal inference of DNA methylation:

- $M_{t_i} \to G \notin E$, i.e., if there is an edge between the genome and the methylation, then it should start at the genome and end at the methylation,
- $M_{t_i} \to M_{t_j} \in E \Rightarrow i < j$, i.e., it is natural that if there are dependencies between methylation values at different points in time, the direction of the edge should be from the older methylation value to the newer one.

This external knowledge gives us a DAG with possible edges as depicted in Figure 1(b), resulting in eight possible networks. The considered random variables V limit the total number of possible independencies to six.

We test all of these independencies on the data, resulting in $\mathcal{I}(P)$. This set of independencies is then given to our algorithm together with the external knowledge κ, resulting in a graph structure that best represents $\mathcal{I}(P)$.

4.4. Parameter Learning

After learning the structures of all $|\mathcal{Q}|$ Bayesian networks, the next step is to learn the parameters for each network. In our paper, we combine two different methodologies to estimate the parameters: Some of the parameters are given by external knowledge, while we use a maximum likelihood estimation (MLE) on our data for the others.

Since there exist numerous population statistics on the probability of specific genomic variants, even for ethnical subgroups, we can leverage this knowledge to model the distribution $P(G^i)$ for any Bayesian network.[1] More precisely, population statistics give us the minor allele frequency MAF_i for each SNP. Let $p_k = P(G^i = k)$, then we can calculate the vector (p_0, p_1, p_2) from the minor allele frequency as $((1 - \mathsf{MAF}_i)^2, 2\mathsf{MAF}_i \cdot (1 - \mathsf{MAF}_i), \mathsf{MAF}_i^2)$.

Modeling the distributions for DNA methylation data, however, requires us to learn the methylation related distributions from our data. While DNA methylation is captured by a real value and thus follows a continuous distribution, the underlying distribution $P(M^r)$ can be considered to be generally multimodal. Therefore, following the general methodology in biomedical applications [29], we discretize the methylation values into a set of bins $B^r = \{B_1, \ldots, B_l\}$, such that $\bigcup_{i=1}^{l} B_i = [0, 1]$ and $\forall i, j \in \{1, \ldots, l\} : B_i \cap B_j = \emptyset \Leftrightarrow i \neq j$.

As stated before, we rely on MLE to learn the remaining distributions in our networks. Let $A \subseteq \mathcal{A}$ be a (sub)set of individuals and $\mathbf{Z}$ be a set or vector of conditions over random variables and $\mathbf{z}$ be an assignment of values to those random variables. Furthermore, we use $\mathbf{z}_a$ to denote the values of an individual $a \in A$ for the corresponding random variables in $\mathbf{Z}$. Then, we estimate any conditional methylation distribution as follows:

$$P(M \in B_j \,|\, \mathbf{Z} = \mathbf{z}) = \frac{|\{m_a \,|\, a \in A \wedge m_a \in B_j \wedge \mathbf{z}_a = \mathbf{z}\}|}{|\{m_a \,|\, a \in A \wedge \mathbf{z}_a = \mathbf{z}\}|}. \quad (1)$$

[1] Note that this is valid for the mother's G^i's in the mother-child network, and for all G^i's in the temporal network, as these do not have any parent in the graph (and thus $P(G^i)$ is not conditioned on any other variable).

		$G_{\mathcal{M}}$		
		0	1	2
$G_{\mathcal{C}}$	0	$p_0 + 0.5p_1$	$0.5p_1 + p_2$	0
	1	$0.5p_0 + 0.25p_1$	0.5	$0.25p_1 + 0.5p_2$
	2	0	$p_0 + 0.5p_1$	$0.5p_1 + p_2$

TABLE 1. THE PROBABILITY DISTRIBUTION $P(G_{\mathcal{C}} \mid G_{\mathcal{M}})$ BASED ON THE LAWS OF MENDELIAN INHERITANCE GIVEN POPULATION STATISTICS OF $p_g = P(G = g)$.

Intuitively, this corresponds to counting all samples in A for which the methylation value is in the bin B_j and for which all conditions specified by $\mathbf{Z} = \mathbf{z}$ hold. Then, this number is divided by the number of samples for which the conditions in $\mathbf{Z} = \mathbf{z}$ hold regardless of the bin the methylation value belongs to. Note that MLE might have to be smoothed in order to compensate for missing data. We will address these issues in Section 6.2.

Mother-Child Networks. Estimating the parameters of our mother-child networks additionally requires to model the distribution $P(G^i_{\mathcal{C}} \mid G^i_{\mathcal{M}})$. Once more, leveraging genetic knowledge, we can rewrite this probability as:

$$P(G^i_{\mathcal{C}} = g^i_{\mathcal{C}} \mid G^i_{\mathcal{M}} = g^i_{\mathcal{M}}) = \sum_{g^i_{\mathcal{P}} \in \{0,1,2\}} P(G^i_{\mathcal{P}} = g^i_{\mathcal{P}}) P(G^i_{\mathcal{M}} = g^i_{\mathcal{M}}) \cdot P(G^i_{\mathcal{C}} = g^i_{\mathcal{C}} \mid G^i_{\mathcal{M}} = g^i_{\mathcal{M}}, G^i_{\mathcal{P}} = g^i_{\mathcal{P}}), \quad (2)$$

where $g^i_{\mathcal{P}}$ denotes the genotype of the father at position i and $G^i_{\mathcal{P}}$ denotes the corresponding random variable. Generally, we will estimate the probability of a certain genotype independent of the gender or subgroup the individual is in and write $P(G^i)$ instead of $P(G^i_{\mathcal{M}})$, $P(G^i_{\mathcal{C}})$ and $P(G^i_{\mathcal{P}})$. While $P(G^i)$ – as stated before – is calculated from population statistics, $P(G^i_{\mathcal{C}} \mid G^i_{\mathcal{M}}, G^i_{\mathcal{P}})$ is exactly specified by the laws of Mendelian inheritance. Combining these finally results in the probability distribution as shown in Table 1.

Temporal Inference. Except for $P(G)$, the parameters of the temporal inference network are learned by applying MLE, similarly to the mother-child network.

4.5. Bayesian Inference

For inferring the probabilities of unobserved random variables conditioned on observed ones, typically the marginal distributions need to be computed. In our case, we rely on variable elimination, an exact inference algorithm for Bayesian networks [24]. While the algorithm, in general, has an exponential time complexity, the simple structure of our Bayesian networks allows the algorithm to be efficient enough in our case. There also exist polynomial-time algorithms for exact or approximate inference, such as junction tree [30] or (loopy) belief propagation algorithms [31], that can be applied for larger or more complex Bayesian networks.

Variable elimination generally works by collecting all factors required for the inference of any marginal distribution $P(X_i \mid \mathbf{E} = \mathbf{e})$, where X_i belongs to the query variables $\mathbf{X}$ and $\mathbf{E}$ is the observed evidence. Then, for a Bayesian network containing the nodes V, all variables in $V \setminus (\mathbf{X} \cup \mathbf{E})$ are eliminated one by one using marginalization (which corresponds to summing out variables $V \setminus (\mathbf{X} \cup \mathbf{E})$ in our discrete scenario), resulting in the marginal probability distributions of interest.

4.6. Privacy Metrics

For the purpose of quantifying the impact of the considered inference attacks, we rely on two privacy metrics: *expected estimation error* and *entropy* [27], [28].

Expected estimation error has already been introduced in the context of genomic data by Humbert et al. [27]. For our setting, we generalize this notion, so that it can also be applied to other types of data, such as DNA methylation values specifically. The estimation error quantifies the expected distance between the adversary's estimate of a value $\hat{x}$ and the true value x. The Bayesian inference step outputs the probability distribution $P(\hat{x} \mid \mathbf{Z} = \mathbf{z})$ given some observed genomic and/or epigenomic data $\mathbf{Z}$, where $\hat{x}$ can take values within a set $\mathcal{X}$ of finite size. Then, we define the expected estimation error as follows:

$$E_x(X \mid \mathbf{Z} = \mathbf{z}) = \sum_{\hat{x} \in \mathcal{X}} P(X = \hat{x} \mid \mathbf{Z} = \mathbf{z}) ||\hat{x} - x||, \quad (3)$$

where $|| \cdot ||$ represents any distance metric, such as the L_1-norm or the Euclidean distance. In our evaluation, we rely on the former. In the context of our study, this definition can be applied to those cases where we aim at quantifying the genomic privacy of an individual. When considering the privacy of methylation points in a region r, however, we have to specify the handling of the bins further. We define the mean value of a bin $B \in B^r$ as $\mu(B) = \frac{\sup(B) - \inf(B)}{2}$. Then, from the probability distribution given by the Bayesian network model, $P(\hat{B} \mid \mathbf{Z} = \mathbf{z})$, and the true methylation value m being part of a bin B, the estimation error is calculated as follows:

$$E_B(M^r | \mathbf{Z} = \mathbf{z}) = \sum_{\hat{B} \in B^r} P(M^r \in \hat{B} \,|\, \mathbf{Z} = \mathbf{z}) ||\mu(\hat{B}) - \mu(B)||. \quad (4)$$

he second metric (i.e., entropy) quantifies the *uncertainty* of the adversary [32], [33] and is defined as:

$$H_x(X | \mathbf{Z} = \mathbf{z}) = -\sum_{\hat{x} \in \mathcal{X}} P(X = \hat{x} | \mathbf{Z} = \mathbf{z}) \log P(X = \hat{x} | \mathbf{Z} = \mathbf{z}). \quad (5)$$

It holds that the higher the entropy is, the higher the adversary's uncertainty is, and the higher the privacy is.

5. Dataset

The dataset we use contains genotypes and DNA methylation values of 75 individuals, 42 of which have parental relations (21 mother-child pairs). For 67 out of 75 individuals, samples collected at the birth of the child, referred to as t_0, were available. Samples one year later (t_1) and four years later (t_4) were also available for 16 individuals.

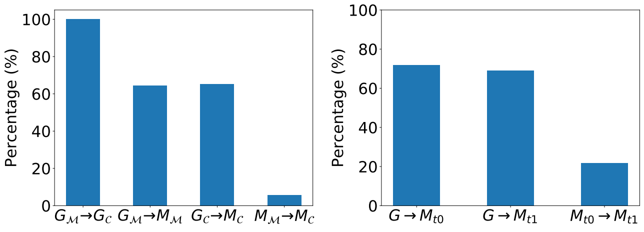
Both, the longitudinal dimension of the dataset and the fact that it contains individuals with parental relations make this dataset a unique and extremely precious data source in the biomedical community. At the time of this writing, the dataset can be considered to be one of the largest – if not the largest – dataset of its kind. Moreover, collecting multiple types of biomedical data from related individuals in such regular intervals involves a tremendous amount of money and time. Note that this dataset is not yet publicly available, but it will be released to other researchers soon.

The DNA methylation was determined using a process called whole genome bisulfite sequencing (WGBS), measuring the methylation levels for all 28 million CpG dinucleotides based on samples taken from the whole blood. In order to determine the methylation levels from the bisulfite-treated sequencing data, the reads (short sequences of the genome) were aligned, followed by a quality assessment and methylation calling. Then, the genotype was determined at known SNP positions as listed in the dbSNP database (version 141). To accomplish the task of determining the genotype from WGBS data, the Bis-SNP tool was used [34].

Next, we selected a set of 4,681,414 pairs of SNPs and methylation regions. This set was determined using a Spearman rank correlation test [35] and a false discovery rate threshold for all SNPs located within 50 kb (kilobases) up-/downstream of methylation regions. The false discovery rate threshold was set to 1% after Benjamini-Hochberg correction [36].

For further analysis and the construction of the Bayesian networks, we assume the selected pairs of SNP and methylation region to be pairwise independent of each other. Therefore, we randomly sample a subset $\mathcal{Q}$ of 31,586 pairs such that the distance between adjacent SNPs and adjacent methylation regions is at least 50 kb. It is well-known that the linkage disequilibrium (i.e., dependencies between SNPs) decays with the distance between the SNPs. While several thresholds have been proposed, the choice of 50 kb is a sufficient threshold to assume independence, given the origin of the population we use [37]. In order to further justify this threshold, we calculated the Spearman's rank correlation coefficient between the next 20 neighbouring methylation regions and SNPs of the resulting pairs (to either side). In both cases, the correlation was below 0.2 for more than 81% of our tests and below 0.4 for more than 97% of our tests.

We also inspected the Spearman's rank correlation coefficient between the methylation value and the genotype for each pair $(i, r) \in \mathcal{Q}$. For about 67% of the pairs, the correlation coefficient lies above 0.6, indicating a strong relationship between methylation and genotype for these pairs. Indeed, this percentage is also reflected in the number of edges between methylation and genotype we will learn in the following section. It is also worth noting that, conversely, our dataset is also diverse enough to contain also about 33% of pairs for which the relationship between the two types of biological data is relatively weak. This makes our dataset representative of the whole genome.



(a) (b)

Figure 2. Distribution of edges after structure learning: (a) in the mother-child setting, (b) for the temporal inference of methylation data.

6. Evaluation

In this section, we first apply our structure learning algorithm to construct the Bayesian networks. Then, we learn the networks' parameters, before quantifying the privacy risks by performing inference under various scenarios.

6.1. Structure Learning

Given the set $\mathcal{Q} \subseteq \mathcal{S} \times \mathcal{R}$ of SNP-methylation pairs as determined in Section 5, for each pair, we apply the algorithm presented in Section 4.3 for both settings we are interested in. Independence is tested using the χ^2-test at a significance level of $\alpha = 0.05$.

Fig. 2(a) shows the percentage of networks containing a specific edge for the mother-child networks. Following the external knowledge, the predefined edge between the mother's and the child's genotype appears in every network. Another interesting observation is that, in most cases, the methylation of the mother does not seem to directly affect the methylation of the child much. An indirect influence through the genomes is much more common. Furthermore, the percentage of edges between the genomes and methylation is roughly similar to the fraction of highly correlated SNP-methylation pairs our dataset contains (cf. Section 5).

Fig. 2(b) depicts the presence of edges for the Bayesian networks in the temporal setting. The main observation here is that the percentage of edges between genome and methylation is more or less consistent with the one in the mother-child networks. Moreover, the DNA methylation of the same individual at different points in time shows more direct dependencies than the methylation of related individuals in the mother-child networks.

6.2. Parameter Learning

We obtain the parameters of the Bayesian networks by relying on: (i) external knowledge (population statistics) and (ii) maximum likelihood estimation on a training set.

We build the population statistics using Kaviar [38], a compilation of 162 million positions of the human genome. Kaviar contains data from 77,781 individuals. Using Kaviar, we estimate the prior probability of an individual carrying a specific variant $P(G^i)$, and also calculate $P(G^i_{\mathcal{C}} \mid G^i_{\mathcal{M}})$ given the laws of Mendelian inheritance.

For the remaining random variables, we rely on our training data to learn their conditional probabilities. More specifically, given all samples in our dataset for which the required data is available, we split the samples into a training set and a testing set. We randomly allocate 70% of the samples to the training set, while the remaining 30% are allocated to the testing set used for inference in Section 6.3. For all of our experiments, we repeat this process 5 times and average over the results, effectively applying a repeated random sub-sampling validation. To discretize the methylation values, we choose five uniformly distributed bins $B^r = \{B_1, \dots, B_5\}$.

In both considered settings, we have specific requirements for the samples. For example, the mother-child networks require both the mother and the corresponding child to be present in the dataset, narrowing down the number samples that we can train and test on. When learning conditional distributions using MLE, we cannot be sure that we have enough samples to estimate the probability of every combination for the random variables due to very low frequencies for some of these combinations. Therefore, we apply Laplace smoothing [39], which mitigates the problem of assigning 0 probabilities to rare methylation values by artificially adjusting the probability. More precisely, Laplace smoothing gives us the following probability estimate:

$$\hat{P}(M \in B_j | \mathbf{Z}=\mathbf{z}) = \frac{|\{m_a \,|\, a \in A \wedge m_a \in B_j \wedge \mathbf{z}_a = \mathbf{z}\}| + \gamma}{|\{m_a \,|\, a \in A \wedge \mathbf{z}_a = \mathbf{z}\}| + \gamma |B^r|}. \quad (6)$$

Based on cross-validation, we found $\gamma = 0.01$ to generally yield the best results.

6.3. Probabilistic Inference

Given the trained Bayesian networks, we conducted a thorough evaluation: inferring unknown (hidden) variables while observing a subset of the remaining variables.

For each individual in the testing sets, we inferred the variables of interest given the considered observations for each of the approximately 32,000 SNP-methylation pairs. Then, we computed the proposed privacy metrics on the outcomes and averaged the results over all runs for each pair separately. The resulting values are then plotted as a cumulative distribution function (CDF), depicting the fraction of variables for which the privacy metrics are less or equal than a particular value. As a baseline, all of these figures also show the estimation error and entropy when predicting the variables based on the prior probabilities only. For the genome, this prior is computed from the population statistics while, for the methylation, it is learned from the training data.

Mother-child inference. In the mother-child networks, our primary focus is to infer an individual's methylation or genome given various observed evidence. Since plotting all inferences in one graph would prove to be counterproductive, we focus hereafter on the most interesting results.

We begin with an analysis of the estimation error. Figure 3(a) and Figure 3(b) depict the CDFs of the privacy metrics for the methylation inference of the mother and the child, respectively. Analogously, Figure 3(c) and Figure 3(d) depict the CDFs for the privacy metrics induced by inferring the genomes.

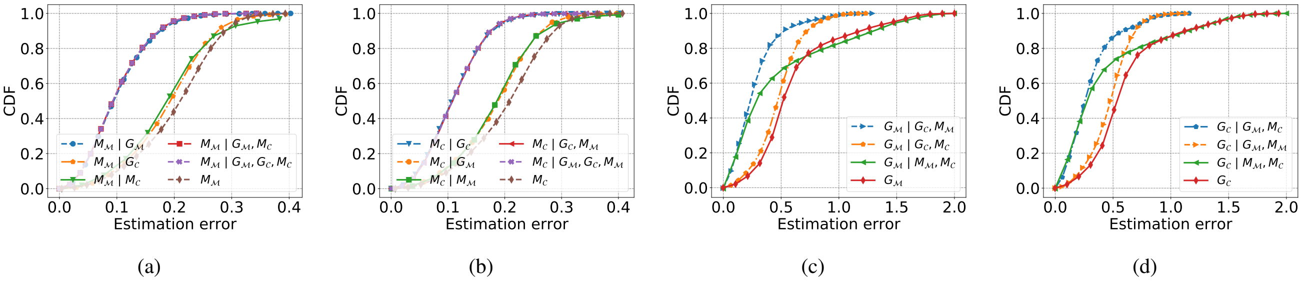
In general, all predictions achieve a strong performance with small estimation error. In almost all cases, the inferences observing at least some variables – and thus leveraging the structure of the Bayesian networks – outperform the baseline model, i.e., considering the prior probabilities. One of the best methylation predictions, i.e., $P(M_{\mathcal{M}} \mid G_{\mathcal{M}}, G_{\mathcal{C}}, M_{\mathcal{C}})$ or $P(M_{\mathcal{M}} \mid G_{\mathcal{M}})$ results in less than 0.1 estimation error for almost 60% of the variables, while the same estimation error for the prior ($P(M_{\mathcal{M}})$) is only achieved in 10% of the networks (Fig. 3(a)). Hence, the percentage of methylation regions that are highly at risk is multiplied by six when considering the observed evidence in this case. This demonstrates the severe privacy risk when combining multiple pieces of evidence across biological layers. Moreover, we notice that observing relatives' data is more helpful when inferring the genome than when inferring the methylation data. Finally, we note that children and mother inference results are very similar.

Analogously to the previous figure for the estimation error, Fig. 4 shows the entropy for the different inference tasks. Here, the advantage of leveraging the Bayesian network with observed variables over the simple baseline prior becomes more apparent. First, observing any other variable as evidence always makes the inference outperform the baseline regarding the entropy. For example, when inferring $G_{\mathcal{M}}$ given $G_{\mathcal{C}}$ and $M_{\mathcal{M}}$, almost 90% of the variables provide a prediction entropy of less than 0.4, while only 7% of the variables result in a similar entropy when using the prior probability for prediction.

By further analyzing the results, some more interesting observations can be made. For instance, to infer a child's methylation, the best predictor uses the child's genome as observed evidence. Interestingly, although one may naively believe that observing more variables should improve the result of the inference, this does not necessarily hold true. For instance, the estimation error for the prediction tasks $P(M_{\mathcal{C}} \mid G_{\mathcal{C}}, M_{\mathcal{M}})$ and $P(M_{\mathcal{C}} \mid G_{\mathcal{M}}, G_{\mathcal{C}}, M_{\mathcal{M}})$ are *equal* due to the d-separation properties. This makes sense as the child's methylation is not influenced by the mother's genome directly, and all related variables are known.

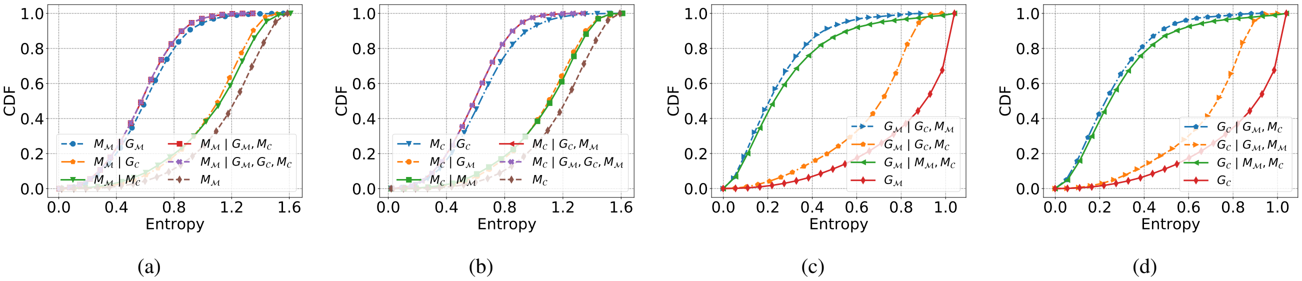
Moreover, the estimation error of these prediction tasks is very similar to the estimation error of the prediction task $P(M_{\mathcal{C}} \mid G_{\mathcal{C}})$, an observation which does not hold true for the entropy. In the same example as above, the entropies $H(M_{\mathcal{C}} \mid G_{\mathcal{C}}, M_{\mathcal{M}})$ and $H(M_{\mathcal{C}} \mid G_{\mathcal{M}}, G_{\mathcal{C}}, M_{\mathcal{M}})$ are smaller than $H(M_{\mathcal{C}} \mid G_{\mathcal{C}})$.

Similarly, – when inferring the methylation of a mother – we observe that there is no difference in the estimation error and entropy of inferring $M_{\mathcal{M}}$ given $G_{\mathcal{M}}, M_{\mathcal{C}}$ and the case where $G_{\mathcal{C}}$ is additionally given. In fact, the plotted lines of the first case are hidden beneath the lines of the latter. From this, we conclude that giving the genome of the child as additional knowledge when the methylation of the child and the genome of the mother are already known,



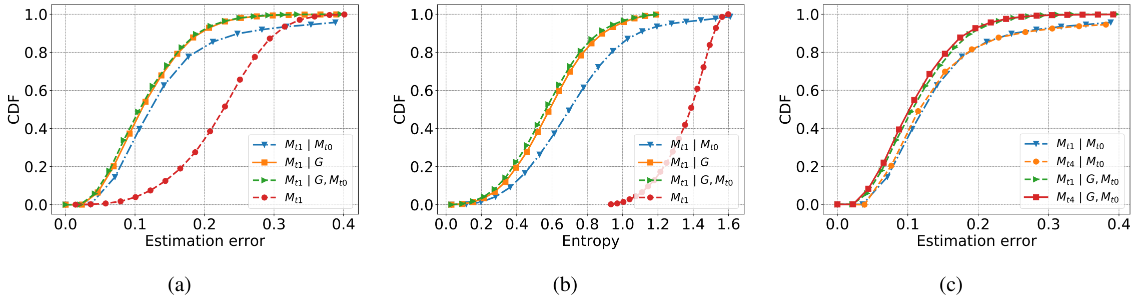
(a) (b) (c) (d)

Figure 3. Estimation error when inferring the methylation of (a) mother and (b) child; the genome of (c) mother and (d) child given various observed data and individuals.



(a) (b) (c) (d)

Figure 4. Entropy when inferring the methylation of (a) mother and (b) child; the genome of (c) mother and (d) child given various observed data and individuals.



(a) (b) (c)

Figure 5. (a) Estimation error and (b) entropy when inferring the methylation at M_{t_1}; (c) comparison of estimation error between inferring M_{t_1} and M_{t_4}.

does not significantly improve the estimation error or the entropy. This behaviour is again due to the structure of our Bayesian networks and its properties. In this case, the additional observation can only affect our inference through the edge between mother's and child's methylation nodes, because $G_{\mathcal{M}}$ is observed. However, as there are less than 6% of such edges in all pairs, it almost has no impact on the inference performance.

Some SNPs are associated with certain diseases, which makes them more privacy-sensitive than others. As an example, we further investigate our inference attack performance at SNP rs17221417 which is known to be linked with Crohn's disease. By applying our framework for inferring $P(G_{\mathcal{M}} \mid M_{\mathcal{M}})$, we obtain an estimation error of 0.025 at this SNP, while the error given the prior $P(G_{\mathcal{M}})$ is of 0.679. Note also that the average error over all the 32,000 SNPs is 0.215. These results demonstrate that our framework can be particularly effective on inferring the disease-related information from observed epigenomic data only.

Concerning the privacy implications, we observe that interdependencies between genomic and epigenomic data, and also between family members have to be taken very seriously, since they may pose a considerable privacy threat when multiple pieces of evidence are collected and combined by an adversary.

Temporal inference. When considering the temporal inference of DNA methylation, we first concentrate on predicting the methylation one year after the first sample was taken.

Fig. 5(a) shows that the target's genome is the best predictor on his future methylation: for 90% of the SNP-methylation pairs the resulting estimation error is less than 0.2, compared to only 40% when considering the prior probability. A similar observation applies to the entropy metric (Fig. 5(b)).

However, the genome is not the only strong evidence for the methylation. The target's methylation in the past can also serve as a strong indicator for the future DNA methylation profile, exhibiting an estimation error of less than 0.2 for approximately 82% of the SNP-methylation pairs. From a privacy point of view, this again clearly demonstrates the strong interdependencies of biomedical data, not only across different layers of the biological stack but also along the temporal dimension.

In order to examine whether the time span between the sample we want to predict and the sample we observe affects the prediction, we also construct Bayesian networks using each individual's methylation at time point 0 to predict her methylation at time point 4 (four years after the first sample was taken). Fig. 5(c) shows the estimation error of both predicting the DNA methylation at t_1 and at t_4. The result strongly suggests that the prediction remains stable even for longer time spans.

7. Case Study: Mother-Child Linking

So far, we have demonstrated that our Bayesian framework is capable of inferring the methylation and the genome of an individual, given some evidence. The role of the Bayesian network, however, is not limited to inference attacks only. The Bayesian network can also serve as a building block for more complex attacks. In order to demonstrate one possible application, we study the possibility of linking methylation profiles of a mother or a child to the methylation profiles of the corresponding child or mother, respectively. This application is especially sensitive as it can reveal paternity information (maternity in our data case) between two samples using only methylation profiles. However, we stress that this is only one possible application and that other use cases can be built upon our framework as well, which we leave for future work.

We assume that the adversary observes a single DNA methylation profile of his (observed) victim v_o and a database $\mathcal{D}$ of other methylation profiles. Then, the adversary's goal is to identify the observed victim's mother or child, denoted as the targeted victim v_t, among the other methylation profiles. By leveraging our Bayesian network, we can use the learned dependencies between genome and methylation to perform this linking, even though no genomic data is observed.

For the sake of simplicity, let us first describe the attack when the adversary aims at finding the child of v_o. As we have demonstrated in Section 6.3, the adversary is already able to predict the methylation profile of the observed victim's child with a small error. Conversely, for most SNP-methylation pairs $(\cdot, r) \in \mathcal{Q}$, the real child's methylation value m_{v_t} should ideally fall into the bin providing the largest probability among all methylation bins, i.e., $P(M^r_{\mathcal{C}} = m^r_{v_t} \mid M^r_{\mathcal{M}} = m^r_{v_o})$ is maximal. This, however, does not have to be true for all pairs, and it might be beneficial for an adversary to only use a subset $\mathcal{Q}' \subseteq \mathcal{Q}$ of all available pairs.

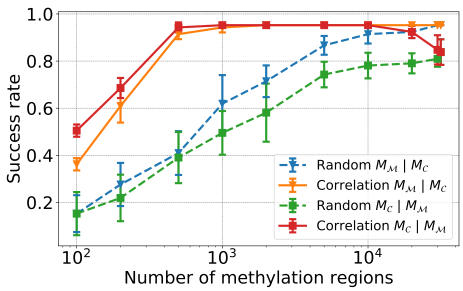


Figure 6. Success rate for discovering mother/child of each observed victim.

For each $a \in \mathcal{D}$ and each methylation region r, we estimate the probability of the child having the methylation value m^r_a as $w_{r,a} = P(M^r_{\mathcal{C}} = m^r_a \mid M^r_{\mathcal{M}} = m^r_{v_o})$. Given a specific a, this still leaves the adversary with a set of probabilities over all considered pairs $(\cdot, r)$ in $\mathcal{Q}'$. Since the adversary is interested in finding a choice a that maximizes $w_{r,a}$ for most regions r, we consider the average or equivalently the sum over all these probability scores instead:

$$\hat{v}_t = \operatorname*{argmax}_{a \in \mathcal{D}} \sum_{(\cdot,r) \in \mathcal{Q}'} w_{r,a} = \operatorname*{argmax}_{a \in \mathcal{D}} \sum_{(\cdot,r) \in \mathcal{Q}'} P(M^r_{\mathcal{C}} = m^r_a \mid M^r_{\mathcal{M}} = m^r_{v_o})$$

The case of finding the mother of a child works analogously.

As already stated before, the choice of $\mathcal{Q}'$ may have a significant impact on the performance of the adversary in this kind of attack. Therefore, we will also evaluate a heuristic to choose a subset of these pairs $\mathcal{Q}'$ from our original set $\mathcal{Q}$. We aim at choosing those pairs that maximize the adversary's success.

To this end, our heuristic should choose the pairs that provide the highest correlation among the methylation of mothers and their children. Since the analysis in Section 6.1 showed that a direct link between the methylation profiles of a mother and her child is rare, we instead focus on the information flowing through the Bayesian network via the genome-methylation link. Hence, we rely on the Spearman's rank correlation coefficient between G and M and only choose the top K SNP-methylation pairs with regard to their correlation coefficient, where K is subject to our analysis. We compare this heuristic with an approach that randomly chooses a subset of size K from $\mathcal{Q}$.

Experimental setup. To evaluate this linking attack on our dataset, we split the mother-child pairs into a training set (70%) and a testing set (30%). After learning the parameters of the Bayesian network on the training data, we pick a mother v_o (or child) and choose $\mathcal{D}$ to contain all remaining samples from the test set, plus all samples from time point 0 that do not have the corresponding child (or mother) available. This results in a database of $|\mathcal{D}| = 40$ samples, which further complicates the linking task. We perform the attack for all 21 mothers (and children) in our dataset.

We compute the success rate over all observed victims for the evaluation. The success rate is computed as the number of correct matches between mother and child, divided by 21 (the total number of observed victims). We emphasize that the metric we use is very strict compared to those used in other domains, such as recommendation systems, since the metrics used there usually allow the correct individual to be present within the top k matches.

Experimental results. Fig. 6 shows our experimental results for varying numbers K of SNP-methylation pairs we consider, ranging from 100 up to $31,586$.

Generally, we are able to achieve an excellent prediction: At the best K, we successfully match 20 out of 21 samples to the corresponding mother/child, given a database of 40 different choices. This makes a best success rate of 95.23%. When comparing the randomly chosen subsets from $\mathcal{Q}$ with our advanced heuristic, it becomes apparent that the randomly chosen subsets are significantly outperformed by our heuristic. Using the top 500 SNP-methylation pairs with the highest correlations enables us to reach the maximum success rate, while the success rate for a randomly chosen subset of size 500 is merely around 50%.

Another interesting observation is that using the whole set of pairs $\mathcal{Q}' = \mathcal{Q}$ may result in a worse performance, compared to the best possible subset. This at least is the case when identifying the child, given the mother's methylation.

8. Related Work

Since the plummeting costs for molecular profiling have caused a tremendous increase in availability of biomedical data, a new research field has emerged, studying the privacy threats induced by the vast amount of biomedical data. So far, most of the research has focused on quantifying and mitigating the threats concerning genomic data specifically, well summarized in recent surveys [40], [41], [42].

Attacks. We start by mentioning the line of work closest to our paper, i.e., approaches relying on graphical models to perform inference and quantify genomic privacy. Humbert et al. analyze the implications of familial relations on kin genomic privacy [27], [28]. Leveraging Bayesian networks and factor graphs, they model the familial dependencies and infer the genomes of the relatives of an individual whose genome or phenotype is observed by an adversary. Similar to our approach, Humbert et al. make assumptions on the independence of the SNPs for their Bayesian network model to be separated into smaller disjoint networks. In contrast to this, Backes et al. use Bayesian networks at scale to model the familial relations of several generations [43]. Based on a large network, they predict the genomic privacy for future generations, simulating various scenarios about how many people of each generation will share their genetic data publicly. We differentiate from the aforementioned works by the various types of biomedical data and the temporal dependencies between them that we consider. Conceptually, we also propose a method for learning the structure and the parameters of the Bayesian networks, which were already given by expert knowledge in previous works. We also instantiate our inference framework on a more concrete parent-child linking attack.

Several papers have studied privacy risks related to various sorts of biomedical data, other than the genome. For instance, Schadt et al. propose a Bayesian method for predicting and linking genotypes and RNA-expression profiles [8]. Backes et al. investigate a similar method for matching DNA methylation profiles to genotypes [11]. They further present a cryptographic mechanism to privately classify brain tumors based on methylation data. Dyke et al. have identified DNA methylation sites in the human body which are closely linked to the genome [6]. Based on their observations, they also propose high-level guidelines for disclosing DNA methylation data. Also considering additional side-channels and external knowledge, Philibert et al. also study the risks of inferring parts of the genome as well as alcohol consumption and smoking behavior from certain methylation data [9]. However, they do not attempt to quantify the success of such any attack in a principled manner. Recently, Backes et al. have studied the extent to which microRNA profiles can be linked over longer time spans [7], and further show that datasets based on microRNAs are prone to membership inference attacks by just relying on average statistics [10]. Both papers present differentially private mechanisms to counter their attacks.

Wang et al. describe a membership inference attack on statistics as published in genome wide association studies [44]. Moreover, they present a second attack, identifying individuals and their SNPs from the same set of statistics. Gymrek et al. demonstrate the possibility of re-identifying genomes by querying genealogy databases containing surnames [22]. They combine the inferred surnames with other types of knowledge, such as age and state, in order to successfully track back the identities of contributors in public datasets. Humbert et al. show that even online social networks can be leveraged as a side-channel by first inferring phenotypic traits (e.g., eye color or blood type) from the genome and then mapping this data to profiles in those networks. Considering side knowledge increases the success of an adversary and thus the privacy risks inherent to the particular types of medical data.

Defenses. On the mitigation side, most effort has been put into designing cryptographically provably secure protocols for many of the applications of genetic data. Some of the most recent work is especially well suited for the DTC area as, for example, a paper by Cristofaro et al. [45], which enables a privacy-preserving genetic relatedness test. In the same vein, Baldi et al. propose techniques for paternity tests, personalized medicine, and genetic compatibility tests based on private set operations [46]. Another recent topic in the field focuses on protocols for similar patient queries [47]. Finally, Karvelas et al. present a novel mechanism for the private processing of whole genomic sequences which is flexible and supports a wide range of queries [48].

Other countermeasures rely on differential privacy techniques. For instance, Johnson et al. have presented a set of privacy-preserving data mining algorithms, facilitating genome-wide association studies while guaranteeing differential privacy [49]. Fredrikson et al. study so-called model inversion attacks, in which an adversary, given a machine learning model and demographic information, predicts a patient's genetic information [50]. They demonstrate that, although differential privacy is able to prevent this kind of attacks, it would simultaneously expose patients to an increased risk of mortality.

9. Conclusion and Future Work

In this paper, we have proposed a generic framework for quantifying privacy risks of any interdependent biomedical data. This model aims to help better assess and anticipate privacy risks arising from the sharing of an ever-increasing variety and amount of biomedical data. Our framework relies on a Bayesian network that allows us to capture and quantify privacy implications, due to correlations between different types of biomedical data, along the temporal dimension, and between related individuals. We propose a general algorithm to learn the structure of the underlying Bayesian networks by combining data with external knowledge. Then, based on our Bayesian networks, we run an extensive set of experiments, considering the familial relationships and the temporal dimension separately. In both scenarios, we demonstrate that our Bayesian network model is able to achieve a strong prediction performance.

For instance, predicting the DNA methylation of a mother given her genome results in an estimation error less than 0.1 for 60% of the methylation regions. For the prior probabilities, this estimation error or smaller is only achieved for 10% of the methylation regions, demonstrating that the percentage of methylation regions that are highly at risk is multiplied by 6 when observing the genome. Moreover, when predicting the genome given the methylation profiles of the mother and the child, we achieve an estimation error of less than 0.4 for around 80% of the genomic positions, compared to smaller than 10% of the genomic positions when using the prior probabilities only. Lastly, analyzing the temporal interdependencies, we found that the prediction of methylation based on a past methylation profile is as successful with a one-year shift as with a four-year shift.

Besides predicting hidden parts of various biomedical profiles, our Bayesian network model can also serve as a fundamental building block for other attacks. To this end, we are the first to propose an attack matching DNA methylation profiles across family members. Building upon our Bayesian network's posterior probabilities, and proposing a heuristic that limits the number of DNA methylation positions to consider, our linking attack is able to achieve a 95% success rate. This further shows the generality of our framework.

In total, our evaluation strikingly proves all three kinds of interdependencies – cross-layer, familial, and temporal – to have a severe impact on the privacy of individuals. An adversary combining information about his victims is able not only to breach the privacy of the victims but also significantly increases his certainty about the outcome. Therefore, we suggest that careful considerations have to be made when releasing any biomedical data and that we are in a strong need for privacy-preserving technologies for securing biomedical data.

We leave it to future research to extend our framework to incorporate more layers, i.e., other data types, and interdependencies between layers. Another complementary extension of our framework is the analysis and handling of intra-genome, and intra-methylation dependencies.

Acknowledgment

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0656) and the German Research Foundation (DFG) via the collaborative research center "Methods and Tools for Understanding and Controlling Privacy" (SFB 1223), project A5. The authors would like to thank Dr. Carl Herrmann for explaining the subtleties of biomedical data.

References

[1] "23andme," https://www.23andme.com.

[2] M. Esteller and J. G. Herman, "Cancer as an Epigenetic Disease: DNA Methylation and Chromatin Alterations in Human Tumours," *The Journal of Pathology*, vol. 196, no. 1, pp. 1–7, 2002.

[3] P. M. Das and R. Singal, "DNA Methylation and Cancer," *Journal of Clinical Oncology*, vol. 22, no. 22, pp. 4632–4642, 2004.

[4] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer Genome Landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[5] M. A. Rothstein, Y. Cai, and G. E. Marchant, "The Ghost in Our Genes: Legal and Ethical Implications of Epigenetics," *Health Matrix (Cleveland, Ohio: 1991)*, vol. 19, p. 1, 2009.

[6] S. O. Dyke, W. A. Cheung, Y. Joly, O. Ammerpohl, P. Lutsik, M. A. Rothstein, M. Caron, S. Busche, G. Bourque, L. Rönnblom *et al.*, "Epigenome Data Release: A Participant-centered Approach to Privacy Protection," *Genome Biology*, vol. 16, pp. 1–12, 2015.

[7] M. Backes, P. Berrang, A. Hecksteden, M. Humbert, A. Keller, and T. Meyer, "Privacy in Epigenetics: Temporal Linkability of MicroRNA Expression Profiles," in *Proceedings of the 25th USENIX Security Symposium (Security)*. USENIX Association, 2016, pp. 1223–1240.

[8] E. E. Schadt, S. Woo, and K. Hao, "Bayesian Method to Predict Individual SNP Genotypes from Gene Expression Data," *Nature Genetics*, vol. 44, no. 5, pp. 603–608, 2012.

[9] R. A. Philibert, N. Terry, C. Erwin, W. J. Philibert, S. R. Beach, and G. H. Brody, "Methylation Array Data Can Simultaneously Identify Individuals and Convey Protected Health Information: An Unrecognized Ethical Concern," *Clinical Epigenetics*, vol. 6, p. 28, 2014.

[10] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, "Membership Privacy in MicroRNA-based Studies," in *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2016, pp. 319–330.

[11] M. Backes, P. Berrang, M. Bieg, R. Eils, C. Herrmann, M. Humbert, and I. Lehmann, "Identifying Personal DNA Methylation Profiles by Genotype Inference," in *Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2017, pp. 957–976.

[12] P. A. Jones, "Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond," *Nature Reviews Genetics*, vol. 13, no. 7, pp. 484–92, 2012.

[13] D. Schübeler, "Function and Information Content of DNA Methylation," *Nature*, vol. 517, no. 7534, pp. 321–326, 2015.

[14] T. Bauer, S. Trump, N. Ishaque, L. Thürmann, L. Gu, M. Bauer, M. Bieg, Z. Gu, D. Weichenhan, J.-P. Mallm *et al.*, "Environment-induced Epigenetic Reprogramming in Genomic Regulatory Elements in Smoking Mothers and Their Children," *Molecular Systems Biology*, vol. 12, no. 3, pp. 861–861, 2016.

[15] S. Trump, M. Bieg, Z. Gu, L. Thürmann, T. Bauer, M. Bauer, N. Ishaque, S. Röder, L. Gu, G. Herberth *et al.*, "Prenatal Maternal Stress and Wheeze in Children: Novel Insights into Epigenetic Regulation," *Scientific Reports*, vol. 6, p. 28616, 2016.

[16] J. Van Dongen, M. G. Nivard, G. Willemsen, J.-J. Hottenga, Q. Helmer, C. V. Dolan, E. A. Ehli, G. E. Davies, M. Van Iterson, C. E. Breeze *et al.*, "Genetic and Environmental Influences Interact with Age and Sex in Shaping the Human Methylome," *Nature Communications*, vol. 7, p. 11115, 2016.

[17] L. G. Tsaprouni, T.-P. Yang, J. Bell, K. J. Dick, S. Kanoni, J. Nisbet, A. Viñuela, E. Grundberg, C. P. Nelson, E. Meduri *et al.*, "Cigarette Smoking Reduces DNA Methylation Levels at Multiple Genomic Loci but the Effect is Partially Reversible upon Cessation," *Epigenetics*, vol. 9, no. 10, pp. 1382–1396, 2014.

[18] A. L. Teh, H. Pan, L. Chen, M.-L. Ong, S. Dogra, J. Wong, J. L. MacIsaac, S. M. Mah, L. M. McEwen, S.-M. Saw *et al.*, "The Effect of Genotype and in Utero Environment on Interindividual Variation in Neonate DNA Methylomes," *Genome Research*, vol. 24, no. 7, pp. 1064–1074, 2014.

[19] J. L. McClay, A. A. Shabalin, M. G. Dozmorov, D. E. Adkins, G. Kumar, S. Nerella, S. L. Clark, S. E. Bergen, C. M. Hultman, P. K. Magnusson *et al.*, "High Density Methylation QTL Analysis in Human Blood via Next-generation Sequencing of the Methylated Genomic DNA Fraction," *Genome Biology*, vol. 16, no. 1, p. 291, 2015.

[20] T. R. Gaunt, H. A. Shihab, G. Hemani, J. L. Min, G. Woodward, O. Lyttleton, J. Zheng, A. Duggirala, W. L. McArdle, K. Ho *et al.*, "Systematic Identification of Genetic Influences on Methylation Across the Human Life Course," *Genome Biology*, vol. 17, no. 1, p. 61, 2016.

[21] "Snpedia," https://www.snpedia.com/index.php/SNPedia.

[22] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying Personal Genomes by Surname Inference," *Science*, vol. 339, pp. 321–324, 2013.

[23] M. Humbert, K. Huguenin, J. Hugonot, E. Ayday, and J.-P. Hubaux, "De-anonymizing Genomic Databases Using Phenotypic Traits," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 2, pp. 99–114, 2015.

[24] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

[25] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic Privacy and Limits of Individual Detection in a Pool," *Nature Genetics*, vol. 41, no. 9, pp. 965–967, 2009.

[26] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "On non-cooperative genomic privacy," in *International Conference on Financial Cryptography and Data Security*. Springer, 2015, pp. 407–426.

[27] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Addressing the Concerns of the Lacks Family: Quantification of Kin Genomic Privacy," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2013, pp. 1141–1152.

[28] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti, "Quantifying Interdependent Risks in Genomic Privacy," *ACM Transactions on Privacy and Security*, vol. 20, no. 1, p. 3, 2017.

[29] W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt, "Predicting Genome-wide DNA Methylation using Methylation Marks, Genomic Position, and DNA Regulatory Elements," *Genome Biology*, vol. 16, no. 1, p. 14, 2015.

[30] F. V. Jensen and F. Jensen, "Optimal Junction Trees," in *Proceedings of the 10th International Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann Publishers Inc., 1994, pp. 360–366.

[31] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 2014.

[32] C. Diaz, S. Seys, J. Claessens, and B. Preneel, "Towards Measuring Anonymity," in *Proceedings of the 2nd International Workshop on Privacy Enhancing Technologies*. Springer, 2002, pp. 54–68.

[33] A. Serjantov and G. Danezis, "Towards an Information Theoretic Metric for Anonymity," in *Proceedings of the 2nd International Workshop on Privacy Enhancing Technologies*. Springer, 2002, pp. 41–53.

[34] Y. Liu, K. D. Siegmund, P. W. Laird, and B. P. Berman, "Bis-SNP: Combined DNA Methylation and SNP Calling for Bisulfite-seq Data," *Genome Biology*, vol. 13, no. 7, p. R61, 2012.

[35] C. Spearman, "The Proof and Measurement of Association between Two Things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[36] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[37] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward *et al.*, "Linkage Disequilibrium in the Human Genome," *Nature*, vol. 411, no. 6834, pp. 199–204, 2001.

[38] G. Glusman, J. Caballero, D. E. Mauldin, L. Hood, and J. C. Roach, "Kaviar: An Accessible System for Testing SNV Novelty," *Bioinformatics*, vol. 27, no. 22, pp. 3216–3217, 2011.

[39] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge university press Cambridge, 2008.

[40] Y. Erlich and A. Narayanan, "Routes for Breaching and Protecting Genetic Privacy," *Nature Reviews Genetics*, vol. 15, no. 6, pp. 409–421, 2014.

[41] E. Ayday, E. De Cristofaro, J.-P. Hubaux, and G. Tsudik, "Whole Genome Sequencing: Revolutionary Medicine or Privacy Nightmare?" *Computer*, pp. 58–66, 2015.

[42] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang, "Privacy in the Genomic Era," *ACM Computing Surveys*, vol. 48, p. 6, 2015.

[43] M. Backes, P. Berrang, M. Humbert, X. Shen, and V. Wolf, "Simulating the Large-scale Erosion of Genomic Privacy Over Time," in *Proceedings of the 3rd International Workshop on Genome Privacy and Security (GenoPri)*, 2016.

[44] 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," in *Proceedings of the 16th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2009, pp. 534–544.

[45] E. De Cristofaro, K. Liang, and Y. Zhang, "Privacy-Preserving Genetic Relatedness Test," in *Proceedings of the 3rd International Workshop on Genome Privacy and Security (GenoPri)*, 2016.

[46] P. Baldi, R. Baronio, E. De Cristofaro, P. Gasti, and G. Tsudik, "Countering GATTACA: Efficient and Secure Testing of Fully-sequenced Human Genomes," in *Proceedings of the 18th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2011, pp. 691–702.

[47] X. S. Wang, Y. Huang, Y. Zhao, H. Tang, X. Wang, and D. Bu, "Efficient Genome-wide, Privacy-preserving Similar Patient Query based on Private Edit Distance," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2015, pp. 492–503.

[48] N. Karvelas, A. Peter, S. Katzenbeisser, E. Tews, and K. Hamacher, “Privacy-preserving Whole Genome Sequence Processing through Proxy-aided ORAM,” in *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2014, pp. 1–10.

[49] A. Johnson and V. Shmatikov, “Privacy-preserving Data Exploration in Genome-wide Association Studies,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2013, pp. 1079–1087.

[50] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, “Privacy in Pharmacogenetics: An End-to-end Case Study of Personalized Warfarin Dosing,” in *Proceedings of the 23rd USENIX Security Symposium (Security)*. USENIX Association, 2014, pp. 17–32.

Appendix A. Proof of Correctness

In this section, we prove the correctness of Algorithm 1. We begin recalling the definition of a Markov blanket in a Bayesian network, as stated by Koller and Friedman [24].

We begin with recalling the definition of a Markov blanket in a Bayesian network.

Proposition 1 (Markov Blanket)**.** *Given a node X in a Bayesian network G, by $MB_G(X)$ we denote the smallest set of nodes $\mathbf{U}$ required to render X independent of all other nodes in the network. $MB_G(X)$ consists of X's parents, X's descendants, and other parents of X's descendants.*

Removal of Edges. Using this definition, we provide and prove the following useful proposition. It can also be leveraged to improve the algorithm performance as discussed in Section 4.3. The proposition states that removing an edge from a graph only introduces new independencies.

Proposition 2. *Given $G = (V, E)$, removing any edge $e \in E$ from G – yielding $G' = (V, E \setminus \{e\})$ – implies that $\mathcal{I}(G) \subsetneq \mathcal{I}(G')$.*

Proof of Proposition 2. Without loss of generality, let $e = X \to Y$. First, removing an edge can only destroy trails in the graph and not introduce new trails. Thus, it also does not introduce new active trails, and we can conclude that $\mathcal{I}(G) \subseteq \mathcal{I}(G')$. In the rest of the proof, we thus focus on $\mathcal{I}(G) \neq \mathcal{I}(G')$, and distinguish two cases: (1) $X \to Y$ is the only active trail between X and Y given $\emptyset$, and (2) there exist active trails between X and Y given $\emptyset$ other than $X \to Y$.

In the first case, the active trail $X \to Y$ shows us that $(X \perp Y) \notin \mathcal{I}(G)$ by the definition of d-separation. Removing this edge from G, however, will cause $(X \perp Y) \in \mathcal{I}(G')$ to become true because the only active trail between X and Y has been removed by removing e. Hence, $\mathcal{I}(G) \neq \mathcal{I}(G')$.

In the second case, we need to find a $\mathbf{Z}$, such that $X \to Y$ is the only active trail given $\mathbf{Z}$ in G. Then, we could deduce that $(X \perp Y \mid \mathbf{Z}) \in \mathcal{I}(G')$, but $(X \perp Y \mid \mathbf{Z}) \notin \mathcal{I}(G)$, which again proves our claim.

If Y is not the parent of a child of X, the Markov Blanket $MB_{G'}(X)$ satisfies our constraint, since it implies $(X \perp Y \mid MB_{G'}(X)) \in \mathcal{I}(G')$, and $Y \notin MB_{G'}(X)$ by definition of the Markov Blanket. Thus, this independence holds in G', but not in G where there exists a direct edge between X and Y.

If X and Y, however, have common descendants, this yields v-structures of the form $X \to X' \leftarrow Y$. Fortunately, there cannot be any active trail between X and Y passing through X' given any set of nodes other than X' or its descendants, as this would result in a cycle in the graph contradicting the DAG properties. Hence, it is safe to remove Y and the descendants X and Y have in common from $MB_{G'}(X)$. Consequently, for $\mathbf{Z} = MB_{G'}(X) \setminus (\{Y\} \cup \mathsf{CommonDescendants}_{G'}(X, Y))$, it holds that $(X \perp Y \mid \mathbf{Z})$ is in $\mathcal{I}(G')$, but not in $\mathcal{I}(G)$. □

Main Proof. Leveraging the proposition from above, we are now able to prove the correctness of our structure learning algorithm as depicted in Algorithm 1.

Proof of Theorem 1. We prove this theorem in three steps. First, we prove that the algorithm only returns None if there is no I-map for $\mathcal{I}$ over V given κ. Second, we prove that if the algorithm returns a DAG G^*, $\mathcal{I}(G^*) \subseteq \mathcal{I}$ and $G \vDash \kappa$. Then, we prove that the removal of any edge would result in either not fulfilling the external knowledge κ or rendering the graph not an I-map, i.e., for any $e \in E$ either $G' = (V, E \setminus \{e\}) \nvDash \kappa$ or it holds that $\mathcal{I}(G') \nsubseteq \mathcal{I}$.

Let us assume that the algorithm returns None, although there is an I-map for $\mathcal{I}$ given κ. That is, there is a G, such that $G \vDash \kappa$ and $\mathcal{I}(G) \subseteq \mathcal{I}$. However, if there is such a G, then $G \in \mathbf{G}$, and hence we will execute the loop in line 3 also with this G. Clearly, G passes the condition in line 5 and – since we assume the algorithm to return None – would also pass the condition in line 6. As this would set $G^* = G$ in line 7, and there is no chance to set G^* back to None, this clearly contradicts our assumption of returning None. Thus, our assumption must have been wrong, and the original claim is proven by contradiction.

Next, we assume that the algorithm does return a DAG G^* and prove that G^* is a valid I-map consistent with κ. Line 1 of the algorithm ensures the consistency: Every graph considered by the algorithm must be consistent with the external knowledge. Line 5 of the algorithm ensures that only such graphs are further considered for which $\mathcal{I}(G^*) \subseteq \mathcal{I}$. Thus, G^* is an I-map for $\mathcal{I}$, which is consistent with κ.

Finally, we must prove that the removal of any edge from $G^* = (V, E)$ either results in not being consistent with the external knowledge κ or rendering G not an I-map for $\mathcal{I}$.

We prove this by contradiction and assume there is an edge $e \in E$ for which none of the two cases above holds true. Since we assume $G' = (V, E \setminus \{e\}) \vDash \kappa$, we know that $G' \in \mathbf{G}$ in line 1. By Proposition 2, we know that for G', it holds that $|\mathcal{I}(G')| > |\mathcal{I}(G^*)|$. Moreover, we know by assumption that $\mathcal{I}(G') \subseteq \mathcal{I}$, because G' is an I-map for $\mathcal{I}$. But then, G' would be considered in the loop in line 3, pass the condition in line 5 and also the condition in line 6. This would mean that G^* is set to G' at some point and there is no way for the original G^* to pass the test in line 6 anymore. Since this makes it impossible to return the original G^*, it contradicts our assumption and proves the actual claim. □