Saarland University
Faculty of Mathematics and Computer Science
Department of Computer Science

Quantifying and Mitigating Privacy Risks in Biomedical Data

Dissertation
zur Erlangung des Grades
des Doktors der Ingenieurwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von
Pascal Berrang

Saarbrücken,
November 2017

Tag des Kolloquiums:	25. Juli 2018
Dekan:	Prof. Sebastian Hack

Prüfungsausschuss:

Vorsitzender:	Prof. Christoph Sorge
Berichterstattende:	Prof. Michael Backes
	Prof. Christian Rossow
	Prof. Jean-Pierre Hubaux
Akademischer Mitarbeiter:	Dr. Yang Zhang

Zusammenfassung

Die stetig sinkenden Kosten für molekulares Profiling haben der Biomedizin zahlreiche neue Arten biomedizinischer Daten geliefert und den Durchbruch für eine präzisere und personalisierte Medizin ermöglicht. Die Veröffentlichung dieser inhärent hochsensiblen und miteinander verbundenen Daten stellt jedoch eine neue Bedrohung für unsere Privatsphäre dar. Während die IT-Sicherheitsforschung sich bisher hauptsächlich auf die Auswirkung *genetischer* Daten auf die Privatsphäre konzentriert hat, wurden die vielfältigen Risiken durch andere Arten biomedizinischer Daten – *epigenetischer* Daten im Speziellen – größtenteils außer Acht gelassen.

Diese Dissertation stellt Verfahren zur Messung und Abwehr solcher Privatsphärerisiken vor. Neben dem Genom konzentrieren wir uns auf zwei der wichtigsten gesundheitsrelevanten epigenetischen Elemente: microRNAs und DNA-Methylierung. Wir quantifizieren die Privatsphäre für die folgenden realistischen Angriffe: (1) Verknüpfung von Profilen über die Zeit, Verknüpfung verschiedener Datentypen und verwandter Personen, (2) Feststellung der Studienteilnahme und (3) Inferenz von Attributen. Unsere Resultate bekräftigen, dass die Privatsphärerisiken solcher Daten ernst genommen werden müssen. Zudem präsentieren und evaluieren wir Lösungen zum Schutz der Privatsphäre. Sie reichen von der Anwendung von Differential Privacy unter Berücksichtigung des Nutzwertes bis zu kryptographischen Protokollen zur sicheren Auswertung eines Random Forests.

Abstract

The decreasing costs of molecular profiling have fueled the biomedical research community with a plethora of new types of biomedical data, allowing for a breakthrough towards a more precise and personalized medicine. However, the release of these intrinsically highly sensitive, interdependent data poses a new severe privacy threat. So far, the security community has mostly focused on privacy risks arising from *genomic* data. However, the manifold privacy risks stemming from other types of biomedical data – and *epigenetic* data in particular – have been largely overlooked.

In this thesis, we provide means to quantify and protect the privacy of individuals' biomedical data. Besides the genome, we specifically focus on two of the most important epigenetic elements influencing human health: microRNAs and DNA methylation. We quantify the privacy for multiple realistic attack scenarios, namely, (1) linkability attacks along the temporal dimension, between different types of data, and between related individuals, (2) membership attacks, and (3) inference attacks. Our results underline that the privacy risks inherent to biomedical data have to be taken seriously. Moreover, we present and evaluate solutions to preserve the privacy of individuals. Our mitigation techniques stretch from the differentially private release of epigenetic data, considering its utility, up to cryptographic constructions to securely, and privately evaluate a random forest on a patient's data.

Background of this Dissertation

This dissertation is based on the papers mentioned in the following. I contributed to all papers as one of the main authors.

The initial idea for our first work on linking microRNA expression profiles over time [P1] originated during a joint discussion of Pascal Berrang, Mathias Humbert, and Andreas Keller. Pascal Berrang and Mathias Humbert subsequently laid the theoretical foundations for the attack and the defenses against it. The implementation and evaluation were done by Pascal Berrang. Andreas Keller provided valuable feedback and insights from the biomedical point of view. Anne Hecksteden, Andreas Keller, and Tim Meyer were involved in collecting and providing the biomedical data. In general, all authors reviewed the paper.

The general idea for our second work [P2] was contributed by Mathias Humbert. The design of the attack was carried out by Pascal Berrang and Mathias Humbert. Pascal Berrang was responsible for the implementation and evaluation of the membership attack and the mitigations thereof. The theoretical comparison of the two differentially private mechanisms was mainly carried out by Praveen Manoharan with the support of Pascal Berrang. All design decisions were discussed by Pascal Berrang, Mathias Humbert, and Praveen Manoharan. All authors reviewed the paper.

The idea for the first part of our work on identifying DNA methylation profiles by genotype inference [P3] vaguely came up during a discussion of Pascal Berrang and Mathias Humbert and was concretized during a meeting with Mathias Bieg, Roland Eils, and Carl Herrmann from the German Cancer Research Center (abb. DKFZ) in Heidelberg. Pascal Berrang and Mathias Humbert both contributed to the design of the attack. Pascal Berrang was also responsible for the implementation and evaluation of the attack. Mathias Bieg and Carl Herrmann further contributed to this work by providing the necessary insights into the data collected by the DKFZ and Irina Lehmann. Moreover, Carl Herrmann contributed by writing a general explanation of DNA methylation data and providing feedback. The idea for the second part of the work, introducing a cryptographic protocol for evaluating random forests, originated from a discussion of Pascal Berrang and Mathias Humbert. The design, implementation, evaluation, and proofs were carried out by Pascal Berrang. All authors reviewed the paper.

Both Pascal Berrang and Mathias Humbert contributed to the idea of our general framework for quantifying privacy [P4] as a generalization and extension of our previous works. Irina Lehmann and Roland Eils contributed by providing us the dataset for our evaluations. Yang Zhang carried out large parts of the implementation and evaluation with the support of Pascal Berrang. The framework and experiments were designed in close collaboration of Pascal Berrang, Mathias Humbert, and Yang Zhang. Pascal Berrang further contributed by proposing the structure learning algorithm and proving its correctness. All authors reviewed the paper.

[P1] Backes, M., Berrang, P., Hecksteden, A., Humbert, M., Keller, A., and Meyer, T. Privacy in epigenetics: temporal linkability of microRNA expression profiles. In: *Proceedings of the 25th USENIX Security Symposium (Security)*. USENIX Association, 2016, 1223–1240.

[P2] Backes, M., Berrang, P., Humbert, M., and Manoharan, P. Membership privacy in microRNA-based studies. In: *Proceedings of the 23rd ACM Conference on Computer and Communication Security (CCS)*. ACM, 2016, 319–330.

[P3] Backes, M., Berrang, P., Bieg, M., Eils, R., Herrmann, C., Humbert, M., and Lehmann, I. Identifying personal DNA methylation profiles by genotype inference. In: *Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2017, 957–976.

[P4] Berrang, P., Humbert, M., Zhang, Y., Lehmann, I., Eils, R., and Backes, M. Dissecting privacy risks in biomedical data. In: *Proceedings of the 2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2018.

Further Contributions of the Author

[S1] Backes, M., Berrang, P., Goga, O., Gummadi, K., and Manoharan, P. Profile linkability despite anonymity in social media systems. In: *Proceedings of the 15th ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2016.

[S2] Backes, M., Berrang, P., Hecksteden, A., Humbert, M., Keller, A., and Meyer, T. On epigenomic privacy: tracking personal microRNA expression profiles over time. In: *Workshop on Understanding and Enhancing Online Privacy (UEOP), affiliated with NDSS*. 2016.

[S3] Backes, M., Berrang, P., Humbert, M., Shen, X., and Wolf, V. Simulating the large-scale erosion of genomic privacy over time. In: *3rd International Workshop on Genome Privacy and Security (GenoPri), Selected for publication in IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2016.

[S4] Backes, M., Berrang, P., and Manoharan, P. From zoos to safaris – from closed-world enforcement to open-world assessment of privacy. In: *Foundations of Security Analysis and Design VIII*. Springer-Verlag, 2016, 87–138.

Acknowledgments

First and foremost, I would like to thank my advisor Michael Backes for giving me the opportunity to pursue my Ph.D. studies in the Information Security & Cryptography Group/CISPA at Saarland University. I am very grateful for his precious advice during the time of my Ph.D. studies, and his confidence in me. It has been a great pleasure working with him and learning from him.

I would like to express my gratitude towards my thesis committee members Michael Backes, Jean-Pierre Hubaux, and Christian Rossow for their time and effort spent reviewing this dissertation.

Naturally, I am very thankful to my collaborators and co-authors for our inspiring discussions and their contributions to this thesis. While my closest collaborator Mathias Humbert, who accompanied me throughout most of my Ph.D., certainly deserves my particular thanks, I do not wish to value my other co-authors any less: Praveen Manoharan and Yang Zhang had their fair share in making this thesis possible. Similarly, I would like to thank my external collaborators for their helpful input and ideas: Mathias Bieg, Roland Eils, Oana Goga, Krishna Gummadi, Carl Herrmann, Andreas Keller, Xiaoyu Shen, and Verena Wolf. Many thanks also to the CISPA administration, our secretaries, and system administrators for their great support during that time. I am especially grateful to Bettina Balthasar, Curd Becker, Sebastian Gerling, Joachim Lutz, and Sabine Nermerich, whom I could always count on regarding administrative and technical matters. Additionally, I would like to thank all of my colleagues and friends in the IS&C group and at CISPA for providing such a comfortable working environment.

I am also thankful to all my friends outside CISPA, who supported me throughout my studies and with whom I spent a memorable time: Amelie, Carolyn, Christian K., Christian R., Curd, Fabian, Janine, Jeanette, Jenni, Johannes, Jonas A., Jonas L., Jonas S., Marie, Max, Nadja, Niklas, Oliver, and Simon. This list is certainly not exhaustive.

I would like to acknowledge Carolyn Guthoff, Mathias Humbert, Andrea Ruffing, and Yang Zhang for reviewing parts of this thesis and providing valuable feedback and advice.

Last but not least, I am very grateful to my (step)family and my girlfriend for their support, their faith in me, and their invaluable advice in many situations.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

Since the first whole-genome sequencing in 2001, the cost of molecular profiling has been plummeting, paving the way for significant progress in biomedical science and the rise of precision medicine [101]. This scientific breakthrough is triggered by the increasing availability of biomedical data, whose main negative counterpart is the new threat towards health privacy. The genome is especially privacy sensitive as it uniquely identifies someone, is very stable over our entire lifetime, and is correlated among relatives [61]. This may explain why the security community has been, so far, focusing primarily on enhancing the privacy of genomic data. The extent of this threat and mechanisms to mitigate it have been extensively studied regarding genomic data. The various attack vectors and protection techniques have already been well surveyed and categorized back in 2014 [35].

However, our genome is not the only element of the human body that is heavily relied upon by the biomedical community. Environmental factors (e.g., pollution, diet, lifestyle, etc.) often play a crucial role in the development of most common diseases. Epigenetics (or epigenomics), transcriptomics, and proteomics aim to bridge the gap between our genome and our health status. Multi-omics research is a complementary step to genome sequencing: the DNA sequence tells us what the cell could possibly do, while the epigenome and transcriptome tell what it is actually doing at a given point in time. Using a computer analogy, if the genome is the hardware, then the epigenome is the software [20].

Despite the growing importance of epigenetics in the biomedical community, privacy concerns stemming from epigenetic data have received little to no attention so far. One reason that might explain this fact is that – contrary to the genome – epigenetic data may vary significantly over time, mainly due to the environmental influences. However, with the increasing understanding of epigenetics, it becomes clear that epigenetic data contains a vast amount of additional sensitive information and can thus raise potential privacy risks. For example, a large number of severe diseases (such as cancers, diabetes, or Alzheimer's [127, 70, 102, 37]) are already identified to be affected by epigenetic changes, and a recent study found that epigenetic alterations could even affect sexual orientation [93]. 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) [105, 33].

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. More specifically, many epigenetic datasets are released (without identifiers) on open online platforms with nonrestricted access. In the light of a multibillion-dollar business selling private medical data, and brokers linking patients' pseudonymised data across multiple sources [129, 89, 117], it becomes unquestionable that we are in need of both quantitative assessments of the inherent privacy risks and proper mitigations to counter these risks. Specifically, it is essential to provide the means for a quantitative assessment of the privacy risks induced by sharing or leaking medical data. Potential privacy risks include linkage, identification and inference attacks against the patients' data, carried out by a multitude of possible adversaries. Furthermore, besides identifying and quantifying these risks, mitigation measures have to be designed.

On the one hand, it is crucial to adjust those measures with the close collaboration of biomedical experts to fit their application scenarios and needs. On the other hand, the mitigation measures have to provide a sufficient amount of privacy, giving the patient control over her data. Hence, a major challenge while designing those measures is to strike a balance between privacy risks, the utility of the resulting data, and ease of use. Not respecting any combination of these will result in the mitigation measure not being adopted for real use, or in the worst case even causing harm to a patient by providing inaccurate statements in return for a higher privacy.

In this thesis, we will outline both qualitative assessments and mitigation measures specifically designed to fit real-life use-cases. Besides the genome, we focus on two of the most important epigenetic elements influencing human health: microRNAs and DNA methylation.

MicroRNAs (abbreviated miRNAs) were discovered in the early 1990s. MiRNAs are small RNA molecules that regulate the majority of human genes. Studies of miRNA expression profiles have shown that dysregulation of miRNA is linked to neurodegenerative diseases, heart diseases, diabetes, and the majority of cancers [84, 127, 70, 102, 37].[1] Therefore, miRNA expression profiling is a very promising technique that could enable more accurate, earlier and minimally invasive diagnosis of major severe diseases. As a consequence, it will certainly be increasingly used in medical practice.

DNA methylation is one of the best understood epigenetic elements. It is an essential regulator of gene transcription. As a consequence, aberrant DNA methylation patterns (such as hypermethylation and hypomethylation) have been associated with a large number of cancer types [36, 23, 124].

The solutions we present in the next chapters can be classified into the following three areas: (1) assessing the privacy risks for biomedical data, (2) provide mitigation measures by perturbing the data, and (3) provide mitigation measures by relying on cryptographic constructions. While the second research direction is more relevant for the release of public medical datasets and statistics, the third research direction allows for secure storage and analysis of medical data without losing utility.

Our work on privacy risks in biomedical data stretches across the following peer-reviewed publications [P1, P2, P3, P4], which each contributed to the progress in understanding and mitigating privacy risks in epigenetic data specifically, as presented in this dissertation:

Linkability of miRNA Expression Profiles. Our work [P1] is one of the first to study the privacy risks of epigenetic data and the first to study linkability attacks on microRNA expression profiles in particular. We analyze the *temporal linkability* of personal miRNA expression profiles by presenting and thoroughly evaluating two different types of attacks. Namely, we present an identification attack, which pinpoints a specific miRNA expression profile in a database of multiple expression profiles by knowing the targeted profile at another point in time, and we present a matching attack, which tracks a set of miRNA expression profiles over time. In our experiments, we show that two blood-based miRNA expression profiles taken with a time difference of one week from the same person can

[1] Known relations between miRNA and human pathologies can be found at `http://www.cuilab.cn/hmdd`.

be matched with a success rate of 90%. We furthermore observe that this success rate stays almost constant when the time difference is increased from one week to one year. In order to mitigate the linkability threat, we propose and evaluate two mitigation measures, which aim at perturbing the dataset: (1) hiding a subset of the miRNA expressions, e.g., those that are irrelevant for a given medical use-case and (2) disclosing noisy miRNA expression profiles in a differentially private and distributed manner. Our experiments show that the second technique provides a better trade-off between privacy and disease-prediction accuracy. By applying our differentially private mechanism, it is possible to decrease linkability by at least 50% for almost no loss of accuracy ($< 1\%$).

Membership Privacy for miRNA Expression Profiles. In our next work [P2], we investigate the threat of *membership attacks* on the privacy of individuals contributing their epigenetic profiles to scientific studies. Our results on public microRNA expression data demonstrate that disease-specific datasets are especially prone to membership detection, offering a true-positive rate of up to 77% at a false-negative rate of less than 1%. We present two attacks: (1) one relying on the L_1 distance, and (2) the other based on the likelihood-ratio test. We show that the likelihood-ratio test provides the highest adversarial success and we derive a theoretical limit on this success. In order to mitigate the membership inference, we propose and evaluate two perturbation mechanisms: a differentially private mechanism and a hiding mechanism. We also consider two types of prior knowledge for the differentially private mechanism and show that, for relatively large datasets, this mechanism can protect the privacy of participants in miRNA-based studies against strong adversaries without degrading the data utility too much. Based on our findings and given the current number of miRNAs, we recommend releasing summary statistics only for datasets containing at least a couple of hundred individuals.

Genotype Inference from DNA Methylation Profiles. While our previous work focuses on a single type of biomedical data, the next work [P3] investigates the dependencies between two different types of data. Specifically, we show that releasing one's DNA methylation data causes privacy issues akin to releasing one's actual genome. We show that already a small subset of methylation regions influenced by genomic variants is sufficient to infer parts of someone's genome and to map this DNA methylation profile to the corresponding genome. Notably, we show that such re-identification is possible with 97.5% accuracy, relying on a dataset of more than 2500 genomes and that we can reject all wrongly matched genomes using an appropriate statistical test. We provide means for countering this threat by proposing a novel cryptographic scheme for privately classifying tumors that enables a privacy-respecting medical diagnosis in a typical clinical setting. The scheme relies on a combination of random forests and homomorphic encryption and is proven secure in the honest-but-curious model. We evaluate this scheme on real DNA methylation data and show that we can keep the computational overhead at acceptable values for our application scenario.

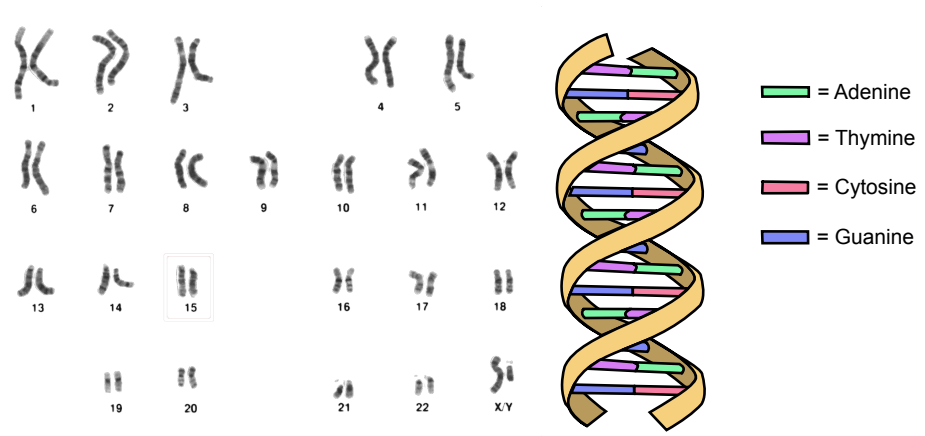
Privacy Risks of Interdependent Biomedical Data. In the work constituting the last part of this thesis [P4], we make a first step towards a more holistic view by proposing a generic framework for quantifying the privacy risks in biomedical data on a large

scale. Specifically, we propose a Bayesian network model that encompasses genomic and epigenomic data (DNA methylation in particular) from mothers and their children, at different points in time. We also introduce a generic algorithm for learning the structure of a Bayesian network by combining data with external expert knowledge. Then, we carry out a thorough evaluation, quantifying privacy risks in our interdependent model with well-established privacy metrics such as estimation error and entropy. The strong performance of our various *inference* tasks confirms that the privacy risks induced by interdependent biomedical data have to be taken very seriously. Besides the effective inference, we further demonstrate that our Bayesian network model can also serve as a fundamental building block to quantify the privacy in light of other attacks. To this end, we study the adversary's success in *linking* DNA methylation profiles of mothers to their children's, corroborating our privacy concerns with a success rate of 95%.

1.1 Structure

The remainder of this dissertation is structured as follows. In Chapter 2, we provide the necessary biomedical background. We present the related work in Chapter 3. In Chapter 4 we then analyze the temporal linkability of miRNA expression profiles and present mitigations thereof. We investigate the threats against membership privacy and proper mitigations in Chapter 5. Next, we discuss the privacy threats implied by combining genome and DNA methylation data in Chapter 6 and we also present a cryptographic protocol to securely evaluate random forests on biomedical data. In Chapter 7, we also consider other background knowledge of an adversary, such as familial relationships, and take the first step towards a holistic framework analyzing privacy risks of interdependent biomedical data. We conclude this dissertation in Chapter 8.

2
Biomedical Background



(a) (b)

Figure 2.1: (a) A set of chromosomes of a human male. (b) The structure of the DNA double-helix.

In this section, we briefly introduce the relevant genetic principles and provide more background on microRNAs and DNA methylation.

2.1 Genomics

In each human cell, the DNA molecule containing the hereditary information is packaged into 23 pairs of chromosomes. Each pair consists of one chromosome inherited from the father and one inherited from the mother. One pair of chromosomes controls the inheritance of the sex: in males, the two sex chromosomes are different (X, Y), whereas, in females, the two sex chromosomes are the same (X, X). Figure 2.1(a) depicts a set of chromosomes of a human male.

Chromosomes are further organized into segments of DNA, which are called genes. A gene is the basic physical unit of heredity.

The DNA is a double-helix structure consisting of complementary polymer chains. Genetic information is encoded on each of these chains as a sequence of nucleotides: Adenine, Thymine, Cytosine, and Guanine (A, T, G, C). Each nucleotide on one strand of the DNA has a corresponding nucleotide on the opposite strand of the DNA. An A nucleotide on one strand is always paired with a T nucleotide on the opposite strand. Similarly, a G nucleotide is always paired with a C nucleotide. Figure 2.1(b) exemplarily shows a DNA double-helix structure.

Since 99.5% of the human DNA of two different individuals is identical [79], the interesting parts are the remaining 0.5% of the positions. These positions that may vary throughout a population are referred to as *single nucleotide polymorphisms* (SNP). Figure 2.2 depicts genes of two different individuals varying in one SNP.

In order to refer to a specific position on the genome, usually a tuple of the chromosome, the position on the chromosome, and the strand is given. The two strands of the genome are called forward strand, abbreviated by $+$, and backward strand,

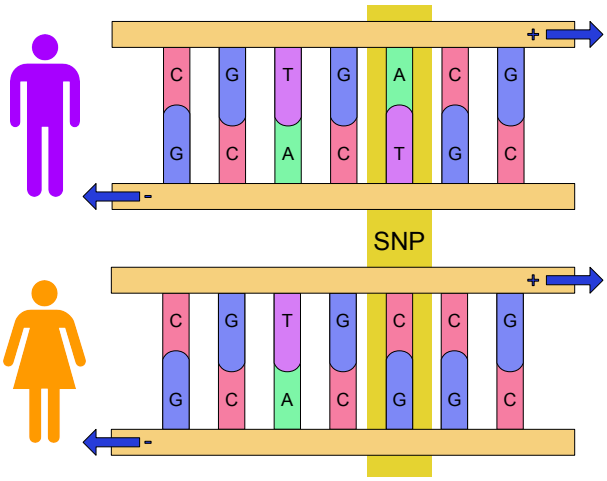


Figure 2.2: A single nucleotide polymorphism.

abbreviated by $-$, as also shown in Figure 2.2. If no strand is specified, the forward strand is assumed.

Generally, two possible nucleotides can be observed at a given SNP (with respect to the forward strand). 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* and 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, since there are always pairs of chromosomes, each SNP position has 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 for the rest of this thesis (specifically in Chapter 6 and 7).

The prior probability of the minor allele is also called minor allele frequency (MAF) and can be retrieved from population statistics databases, such as dbSNP [111, 24], or Kaviar [47].

Mendel's First Law 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 Epigenetics

The term *epigenetics* etymologically comes from the combination of *epi*, which means "above", "over" in Ancient Greek, and *genetics*, which means "origin". This term broadly refers to the study of cellular and phenotypic trait variations stemming from other causes than changes in the genotype. These external factors are for example the in-utero or childhood development, environmental chemicals, aging or diet. Epigenetics can also refer to the changes themselves, such as DNA methylation and histone modification, which determine how genes are expressed without modifying the genome.

2.3 miRNAs

MicroRNAs (miRNAs) are epigenetically regulated mechanisms discovered in the early 1990s. MiRNAs are small non-coding RNA molecules that regulate gene expression in plants and animals. It has been shown that 60% of genes coding human proteins are regulated by miRNAs [42]. Currently, there are more than 5,000 miRNAs known in human beings [94], and this number will undoubtedly keep increasing [83]. Whereas a miRNA is a RNA molecule containing around 22 nucleotides, *miRNA expression* is a real-valued number quantified in a two-step polymerase chain reaction (PCR) process. Different sets of miRNAs are expressed in different cell types and tissues. The miRNA expression level (or value) of a specific miRNA captures how often this miRNA is expressed in the given sample of cells and tissue. A miRNA expression profile represents the set of miRNA expressions of an individual measured from a sample taken at one point in time.

Biomedical research is notably interested in discovering how miRNA expression affects physiological and pathological processes.[1] Studies of miRNA expression profiling have demonstrated that dysregulation of miRNA is linked to neurodegenerative diseases (Alzheimer's and Parkinson's), heart diseases, diabetes, and the majority of cancers [84, 127, 70, 102, 37]. MiRNA expression profiling is hence a very promising technique that could enable more accurate, earlier and minimally invasive diagnosis of severe diseases. Especially when taken from *blood samples*, miRNAs represent a non-invasive diagnosis and have been shown to help identify severe diseases such as cancers or Alzheimer's [72, 77]. A summary of the relation between miRNA and human pathologies can be found in the Human miRNA Disease Database [60].

In Chapter 4 and 5, we will rely on datasets from miRNA-expression-based studies.

2.4 DNA Methylation

One of the most important epigenetic modifications in the DNA is the methylation. Its consequences affect the structure and the activity of the DNA molecule [69, 108].

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

[1] Strictly speaking, miRNA is part of the epigenome while miRNA expression is considered more as part of the transcriptome. In this thesis, we use the term epigenetics in its broader meaning, thus including miRNA expression.

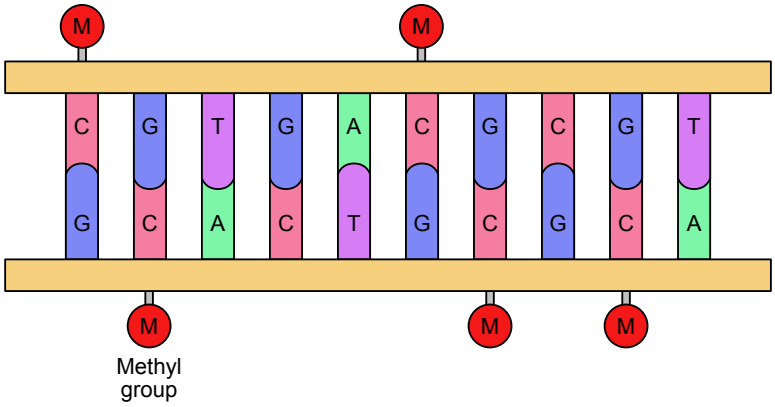


Figure 2.3: Methyl groups attached to a human DNA.

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: it can be methylated or not. Figure 2.3 shows a DNA segment with some CpG-dinucleotides being methylated.

However, since DNA methylation in principle can vary between strands and 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 [36]. 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 [9, 120, 27, 121].

Besides these external factors, the genotype of an individual can also affect the methylation of some regions [116, 87, 43]. Carrying particular alleles at certain SNPs can cause specific DNA methylation patterns at other positions or regions. Such SNPs influencing the DNA methylation are also called methylation quantitative trait loci (meQTLs). In Chapter 6 and 7, we rely on such meQTLs and their strong influence on DNA methylation patterns.

DNA methylation mostly has a repressive effect on gene expression, i.e., which parts of the genome are active in a given cell. Hence, DNA methylation at the promoter of genes can silence specific genes during development, for example, to maintain the pluripotent state of stem cells [110].

An apparent dependency occurs when a SNP affects a CpG dinucleotide. If the polymorphism changes the cytosine (C) or the guanine (G), the CpG dinucleotide may be removed, and no methylation can occur at this position.

3
Related Work

Here, we present the previous work on privacy in genomics and epigenomics and how it relates to our work. We start with a broad overview of privacy in genomics and continue with a more detailed discussion of the topics relevant to our work.

3.1 General Overview

Since the plummeting costs of 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 in particular, well summarized in recent surveys [35, 4, 91]. Recent advances also highlight the role of direct-to-consumer (DTC) genetics, which further increases the chances of genomic data to be made available in less regulated environments like the internet.

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 publications are exceptionally well suited for the DTC area as, for example, a paper by Cristofaro et al. [25], which allows for 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 [8]. Another recent topic in the field focuses on protocols for similar patient queries [126]. 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 [71].

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 [68]. 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 [40]. 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. Thus, it is unacceptable to apply differential privacy in this case, which stands in contrast to our results in Chapter 4, thus proving that, in some cases, differentially-private mitigations are a viable option.

3.2 Linkability and Inference

Recently, some researchers also started to explore new privacy issues stemming from various other types of biomedical data. Similarly to our work on genotype inference from DNA methylation profiles, Schadt et al. have shown that RNA expression data could be used to accurately predict genotypes [107]. The authors present a Bayesian framework that relies on the association existing between expression levels of thousands of genes and genomic variations called expression quantitative trait loci (eQTLs). In the same vein, Philibert et al. demonstrate how methylation array data can be used to construct individually identifying genetic profiles (around 1,000 positions), and to infer substance-use histories, such as alcohol or smoking [99]. They warn that such a genotype inference could reveal personally identifying information, but they do not

study further how genotypes could be matched to methylation profiles. They also do not quantify with what success such an attack could be carried out, and under which conditions. Besides also identifying CpGs correlated with genomic variants, Dyke et al. propose high-level guidelines for methylation data disclosure that preserves privacy [33]. They notably mention the restriction of access to methylation data that are highly correlated with the genotype. However, a concrete scenario to evaluate the extent of the threat as well as the protection provided by their countermeasure is missing.

Franzosa et al. study whether individuals possess microbial patterns that could be used to uniquely identify them [39]. Their results demonstrate that more than 80% of individuals can still be uniquely identified among a population of hundreds of individuals, up to one year later in the case of the gut microbiome. Fierer et al. had already provided some evidence on the feasibility of linking skin bacterial communities back in 2010, but with very few individuals [38].

Gymrek et al. show that genotypes can be re-identified by querying genetic genealogy databases (containing surnames) with short tandem repeats on the Y chromosome [53]. By combining the inferred surnames with other types of metadata, such as age and state, they are able to trace back with high success the identities of multiple contributors in public databases. Humbert et al. show that single nucleotide polymorphisms (SNPs), which are more commonly available online, can also be exploited to infer various phenotypic traits, such as eye color or blood type, in order to further re-identify anonymous genotypes, by typically using side channels such as online social networks [64]. Both of these works illustrate that, once the genotype corresponding to an epigenetic profile, such as a DNA methylation profile, has been identified, it becomes relatively straightforward to recover the real identity of the owner of this methylation profile.

Lastly, there are several papers explicitly relying on graphical models to perform inference and quantify genomic privacy. Humbert et al. analyze the implications of familial relations on kin genomic privacy [61, 63]. 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 in Chapter 7, 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 [S3]. 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. Our approach in Chapter 7 differs from the works mentioned above by the various types of biomedical data and the temporal dependencies between them that we take into account. 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.

3.3 Membership Privacy

Homer et al. were the first to present a membership attack by relying upon allele frequencies (i.e., means of genomic variants' values) and the L_1 distance between those

and the actual genomic data of the victim [59]. Wang et al. extend this attack by making use of the correlations among the different positions in the genome [125]. This improvement on the attack allows them to use the statistics related to only a few hundred genetic variants. Zhou et al. further analyze the theoretical complexity of membership and recovery attacks based on summary statistics [133]. Sankararaman et al. show empirically that the likelihood-ratio test is more powerful than the L_1 distance attack proposed by Homer et al. [106]. Moreover, they derive a theoretical bound on the LR test that provides an excellent approximation of the empirical LR test. Our work on membership privacy in Chapter 5 confirms that, for miRNA expression data, the empirical LR test is better than the L_1 distance attack. Our theoretical relation shows that, in the miRNA case, for a successful attack, the number of miRNAs m has to scale with the square of the number n of participants in the pool. However, our relation is less accurate than theirs with respect to the empirical evaluation, especially when the pools contain individuals carrying a specific disease. This discrepancy can be explained by two facts: (1) the dimensions of both m and n are relatively small compared to those in the genomic setting considered in [106], typically an order of magnitude lower for both, and (2) miRNAs are certainly more affected by diseases than the genome is (as the latter is very stable and only has a few out of millions of variants associated with a given disease). Im et al. show that, if the victim's phenotype is rather extreme or if multiple phenotypes are available, regression coefficients can reveal the victim's participation in a genome-wide association study as much as allele frequencies [66].

3.4 Differentially Private Mitigations

On the defense side, various publications have studied how to properly apply noise to summary statistics for protecting the privacy of GWAS participants. Johnson and Shmatikov propose and implement algorithms for accurate and differentially private computation of various statistics of interest, such as the location of the most significant genomic variants, or the p-values of statistical tests between a given variant and the associated diseases [68]. Uhler et al. have also proposed to rely on differential privacy for sharing GWAS results privately. They present methods for privately disclosing allele frequencies, chi-square statistics, and p-values [122]. Yu et al. extend these methods by allowing for an arbitrary number of cases and controls, assess their performance and compare it with the mechanism proposed by Johnson and Shmatikov [130]. Moreover, Yu et al. present a differentially private mechanism for logistic regression and show how it can be applied to the analysis of GWAS data [131]. In the pharmacogenetics context, Fredrikson et al. show that differential privacy mechanisms can induce bad warfarin dosing, thus expose patients to an increased risk of stroke, bleeding events, and mortality [40]. Many of these previous works also highlight that the amount of noise to be added to the summary statistics is non-negligible, and thus can lead to an unacceptable loss for research utility.

Tramèr et al. [119] investigate how a relaxation of differential privacy that considers weaker adversary can help reach a better privacy-utility trade-off for releasing differentially private chi-square statistics in GWAS. In Chapter 5, we show that, given the structure of miRNA expression data, the same relaxation does not help much to improve

utility in our context, and we thus deduce that the traditional differential privacy model is better suited to release miRNA expression statistics. Finally, Dwork et al. analyze the robustness of the membership attack on noisy summary statistics, and briefly present a generalization to real-valued data [32].

Our work on temporal linkability in Chapter 4 differs from those above in that our differentially private protection mechanisms directly applies noise on the raw miRNA data to guarantee a certain degree of indistinguishability between them, instead of adding noise to summary statistics. Our second defense technique relies on sharing a subset of miRNA data, which is closer to what Humbert et al. have developed in the genomic-privacy context. In particular, Humbert et al. propose an optimization algorithm that allows for sharing raw genomic variants (rather than summary statistics), e.g., for research, satisfying the genomic privacy requirements of all individuals in a family [62]. More generally, our work aims to protect real-valued miRNA expression vectors, which vary over time much more than DNA data.

Comparing the work of Dwork et al. as presented above with our approach on membership privacy in Chapter 5, our work has fewer restricting assumptions (such as the range of the means bounded between -1 and 1 in their work). We consider a reference population containing a substantially higher number of individuals than in the pool, and we provide an experimental validation of our analytical results with real data. Our theoretical relation confirms their finding, i.e., that the dimensionality of the data (referred to as m in this work, d in theirs) for a successful attack scales with n^2. However, our empirical results demonstrate that these theoretical bounds should be taken very cautiously, depending on the application context.

3.5 Cryptographic Solutions

Finally, there have been several works on privacy-preserving disease prediction by relying on encrypted genomic data. Bost et al. develop three main private classification protocols (including decision trees) that protect both the patients' data and the classifier model [13]. They prove their protocols to be secure in the honest-but-curious adversarial model and evaluate its performance on real medical datasets. We build upon their constructions for our own private random forest classifier. Wu et al. propose a novel protocol for the private evaluation of decision trees and also provide an extension to random forests [128]. In contrast to our work, their extension only applies to affine aggregation functions, which are suitable for regression problems, but not for general classification tasks. Thus, their protocol is not able to return the plurality vote. Duverle et al. propose a new protocol that allows for privately computing statistical tests on patients' data by relying on exact logistic regression [28]. Their performance evaluation shows that they can perform statistical tests with more than 600 SNPs across thousands of patients in several hours.

Ayday et al. have developed schemes for private disease susceptibility tests by using homomorphic encryption and proxy-encryption [5, 6]. The considered tests are based on linear combinations of the SNPs (and other environmental and clinical factors in [6]) contributing to a given disease and do not involve complex machine-learning classifiers. Danezis and De Cristofaro improve upon the protocol of [5] by using an alternative

SNP encoding and make the patient-side computation more efficient [21]. McLaren et al. use a similar security architecture as the one initially proposed by Ayday et al. to develop a practical privacy-preserving scheme of genome-based prediction of HIV-related outcomes [88]. All these papers – with the exception of the work of Wu et al. – assume an honest-but-curious adversary, which is considered realistic in the healthcare environment.

4

Linkability of miRNA Expression Profiles

Temporal Linkability Attacks against miRNA Expression Profiles and Practical Mitigations

4.1 Motivation

In contrast to the DNA sequence, which mostly stays constant over time, most other types of biomedical data are variable. This especially holds for epigenetics (or epigenomics), transcriptomics, and proteomics, which aim to bridge the gap between the genome and our health status. These are also influenced by environmental factors (e.g., pollution, diet, lifestyle) and play a crucial role in the development of most common diseases.

Although these types of data are closely linked to a person's health status, their growing importance leads biomedical researchers to publish their study results and datasets in *publically available* biomedical databases, such as the Gene Expression Omnibus (GEO) [44] and the ArrayExpress [3] databases. These databases contain millions of biomedical samples in a pseudonymized form.

Given the easy accessibility of such data, an adversary might attempt to link samples belonging to the same person – despite the temporal variability – in order to obtain a complete profile of his victims. In this chapter, we focus on the temporal linkability of microRNAs (abbreviated miRNAs), one of the most important elements of the epigenome discovered in the early 1990s. Studies of miRNA expression profiles have shown that dysregulation of miRNA is linked to neurodegenerative diseases, heart diseases, diabetes and the majority of cancers [84, 127, 70, 102, 37].

Prior to our publication, it was often believed in the biomedical community that the miRNA expression levels are varying sufficiently to invalidate any linkability attempts over time, thus naturally protecting personal privacy. Our evaluation, however, showed the contrary: despite their temporal variability, microRNA expression profiles are still identifiable and linkable after time periods of several months.

4.2 Contributions

In this chapter, we study the temporal linkability of personal miRNA expression profiles, by presenting and thoroughly evaluating different attacks and proposing defense mechanisms to enhance unlinkability.

In particular, we first study an identification attack, which pinpoints a specific miRNA expression profile in a database of multiple expression profiles by knowledge of the targeted profile at another point in time. Second, we study a matching attack, which tracks a set of miRNA expression profiles over time. We rely on principal component analysis to pre-process the miRNA expression levels, and on a minimum weight assignment algorithm for the matching attack. We thoroughly evaluate these linkability attacks by using three different longitudinal datasets: (1) the blood-based miRNA expression levels of athletes at two time points separated by one week, (2) the plasma-based miRNA expression levels of the same athletes at two time points separated by one week, and (3) the plasma-based miRNA expression levels of patients with lung cancer over more than 18 months and eight time points. Our experimental results show that blood miRNA expression profiles are about twice as easy to track over time than plasma miRNA profiles. Furthermore, the matching attack is more successful than the identification attack: We reach a success rate of 90% with blood and a success rate of 48% with plasma miRNAs in the matching attack whereas, in the identification attack,

we reach a success rate of 76% with blood and 28% with plasma miRNAs. Moreover, we demonstrate that 10% of the miRNAs are already sufficient to achieve similar success rates as with all miRNAs. With the third dataset, we also observe that the attack achieves a similar success up to 12-month time periods.

We present two protection mechanisms to improve the unlinkability of miRNA expression profiles: (1) hiding a subset of the miRNA expressions, e.g., those that are not relevant for medical practice, and (2) disclosing perturbated miRNA expression profiles by adding noise in a differentially private and distributed manner. While the first countermeasure is useful especially in a clinical setting, in which the disease-relevant miRNAs are already known, the second countermeasure is intended to be better suited for the biomedical research community. In this context, as one of the objectives is to discover associations between miRNAs and diseases, it is impossible to restrict the released data to only a few miRNAs.

We evaluate our protection mechanisms with the first aforementioned blood-based miRNA profiles of athletes and a fourth, also blood-based, miRNA dataset of more than 1,000 participants that includes information about 19 diseases (at a single point in time). The former is used to measure how temporal linkability is reduced with our countermeasures, whereas the latter helps us evaluate the evolution of accuracy (i.e., utility) in predicting patients' diseases from their miRNA expressions. The experiments show that it is possible to decrease linkability by at least 50% for almost no loss of accuracy ($< 1\%$) for the majority of diseases with the perturbation mechanism. Moreover, our results demonstrate that the perturbation mechanism provides better privacy-utility trade-offs than the hiding method in 17 out of 19 of diseases while allowing more flexibility in the data usage for biomedical researchers. This finding is reinforced by the fact that an adversary could use correlations between miRNA expressions to infer more miRNA expressions than those shared by our first countermeasure.

4.3 Threat Model

We assume the adversary gets access to miRNA expression profiles of individuals at different points in time. Such epigenetic data is increasingly available in public research databases, such as the Gene Expression Omnibus (GEO) [44] or ArrayExpress [3] databases. Moreover, such data could be leaked through a major security breach, e.g., of a hospital server. Health data is also increasingly available on the black market. For instance, cyber attacks against healthcare companies have increased by 72% from 2013 to 2014 [117]. Also, 91% of healthcare companies have experienced a violation of their databases over the last two years, and only 32% feel they have adequate resources to defeat these incidents [89]. Real-world cyber attacks show us that health data can be hacked en masse [56, 100] or that attacks can be more targeted towards high-profile victims [123]. Very sensitive medical data of thousands of patients can also end up online, due to a human mistake [11].

In a typical scenario, the adversary would get access to miRNA expression levels of one or multiple individuals from a (private) health insurance or hospital database and wants to match them with a (public) research dataset of miRNA expression levels at another point in time. A particularly sensitive scenario would be the matching of

non-anonymized healthy miRNA samples with miRNA profiles that are known to be associated with diseases. Also, note that researchers have demonstrated that RNA expression profiles could be matched to genotypes by relying on expression quantitative trait loci (eQTLs) [107]. Therefore, if the adversary can also access the genotypes of the victims, these genotypes provide him with further means for de-anonymizing the corresponding (micro)RNA expression profiles [53, 64]).

4.4 Linkability Attacks

We study the extent of the linkability threat (as described in Section 4.3) by means of two attacks. In this section, we describe the mathematical principles our attacks are based on.

The first attack, called *identification attack*, refers to a scenario in which the adversary knows the miRNA expression profile of a targeted individual and aims at finding the corresponding miRNA expression profile in a database of n miRNA expression profiles, e.g., later in time. The second attack, called *matching attack*, refers to the case where the attacker has access to two databases of miRNA profiles collected at different points in time and wishes to match their elements.

4.4.1 Pre-processing

For both of our attacks, since there are more than 1000 known miRNAs with real-valued expression levels, we apply a pre-processing step using principal component analysis (PCA) with whitening. In particular, we apply the probabilistic PCA model proposed by Tipping and Bishop [118], which relies on singular value decomposition. This PCA step projects the high-dimensionality miRNA expression vectors to smaller-dimensionality uncorrelated components. The whitening step divides the resulting PCA components by the number of samples multiplied by the singular values in order to provide uncorrelated expression vectors of unit variance. We then make use of the Euclidean distance between the miRNA expression vectors projected on the first c principal components.

4.4.2 Identification Attack

In the identification attack, we assume the adversary has had access to the miRNA profile $\mathbf{x}_k^{t_1}$, a vector containing the miRNA expressions of an individual k at time t_1, and he wishes to identify this individual in a database of n miRNA expression profiles $\{\mathbf{x}_i^{t_2}\}_{i=1}^n$ collected at time $t_2 \neq t_1$. After having extracted the c principal components from the whole dataset by using PCA, the adversary ranks the n profiles (projected on the c components) $\{\bar{\mathbf{x}}_i^{t_2}\}_{i=1}^n$ by decreasing distance to the targeted miRNA profile (also projected on the c components) $\bar{\mathbf{x}}_k^{t_1}$ and picks the profile with minimum distance to the targeted profile. Formally, the adversary will select the profile $\bar{\mathbf{x}}_{i^*}^{t_2}$ where

$$i^* = \arg\min_i \left\| \bar{\mathbf{x}}_i^{t_2} - \bar{\mathbf{x}}_k^{t_1} \right\|_2 .$$

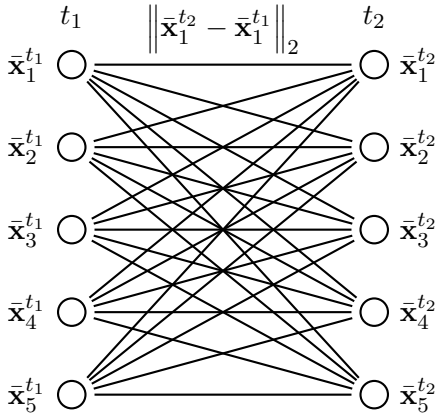


Figure 4.1: The bipartite graph representation of a matching attack.

4.4.3 Matching Attack

In the matching attack, the adversary has access to two databases of miRNA expression profiles at two different points in time t_1 and t_2. We assume that the databases are of sizes n_1 and n_2, both strictly greater than 1. First, if $n_1 = n_2 = n$, the adversary will assign one miRNA profile at time t_1 to exactly one profile at time t_2. In this case, the best assignment σ^* is the one that minimizes the sum of the distances between every matched pair:

$$\sigma^* = \arg\min_{\sigma} \sum_{i=1}^{n} \left\|\bar{\mathbf{x}}_{\sigma(i)}^{t_2} - \bar{\mathbf{x}}_i^{t_1}\right\|_2.$$

This problem boils down to finding a perfect matching on a weighted bipartite graph with n vertices on both sides representing the miRNA profiles and a weight on each edge representing the Euclidean distance between any pair of miRNA profiles (vertices) projected on the first c principal components. Figure 4.1 exemplarily depicts the representation of the matching attack as a bipartite graph problem. We want to find the matching among $n!$ possible assignments that minimizes the sum of the weights between vertices. Fortunately, in literature, there exist several algorithms that find the minimum weight assignment in polynomial time. We use the Blossom algorithm [34], because it only has a complexity of $O(n^3)$ and it can also be applied to general graphs.

If $n_1 \neq n_2$, we fill the smaller side of the bipartite graph with dummy vertices, so that both sides have an equal number of vertices. Then we assign infinite weight to all edges from actual vertices to these dummy vertices in order to ensure that the dummy vertices will be the least likely assigned vertices.

4.5 Dataset Description

Unlike in other fields of privacy research, where large amounts of data can be collected in a small amount of time and at low cost, in the health-privacy field, we face the exact opposite: measuring the miRNA expression levels of one single sample already costs several hundred dollars. Longitudinal epigenetic data are particularly valuable, since

patients have to regularly provide their biological samples over a long period of time. Therefore, the four datasets used throughout the chapter, and described hereunder, represent very rich data.

We start by describing our three longitudinal datasets. The first dataset contains the blood-based miRNA expression levels of 29 well-trained male athletes (15 endurance athletes and 14 strength athletes) at two points in time, while the second dataset contains the plasma-based miRNA expression levels of those athletes at the same points in time.[1] None of the athletes is known to be affected by a disease. The samples were taken prior and post exercising (period of one week), similar to the data previously presented in [7]. The athletes followed a 6-day training with two training sessions a day, except at day 4 when only one session was scheduled. The tests were conducted at Saarland Unversity (Germany) for the endurance athletes, and at Ruhr University Bochum (Germany) for the strength athletes. These datasets are part of a publication of Hecksteden et al. [57] and are publicly available as an additional file.

In order to confirm our results, we make use of a third, independent dataset. This dataset contains the miRNA expression data of plasma of 26 lung-cancer patients (9 females and 17 males) over a period of more than 18 months [78]. The samples were taken at eight points in time: before surgery (tumor resection), two weeks after surgery (abbreviated A.S. in the graphs), and 3, 6, 9, 12, 15, and 18 months after surgery.[2] The patients' ages range from 47 to 79. This dataset is freely available in the GEO database (see accession number GSE68951).

All three longitudinal datasets include the expression levels of 1,189 miRNAs for each individual at every time point.

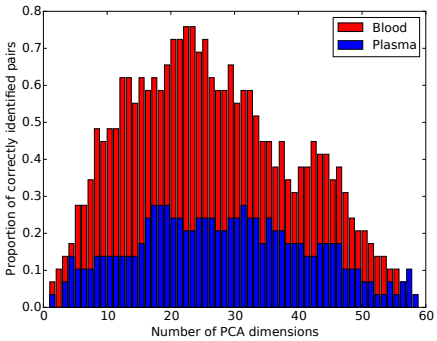
Our fourth dataset contains the expression levels for 848 miRNAs collected from blood samples for each of 1,049 individuals [73] at only one time point. 94 of these individuals are considered to be healthy and are used as a control group in Section 4.7. Most of the rest represent cases, i.e., individuals carrying one out of the following 19 different diseases: 124 have Wilms tumor, 73 lung cancer, 65 prostate cancer, 62 myocardial infarction, 47 chronic obstructive pulmonary disease (COPD), 45 sarcoidosis, 45 ductal adenocarcinoma, 43 psoriasis, 37 pancreatitis, 35 benign prostate hyperplasia, 35 melanoma, 33 non-ischaemic systolic heart failure, 29 colon cancer, 24 ovarian cancer, 23 multiple sclerosis, 20 glioma, 20 renal cancer, 18 periodontitis, and 13 stomach tumor. The dataset can also be found in the GEO database under accession number GSE61741.

Note that a miRNA expression generally takes values between 0 (meaning the miRNA is not expressed at all) and tens of thousands. As we will mention later, we typically filter out miRNA whose median expressions among all individuals are smaller than 50, since these are non-expressed or not expressed enough to be significant.

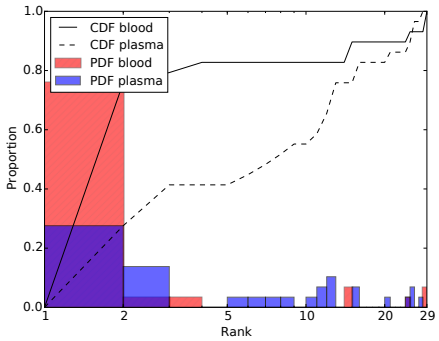
An overview and comparison with the datasets in the other chapters can be found at the end of this thesis.

[1] We selected blood and plasma since these two body fluids are likely candidates as source for biomarkers in future applications.

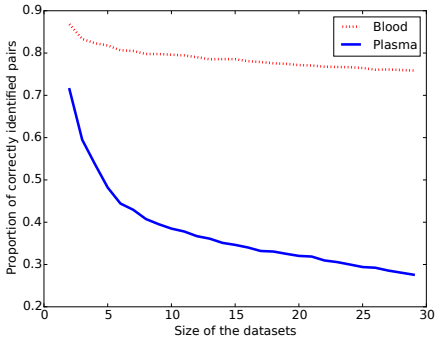
[2] Note that for the last two points in time, we have the miRNA profiles of 25 and 22 patients, respectively.



(a)



(b)



(c)

Figure 4.2: Success rate of the identification attack for the athletes dataset. (a) Proportion of successfully identified pairs plotted against the number of PCA dimensions (in $\{1, \ldots, 58\}$). (b) Probability density function (PDF) and cumulative distribution function (CDF) of obtained ranks. (c) Proportion of successfully identified pairs plotted against the number of miRNA expression profiles.

4.6 Experimental Results

We evaluate how successful both our attacks are in breaking the privacy of our three longitudinal datasets. We implement the attacks in Python and make use of the libraries Scikit-learn [98, 15] (for PCA) and NetworkX [54] (for the graph matching).

4.6.1 Identification Attack

In this subsection, we evaluate the success of an adversary who aims at identifying the miRNA profile of a targeted individual in a longitudinal dataset. As mentioned in Section 4.5, the first two longitudinal datasets contain miRNA expression levels of 29 individuals collected at a time interval of one week.

First, we compare the success rate for correctly identifying samples between these two longitudinal datasets. Figure 4.2(a) indicates that the blood-based miRNA expression levels are easier to identify over time than the plasma-based miRNA expression levels.

When identifying samples by their blood miRNA expression levels, we can reach a maximum success rate of 76% for the blood with 22 or 23 PCA dimensions. The maximum success rate for the plasma is 28% with 17, 18, 19 or 31 PCA dimensions. Note that both achieve their highest success with a number of PCA dimensions around 20.

Next, we rank the miRNA profiles at time t_2 in order of increasing distance to the targeted profile $\mathbf{x}_k^{t_1}$. Figure 4.2(b) shows the rank of the correct sample $\mathbf{x}_k^{t_2}$ by using 22 PCA dimensions for the blood and 18 PCA dimensions for the plasma. The correct profile is ranked within the top 2 profiles in more than 40% of the cases for the plasma, whereas the correct sample is ranked within the top 2 samples in 80% of the cases for the blood.

In order to get an impression of the attack performance on larger datasets, we also analyze the success of the identification attack with respect to the number of participants in the dataset, i.e., we vary the number of profiles among which the attacker has to identify the targeted miRNA profile, again using 22 PCA dimensions for the blood and 18 PCA dimensions for the plasma. Intuitively, when the number of miRNA samples increases, the success rate of the attacker should decrease. In this experiment, we adjust the number n of miRNA profiles between 2 and 29 and evaluate the attacker's success on a subset of our datasets. In particular, for each number of profiles n, we randomly choose 1000 different combinations (or fewer if necessary) of n out of 29 miRNA profiles and run the identification attack on every sample within this subset. Figure 4.2(c) depicts the average success rates for each number of profiles n. As expected, the success rate monotonically decreases with the number of participants for blood and plasma samples. For plasma, however, this decrease is much sharper, confirming that the blood's miRNA expression levels provide means for easier identification. From the curves' slopes, we can predict that, for larger datasets, blood-based samples will still be subject to a relatively high identification success.

In order to validate our findings, we also evaluate our experiments on our other longitudinal, independent dataset containing plasma miRNA profiles from 26 individuals with lung cancer collected over up to eight different points in time.

First, we evaluate the attacker's success with respect to a varying number of PCA dimensions. Figure 4.3(a) depicts the minimum, average and maximum success rate of an attacker when identifying the samples between different points in time, irrespective of the time period between them. The maximum success rate for the identification attack is 42% and is achieved for 25 and 39 PCA dimensions. The usage of 22 PCA dimensions yields the highest average success rate of 22%. The highest minimal success rate in the dataset is achieved for 17 PCA dimensions (12%).

These results are similar to what we obtained in our experiments for the plasma-based athletes dataset: The best results are achieved for a number of PCA dimensions around 20 in both datasets. The highest average success rate lies 6 points below the best success rate for the athletes dataset. This could be explained by longer time periods in this dataset. However, for some time periods, we can achieve one and a half the success rate of the first dataset. When comparing the top 10 miRNAs contributing to the first PCA dimension in this dataset and in the athletes' plasma dataset, we find an overlap of 80% between these miRNAs. This overlap indicates that approximately the

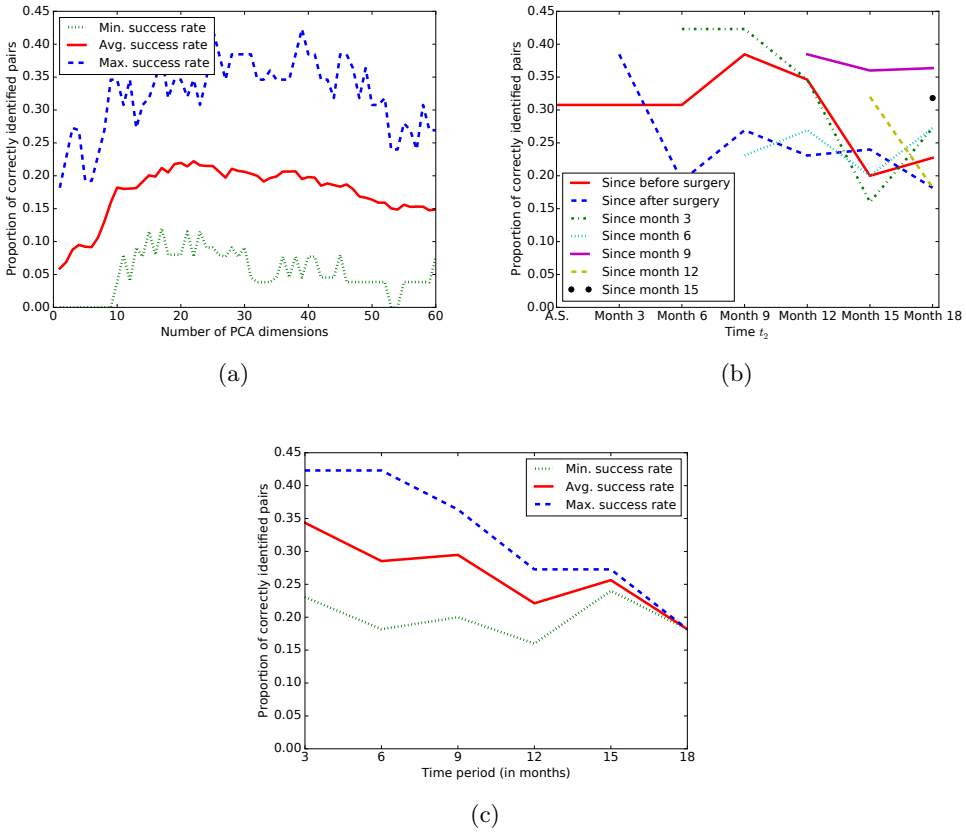


Figure 4.3: Success rate of the identification attack for the lung cancer dataset. (a) Success rate aggregated over all identifications between any t_1 and t_2 plotted against the number of PCA dimensions. (b) Success rate of identifying the miRNA profiles between time pairs t_1 and t_2. (c) Success rate plotted against the time period between t_1 and t_2.

same set of miRNAs can be used to differentiate plasma expression profiles between individuals in both datasets. Thus, we can conclude that, while miRNA expression levels are directly linked to health status, the health status only affects a subset of the miRNAs, which has only little effect on the temporal linking.

To further investigate the effect of different time periods on the attacker's success, we plot the maximum success rates between all possible, ascending combinations of time points in Figure 4.3(b). With only a few exceptions, the best success rates are achieved for consecutive time points. The only two exceptions are found for t_1=*before the surgery* and t_1=*the sixth month after the surgery*. In general, however, we notice a tendency of slight decrease in success over an increasing time period.

In order to verify this finding, we group the results by the period between t_1 and t_2 (Figure 4.3(c)). Note that, since we do not know the time period between before the surgery and after it, we leave out all results that use samples collected before the surgery. Clearly, the best achievable success rate drops for increasing time periods.

This decrease over larger periods of time can partially explain the lower average success rate in this dataset compared to the athletes' dataset (considering a much smaller time period).

Next, we computed the guessing entropy [86, 16] for the identification attack. The guessing entropy $\mathrm{E}[G(X)]$ is the expected number of guesses an adversary would need to identify the correct sample at a different point in time. For the identification attack it is given by $\mathrm{E}[G(X)] = \sum_{i=1}^{n} i \cdot \Pr[X = i]$, where X denotes the rank of the correct sample at time t_2 and $\Pr[X = i]$ denotes the empirical probability that the correct sample is ranked at the i^{th} position.

For blood-based samples of our athletes dataset, the attack can achieve a guessing entropy just below 4, clearly outperforming random guesses, which would yield an entropy of 15 guesses on average. For plasma-based samples of the same dataset, the attack yields an entropy of approximately 9 guesses. This result is consistent with the results on the lung cancer dataset, where, on average, an adversary would need just fewer than 9 guesses (compared to a guessing entropy of 13.5 for random guesses). Moreover, for some t_1 and t_2, the attack is even able to achieve a guessing entropy smaller than 6.

4.6.2 Matching Attack

We evaluate here the success of the adversary, who tries to link all participants over time, again for the three aforementioned longitudinal datasets. Starting with the athletes' datasets, we compare the success rate of matching the blood and the plasma over all possible PCA dimensions for 29 participants. In Figure 4.4(a), we notice the same behavior as in the identification attack: The blood-based miRNA expression levels are much easier to link over time than the plasma-based levels. We even reach a higher maximum absolute success rate than in the identification attack: 90% with 39 or 40 PCA dimensions for the blood and 48% success with 34 PCA dimensions for the plasma samples.

The lower success rate of the identification attack is due to the fact that it is evaluated for each sample individually, thus allowing multiple samples at t_1 to be linked to the same (potentially wrong) sample at t_2. Since our matching attack rules out those cases by forcing each profile at t_2 to be matched to exactly one profile from t_1, it also decreases the number of wrongly matched samples.

Next, we also analyze the success of the attack with respect to the number of participants to be matched together. We suppose that the more miRNA profiles there are, the more challenging it should be for the adversary to match them at different time points. Again, we vary the number of participants n between 2 and 29 at both time points, again randomly sampling 1000 combinations (or fewer, if there are fewer than 1000 combinations) and averaging the result. Figure 4.4(b) shows the expected trend of decreasing success for the blood miRNA samples. The plasma scenario monotonically decreases between 2 and 25 participants and then slightly increases until 29. This artifact could be explained by the smaller number of random combinations, and thus experiments, when $n > 26$. We also find that the blood attack faces a rather linear decrease in success, whereas the plasma success rate decreases much faster. By extrapolating this linear

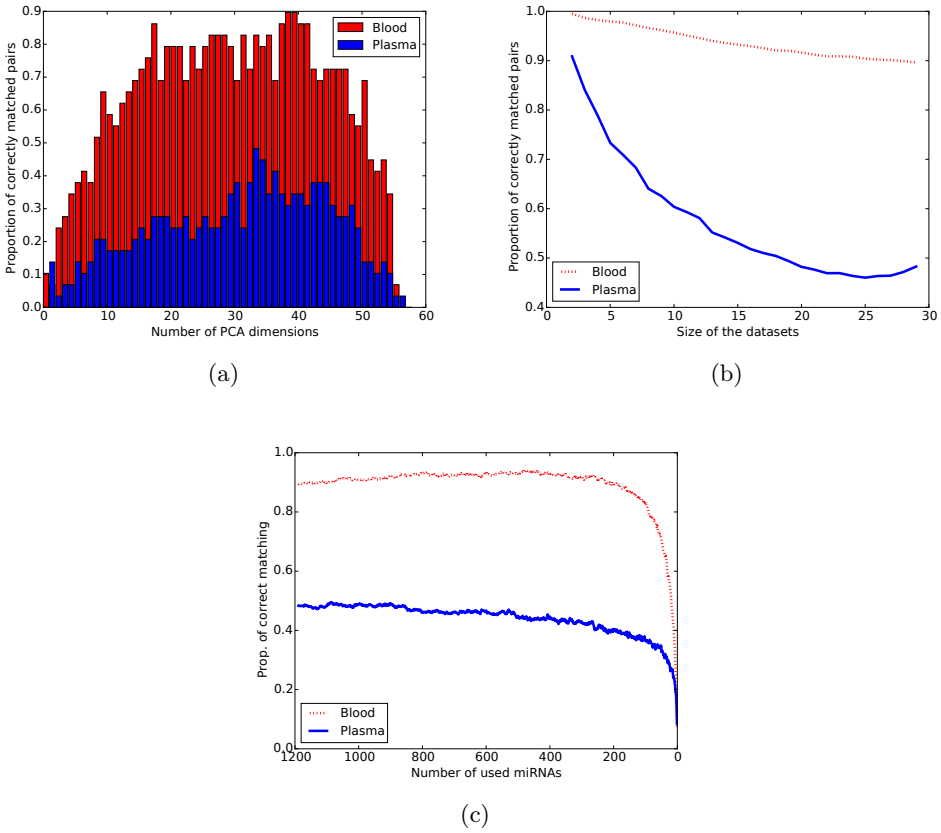


Figure 4.4: Success rate of the matching attack for the athletes dataset. (a) Proportion of successfully matched pairs plotted against the number of PCA dimensions. (b) Proportion of successfully matched pairs plotted against the number of miRNA profiles. (c) Proportion of successfully matched pairs plotted against the number of revealed miRNAs.

trend, we can expect a success rate as high as 60% with 120 participants in the datasets. Therefore, we conclude again that the blood has miRNA expression levels that allow for much easier tracking over time than the plasma, which is consistent with the results of the identification attack.

Figure 4.4(c) investigates how the attack success evolves when revealing only a subset of the miRNA expression levels. We gradually drop individual miRNAs in random order and compute the attack success. The figure shows the success rate (for each possible number $m \in \{1189, 1188, \ldots, 2, 1\}$ of miRNAs) averaged over 50 randomly chosen orderings of miRNAs. We notice that the attack success is very stable, especially for the blood samples, from 1189 to 200 miRNAs. For the blood, the success decreases below 80% the first time when there are fewer than 100 miRNAs available to the adversary. We further study the implications of this robustness in the context of our countermeasures in Section 4.7.

We also made use of our third longitudinal dataset containing plasma miRNA

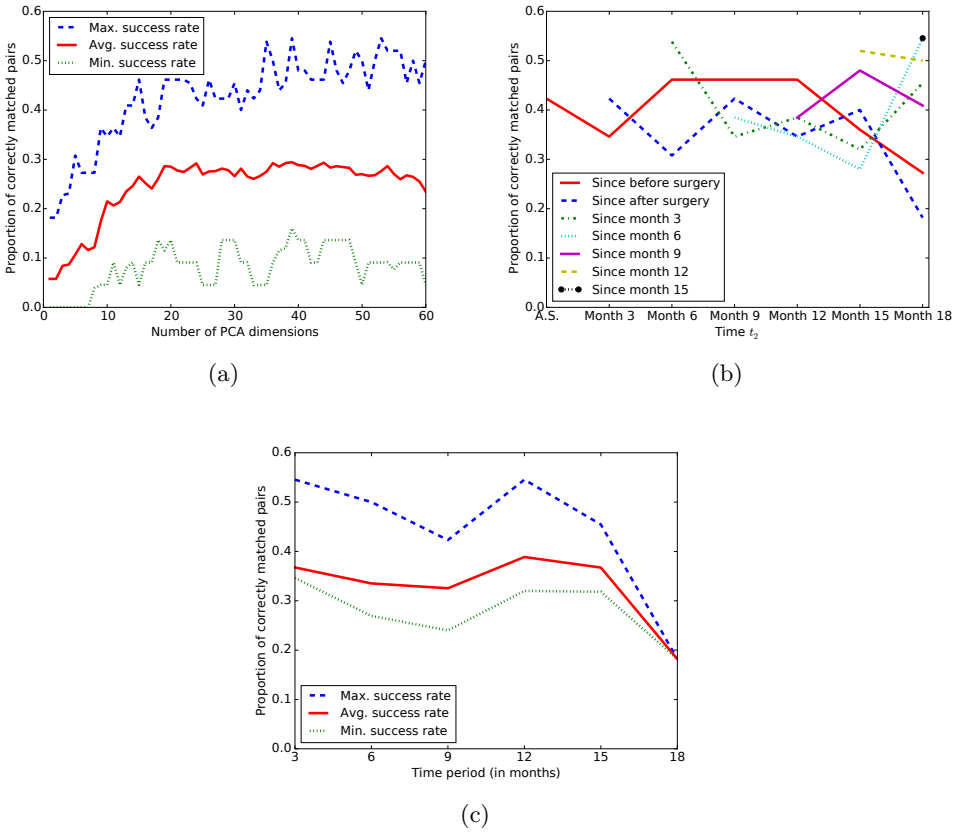


Figure 4.5: Success rate of the matching attack for the lung cancer dataset. (a) Success rate aggregated over all matchings between any t_1 and t_2 plotted against the number of PCA dimensions (in $\{1, \ldots, 60\}$). (b) Success rate of matching the miRNA profiles between time pairs t_1 (various curves), t_2 (x-axis value). (c) Success rate plotted against the distance between t_1 and t_2.

expression profiles of 26 individuals over up to eight different time points (cf. Section 4.5). In Figure 4.5(a), we see that the average success rate reaches its maximum at a number of PCA dimensions very close to the number of dimensions for the athletes dataset, i.e., 34. However, this maximum is approximately 30%, which is smaller than the 48% reached for the first dataset. A greater period between time points could explain this behavior, and we also see that we still can reach a maximum success rate of 55% between some time points, with 39 PCA dimensions. We explore the time effect in more detail in the following figures.

Figure 4.5(b) depicts the maximum success rate between any pair of time points t_1, t_2. For instance, the solid red line shows the success rates between t_1=*before surgery* and all others. It is difficult to detect any trend with respect to the time period in the different curves, except a slight decrease when the time period is higher or equal to 15 months. This is confirmed by Figure 4.5(c) that depicts the maximum, average, and minimum success rate with respect to the period between t_1 and t_2. We notice a

decreasing rate between 3 and 9 months, an increase to 12 months, and finally a clear decrease towards 15 and 18 months.

4.7 Mitigations

In this section, we propose and evaluate two main protection mechanisms for preventing miRNA expression data from being tracked over time. The proposed techniques are based on well-established privacy-enhancing methods, previously applied in other privacy contexts, such as location privacy. The first approach relies on a quite straightforward technique: releasing only a subset of the miRNAs. We can already see from Figure 4.4(c) of Section 4.6 that the matching attack is quite robust to a decrease in the number of miRNAs. Nevertheless, we show hereafter how we can keep a high utility in combination with unlinkability of expression profiles over time by revealing a small subset of miRNAs. The second countermeasure consists in adding noise to the released miRNA expression vectors, independently for every individual. This method shows very promising results, reaching an even better privacy-utility trade-off than the hiding mechanism. Furthermore, we also investigate the effect of correlations between miRNA expression levels and present the privacy evolution when the adversary can infer missing miRNAs by using these correlations.

For evaluating the privacy provided by our protection mechanisms, we focus on the matching attack against blood-based miRNAs, as this constitutes the worst-case attack from a privacy perspective, as shown in Section 4.6. Moreover, we assume the attacker is able to select the number of PCA dimensions that maximize his success. This provides us with a conservative measure of privacy, showing the worst-case privacy levels individuals can expect.

4.7.1 Baseline Utility

Before presenting the proposed countermeasures and their efficiencies, we must carefully define the context in which they should apply. Indeed, we can rarely have both perfect privacy and maximum utility, so that we often aim for a trade-off between these two. Therefore, the efficiency of the defense mechanism cannot only be judged based on the privacy metric, but must also relate to the utility brought in the context in which the data is used.

According to biomedical experts, miRNA expression profiles have strong potential to help predict various severe diseases, from cancer to Alzheimer's disease. Biomedical researchers typically rely on standard machine learning algorithms to identify which miRNAs are playing a significant role in the disease of interest. They are dealing with binary classification between cases (carrying the disease) and controls (healthy) and most often rely on support vector machines (SVMs). In particular, they typically use radial basis function SVMs and select a subset of features by subsequently adding miRNAs in order of their significance values (e.g., p-values computed by the Wilcoxon-Mann-Whitney (WMW) test) [77] or equivalently in order of their area under the ROC curve (AUC). Given samples of cases and controls, the accuracy is then defined as the number of correctly classified samples divided by the total number of samples. Note

Disease	Maximum accuracy with the best subset of expressed miRNAs (# miRNAs)	Accuracy with all expressed miRNAs
Periodontitis	0.941 (37)	0.88
Renal cancer	0.988 (32)	0.962
Wilms' tumor	0.95 (150)	0.937
Benign prostate hyperplasia	0.921 (105)	0.883
Chronic obstructive pulmonary disease	0.932 (70)	0.886
Colon cancer	1.0 (30)	0.997
Ductal carcinoma	0.938 (55)	0.92
Glioma	0.927 (19)	0.83
Lung cancer	0.899 (60)	0.848
Melanoma	0.996 (185)	0.992
Multiple sclerosis	0.992 (40)	0.979
Myocardial infarction	0.893 (400)	0.884
Nonischaemic systolic heart failure	0.9 (135)	0.871
Ovarian cancer	0.919 (18)	0.876
Pancreatitis	0.941 (130)	0.899
Prostate cancer	0.923 (90)	0.91
Psoriasis	0.914 (350)	0.902
Sarcoidosis	0.977 (200)	0.97
Tumor of stomach	0.969 (160)	0.89

Table 4.1: Accuracy of the SVM algorithm in classifying individuals between cases (carrying the disease) and controls (healthy), for 19 diseases, without countermeasure.

that we compute the average accuracy over a repeated k-fold cross-validation.

In this chapter, we define the utility as the accuracy of the SVM classifier, as defined above. We use a 10-fold cross-validation with 5 repeats (using R [103] and the caret [76] library) and determine the miRNAs' p-values by using the WMW test and adjusting the significance values for multiple tests using the Benjamini-Hochberg adjustment. The WMW test statistic is applied for each miRNA individually in order to test whether this miRNA has similar expressions between cases and controls (null hypothesis). The p-values then provide us with the relevance of the miRNA to the disease of interest. In contrast to the t-test, the WMW test can be applied to unknown distributions. This way, we follow the standard procedure of biomedical research. Table 4.1 shows the accuracy of our SVM algorithm applied to our 1000+ participants dataset to predict 19 diseases, without any obfuscation. The maximum accuracy here is what we refer to as the baseline utility in our evaluation.

Note that, before running the SVM algorithm, we filter out non-expressed miRNAs, i.e., those with a median level of expression smaller than 50 over the 1000+ individuals, which leaves us with 446 expressed miRNAs.

4.7.2 Hiding MicroRNA Expressions

The first countermeasure that we study is miRNA expression hiding. This obfuscation technique has the advantage to be non-pertubative, i.e., to preserve the correct values of all revealed miRNA expressions. However, as we have seen in Section 4.6, the attacks are extremely robust against removal of miRNAs. In the following, we want to find an optimal trade-off between the diagnosis accuracy, i.e., the utility, and the unlinkability of the data, i.e., the privacy. To this end, we make use of both our blood-based datasets, the 1000+ dataset with blood-based miRNA expressions to run our SVM algorithm and the athletes' dataset with blood-based miRNAs to evaluate the level of privacy. Note that we filter both datasets' miRNAs in order to obtain the same set of 446 miRNAs in both cases. While we measure the utility in terms of accuracy of the SVM, the privacy is measured in terms of the maximum achievable success rate (over all possible PCA dimensions) of our matching attack.

Figure 4.6 shows the evolution of privacy and utility for a range of 1 to 100 disclosed miRNAs, for 6 different severe diseases.[3] We focus on this range of miRNAs as: (1) for more than 100 miRNAs, the attack success rate is approximately the same as the one without countermeasure, and (2) the SVM can already achieve very high accuracy with up to 100 miRNAs. We gradually reveal the miRNAs in decreasing order of significance (based on p-values), as computed in Section 4.7.1.

Figure 4.6 demonstrates that there exists a trade-off between the utility of miRNA expressions and the privacy of the contributors' data. Note that we also depict the relative decrease in accuracy compared to the maximum SVM accuracy computed in Section 4.7.1 and the relative decrease in the attack success (increase in privacy) compared to the attack success with all miRNAs, i.e., 90%. We see that the relative decrease in accuracy is almost always smaller than 10%. The only exceptions to this are pancreatitis and melanoma, for fewer than 3 disclosed miRNAs. Moreover, regarding the privacy, the figures show that we can never reduce the attack success by more than 50% when revealing more than 20 miRNAs. Nevertheless, within the range of 3 to 20 disclosed miRNAs, we can find, for all diseases, a satisfactory trade-off between utility and privacy.

In particular, for glioma, we can decrease the success of the linkability attack and thus improve the privacy by 80.8% when using 4 miRNAs, while reducing the classification accuracy by only 1.1%. Similarly, for multiple sclerosis, 7 miRNAs provide an increase in privacy of 53.8%, while the decrease in accuracy only amounts to 0.9%. For renal cancer and 10 miRNAs, we are able to achieve an improvement in privacy of 69.2% and a decrease of accuracy of only 1.7%. There are only two diseases for which it is very difficult to have both unlinkability and very high utility: melanoma and pancreatitis. For melanoma, we notice that the success of the matching attack exhibits a fast increase with very few miRNAs, and already exceeds 50% starting with only 7 miRNAs. For pancreatitis, the SVM's accuracy is relatively low (compared to the maximum) for the first 20 miRNAs. Thus for both diseases, either privacy or utility would have to be sacrificed for the other.

[3] These are representative of the behavior of all 19 diseases we tested our privacy-preserving mechanisms on.

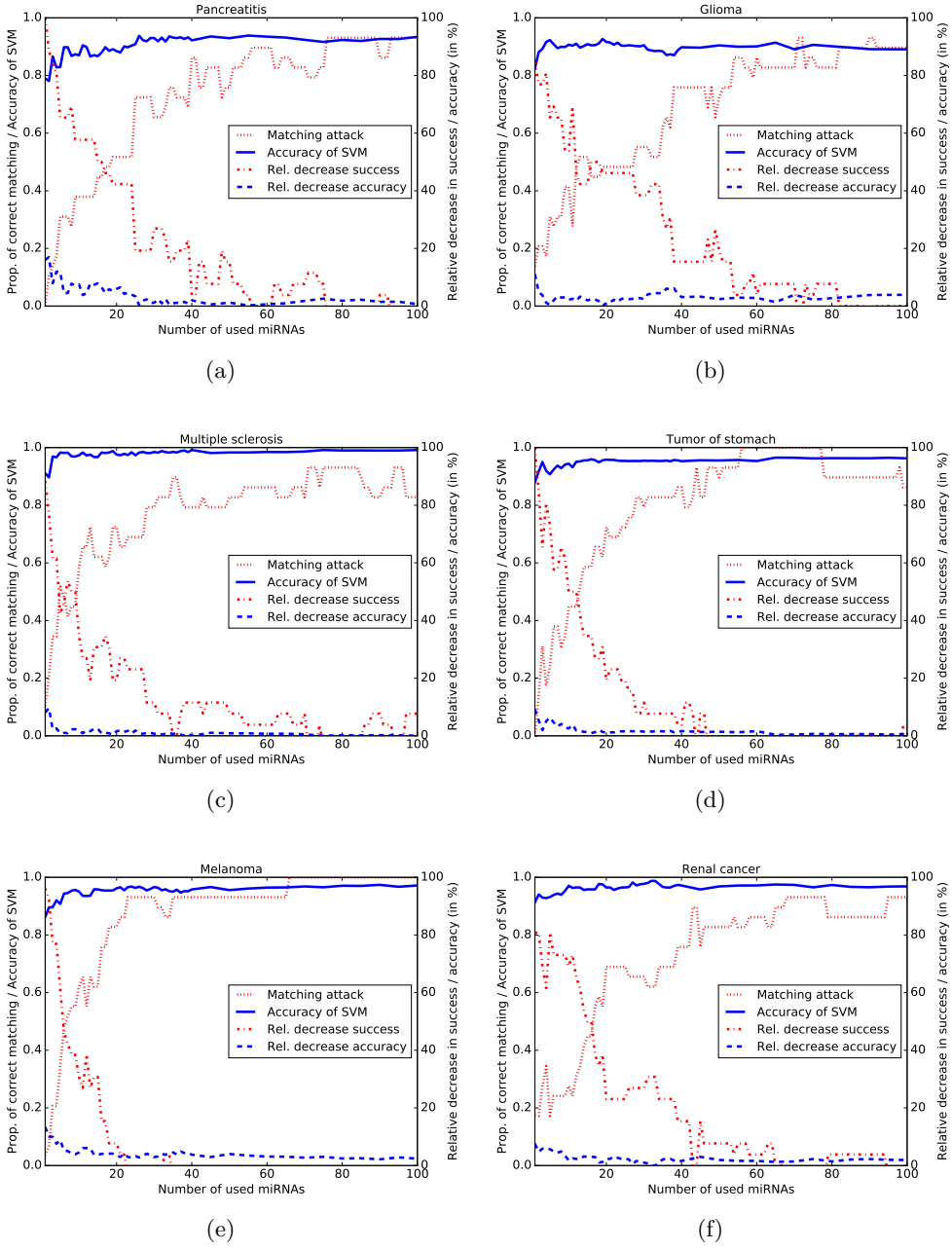


Figure 4.6: Evolution of privacy (unlinkability) and utility (classifier accuracy) plotted against the number of released miRNAs for the following diseases: (a) Pancreatitis, (b) Glioma, (c) Multiple sclerosis, (d) Tumor of stomach, (e) Melanoma, (f) Renal cancer. The *relative decrease success* curve refers to the decrease in success of the matching attack compared to the success without countermeasure. Similarly, the *relative decrease accuracy* curve refers to the decrease in accuracy of the SVM classifier with respect to the case without the protection mechanism.

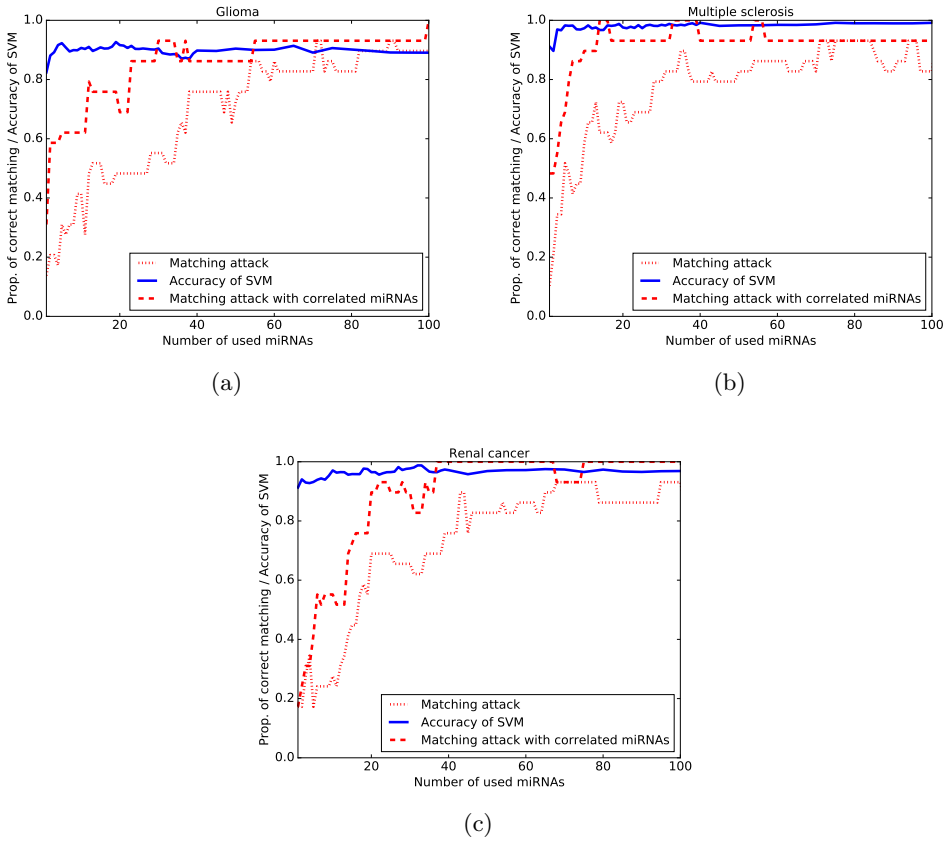


Figure 4.7: Correlations between miRNAs. Evolution of privacy and utility, when miRNAs correlated with the revealed miRNAs are taken into account for the attack. This provides an upper bound on the best linkability of miRNA expression profiles, i.e., worst-case privacy level. (a) Glioma, (b) Multiple sclerosis, (c) Renal cancer.

MiRNA Co-expression. Like between positions in the genome, there exist correlations between miRNA expressions: Around 40% of miRNAs are not independently expressed [90]. This means that the adversary, by knowing these correlations, could increase his knowledge about the non-disclosed miRNA expressions. In order to evaluate the importance of such correlations, we first compute the Pearson's correlation coefficients and their corresponding p-values in all 99,235 pairs of the 446 expressed miRNAs in our fourth dataset. Filtering out all correlations with p-values greater than 0.001 (after Bonferroni correction for multiple correlations testing) or correlation coefficient smaller than 0.5 leaves us with 47% of miRNAs not independently expressed. Figure 4.7 shows the updated attack success by taking into account all significant correlations as defined above. In our experiments, we take a quite conservative approach: We assume that the adversary can perfectly infer the miRNAs correlated with those that are gradually disclosed. The dotted curve provides an upper bound estimate on the success rate. A tighter bound could be derived by knowing more precisely the probabilistic dependencies between miRNAs. This is left for future work.

For Figure 4.7, we make use of the three diseases of Figure 4.6 that gave the best trade-off between privacy and utility, i.e., glioma, multiple sclerosis and renal cancer. We observe that the success rate when the adversary knows miRNAs correlated with disclosed miRNAs is much higher than without them, except for the very first miRNAs in Figure 4.7(c). It shows that the most significant miRNAs for the SVM classification are co-expressed with others, which penalizes privacy significantly. Making use of the best subsets of miRNAs found above without correlations, containing 4 miRNAs for glioma, 7 for multiple sclerosis, and 10 for renal cancer, we evaluate the new privacy levels when miRNA correlations are taken into account. For glioma, instead of improving unlinkability by 80.8%, the 4 miRNAs and their correlated miRNAs yields an improvement in privacy of 34.6%. For renal cancers, the privacy enhancement drops from 69.2% to 38.5% and, for multiple sclerosis, using 7 miRNAs and their co-expressed miRNAs yield an attack success rate almost equal to the highest rate with the full set of miRNAs. However, we can find new, better trade-offs: e.g., disclosing 5 miRNAs for multiple sclerosis still provides the same high SVM accuracy (decrease of 0.9% compared to the baseline) while reducing the attack success by 23%. Note that we do not make use of the correlated miRNAs for the SVM algorithm as we are not certain about how they correlate with the disclosed ones.

4.7.3 Noise Mechanism

As we have noticed in the first protection mechanism, it is possible to hide the vast majority of miRNAs while retaining a fair level of prediction accuracy. This is typically very useful in the clinical setting where medical practitioners already know the miRNAs to test for predicting a specific disease. However, such a privacy-preserving mechanism could dramatically jeopardize miRNA utility for biomedical research. Indeed, as we have seen in our previous experiments, the majority of miRNAs need to be masked in order to gain a significant amount of unlinkability, which is not possible if researchers want to test for associations between miRNAs and diseases. Therefore, we additionally present and study a countermeasure where contributors of miRNA expressions directly apply random noise to their vectors of expression levels before providing them to the research community (possibly online), in a fully distributed manner (i.e., independent from other contributors).

The idea behind adding noise to the raw expression data is to provide indistinguishability between different expression vectors and consequently reduce the tracking capabilities of the adversary. Following the generalized notion of differential privacy [19] previously applied to location privacy [2], we state that a mechanism $\mathcal{A}$ achieves *epigeno-indistinguishability* if and only if, for all m-miRNA expression vectors $\mathbf{x}_1$, $\mathbf{x}_2$,

$$\Pr\left[\mathcal{A}(\mathbf{x}_1) \in S\right] \leq exp(\epsilon d_2(\mathbf{x}_1, \mathbf{x}_2)) \times \Pr\left[\mathcal{A}(\mathbf{x}_2) \in S\right],$$

where S is any subset of the set of possible responses and $d_2(\cdot, \cdot)$ denotes the Euclidean distance. In the following, we assume that the set of possible responses lies in the same m-dimensional real-valued space $\mathbb{R}^m$ as the set of original expression vectors. Before defining our mechanism $\mathcal{A}(\cdot)$ for achieving epigeno-indistinguishability, let us first give some intuition about the mechanism. The noise mechanism is such that the probability

of reporting a noisy expression vector $\mathcal{A}(\mathbf{x})$ differs by at most a factor $exp(\epsilon d_2(\mathbf{x}_1, \mathbf{x}_2))$ when the actual, non-obfuscated miRNA expression vectors are $\mathbf{x}_1$ and $\mathbf{x}_2$. This can be achieved by relying on the multivariate Laplacian mechanism that adds noise $\mathbf{y}$ according to the following probability density function $g(\mathbf{y}) = \frac{1}{\alpha} e^{-\epsilon \|\mathbf{y}\|_2}$, where α is a normalization factor ensuring that the integral over all $\mathbf{y} \in \mathbb{R}^m$ equals one.

Sampling noise from the distribution $g(\mathbf{y})$ can be carried out efficiently by generalizing the method used for the planar Laplacian mechanism in [2]. First, we sample the magnitude $\|\mathbf{y}\|_2$ of the noise from a gamma distribution with shape m and scale $1/\epsilon$. Second, we randomly generate the direction $\hat{\mathbf{y}} = \mathbf{y}/\|\mathbf{y}\|_2$ of the noise by uniformly sampling points on the surface $\mathbb{S}^{m-1}$ of a hypersphere [75]. To do so, we can generate m independent Gaussian random variables z_1, z_2, ..., z_m, and let $\hat{z}_i = z_i / \sqrt{z_1^2 + \cdots + z_m^2}$ for $i = 1, \ldots, m$. Then the distribution of the vector $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_m)$ is uniform over the surface $\mathbb{S}^{m-1}$, and thus we can set the direction $\hat{\mathbf{y}} := \hat{\mathbf{z}}$. Each person i contributing his miRNA expression profile $\mathbf{x}_i$ will then share, instead of the actual expression data, the noisy vector $\mathcal{A}(\mathbf{x}_i) = \mathbf{x}_i + \mathbf{y}$, where $\mathbf{y}$ is independently generated for all participants $i = 1, \ldots, n$.

Following this approach, in our evaluation, we first add noise to our dataset of 1000+ individuals (considering only the 446 miRNAs as before). Then, in the second step, we calculate the p-values on the noised data (since the researchers would be provided with exactly this data) and train the SVM as previously by subsequently adding miRNAs in the order of their p-values. Similarly, we evaluate the success of our attack on the athletes' dataset, when considering the same 446 miRNAs, but after adding noise. Moreover, we repeat both our experiments 50 times and average the results over all runs.

Figure 4.8 shows the evolution of the SVM accuracy and linkability (success of the attack), with respect to the amount of noise, tuned by ϵ, that is added to each contributor's miRNA expression profile. As privacy is measured on the same dataset for all six figures, it evolves in a very similar way. Even if the noise is randomly generated, the differences average out with the Monte Carlo method we use. We clearly see that with $\epsilon = 1$, there is almost no privacy gain compared to the attack without countermeasure, whereas for $\epsilon = 0.001$, the attack success drops by almost 90%. As for the first countermeasure, there is a utility-privacy trade-off to be found between these two extreme values.

In Figure 4.8(a), we can observe that, for pancreatitis, $\epsilon = 0.075$ is a good trade-off, with an accuracy decrease of only 0.8% and an unlinkability improvement of 40%. For glioma (Figure 4.8(b)), the best trade-off is certainly at $\epsilon = 0.05$, with an accuracy decrease of 1.2% and an unlinkability improvement of 51%. For multiple sclerosis, we reach the best trade-off at $\epsilon = 0.025$ with an accuracy decrease of 0.65% and an unlinkability improvement of 63%. For tumor of stomach, we can reach an accuracy decrease of only 0.2% and still improve the unlinkability by as much as 70% with $\epsilon = 0.01$. For renal cancer, we have to sacrifice a bit more of utility, 2.3%, for a privacy increase of 61%, with $\epsilon = 0.025$. The only disease for which it is quite difficult to get both satisfactory unlinkability and excellent accuracy is melanoma (Figure 4.8(e)). This is consistent with the hiding mechanism presented in Section 4.7.2, where we observed (in Figure 4.6(e)) a fast and sharp increase in the attack success.

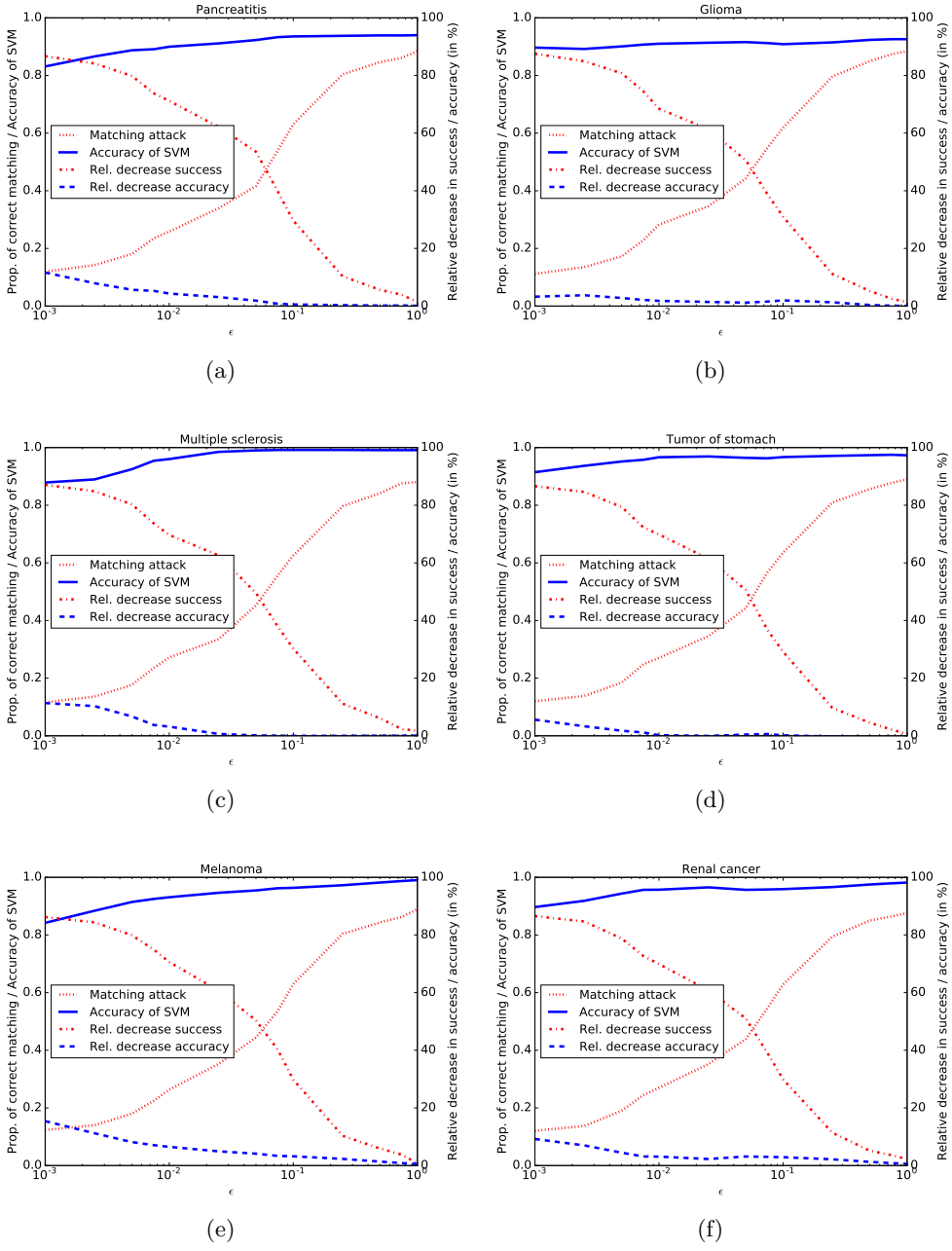


Figure 4.8: Evolution of privacy and utility (classifier accuracy) plotted against the noise (tuned by ϵ) added to the individual miRNA expression profiles, for the following diseases: (a) Pancreatitis, (b) Glioma, (c) Multiple sclerosis, (d) Tumor of stomach, (e) Melanoma, (f) Renal cancer.

4.7.4 Comparison of Protection Mechanisms

In order to compare both approaches, we decide upon a utility or a privacy requirement, fix it, and then evaluate the best privacy, respectively utility, achieved with both coun-

termeasures. We carry out this evaluation on all 19 diseases for different requirements of utility and privacy.

First, we start by fixing the utility, more precisely the relative accuracy decrease compared to the baseline accuracy. The privacy is measured in terms of the decrease in the success of the matching attack. For a given maximal decrease in accuracy $\Delta_{\mathsf{acc}}^{\mathsf{max}}$, we select the optimal number of miRNAs m^* and the optimal amount of noise ϵ^* that maximize the privacy increase $\Delta_{\mathsf{priv}}^{m}$ and $\Delta_{\mathsf{priv}}^{\epsilon}$. In case of the hiding mechanism, we select $m^* = \arg\max_m \Delta_{\mathsf{priv}}^{m}$ such that $\Delta_{\mathsf{acc}}^{m} < \Delta_{\mathsf{acc}}^{\mathsf{max}}$. In case of the noise mechanism, we select $\epsilon^* = \arg\max_\epsilon \Delta_{\mathsf{priv}}^{\epsilon}$ such that $\Delta_{\mathsf{acc}}^{\epsilon} < \Delta_{\mathsf{acc}}^{\mathsf{max}}$, respectively.

Considering $\Delta_{\mathsf{acc}}^{\mathsf{max}} \in \{0.5\%, 1\%, 2\%, 3\%, 4\%, 5\%\}$ for all 19 diseases, we mostly experience that the noise mechanism provides a better privacy improvement compared to hiding a subset of miRNAs (all results are listed in Table 4.2 and 4.3, Section 4.10). In particular, 90 out of 114 cases (combinations of disease and $\Delta_{\mathsf{acc}}^{\mathsf{max}}$) yield a better privacy with the noise mechanism. When examining a maximal decrease in accuracy of 2%, the hiding technique provides a better privacy for only 2 diseases, namely glioma and renal cancer. Interestingly, these two diseases stand out also for other values of the maximal accuracy decrease, providing better privacy with the hiding technique in 10 out of 12 cases. However, for all other diseases, adding noise in a distributed manner to individual expression profiles provides better utility for similar levels of privacy. For example, for lung cancer, we are able to achieve an increase in privacy of 79.3% while maintaining a decrease in accuracy of 0.8% using noise with $\epsilon = 0.005$. The best we can achieve for the hiding technique here is either a decrease in accuracy of 0.97% and an increase in privacy of only 46.2% or a larger decrease in accuracy of 1.9% and a privacy improvement of only 50%.

Next, we discuss the results for a fixed minimal improvement of the privacy and compare the corresponding minimal decrease in accuracy in both countermeasures. We now fix the minimal increase in privacy (i.e., the minimal decrease in the attack success) $\Delta_{\mathsf{priv}}^{\mathsf{min}}$ and minimize the decrease in accuracy: $\arg\min_m \Delta_{\mathsf{acc}}^{m}$ such that $\Delta_{\mathsf{priv}}^{m} > \Delta_{\mathsf{priv}}^{\mathsf{min}}$ and $\arg\min_\epsilon \Delta_{\mathsf{acc}}^{\epsilon}$ such that $\Delta_{\mathsf{priv}}^{\epsilon} > \Delta_{\mathsf{priv}}^{\mathsf{min}}$, respectively. We run experiments for values of $\Delta_{\mathsf{priv}}^{\mathsf{min}}$ from 10% up to 90%, in steps of 10% (all results are provided in Table 4.4 and 4.5, Section 4.10).

We again observe that, for most of the evaluated cases, the achieved accuracy is better when adding noise compared to when hiding miRNAs. In particular, this holds true for 143 out of 171 cases, clear exceptions being again glioma and renal cancer. For those two diseases, the hiding technique provides better accuracy than the noise mechanism in 87.5% of the cases. When fixing the minimal increase in privacy to 70%, only these two diseases provide better results with the hiding technique. For instance, with renal cancer, we achieve 60.8% improvement in privacy with a decrease in accuracy of 2.3% using noise with $\epsilon = 0.025$, whereas we can obtain an increase in privacy of 69.2% and a decrease in accuracy of only 1.7% when using the hiding technique. For the majority of diseases, however, it is clearly the noise mechanism that provides much higher utility. For example, for lung cancer, an increase in privacy of at least 70% is achievable with a decrease in accuracy of only 0.2% with the noise mechanism, while the hiding technique yields a decrease in accuracy of 11.2%.

In summary, we find that the noise mechanism presented in Section 4.7.3, providing

epigeno-indistinguishability, is able to achieve a better privacy-utility trade-off than the hiding mechanism for the vast majority of studied diseases (17 out of 19). We have also shown in Section 4.7.2 that the privacy improvement with the hiding mechanism could actually be too optimistic due to the correlations existing between miRNAs. This is another argument to favor the noise mechanism rather than the hiding technique. Moreover, the p-values used to rank the miRNAs in the hiding mechanism actually require that, at some point in time, some entity gets access to the full set of miRNAs of a significant number of individuals in order to measure these p-values. The noise mechanism is fully distributed and does not need to rely on a trusted entity at any point in time. Finally, it allows for more flexibility as it enables, e.g., the biomedical research community to access all miRNA expression levels of contributors.

4.8 Limitations

Unlike other fields of privacy research, where large amounts of data can be collected in a small amount of time for a low amount of money, here we face the exact opposite: measuring the miRNA expression levels for one single sample already costs several hundred dollars. This means that the total value of our datasets is of more than half a million dollars. Moreover, in order to gather longitudinal epigenetic data, real patients have to regularly provide their blood samples over a long period of time. For instance, in the case of our lung cancer dataset, patients' blood had to be collected every 3 months over a period of 18 months.

Although the number of participants in our three longitudinal datasets may look small at first sight, this number is currently certainly very substantial, given that these datasets must contain expression profiles of individuals at multiple points in time. Other datasets that include a much larger number of participants, such as our 1000+ dataset, do not fulfill the requirement of multiple time points. We counterbalance the relatively small number of participants in our longitudinal datasets by making use of different datasets that show very similar behavior (plasma-based miRNAs).

4.9 Conclusion

To the best of our knowledge, we are the very first to demonstrate that personal miRNA expression profiles can be successfully tracked over time. Our study sheds light on a widely overlooked problem, namely privacy risks stemming from epigenetic data, and brings this issue to the attention of both the biomedical and computer security research communities. In addition to the in-depth evaluation of the temporal linkability of miRNA expression profiles, we propose two defense mechanisms based on well-established privacy-enhancing methods: (1) hiding a subset of the expression data, and (2) adding noise to the released expression profiles. We thoroughly evaluate the impact of these countermeasures on biomedical utility by studying how much accuracy decrease they induce in a typical machine-learning algorithm for predicting diseases. We observe that, for the majority of the 19 diseases studied in our experiments, the noise mechanism provides a better privacy-utility trade-off than the hiding method. Moreover, we highlight that the noise mechanism can be applied directly by the data

$\Delta_{\text{acc}}^{\text{max}}$	0.5%		1.0%		2.0%	
Disease	Δ_{priv}^{m}	$\Delta_{\text{priv}}^{\epsilon}$	Δ_{priv}^{m}	$\Delta_{\text{priv}}^{\epsilon}$	Δ_{priv}^{m}	$\Delta_{\text{priv}}^{\epsilon}$
Periodontitis	26.9%	74.1%	26.9%	79.2%	50.0%	79.2%
Renal cancer	30.8%	-	30.8%	3.6%	69.2%	5.2%
Wilms tumor	3.8%	6.4%	7.7%	9.5%	7.7%	40.1%
Benign prostate hyperplasia	−3.8%	10.6%	3.8%	70.5%	11.5%	72.2%
Chronic obstructive pulmonary disease (COPD)	0.0%	2.7%	0.0%	5.5%	0.0%	12.5%
Colon cancer	19.2%	11.4%	30.8%	30.5%	30.8%	60.2%
Ductal adenocarcinoma	0.0%	50.6%	3.8%	50.6%	7.7%	62.5%
Glioma	65.4%	5.2%	65.4%	5.2%	80.8%	68.5%
Lung cancer	11.5%	74.1%	46.2%	79.3%	50.0%	79.3%
Melanoma	0.0%	-	0.0%	3.8%	0.0%	5.9%
Multiple sclerosis	19.2%	49.5%	53.8%	62.6%	53.8%	62.6%
Myocardial infarction	3.8%	52.4%	3.8%	52.4%	3.8%	60.5%
Non-ischaemic systolic heart failure	−3.8%	80.0%	0.0%	80.0%	46.2%	80.0%
Ovarian cancer	26.9%	78.5%	26.9%	78.5%	42.3%	78.5%
Pancreatitis	19.2%	10.3%	26.9%	39.8%	26.9%	53.5%
Prostate cancer	−3.8%	-	−3.8%	-	3.8%	4.0%
Psoriasis	0.0%	6.5%	0.0%	31.4%	3.8%	74.0%
Sarcoidosis	0.0%	69.3%	3.8%	74.0%	50.0%	79.8%
Tumor of stomach	15.4%	69.8%	34.6%	69.8%	65.4%	79.4%

Table 4.2: Relative increase in privacy for both defense mechanisms in relation to a fixed maximal decrease in accuracy ranging from 0.5% to 2.0%. For "-" cells, we did not achieve the targeted accuracy with any tested ϵ. A negative value means that the attack success rate could even exceed the success rate incorporating all miRNAs.

contributors, independently of other contributors, and provides more flexibility for the biomedical community. Our work demonstrates that achieving indistinguishability by adding noise is a promising technique that could be applied to other types of biomedical data in the future.

Our results provide enough evidence about the extent of the threat to remove miRNA expression data from publicly accessible databases. Due to the limited number of individuals present in our datasets, we could not rely on supervised learning algorithms, which would undoubtedly further improve the tracking capabilities of the adversary.

4.10 Additional Tables

Tables 4.2, 4.3, 4.4, and 4.5 contain the detailed results of our comparison of protection mechanisms.

$\Delta_{\text{acc}}^{\text{max}}$	3.0%		4.0%		5.0%	
Disease	Δ_{priv}^{m}	$\Delta_{\text{priv}}^{\epsilon}$	Δ_{priv}^{m}	$\Delta_{\text{priv}}^{\epsilon}$	Δ_{priv}^{m}	$\Delta_{\text{priv}}^{\epsilon}$
Periodontitis	88.5%	83.6%	88.5%	83.6%	88.5%	83.6%
Renal cancer	73.1%	60.8%	73.1%	72.7%	80.8%	78.8%
Wilms tumor	7.7%	61.5%	7.7%	70.4%	11.5%	74.3%
Benign prostate hyperplasia	46.2%	79.2%	57.7%	79.2%	65.4%	79.2%
Chronic obstructive pulmonary disease (COPD)	0.0%	12.5%	15.4%	50.3%	23.1%	69.8%
Colon cancer	57.7%	60.2%	73.1%	70.5%	73.1%	73.8%
Ductal adenocarcinoma	42.3%	62.5%	50.0%	69.5%	50.0%	74.2%
Glioma	80.8%	80.8%	80.8%	87.5%	80.8%	87.5%
Lung cancer	50.0%	79.3%	50.0%	79.3%	50.0%	79.3%
Melanoma	3.8%	10.3%	38.5%	40.2%	38.5%	60.7%
Multiple sclerosis	61.5%	62.6%	61.5%	73.7%	61.5%	73.7%
Myocardial infarction	38.5%	60.5%	42.3%	74.6%	42.3%	74.6%
Non-ischaemic systolic heart failure	46.2%	84.7%	46.2%	84.7%	46.2%	84.7%
Ovarian cancer	42.3%	84.3%	50.0%	84.3%	50.0%	86.2%
Pancreatitis	26.9%	53.5%	57.7%	62.2%	65.4%	71.2%
Prostate cancer	42.3%	6.2%	42.3%	10.1%	42.3%	38.5%
Psoriasis	19.2%	80.1%	23.1%	80.1%	61.5%	80.1%
Sarcoidosis	92.3%	79.8%	92.3%	79.8%	92.3%	79.8%
Tumor of stomach	65.4%	79.4%	65.4%	84.6%	65.4%	84.6%

Table 4.3: Relative increase in privacy for both defense mechanisms in relation to a fixed maximal decrease in accuracy ranging from 3.0% to 5.0%. A negative value means that the attack success rate could even exceed the success rate incorporating all miRNAs.

$\Delta_{\text{priv}}^{\text{min}}$	30.0%		40.0%		50.0%	
Disease	Δ_{acc}^{m}	$\Delta_{\text{acc}}^{\epsilon}$	Δ_{acc}^{m}	$\Delta_{\text{acc}}^{\epsilon}$	Δ_{acc}^{m}	$\Delta_{\text{acc}}^{\epsilon}$
Periodontitis	1.9%	−1.9%	1.9%	−1.9%	2.6%	−1.9%
Renal cancer	0.0%	2.3%	1.7%	2.3%	1.7%	2.3%
Wilms tumor	5.2%	1.4%	5.2%	1.7%	5.5%	2.2%
Benign prostate hyperplasia	2.7%	0.5%	2.7%	0.6%	3.5%	0.6%
Chronic obstructive pulmonary disease (COPD)	7.9%	3.3%	12.0%	3.3%	12.0%	3.3%
Colon cancer	0.7%	0.8%	2.4%	1.3%	2.4%	1.3%
Ductal adenocarcinoma	2.8%	0.1%	2.8%	0.4%	5.2%	0.4%
Glioma	0.0%	1.2%	0.0%	1.2%	0.4%	1.2%
Lung cancer	0.7%	−1.5%	1.0%	−1.5%	6.6%	−1.5%
Melanoma	3.7%	3.4%	5.0%	3.4%	7.5%	4.1%
Multiple sclerosis	0.9%	−0.0%	0.9%	0.1%	0.9%	0.7%
Myocardial infarction	2.8%	0.0%	3.6%	0.4%	7.2%	0.4%
Non-ischaemic systolic heart failure	2.0%	−2.6%	2.0%	−2.6%	8.5%	−2.1%
Ovarian cancer	1.3%	−0.7%	1.3%	−0.7%	5.5%	−0.7%
Pancreatitis	3.8%	0.8%	3.8%	1.9%	3.8%	1.9%
Prostate cancer	2.7%	4.8%	2.7%	5.0%	7.6%	5.0%
Psoriasis	4.3%	1.0%	4.3%	1.3%	4.3%	1.3%
Sarcoidosis	1.4%	−0.2%	1.6%	−0.2%	2.2%	−0.2%
Tumor of stomach	0.9%	−0.0%	1.7%	−0.0%	2.0%	−0.0%

Table 4.4: Relative decrease in accuracy for both defense mechanisms in relation to a fixed minimal increase in privacy ranging from 30% to 50%. A negative value means that the accuracy could, in this case, even exceed the baseline accuracy (utility).

$\Delta_{\text{priv}}^{\text{min}}$	60.0%		70.0%		80.0%	
Disease	Δ_{acc}^{m}	$\Delta_{\text{acc}}^{\epsilon}$	Δ_{acc}^{m}	$\Delta_{\text{acc}}^{\epsilon}$	Δ_{acc}^{m}	$\Delta_{\text{acc}}^{\epsilon}$
Periodontitis	2.6%	−1.9%	2.6%	−0.8%	2.6%	2.9%
Renal cancer	1.7%	2.3%	2.5%	3.1%	4.8%	7.0%
Wilms tumor	5.5%	2.8%	8.1%	3.2%	15.5%	11.3%
Benign prostate hyperplasia	5.0%	0.9%	5.6%	0.9%	5.6%	5.5%
Chronic obstructive pulmonary disease (COPD)	15.4%	4.1%	15.6%	5.3%	15.6%	9.0%
Colon cancer	3.3%	1.9%	3.9%	3.3%	7.7%	9.4%
Ductal adenocarcinoma	5.2%	1.8%	6.4%	4.6%	6.4%	6.4%
Glioma	0.4%	1.4%	1.1%	2.1%	1.1%	2.8%
Lung cancer	8.1%	−1.5%	11.2%	0.2%	18.2%	5.5%
Melanoma	7.5%	4.9%	7.5%	6.5%	10.0%	11.2%
Multiple sclerosis	2.3%	0.7%	8.1%	3.8%	8.1%	6.7%
Myocardial infarction	7.3%	1.3%	7.3%	3.3%	11.2%	6.7%
Non-ischaemic systolic heart failure	8.5%	−2.1%	9.3%	−1.5%	9.3%	2.5%
Ovarian cancer	6.7%	−0.7%	9.0%	−0.7%	9.0%	2.5%
Pancreatitis	4.5%	3.1%	7.9%	4.3%	7.9%	7.9%
Prostate cancer	7.6%	5.6%	7.6%	5.6%	11.5%	8.9%
Psoriasis	4.3%	1.4%	5.8%	1.4%	10.0%	2.1%
Sarcoidosis	2.2%	−0.2%	2.2%	0.6%	2.2%	5.3%
Tumor of stomach	2.0%	−0.0%	5.1%	1.1%	5.1%	3.3%

Table 4.5: Relative decrease in accuracy for both defense mechanisms in relation to a fixed minimal increase in privacy ranging from 60% to 80%. A negative value means that the accuracy could, in this case, even exceed the baseline accuracy (utility).

5

Membership Privacy for miRNA Expression Profiles

Considering Membership Privacy in miRNA-based Studies

5.1 Motivation

During the last decades, we have experienced significant achievements in the biomedical field. A necessary condition for such a scientific breakthrough is the availability of large amounts of biological data. However, this availability imposes severe privacy risks for individuals who contribute their biological samples towards improving medicine.

One of the first attacks showing the extent of this threat was proposed by Homer et al. back in 2008 [59]. Specifically, the authors demonstrated that, given (some parts of) the genomic data of an individual and summary statistics of a genome-wide association study (GWAS [45]), it is possible to determine whether this individual participated in the GWAS. Such a *membership attack* can have disastrous privacy implications if the individual happens to be part of the case group (e.g., carrying a sensitive disease). This first attack led to substantial follow-up work aiming to identify the theoretical bounds on the attack success more precisely and to propose defense mechanisms for countering it.

While some biomedical studies publish their whole dataset in publicly available databases online, other studies only release summary statistics of their dataset. In the latter case, temporal linkability attacks such as presented in Chapter 4 are not possible anymore. However, any kind of study publishing summary statistics might still be susceptible to membership attacks.

In this chapter, we will again focus on microRNA expression profiles. Even though biomedical research on miRNAs is far from complete, studies of miRNA expression profiles have already shown that dysregulation of miRNA is linked to neurodegenerative diseases, heart disease, diabetes, and the majority of cancers [84, 127, 70, 102, 37]. Therefore, miRNA expression profiling promises to allow for a more accurate and minimally invasive diagnosis of major severe diseases. On the downside, this also implies that miRNA expressions can tell us much more about whether someone is affected by a disease at a given point in time than the genome, which only informs about the *risk* of getting certain diseases.[1]

5.2 Contributions

In this chapter, we first study whether, and to what extent, membership inference can be successfully carried out against datasets of biomedical data other than the genome. To this end, we focus on miRNA expression profiles for our study. Notable challenges we needed to overcome are that miRNA expressions are real-valued rather than discrete, but of several orders of magnitude lower in dimension and noisier than genomic data. Indeed, whereas a genome typically contains tens of millions of single nucleotide polymorphisms (SNPs), there are currently only around five thousand identified miRNAs.

We present two attacks, one based on the L_1 distance, as proposed by Homer et al. in their seminal work, and another based on the likelihood-ratio (LR) test, which is optimal, in the sense that it achieves maximum attack true-positive rate at a given false-positive level. For the latter attack, we also derive the theoretical relation between

[1] The only exceptions are Mendelian disorders, such as cystic fibrosis, which are primarily determined by our genes.

true-positive rate, false-positive rate, number of miRNAs, and number of individuals in the dataset. This relation is especially valuable as it is independent of the actual individual miRNA expression values and any population-wide statistics.

Our experimental results demonstrate that, in general, the L_1 distance attack performs a bit worse than the LR attack, as expected, and that the LR theoretical relation provides bounds that are slightly lower than the power of the empirical LR test (i.e., the LR attack with actual miRNA expression data). Finally, we show that the membership inference attack is a lot more successful against datasets composed of participants carrying a specific disease than against randomly generated datasets. This is essentially due to the fact that miRNA expressions are profoundly affected by the health status of their owner, much more than genomic data. The latter result tells us that the theoretical relation on the LR test has to be taken very cautiously regarding the privacy levels it provides to miRNA-based studies in practice.

Second, given the extent of the threat to membership privacy, we propose and evaluate both a perturbative, differentially private mechanism and a hiding mechanism for countering the membership attack. More precisely, we first study two variants of the perturbative algorithm assuming different prior knowledge of the attacker. We show that, in our context, it does not make a substantial difference to the membership of a victim whether an attacker assumes bounded or unbounded priors. Then, we evaluate the impact of both protection mechanisms (perturbative and hiding) on mitigating the success of the attacks. For the perturbative noise mechanism, we also thoroughly study the evolution of noise and its impact on utility, as it can lead to a prohibitive loss for research and medical utility. One critical observation is that the differentially private mechanism is able to reduce the attack power to nearly random guessing, whereas the hiding method is not. Moreover, the attack is in general very robust to hiding miRNA means. Finally, we notice that the attack and differentially private mechanism are influenced mostly by the number of individuals in the dataset. Based on our analytical and experimental results, given the current number of miRNAs, we recommend to only release summary statistics of datasets including at least a couple of hundred individuals.

5.3 Threat Model

The adversary's goal is to determine whether a specific person (referred to as *victim*) is a member of a group of study, that we will refer to as a *pool*.

First, we assume the adversary has access to the exact miRNA expression profile $\mathbf{x}_v \in \mathbb{R}^m$ of the victim v. Such data can be easily extracted from a blood sample of the victim, for a few hundred dollars (and the cost will undoubtedly decrease over time). Full individual miRNA expression data are also increasingly available in public research databases, such as the Gene Expression Omnibus (GEO) [44] or ArrayExpress [3] databases. Furthermore, this data could be collected by hacking a healthcare provider server, e.g., a hospital server. Indeed, healthcare companies are facing an increasing number of cyber attacks [89] such as the Anthem's breach, in which the medical records of around 80 million patients were leaked [56].

Also, note that we will assume that the victim's profile to which the adversary has access and the profile the victim contributed to the pool were collected at the same time.

Although miRNA expressions can vary in time, we have shown previously that miRNA expression profiles can be efficiently linked over time frames of up to one year [P1].

Second, we assume the adversary has access to some summary statistics released for the pool. Formally, the pool is defined as a set $\mathcal{P} \in \mathbb{R}^{n \times m}$ containing the miRNA expression profiles of n entities gathered from an underlying population $\mathcal{U}$, where each profile is a vector of m real values representing the expression of every miRNA. Such pools of individuals are typically used by biomedical researchers in order to infer associations between miRNAs and diseases. If significant associations exist, the researchers publish their results in articles (typically available online) along with summary statistics about their pool, such as mean values of miRNA expressions. In this work, we assume mean statistics are available to the adversary, but other statistics could also be accessed, even further increasing the adversary's power.

Finally, we assume the adversary also has access to general miRNA expression statistics of the underlying population $\mathcal{U}$, the so-called reference population. Currently, these statistics have to be estimated by the adversary using a subset of $\mathcal{U}$, but we expect that population-wide statistics will soon become publicly available, as it is the case for genomic data.

5.4 Differential and Membership Privacy

In this work, beyond presenting attacks against membership privacy in miRNA-based studies, we also propose countermeasures, notably relying on differential privacy [29]. We review here the definitions and results concerning differential privacy and positive membership privacy relevant to this work.

Definition 1 (Differential Privacy [29])**.** *A mechanism $\mathcal{A}$ provides ϵ-differential privacy if and only if for any two datasets H_1 and H_2 differing in one element, and any $S \subseteq \mathsf{range}(\mathcal{A})$, it holds that*

$$\Pr\left[\mathcal{A}(H_1) \in S\right] \leq e^{\epsilon} \cdot \Pr\left[\mathcal{A}(H_2) \in S\right]$$

In this work, we will also discuss a relaxed version of differential privacy, membership privacy, that ideally allows for smaller utility loss and at the same time satisfactory privacy guarantees under relaxed adversarial assumptions. Positive membership privacy, proposed by Li et al. [80], potentially allows to bound the change in the adversary's belief regarding an entity's membership in a database after observing some statistics of the database.

Definition 2 (Positive Membership Privacy [80])**.** *A mechanism $\mathcal{A}$ provides $(\gamma, \mathbb{D})$-positive membership privacy (PMP) under a distribution family $\mathbb{D}$, where $\gamma \geq 1$ if and only if for any $S \subseteq \mathsf{range}(\mathcal{A})$, any distribution $D \in \mathbb{D}$ and any entity $u \in \mathcal{U}$, it holds that*

$$\Pr_{D,\mathcal{A}}\left[u \in H \mid \mathcal{A}(H) \in S\right] \leq \gamma \cdot \Pr_{D}\left[u \in H\right] \tag{5.1}$$

$$\Pr_{D,\mathcal{A}}\left[u \notin H \mid \mathcal{A}(H) \in S\right] \geq \frac{1}{\gamma} \cdot \Pr_{D}\left[u \notin H\right] \tag{5.2}$$

In general, $(e^\epsilon, \mathbb{D})$-membership privacy and ϵ-differential privacy are equivalent for arbitrary distribution families $\mathbb{D}$ and thus require the same amount of noise. However, the required amount of noise can be reduced by restricting the distribution families, assuming prior bounds on the probability of membership. In particular, if the membership probability p_u of an entity u to a database is restricted to $p_u \in [a, b] \cup \{0, 1\}$, for $0 < a \leq b < 1$, then achieving weaker differential privacy is sufficient to achieve (positive) membership privacy, as shown by Tramèr et al. [119].

Theorem 1 (Tramèr et al. [119])**.** *A mechanism $\mathcal{A}$ provides $(\gamma, \mathbb{D}_B^{[a,b]})$-PMP for some $0 < a \leq b < 1$, if $\mathcal{A}$ satisfies ϵ-differential privacy for*

$$e^\epsilon = \begin{cases} \min(\frac{(1-a)\gamma}{1-a\gamma}, \frac{\gamma+b-1}{b}) \text{ if } a\gamma < 1, \\ \frac{\gamma+b-1}{b} \text{ otherwise.} \end{cases} \tag{5.3}$$

5.5 Membership Inference Attack

In this section, we first introduce the two test statistics used in our attack, one that is based on the approach proposed by Homer et al. [59] and another that relies on the likelihood ratio test. Then, we evaluate both approaches using a real dataset containing more than 1,000 miRNA expression profiles [73] and compare their performance.

5.5.1 Analytical Results

The mean of miRNA expression values is one of the most frequently released summary statistics in miRNA-based studies. Indeed, for studies which aim to discover associations between dysregulated miRNAs and diseases, it is crucial to disclose the mean of miRNA expression values over all case samples (individuals carrying the disease of interest to the study) and, separately, over all control samples. Another statistic used for the same purpose is the p-value of the t-test. We show, in the following, that, in many cases, the average values of miRNAs are already sufficient to identify participation of a victim in a miRNA-based pool.

The expression value of the miRNA j of the individual i is denoted by $x_i^j \in \mathbb{R}$. $\mathbf{x}_i \in \mathbb{R}^m$ is the vector of all miRNA expression values, also called the miRNA expression profile, of the individual i. Further, μ_j denotes the average expression value of miRNA j in the reference population, while $\hat{\mu}_j$ denotes the average of miRNA j's expression value in the pool.

5.5.1.1 L_1 Distances Difference

In order to determine whether a victim v is part of the pool, extending Homer et al.'s idea to real-valued miRNA expression profiles, one can simply compare the distances between (1) x_v^j and μ_j, and (2) x_v^j and $\hat{\mu}_j$. By computing the difference between these distances we obtain the following statistic:

$$D(x_v^j) = |x_v^j - \mu_j| - |x_v^j - \hat{\mu}_j| \tag{5.4}$$

Under the null hypothesis, if x_v^j is not part of the pool, $D(x_v^j)$ should approach zero. Under the alternative hypothesis, where x_v^j is a member of the pool, it should be greater than zero because the victim's contribution x_v^j to $\hat{\mu}_j$ will shift $\hat{\mu}_j$ away from μ_j. When $D(x_v^j)$ is negative, x_v^j is further away from the pool than from the reference population, and thus even less likely to be part of the pool.

Following from the central limit theorem, if the number of miRNAs is sufficiently high, the sum of $D(x_v^j)$ over all miRNAs j will converge to the normal distribution. Hence, we use the one-sample t-test to determine whether the person of interest v is part of the pool: If the test is strictly greater than a threshold, we assume v is part of the pool and, otherwise, that v is not in the pool.

5.5.1.2 Likelihood-Ratio Test

Although the aforementioned test can be very accurate, there is no known theoretical guarantee on the power[2] of detection it can achieve. Thus, it is possible that another approach could provide better attack power. We therefore also propose and evaluate a test statistic based on the likelihood-ratio test (LR test).

This method has the non-negligible advantage of attaining the maximum achievable power for a given false-positive level, thus providing a theoretical limit on the maximum detection power of the adversary, according to the Neyman-Pearson lemma. This lemma states that the exact LR test achieves the maximum power at a given false-positive level in binary hypothesis testing [92]. Furthermore, in the context of genomic privacy, the LR test has been empirically shown to be more powerful than Homer et al.'s attack, especially for small false-positive levels [106]. Before deriving the exact likelihood-ratio statistic for miRNA expression profiles, we have to impose some assumptions on their characteristics.

First, we assume that miRNAs are independent[3] and that the expression value of each miRNA j is distributed according to a normal distribution (with different parameters for the reference population and the pool). Note that the normal distribution is the distribution that best fits the distributions observed from our miRNA expression dataset. For the reference population, we denote the mean by μ_j and the standard deviation by σ_j. For the pool, we denote them by $\hat{\mu}_j$ and $\hat{\sigma}_j$ respectively. Note that a deviation from the Neyman-Pearson lemma might occur if, for example, the miRNAs are only approximately normally distributed.

Under the null hypothesis that the victim is not part of the pool, this victim's miRNA expressions are drawn from the reference population as defined above, i.e., each miRNA expression j of individual v is drawn with the probability density:

$$f(x_v^j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{x_v^j - \mu_j}{2\sigma_j^2}} \tag{5.5}$$

Similarly, under the alternative hypothesis, following a similar reasoning as in the theoretical analysis of [106], we consider the miRNA expressions of the victim to be

[2] Power refers to the true-positive rate, also called sensitivity.

[3] We make this assumption for tractability reasons, noting that about 60% of miRNAs are independent. Moreover, such assumption leads us to an upper bound on the adversary's power in inferring membership of the victim.

drawn according to the probability distribution of the pool:

$$\hat{f}(x_v^j) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_j} e^{-\frac{x_v^j - \hat{\mu}_j}{2\hat{\sigma}_j^2}} \tag{5.6}$$

We can then derive the following likelihood ratio between the alternative and the null hypotheses:

$$LR = \frac{\sigma}{\hat{\sigma}} e^{\frac{x_v^j - \mu_j}{2\sigma_j^2} - \frac{x_v^j - \hat{\mu}_j}{2\hat{\sigma}_j^2}} \tag{5.7}$$

Hence, the log-likelihood ratio over all miRNAs can then be written as:

$$LLR = \sum_{j=1}^{m} \frac{(x_v^j - \mu_j)^2}{2\sigma_j^2} - \frac{(x_v^j - \hat{\mu}_j)^2}{2\hat{\sigma}_j^2} + \log \frac{\sigma_j}{\hat{\sigma}_j} \tag{5.8}$$

If the adversary has access to the average values $\hat{\mu}_j$ of miRNA expressions in the pool, as assumed in this work, he still has to derive μ_j, σ_j, and $\hat{\sigma}_j$. The reference population's parameters μ_j and σ_j can be approximated by relying on publicly available datasets of miRNA expression levels. In Section 5.5.3, we approximate these parameters with our dataset of miRNA expressions. Finally, the adversary still needs to estimate $\hat{\sigma}_j$. For large n, the standard deviation should be very close to the standard deviation in the reference population because participants in the pool are supposed to come from the same reference population. Hence, $\hat{\sigma}_j \approx \sigma_j$ is the best approximation the adversary can make about $\hat{\sigma}_j$. In our evaluation, we will compute both the LR with the exact standard deviation $\hat{\sigma}_j$ and with $\hat{\sigma}_j = \sigma_j$, and compare the outcomes.

We now present the theoretical approximation on the maximum achievable power given the false-positive rate, the number of miRNAs, and the number of individuals in the pool.

Theorem 2. *Assuming $\forall j : \sigma_j \approx \hat{\sigma}_j$, the relation between the power β, the false-positive rate α, the number of miRNAs m, and the number of individuals n in the pool is*

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{2m}{n^2}}, \tag{5.9}$$

where z_x is the 100(1-x)th percentile of the standard normal distribution.

Proof of Theorem 2. First of all, we need to compute the statistics of the LLR defined in (5.8) under the null and the alternative hypotheses. Focusing on a single miRNA j's expression (i.e., one term of the LLR sum), we have the following mean $\mu_{j,0}$ under the

null hypothesis:

$$\mu_{j,0} := E[LLR_j \mid H_0] = \frac{1}{2\sigma_j^2}\int_{-\infty}^{\infty}(x_v^j - \mu_j)^2 f(x_v^j)dx_v^j \tag{5.10}$$

$$-\frac{1}{2\hat{\sigma}_j^2}\int_{-\infty}^{\infty}(x_v^j - \hat{\mu}_j)^2 f(x_v^j)dx_v^j + \log\frac{\sigma_j}{\hat{\sigma}_j}\int_{-\infty}^{\infty} f(x_v^j)dx_v^j \tag{5.11}$$

$$= \frac{1}{2} - \frac{1}{2\hat{\sigma}_j^2}\int_{-\infty}^{\infty}(x_v^j - \hat{\mu}_j)^2 f(x_v^j)dx_v^j + \log\frac{\sigma_j}{\hat{\sigma}_j} \tag{5.12}$$

$$= \frac{1}{2} - \frac{1}{2\hat{\sigma}_j^2}\int_{-\infty}^{\infty}(x_v^j - \mu_j - \frac{x_v^j - \mu_j}{n})^2 f(x_v^j)dx_v^j + \log\frac{\sigma_j}{\hat{\sigma}_j} \tag{5.13}$$

$$= \frac{1}{2} - \frac{\sigma_j^2}{2\hat{\sigma}_j^2} + \frac{\sigma_j^2}{n\hat{\sigma}_j^2} - \frac{\sigma_j^2}{2n^2\hat{\sigma}_j^2} + \log\frac{\sigma_j}{\hat{\sigma}_j}. \tag{5.14}$$

From (5.12) to (5.13), we assume that the pool is constituted of the victim v and $n-1$ individuals drawn as under the null hypothesis, i.e., $\hat{\mu}_j = \frac{(n-1)\mu_j + x_v^j}{n}$. Using our assumption $\forall j : \hat{\sigma}_j = \sigma_j$, we obtain

$$\mu_{j,0} = \frac{1}{n} - \frac{1}{2n^2} = \frac{2n-1}{2n^2}. \tag{5.15}$$

Following the same reasoning, replacing μ_j by $\frac{n\hat{\mu}_j - x_v^j}{n-1}$, we get the following mean under the alternative hypothesis:

$$\mu_{j,1} := E[LLR_j \mid H_1] = \frac{2n-1}{2(n-1)^2} \tag{5.16}$$

The variances of the LLR under the null and the alternative hypotheses are equal to:

$$\sigma_{j,k}^2 := E[LLR_j^2 \mid H_k] - \mu_{j,k}^2, k \in \{0,1\} \tag{5.17}$$

$E[LLR_j^2 \mid H_k]$ can be derived similarly to the means, by using the central moments ($E[(X - E(X))^c]$) of the normal distribution up to order $c = 4$. We obtain the following standard deviations:

$$\sigma_{j,0} = \frac{2n-1}{\sqrt{2}n^2}, \tag{5.18}$$

$$\sigma_{j,1} = \frac{2n-1}{\sqrt{2}(n-1)^2} \tag{5.19}$$

Note that the mean and variance statistics do not depend on miRNA j's values. Then, for moderately large m, it is known that the exact LLR statistics are approximately Gaussian, which allows us to use the relationship $m\mu_{j,0} + z_\alpha\sqrt{m}\sigma_{j,0} = m\mu_{j,1} - z_{1-\beta}\sqrt{m}\sigma_{j,1}$, where z_α and $z_{1-\beta}$ are the quantiles of level $1-\alpha$ and β of the normal distribution. Thus, we

obtain the following relations:

$$\sigma_{j,0} z_\alpha + \sigma_{j,0} z_{1-\beta} = \sqrt{m}(\mu_{j,1} - \mu_{j,0}) \tag{5.20}$$

$$\frac{1}{\sqrt{2}n^2} z_\alpha + \frac{1}{\sqrt{2}(n-1)^2} z_{1-\beta} = \sqrt{m}\left(\frac{1}{2(n-1)^2} - \frac{1}{2n^2}\right) \tag{5.21}$$

$$(n-1)^2 z_\alpha + n^2 z_{1-\beta} = \sqrt{2m}(n - \frac{1}{2}) \tag{5.22}$$

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{2m}{n^2}} \tag{5.23}$$

□

The theoretical relation does not depend on the average values μ_j, $\hat{\mu}_j$ of the miRNA expressions, nor does it make any assumptions about their values. It only requires m to be relatively large. Theorem 2 shows us that, for a successful attack, the number of exposed miRNAs m has to scale with the square of the number of participants in the study (n^2), which is better from a privacy point of view than with genomic data where it has to scale linearly with n [106]. Nevertheless, this *does not imply* that participants in miRNA-based studies are fully protected against membership inference attacks: First, as we will see in our dataset, the number of participants in pools can be lower than 20 in current practice. Second, biomedical researchers continuously keep discovering new miRNAs and, thereby, implicitly increase the number m of available statistics [94]. Finally, real case groups can have expression means that are further away from reference population means than what we assume in our theoretical analysis. This can be explained by the fact that miRNA expressions are profoundly affected by diseases.

5.5.2 Dataset Description

The dataset was first presented and used by Keller et al. in [73] and is publicly available in the gene expression omnibus (GEO) database under reference GSE61741. It is the same dataset as the one we used in Chapter 4 to evaluate the accuracy of a diagnosis in the presence of our privacy-preserving techniques. It contains the miRNA expression profiles of 1,049 individuals and, hence, can be considered a very rich dataset in the biomedical field. Every profile contains a set of 848 miRNA expressions. 94 of the 1,049 individuals are healthy people whereas the others are affected by one out of 19 diseases: 124 people have Wilms tumor (D1), 73 lung cancer (D2), 65 prostate cancer (D3), 62 myocardial infarction (D4), 47 chronic obstructive pulmonary disease (COPD) (D5), 45 sarcoidosis (D6), 45 ductal adenocarcinoma (D7), 43 psoriasis (D8), 37 pancreatitis (D9), 35 benign prostate hyperplasia (D10), 35 melanoma (D11), 33 non-ischaemic systolic heart failure (D12), 29 colon cancer (D13), 24 ovarian cancer (D14), 23 multiple sclerosis (D15), 20 glioma (D16), 20 renal cancer (D17), 18 periodontitis (D18), and 13 stomach tumor (D19).

Before running our experiments, we filter out non-expressed miRNAs, i.e., those with a median level of expressions over all individuals smaller than 50, which leaves us with 466 expressed miRNAs. This preprocessing phase is standard in the biomedical research field.

5.5.3 Experimental Results

We evaluate our attacks on the aforementioned dataset in two different settings: (1) we randomly pick a varying number n of individuals from the dataset to form a pool, and (2) we consider every case group (carrying a disease) described previously as a pool. The reference population is estimated using the entire dataset, i.e., all 1,049 individuals.

While the first setting allows us to evaluate the attack success independent of any effects that might be caused by diseases, the second setting is actually more realistic. Indeed, biomedical publications usually include the mean values of cases carrying specific diseases.

We evaluate each attack on aforementioned pools, using each of the 1,049 individuals as a potential victim. Given an attack and a pool, we obtain a test statistic T_v for every victim v. We then say v is more likely to be part of the pool than to be part of the reference population if the test statistic is higher than a given threshold th, i.e., $T_v > th$. Depending on whether v is part of the pool or not, we classify the result as true-positive (v is part of the pool and $T_v > th$), false-positive (v is not part of the pool and $T_v > th$), true-negative (v is not part of the pool and $T_v \leq th$) or false-negative (v is part of the pool and $T_v \leq th$). These metrics are then used to compute the true-positive and false-positive rates for varying thresholds.

Random Pools. In the first setting, we randomly select 50 subsets of n different individuals among the 1,049 in our dataset and average the results.

All figures in this section will depict the receiver operating characteristic (ROC) curves that compare the false-positive rate, on the x-axis, with the power of the attack, on the y-axis. We show four different ROC curves for (1) the attack based on the difference of the L_1 distances, (2) the likelihood-ratio attack knowing all the population and pool statistical parameters, i.e., $\mu, \hat{\mu}, \sigma, \hat{\sigma}$ (referred to as LR exact), (3) the LR attack not knowing $\hat{\sigma}$, and approximating it as $\hat{\sigma} \approx \sigma$ (corresponding to our assumed threat model), and (4) the theoretical LR relation derived in Theorem 2 also assuming $\hat{\sigma} \approx \sigma$. The figures are shown with a logarithmic x-axis, representing the false-positive rate in the range $[10^{-3}, 1]$.

In Figure 5.1, we depict three diagrams of randomly constructed pools for $n \in \{35, 65, 124\}$. We select these numbers because they are representative for our dataset and also correspond to the numbers of cases of three disease-specific groups shown in Figure 5.2. For $n = 35$, the power of the LR test is more than 40% for a false-positive rate of 10%. As expected, increasing the size of the pool results in a loss of power. The more participants contribute to the pool's statistics, the more challenging it is to identify whether the victim participated in this pool.

In all cases, the exact LR test performs best, most likely due to the availability of all statistical parameters, followed by the LR test corresponding to our threat model. The L_1 distance test achieves the least power of the empirical tests.

Finally, we observe that the theoretical LR curve is quite close to the empirically evaluated LR curve when $n = 35$, but also that it degrades faster when n increases.

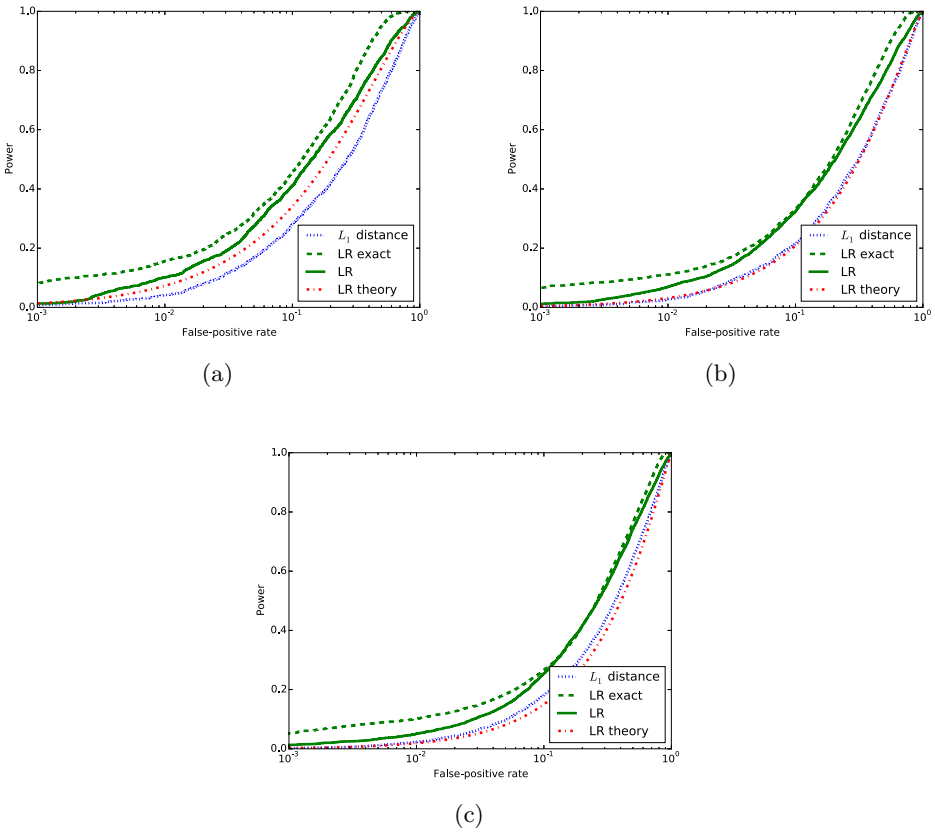


Figure 5.1: ROC curves for pools of n randomly chosen individuals: (a) $n = 35$, (b) $n = 65$, (c) $n = 124$.

Case Groups. Figure 5.2 depicts ROC curves for six different case groups of individuals carrying a specific disease. Specifically, we select six case groups ranging from the smallest (stomach tumor) to the largest (Wilms tumor) number of individuals and use them as pools. Note that these groups are fairly representative for all of the 19 case groups.

We first observe that, as previously, the exact LR test performs best, followed by the realistic LR test and L_1 distance test in most cases. We also notice that the empirically evaluated attacks perform significantly better than the theoretical approximation of the LR test for almost all case groups.

If we compare the performance on randomly constructed pools in Figure 5.1 and on case groups in Figure 5.2 for the same number of individuals n, the attack on case groups yields higher power for the same false-positive level. For instance, we observe a power of around 60% at a false-positive rate of 10% for Wilms tumor (Figure 5.2(f)) against a power of around 25% at the same false-positive rate when the individuals are randomly picked to be part of the pool (Figure 5.1(c)).

Furthermore, as shown by our dataset, it often happens that the case group is very tiny. Then, in the case of stomach tumor, for example, the power reaches 100% at a small false-positive rate of 3.5%, and 77% at a false-positive rate of 0.9% (Figure 5.2(a)).

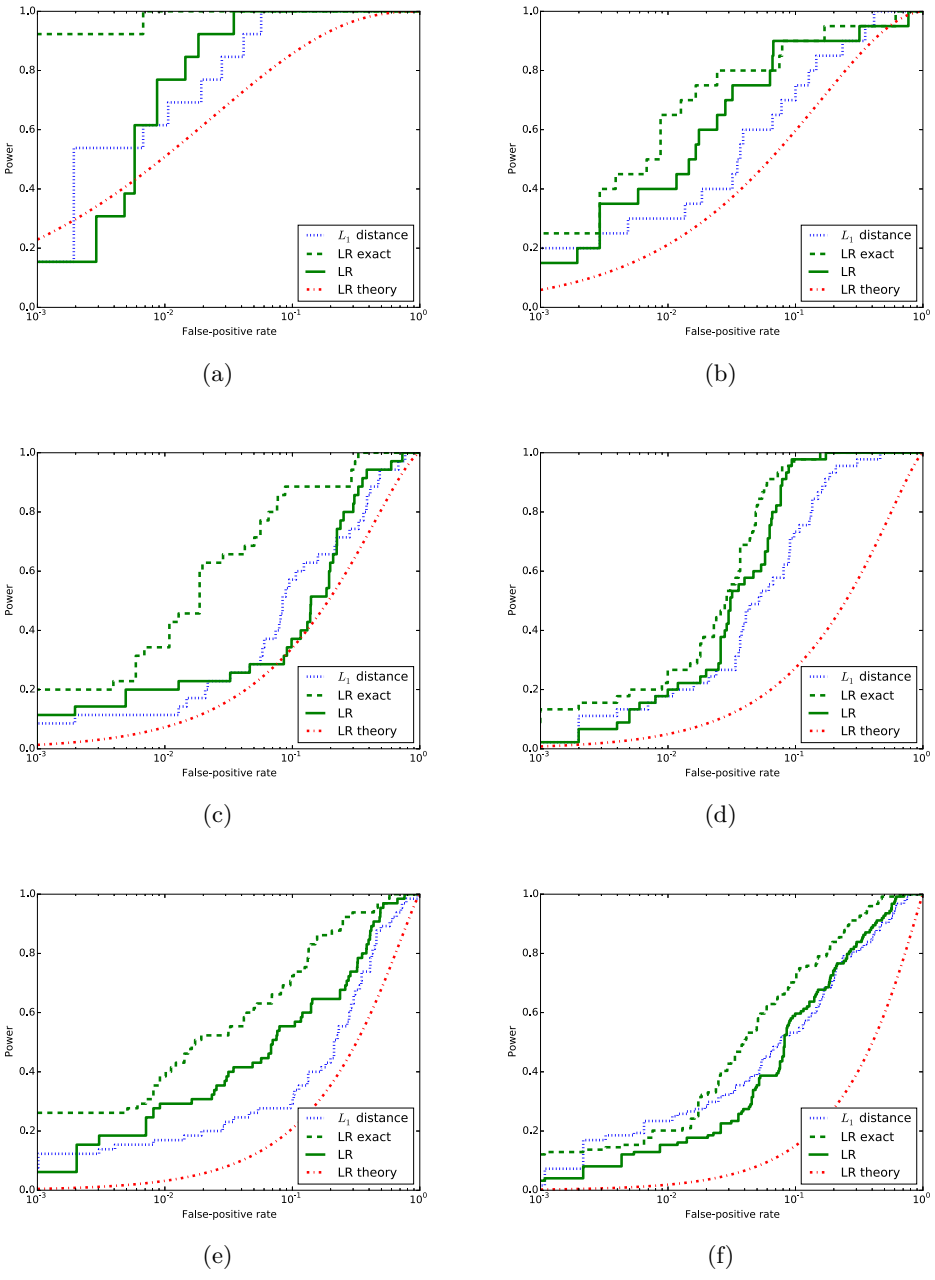


Figure 5.2: ROC curves for case groups of n individuals carrying: (a) stomach tumor (D19, $n = 13$), (b) renal cancer (D17, $n = 20$), (c) benign prostate hyperplasia (D10, $n = 35$), (d) ductal adenocarcinoma (D7, $n = 45$), (e) prostate cancer (D3, $n = 65$), and (f) Wilms tumor (D1, $n = 124$).

This demonstrates that one should be very careful when releasing summary statistics about disease-related case groups in miRNA studies, as attacks against such pools clearly outperform the theoretical LR power. This is certainly due to the fact that

miRNA expressions are highly correlated with the overall health status of their owners, and more precisely with their disease status. Note that while case groups affect the inference success, the inference result cannot be used to classify individuals as healthy or diseased. Bioinformaticians usually carry out such classifications using more advanced techniques such as support vector machines [77].

In any case, we strongly discourage researchers from publishing the exact statistics of disease-specific case groups, at least for pools smaller than a few hundred participants (which we have shown not to be resistant to membership inference attacks). Instead, we suggest applying probabilistic sanitization before disclosing the summary statistics or reducing the number of released means drastically.

Finally, note that an attack aiming at discriminating between two different pools, i.e., classifying whether an individual is part of one of two pools, would be even more successful than ours, as shown in the context of genomic privacy in [106]. For instance, the authors of this paper showed that, if the sizes of both pools were equivalent, then the number of genomic variants needed to achieve a given power and false-positive rate dropped by four compared to the more complex membership attack in which there is no information about the presence of the victim in any of the pools.

5.6 Membership Protection

In this section, we discuss and evaluate the sanitization of miRNA expression statistics, aiming at protecting the membership of any entity in the pool. To this end, we employ two different techniques, namely (1) adding noise to achieve differential privacy and (2) publishing only a subset of miRNA expression statistics.

In particular, we first analytically examine the technique based on adding noise, before we empirically evaluate the effect of both of our techniques on the privacy of contributors and the utility for research.

5.6.1 Analytical Results

For the analytical examination of the differential privacy approach, we first determine a suitable noise distribution for the mean statistic, then present utility bounds based on this noise distribution, and finally evaluate the discrepancy between noise magnitudes under two adversarial assumptions and different parameters.

A standard method to achieve differential privacy for real-valued functions is to add Laplace noise: We replace the original mechanism for computing mean values $f_{\mathsf{avg}} : \mathcal{H} \rightarrow \mathbb{R}^m$ by the sanitized mechanism $f'_{\mathsf{avg}} = f_{\mathsf{avg}} + (y_1, \ldots, y_m)$ that adds noise y_i to each miRNA expression mean distributed by a suitably scaled Laplace distribution $Lap(b)$. As shown by Dwork et al. [30], we achieve ϵ-differential privacy for f_{avg} by adding Laplace noise scaled with $b = \frac{\Delta(f_{\mathsf{avg}})}{\epsilon}$, where $\Delta(f_{\mathsf{avg}})$ is the global sensitivity of f_{avg}, defined as follows.

Definition 3. *For the statistic $f_{\mathsf{avg}} : \mathcal{H} \rightarrow \mathbb{R}^m$ that releases the means of m miRNA expression values over n samples, where the expression value of miRNA i has range δ_i,*

the global sensitivity $\Delta(f_{\mathsf{avg}})$ *is determined by*

$$\begin{aligned}\Delta(f_{\mathsf{avg}}) &= \max_{H_1,H_2\in\mathcal{H}} \|f_{\mathsf{avg}}(H_1) - f_{\mathsf{avg}}(H_2)\|_1 \\ &= \max_{H_1,H_2\in\mathcal{H}} \sum_{i}^{m} |f_{\mathsf{avg},i}(H_1) - f_{\mathsf{avg},i}(H_2)| = \sum_{i}^{m} \frac{\delta_i}{n},\end{aligned}$$

where H_1 *and* H_2 *are two datasets differing in one element.*

Applying this definition, for every miRNA i in $\{1,\ldots,m\}$ and pool containing n individual samples, the noise y_i added to the mean to achieve ϵ-differential privacy is drawn from $Lap(\frac{\sum_{k=1}^{m}\delta_k}{n\epsilon})$.

Note that the range δ_k of miRNA k's expression is the global range of its expression values, not the range within the pool only. In our evaluations, we approximate this range by the difference between the minimum and maximum expression values found in our whole dataset.

One of the main criticisms of differential privacy is that adding noise to the original statistics negates its utility. We now derive a bound for the probability that the most noise added to any element $f_{\mathsf{avg},i}$ of f_{avg} exceeds a value z. Note that, as shown by Ghosh et al. [46], using a geometric noise mechanism can lead to slightly better utility bounds. However, in our specific use case, the high sensitivity of our release mechanism will dominate any practical utility concerns, and we thus stick to the simpler Laplacian mechanism.

Theorem 3. *Let* $f_{\mathsf{avg}} : \mathcal{H} \to \mathbb{R}^m$ *and let* $f'_{\mathsf{avg}} = f_{\mathsf{avg}} + (y_1,\ldots,y_m)$, $y_i \sim Lap(\frac{\Delta(f_{\mathsf{avg}})}{\epsilon})$. *Then,* $\forall z \geq 0$

$$\Pr\left[|f_{\mathsf{avg},i}(H) - f'_{\mathsf{avg},i}(H)| \geq z\right] \leq e^{-\frac{\epsilon n z}{\sum_{k=1}^{m}\delta_k}}$$

Proof of Theorem 3. By Theorem 3.8 in [31], it holds that

$$\Pr\left[|f_i(H) - f'_i(H)| \geq ln\left(\frac{1}{\alpha}\right)\left(\frac{\Delta(f)}{\epsilon}\right)\right] \leq \alpha$$

for some probability $\alpha \in (0,1]$. Note that, instead of considering the L_∞ norm of the whole output of f as in the original result, we bound the difference for any of the output values f_i.

By setting $z = ln\left(\frac{1}{\alpha}\right)\left(\frac{\Delta(f)}{\epsilon}\right)$, replacing $\Delta(f)$ by the formula derived in Definition 3, and solving it for α, we get our upper bound. □

Given that the range of some of the miRNA expressions in our dataset is very high, the sensitivity $\Delta(f_{\mathsf{avg}}) = \frac{\sum_{i}^{m}\delta_i}{n}$ of the mean statistic will be very high too. Figure 5.3(a), which represents empirical complementary cumulative distribution function of the miRNA expression means, helps to understand this behavior. Indeed, it shows that the majority of expression values' means are smaller than 200, but also that some are higher than 10,000. Similar substantial discrepancies occur for the expression ranges δ_i. As the sensitivity is, for every miRNA, by definition, the sum over all miRNA ranges, it affects the noise distribution added to every miRNA. The probability bound on the

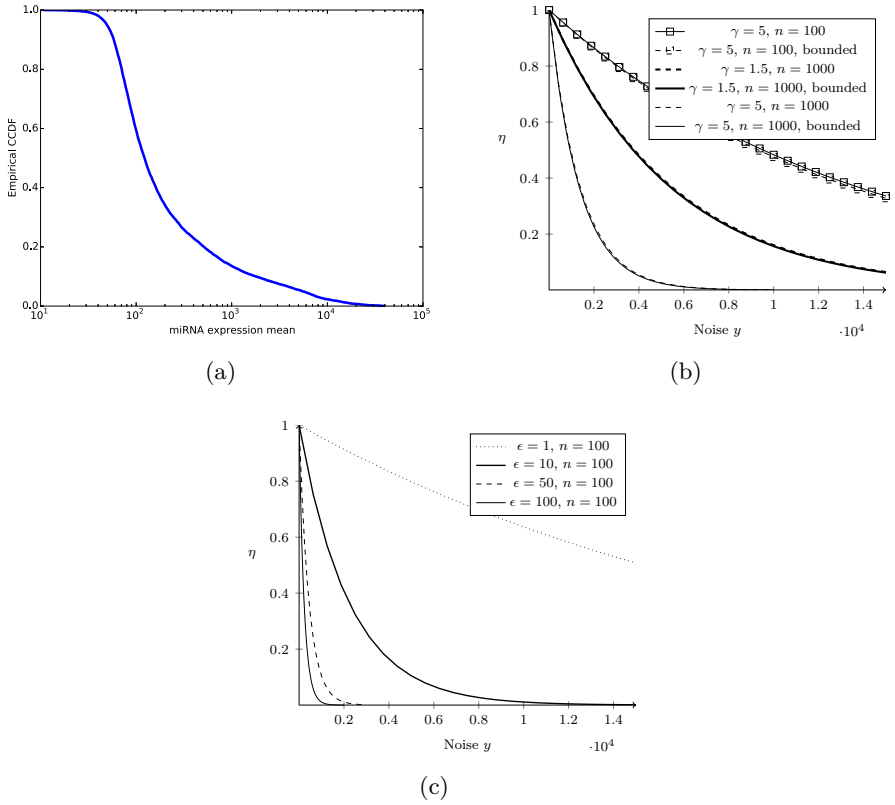


Figure 5.3: Comparison of initial miRNA expression means and typical noise distributions with and without bounded priors. (a) Empirical complementary cumulative distribution function (CCDF), (b) Probability upper bound η that the noise added to our statistic f_{avg} is greater than or equal to y, given the membership-privacy parameter $\gamma_1 = 1.5$ and $\gamma_2 = 5$ and the pool sizes $n_1 = 100$ and $n_2 = 1000$, (c) Probability upper bound η given the differential-privacy parameter $\epsilon \in \{1, 10, 50, 100\}$ and the pool size $n = 100$.

maximum noise added to $f_{\text{avg},i}$ is thus large unless the pool contains a large number of samples n, or ϵ is large.

We now evaluate whether providing $(\gamma, \mathbb{D}_B^{[a,b]})$-positive membership privacy by considering a weaker adversary can help reduce the amount of noise in our context. To achieve membership privacy for bounded prior membership probabilities, we can derive ϵ according to Theorem 1 from γ and the priors a and b. Contrary to the application example in [119], in which the adversary aims to distinguish the membership between a case group of size n and a control group of size $N - n$, our adversary has to determine membership in a pool without knowing a priori that the victim is either in the case group or in the control group. Therefore, our priors are not the probabilities of being in the case group or in the control group knowing that the victim is part of the N individuals contributing their data to the study,[4] but rather the probability that an individual contributed his data to a pool, given that he is part of a given population,

[4]Under this assumption, the probability of being in the case or control group is typically 0.5 [119].

D	a, b	ϵ		
		$\gamma = 1.3$	$\gamma = 1.5$	$\gamma = 5$
D19	0.0003	0.2624	0.4056	1.6104
D17	0.0013	0.2627	0.4061	1.6145
D3	0.009	0.2651	0.41	1.6464

Table 5.1: Privacy parameters for the diseases stomach tumor (D19), renal cancer (D17), and prostate cancer (D3 – male only) achieving γ-membership privacy under prior probability determined from disease prevalence rate in the US (collected on (17)).

much larger than N.

Here, we assume the adversary only knows the country in which the victim lives and relies on the nation-wide disease-prevalence statistics as background knowledge. Table 5.1 presents the prior probabilities, for a victim living in the US and for three cancers present in our dataset, and the resulting values of the privacy parameter ϵ for each disease and typical values of γ. We notice that, for these values of γ, the resulting ϵ values do not differ a lot between different diseases, even though the prevalence rate, or prior, of D3 is 30 times higher than D19's rate/prior. This can be explained by the relatively small absolute priors given by the prevalence rates.

Figure 5.3(b) illustrates the dependence of the utility bound provided in Theorem 3 on the number of individuals in the pool, and on the values of membership privacy parameter γ. We depict the probability upper bound from Theorem 3 (referred to as η in the figure) for the general differential privacy case (i.e., $\epsilon = \ln(\gamma)$) and for the case with bounded priors, given membership-privacy parameters $\gamma_1 = 1.5$ and $\gamma_2 = 5$, and pools of sizes $n_1 = 100$ and $n_2 = 1000$. For the case with bounded priors (so-called "bounded" in the figure), the privacy parameter ϵ corresponding to the membership-privacy parameter γ has been derived from the priors of disease D3, as provided in Table 5.1.

We make the following observations from Figure 5.3(b). First, for prior membership probabilities that are relevant for our use case, using the privacy parameter with bounded priors determined by Theorem 1 does not make a noticeable difference to using traditional differential privacy (with unbounded priors). Using the privacy parameter determined for the other two diseases (D19 or D17) in Table 5.1 leads us to the same conclusion. Therefore, we suggest making use of traditional differential privacy as it provides privacy guarantees against a stronger adversary. For this reason, we focus on traditional differential privacy in our empirical evaluations.

Second, the accuracy of the noised summary statistic increases exponentially with the sample size n. This is consistent with the result of Theorem 2. In other words, the higher n is, the less powerful is the membership attack, and the less noise needs to be added to the summary statistics for guaranteeing differential privacy. We can therefore only encourage biomedical researchers to increase the size of their miRNA pools, which will benefit privacy as well as accuracy and significance of their results.

Finally, for the pool sizes, we observe in our dataset, the expected accuracy of our noisy summary statistic f'_{avg} will be very low unless we significantly increase the privacy parameter ϵ. Figure 5.3(c) shows how our utility bound evolves depending on the parameter ϵ. By comparing the noise values y with the CCDF of Figure 5.3(a), we

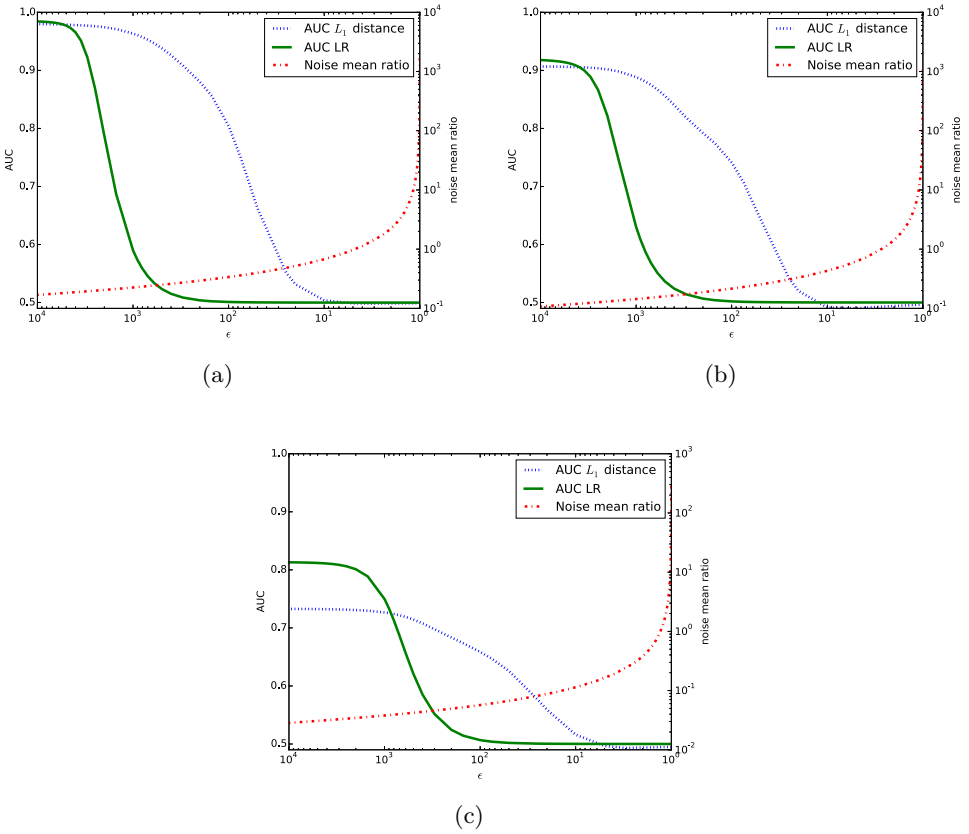


Figure 5.4: Membership inference attacks in the presence of a differentially private mechanism. AUCs and noise-to-mean ratios for three case groups: (a) stomach tumor (D19), (b) renal cancer (D17), (c) prostate cancer (D3).

clearly notice that the noise is too large with respect to most of the miRNA means for the chosen (low) privacy parameters. Since ϵ is a parameter that can be freely chosen by the designer of the sanitization mechanism, we will, in our evaluation in the following section, examine how far we can increase ϵ while at the same time ensuring that the attacks presented in Section 5.5 are countered. In any case, given the sensitivity of the mean statistics of miRNA expressions, we can expect that ϵ will have to be large to reach a level of noise that is not too high. Then, if ϵ is large (and consequently γ is very large), there is again almost no utility difference between providing membership privacy with bounded or unbounded priors (i.e., differential privacy).

5.6.2 Experimental Results

In this section, we evaluate first the impact of the differentially private mechanism on the membership attack and on the utility. Then, we evaluate the effect of hiding a certain number of released miRNA expression means.

Differentially Private Mechanism. We follow the approach presented above for ensuring ϵ-differential privacy. That is, we generate the noise vector $\mathbf{y}$ from m randomly generated Laplacian samples drawn from $L(\frac{\sum_{k=1}^{m}\delta_k}{n\epsilon})$ and add its value to the vector of miRNA expression means: $\hat{\boldsymbol{\mu}}' = \hat{\boldsymbol{\mu}} + \mathbf{y}$.

We repeat this process 1000 times, evaluate each run as presented in Section 5.5.3, and derive the average ROC curve and its resulting area under the curve (AUC) for ϵ between 1 and 10^4. Note that an AUC of 0.5 represents a similar performance as randomly guessing whether the victim is part of the pool or not, meaning best privacy. On the contrary, an AUC of 1 represents the worst outcome from a privacy perspective: 100% power at any false-positive level.

In this subsection, we focus on three case groups related to cancer that represent the groups for which the membership attack was most successful (see Figure 5.2). Figures 5.4(a)–(c) show the AUC of the L_1 distance and LR attacks, and the noise-to-mean ratio $\frac{1}{m}\sum_{i=1}^{m}\frac{|y_i|}{\hat{\mu}_i}$ resulting from the noise mechanism. This ratio can be viewed as an indicator of the utility of the published statistics: A ratio of 0 means that all utility is preserved, whereas a ratio of 1 means that, on average (over all runs and miRNAs), the added noise is equivalent to the initial mean.

First of all, we observe that, for all three depicted case groups, when noise is added to the actual means, the L_1 distance test can perform better than the LR test. In other words, the L_1 distance test is more robust to noise than the LR test. While this observation might seem counter-intuitive at first glance, especially because of the Neyman-Pearson lemma, it becomes more apparent when revisiting the impact of the noise on the tests: The L_1 distance test is influenced by the noise in a linear shift of the distance between the victim and the mean values of the pool. However, for the LR test, this distance is scaled quadratically. Hence, the LR test is more sensitive to noise than the L_1 distance test. Moreover, this observation does not invalidate the Neyman-Pearson lemma but changes the assumptions imposed on the data.

In general, the figures show that there is no ideal ϵ value achieving both membership privacy and full utility. In order to achieve perfect privacy against the membership attack with the L_1 distance, ϵ must be smaller than 10. Choosing the privacy parameter $\epsilon = 10$, however, can significantly decrease the utility of the miRNA expression means, from approximately 100% added noise (compared to the mean) for stomach tumor (Figure 5.4(a)) to around 10% added noise for prostate cancer (Figure 5.4(c)). We observe that the number of participants n in the pool plays a positive role on the privacy-utility trade-off, confirming our analytical findings. Indeed, as already mentioned, a higher value of n reduces the noise for the same ϵ value and reduces the success of the membership attack in general.

Hiding Mechanism. Considering that the differentially private method adds too much noise when n is relatively small (typically smaller than 50, like for the stomach tumor and renal cancer case groups), we also propose a non-perturbative mechanism that discloses only a subset of miRNA expression means. Ideally, this protection mechanism could obfuscate miRNA means irrelevant to the research study, such as miRNAs that are found not to be associated with the disease of interest.

In our experiments, we randomly select the subset of miRNAs to be hidden, in order

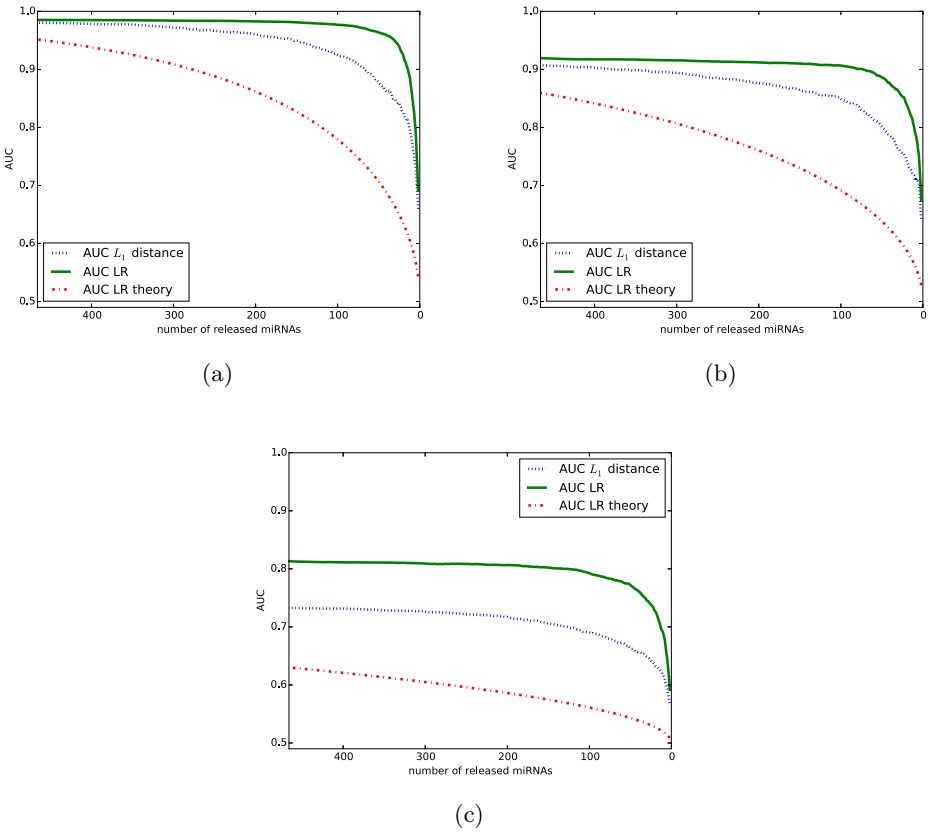


Figure 5.5: Membership inference attacks in the presence of a hiding mechanism. AUCs for three case groups: (a) stomach tumor (D19), (b) renal cancer (D17), (c) prostate cancer (D3).

to have a general idea on the impact of hiding miRNA means. To this end, we first randomly sample 50 different orders of the 466 miRNAs. Then, for each of these 50 ordered sequences, we decrease the number of released miRNA expression means from all miRNAs ($m = 466$) to $m = 1$. Finally, we average the attack results over the 50 samples for every number m. Figures 5.5(a)–(c) show the AUCs of the attacks presented in Section 5.5.

In contrast to the differentially private mechanism, the hiding of miRNA expression means preserves the guarantees of the Neyman-Pearson lemma and the assumptions of our data model. Hence, the LR attack always outperforms the L_1 distance attack. We also observe that the AUC of the theoretical LR test slightly underestimates the success of the attack, as already noticed in Section 5.5.3, due to the disease-specific pool. Moreover, we notice that theoretical AUC curves are shaped like $\sqrt{m}$, as expected from relation (5.9) of Theorem 2. This decreasing success of the attack is also observed in both empirical curves, but in a sharper manner and with a significant decrease with very few miRNAs. The empirical LR curve especially shows almost maximal AUC for $m = 50$. This demonstrates that, in practice, due to the type and behavior of miRNA

data, the LR attack is very robust against a decreasing number of released miRNA means. This robustness should again warn privacy designers about the theoretical relation that underestimates the actual attack success with disease-specific pools.

Concerning the general impact of the hiding mechanism on privacy, we notice that it does not substantially improve the situation if more than 50 miRNA means are disclosed. The number of published miRNA expressions has to be very small in order to achieve low AUCs, typically smaller than 10. In comparison to the differentially private mechanism, the AUCs with hiding never reach a point near random guessing (i.e., 0.5). Hence, while this protection mechanism might be more desirable for biomedical researchers because it does not perturb the released data, it is not able to fully protect membership privacy.

5.7 Conclusion

This work sheds further light on privacy risks stemming from miRNA expression data, showing that it is possible to detect membership in datasets of miRNA-based studies by relying on their published mean statistics. In particular, we present two attacks, one based on the L_1 distance and the other based on the likelihood-ratio test, known to be optimal. The theoretical limit derived for the latter attack has nevertheless to be taken very cautiously: Indeed, miRNA expressions are substantially more affected by the health status than genomic data. Therefore, as miRNA-based studies very often include individuals carrying specific diseases, their statistics are more different from healthy general population statistics. This, in turn, increases the adversary's power to detect membership of a given individual. Our experimental results confirm this by clearly showing that membership is much easier to detect in disease-specific datasets than in random ones.

Moreover, we propose and thoroughly study two protection mechanisms: The first protection mechanism is based on the notion of differential privacy, perturbing the released miRNA expression means, whereas the second technique only releases a subset of the miRNA expression means. We observe that the differentially private mechanism is able to protect the privacy, effectively decreasing the attack success to nearly random guessing. However, the amount of noise introduced by this protection mechanism might render the released statistics useless, in particular for small datasets. In general, we recommend the following approach for ensuring membership privacy for study participants and preserving the biomedical utility of the data: Having a large number of participants, at least a few hundred and, if necessary, slightly perturbing the summary statistics in a differentially private manner.

Possible future directions include the derivations of theoretical bounds on the attack power with noisy statistics. It would also be important to evaluate the impact of correlated miRNAs. Finally, it could be interesting to formally quantify the increased power of the attack when the adversary does not aim to detect membership in one pool but instead wants to detect membership between two pools.

6

Genotype Inference from DNA Methylation Profiles

Quantifying and Mitigating Privacy Concerns with meQTLs

6.1 Motivation

The previous chapters have focused on privacy threats arising from microRNA-expression-based studies specifically, only considering a single type of biological data available to an adversary. Biological organisms, however, typically involve a multitude of highly complex processes, each of which is defined by an interaction between different types of biological data. As a consequence, different types of biomedical data are often correlated, and there exist causal dependencies between them.

In this chapter, we study such an interdependency for two of the most important elements in the human body: the human DNA and DNA methylation. DNA methylation is not only one of the most important, but also the probably best understood epigenetic element influencing human health. It is an essential regulator of gene transcription. Aberrant DNA methylation patterns (such as hypermethylation and hypomethylation) have been associated with a large number of cancer types [36, 23, 124]. Because of its crucial role in human health, DNA methylation data might constitute highly sensitive data as well, whose privacy should be protected using dedicated legal or technical means.

Similarly to other epigenetic data, DNA methylation data vary quite significantly over time, mainly because they are profoundly influenced by environmental factors. This variability may explain why DNA methylation data are simply released (without identifiers) on open online platforms with nonrestricted access. In order to prevent privacy breaches, the genomic data corresponding to the DNA methylation data are generally *not* made publicly available and follow stricter privacy rules.

It is well-known that DNA methylation is also influenced by genetic factors [87]. As a consequence, correlations between DNA methylation and the genome could be exploited in order to re-identify anonymous DNA methylation profiles by using some public genomic database (e.g., OpenSNP [95]). Unfortunately, previous work has only tackled potential re-identification risks and countermeasures from a relatively high-level qualitative perspective (see Chapter 3). In this chapter, we provide the first detailed quantitative assessment of the identification risks inherent to DNA methylation data and, moreover, propose a provably secure technical mechanism to enable privacy-preserving methylation-based diagnosis.

6.2 Contributions

Specifically, we present a Bayesian inference framework to predict part of the genotype from DNA methylation data. We then propose an algorithm that matches DNA methylation profiles to genotypes, maximizing the posterior probabilities, given these methylation profiles. By using a rich methylation-genotype dataset, we show that only a few tens of methylation regions are sufficient to accurately match DNA methylation to genotypes. Furthermore, we present a statistical method that enables us to reject the small fraction of cases where the matching algorithm does not provide 100% accuracy, e.g., when the genotype corresponding to the methylation profile is not part of the genotype dataset. We also observe that, in such cases, if a relative is part of the genotypes' dataset, it is the one (wrongly) matched to the methylation profile. By including all genotypes contained in phase 3 of the 1000 Genome Project, we show that

the attack success is very robust to an increase in the size of the genotype dataset. Accuracy, false-positive rate, and true-positive rate remain constant for a size of the genotype dataset varying from 75 to 2579.

Given the extent of the threat, we propose a novel cryptographic scheme for privately classifying tumors, which allows for a privacy-preserving medical diagnosis in a standard clinical setting. With our method, neither can a curious third-party running the machine-learning algorithm learn the personal DNA methylation data, nor can the data owner (e.g., the patient) learn the detailed machine-learning model. In particular, we adapt existing homomorphic schemes in order to evaluate random forests with encrypted data privately. We prove the resulting scheme secure in the honest-but-curious adversarial model, which constitutes the state-of-the-art adversary model in this problem setting. We evaluate the classifier performance on real methylation data and show that it can precisely classify brain tumors in 9 subtype classes based on 900 methylation levels in less than an hour, which represents an entirely tolerable computational time for the considered application scenario.

6.3 Threat Model

We assume that the adversary gets access to one or multiple individual profiles of genome-wide DNA methylation levels, as well as to a set of genotypes. There are around 28 million CpG sites per individual and about 150 million known genomic variants to which the adversary can potentially have access. Then, we study various scenarios that could occur in practice. A typical example is to map a given anonymized DNA methylation profile to a genotype in order to re-identify it. Indeed, genomic data can facilitate de-anonymization, because there are already many profiles publicly available online with real identifiers, but also because it includes information about phenotypic traits and kinship that can be further matched to side channels such as surname-genome associations databases [53] or online social networks [64]. Moreover, the genome is very stable over our whole lifetime, and thus cannot be revoked.

Note that we assume the adversary to have no prior knowledge about the presence of the target's genotype in the set of genotypes. Thus, the adversary also wants to determine whether the genomic profile that most likely matches to the DNA methylation profile belongs to the same person. In other words, the adversary also tests if the owner of the DNA methylation profile is also part of the genomic dataset. We also study if familial relationships can mislead the adversary about the genotype corresponding to the methylation profile.

In the private classification model, we consider an honest-but-curious adversary as this assumption is standard in previous works on privacy-preserving medical diagnosis in a clinical setting [13, 5, 88, 21]. Indeed, it seems reasonable to assume that involved parties in the healthcare setting, such as hospitals or medical practitioners, will follow the protocol honestly. We leave the strengthening of our protocols to work with active adversaries for future investigations.

6.4 Attack Methodology

We present here our de-anonymization attack from a theoretical perspective. The attack relies upon the matching of one or multiple DNA methylation profiles to their corresponding genotypes. To do so, the adversary first infers the probability of a genotype, given only methylation data, and second maps the methylation profile to the genotype that maximizes the average posterior probabilities between genotypic positions and methylation sites. During these steps, the adversary takes advantage of the correlations between methylation profile and genotype. Once the best matching has been found by the adversary, he also wants to make sure that the methylation and genotypic samples in the matching pair belong to the same person. Indeed, it could be that an individual is part of the DNA methylation dataset, but not of the genotype dataset, or vice versa. To verify this, the adversary relies on a test statistic related to the matching score that provides him with a degree of certainty about whether the matching between methylation data and genotype is significant enough to be considered correct. If there is not enough certainty, the adversary can conclude that the corresponding genotype is most likely not part of the dataset.

6.4.1 Learning the Attack Model

The probabilistic relationships between methylation levels and genotypes are derived by relying on a separate training dataset $\mathcal{W} = \{(\mathbf{m}_i, \mathbf{g}_i)\}_{i=1}^{l}$ containing l pairs of DNA methylation profiles and their corresponding genotypes. In practice, methylation profiles $\mathbf{m}_i$ and genotypes $\mathbf{g}_i$ have tens of millions of different positions. Thus, the training phase aims: (1) to determine the meQTLs (cf. Chapter 2), i.e., the positions q in the genotype influencing the methylation levels in a region r, and (2) to learn the magnitude of this influence.

During this training phase, we want to select a subset of n positions in the genome that are highly correlated with certain regions in the DNA methylation pattern. To this end, for each methylation region m_i^r of an individual i in the region r, we determine the single most correlated genotype g_i^q at position q. In case more than one methylation region is most correlated with the same genotype, we pair the highest correlated methylation region with the given meQTL first, and then pair the other methylation region with the second most correlated meQTL, and so on and so forth. In order to keep the search tractable, we follow standard biomedical practice and restrict the distance between genotype position and methylation region as described later in Section 6.5. Eventually, selecting only the n highest correlations, this provides us with a set of meQTL-methylation region position pairs $\mathcal{Q} = \{(q_j, r_j)\}_{j=1}^{n}$, where $\forall (q_j, r_j), (q_k, r_k) \in \mathcal{Q} : q_j \neq q_k \Leftrightarrow r_j \neq r_k$.

Once we have identified the positions in the genotype that influence most the DNA methylation, we are interested in inferring the posterior probability of every meQTL g_j^i, given the corresponding methylation region m_j^i, $\Pr\left[G^i = g_j^i \mid M^i = m_j^i\right]$. In this probability, G^i denotes the discrete random variable of the meQTL at position q_i, where $g_j^i \in \{0, 1, 2\}$ for any q_i and individual j, and M^i denotes the continuous random variable representing the methylation levels averaged over all CpG sites within region r_i, where

$m_j^i \in [0,1]$ for individual j. The Bayes theorem states that:

$$\Pr\left[G^i = g_j^i \mid M^i\right] = \frac{p(M^i \mid G^i = g_j^i)\Pr\left[G^i = g_j^i\right]}{\sum_{g_j^i} p(M^i \mid G^i = g_j^i)\Pr\left[G^i = g_j^i\right]} \tag{6.1}$$

The prior genotype probabilities $\Pr\left[G^i = g_j^i\right]$ can be retrieved from population statistics databases, such as dbSNP [111, 24]. Alternatively, they can be directly estimated on any dataset of populations with similar ethnicity background. Moreover, we can learn the conditional probability distributions $p(M^i \mid G^i = g_j^i)$, for all $g_j^i \in \{0,1,2\}$, by relying on our training dataset $\mathcal{W}$, focusing only on the meQTL-methylation pairs contained in $\mathcal{Q}$. In this process, we must select the continuous distribution function that best fits the data. We discuss what distribution function fits best in Section 6.6.

6.4.2 Matching Attack

After having trained $p(M^i \mid G^i = g_j^i)$ for all pairs in $\mathcal{Q}$ on the training dataset $\mathcal{W}$, we can predict the posterior probabilities $\Pr\left[G^i = g_j^i \mid M^i\right]$ of the n meQTLs in $\mathcal{Q}$ given methylation profiles in another dataset, referred to as the test set in the following. The test set consists of two independently chosen subsets: (1) a set $\mathcal{G} = \{(\mathbf{g}_u)\}_{u=1}^{n_g}$ containing $n_g \geq 1$ genotypes, and (2) a set $\mathcal{E} = \{(\mathbf{m}_v)\}_{v=1}^{n_m}$ containing $n_m \geq 1$ methylation profiles. These represent the two databases the adversary has access to. Note that individuals in $\mathcal{G}$ and $\mathcal{E}$ may be different, and that the adversary wants to infer the links between $\mathcal{G}$ and $\mathcal{E}$. In this endeavor, the adversary must compute, for all meQTLs in $\mathcal{Y}$ and $n_g \times n_m$ pairs of individuals in the test set, the posterior probabilities of the actual value of the genotypes, given the methylation sites (by using the previously learned probabilities), i.e., $p_{j,k}^i := \Pr\left[G^i = g_j^i \mid M^i = m_k^i\right]$.

We derive a match score $w_{j,k}$ between individuals j and k by averaging the conditional probabilities $p_{j,k}^i$ over all n meQTL-methylation pairs in $\mathcal{Q}$, i.e., $w_{j,k} = \frac{1}{n}\sum_{i=1}^{n} p_{j,k}^i$. We then select the matching α^* over $(\max(n_g, n_m))!\,/(\max(n_g, n_m) - \min(n_g, n_m))!$ possible assignments that maximizes the sum of the individual match scores:

$$\alpha^* = \arg\max_{\alpha} \sum_{j=1}^{\max(n_g,n_m)} w_{j,\alpha(j)}, \tag{6.2}$$

with $w_{j,\alpha(j)} = -\infty$ if $j > n_g$ or $\alpha(j) > n_m$.

This problem boils down to finding the best vertex matching on a weighted bipartite graph, with n_g vertices on one side representing the genotypes of n_g individuals, and n_m on the other side representing the methylation profiles of n_m individuals. Each edge between any two vertices pair (j,k) has a weight equal to $w_{j,k}$. As the number of possible assignments increases with $O(\max(n_g, n_m)^{\min(n_g,n_m)})$, the naive matching approach is computationally intractable if both, n_g and n_m, are big. As in Chapter 4, we rely on the blossom algorithm [34] in our experiments, because it only has a complexity of $O((n_g + n_m)^3)$ and it can also be applied to general graphs. Of course, if $n_m = 1$ or $n_g = 1$, there is no need to use any maximum weight assignment algorithm, as one can simply select the genotype $\mathbf{s}_j$, respectively methylation profile $\mathbf{e}_k$, maximizing $w_{j,1}$, respectively $w_{1,k}$, and the complexity is then linear in n_g, respectively n_m.

6.4.3 Statistical Validation of the Best Match

In order to evaluate the significance of the match score between genotype $\mathbf{g}_j$ and methylation profile $\mathbf{m}_k$, we rely on the z-test and the corresponding z-score, defined as $z_{j,k} = (w_{j,k} - \mu(\mathbf{w}_k)) / \sigma(\mathbf{w}_k)$, where $\mathbf{w}_k$ is the vector of match scores between the methylation profile of individual k, $\mathbf{e}_k$, and all genotypes in $\mathcal{G}$, $\mu(\mathbf{w}_k)$ is its mean, and $\sigma(\mathbf{w}_k)$ is its standard deviation. The z-score can be similarly derived between the genotype $\mathbf{g}_j$ of individual j and all methylation profiles in $\mathcal{E}$. The only requirement is that the cardinality of the set over which we compute the mean and variance is large enough. The z-score allows us to determine, once a methylation profile is mapped to a genotype, whether these two profiles correspond to the same individual. Indeed, the pair that maximizes the match score might not be the one between the profiles of the same individual, especially when the individual's data is not part of one of the sets $\mathcal{E}$ or $\mathcal{G}$. In this case, we should be able to detect that the mapped pair does not include the same individual. This is done by validating the mapped pair for a z-score higher than a given threshold.

If n_m-by-n_g matching becomes computationally infeasible, it is worth noting that it is also possible to map methylation profiles one-by-one to genotypes, i.e., carry out n_m times a one-by-n_g matching, the complexity of which is then linear in $n_m n_g$. Moreover, the adversary may have access to multiple methylation profiles of the same person, but at different points in time. In this case, it can also be beneficial to rely on the one-by-n_g matching, which allows multiple methylation profiles to be mapped to the same genotype, contrary to the bipartite graph matching. In case the adversary is certain that there is only one methylation profile per individual, the n_m-by-n_g matching outperforms the one-by-n_g matching (see Section 6.6), but if he is not sure about the number of methylation profiles per individual, the n_m-by-n_g matching becomes more challenging to use.

6.4.4 Comparison with Previous Chapters

The matching task at hand exhibits considerable parallels to the scenarios described in Chapter 4. A n_m-by-n_g matching corresponds to a matching attack in our previous work, while a one-by-n_g matching corresponds to the scenario of an identification attack. However, in this work, we do not aim at linking profiles of the same datatype over time but provide an approach for linking profiles of different types of biomedical data. Moreover, both our works differ significantly in the methodology applied.

6.5 Dataset Description

The dataset that was used in this study consists of meQTLs determined from a set of 75 individuals, 42 of which have parental relations (21 mother/child pairs) for which whole blood was available. The DNA methylation was determined using whole genome bisulfite sequencing (WGBS), allowing a genome-wide measurement of the DNA methylation levels for all 28 million CpG dinucleotides. The sequencing data were processed using an in-house processing pipeline consisting of alignment of the sequencing reads, quality assessment, and methylation calling. Then, the genotype was determined at known

SNP loci listed in the dbSNP [111, 24] database version 141, using the Bis-SNP tool, which calls SNP genotypes from WGBS data [82]. For the majority of individuals (67 out of 75), samples collected at the birth of the child, referred to as t_0, were available, but also at later times: one year (t_1), up to 8 years (t_8) for some individuals after birth.

Such a longitudinal dataset containing individuals with parental relations represents a unique and valuable data source in the biomedical community. Note that this dataset cannot be released publicly yet, but will be made available to researchers in the near future.

On a subset of these samples, we selected the CpGs based on their high variance across the dataset. CpGs showing a very stable methylation profile across the subset of samples were discarded, as they are not expected to be under the influence of meQTLs. meQTLs were determined using a Spearman rank correlation test [115] (false discovery rate threshold after Benjamini-Hochberg correction [10] of 1%) for all SNPs located within 50 kb (kilobases) up-/downstream of the CpG showing highly variable methylation. This filtering process eventually output 568,103 meQTL-methylation pairs containing 502 methylation regions and 544,762 different SNPs. This implies an average number of approximately 1132 meQTLs per methylation region.

6.6 Attack Evaluation

We present here our main experimental results, using the dataset described in the previous section. As explained in Section 6.4, the training phase relies on two different steps: (1) identify the meQTLs, i.e., the positions in the genotype that influence the methylation levels, and (2) quantify the magnitude of this influence. As we carry out the first step similarly for all experiments, this can also be seen as a data preprocessing step, which filters out non-relevant genotypic positions and methylation regions.

6.6.1 Generic Training Phase

We focus here on the meQTL-methylation pairs with a Spearman rank correlation coefficient larger than 0.49 (FDR threshold after Benjamini-Hochberg correction of 1%). This provides us with 326 methylation regions and 9,532 pairs, i.e., around 29 meQTLs per methylation region. Then, we keep only one most correlated meQTL for each methylation region, resulting in 326 pairs, as expected. Filtering out the meQTLs for which no information was available on dbSNP, we are left with 314 meQTL-methylation pairs. Finally, since we have to compute the variance (see below) of the conditional probability $p(M^i \mid G^i = g^i_j)$ for all possible values of g^i_j, we filter out meQTLs that do not have at least two samples per genotype value g^i_j. This eventually leads us to a total of 293 meQTL-methylation pairs for the whole dataset.

Normal Distribution Function. The first step towards precisely modeling the influence of meQTLs on methylation regions is the selection of the continuous distribution function that best fits the observed data. We rely on the normal distribution which happens to be well suited from both, a visual and statistical perspective. First, in order to evaluate if the normal distribution approximation was statistically significant, we applied the

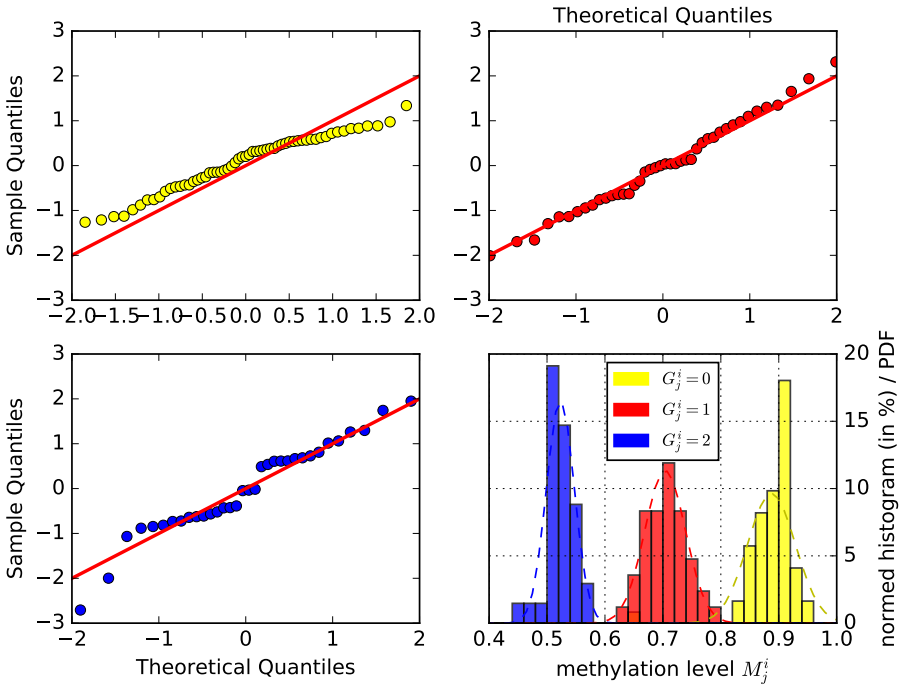


Figure 6.1: Example of the empirical distribution $\hat{p}(M_j^i \mid G_j^i)$ of methylation levels conditioned on genotype values $g_j^i = \{0, 1, 2\}$ for the pair with meQTL rs10928633 (in chromosome 2, position 138625907) and methylation region $[138625907, 138626564]$ in the same chromosome. Yellow color (top-left plot) is $\hat{p}(M^i \mid G^i = 0)$, red color (top-right plot) is $\hat{p}(M^i \mid G^i = 1)$, and blue color (bottom-left plot) is $\hat{p}(M^i \mid G^i = 2)$.

one-sample Kolmogorov-Smirnov test to all 293 meQTL-methylation region pairs and possible genotype values, $g_j^i \in \{0, 1, 2\}$. The null hypothesis (the samples belonging to the normal distribution) was only rejected in a minority of cases at significance level 0.05 (134 out of 879). When we manually examined those few cases by plotting their histogram, we found that all of those cases contained either a very few outliers, or almost all of the methylation levels belonged to the same bin in the histogram and thus were almost the same.

We also visually inspected the empirical conditional distributions $\hat{p}(M^i \mid G^i = g_j^i)$ for $g_j^i \in \{0, 1, 2\}$ and reached the same conclusion. Figure 6.1 exemplarily shows Q-Q plots as well as the empirical distribution of methylation levels, given each possible genotype of a representative pair (q_i, r_i) in our dataset. Moreover, it also displays the corresponding normal distributions induced by the unbiased estimators of the mean and standard deviation. The Q-Q plots depict on the x-axis the theoretical quantiles of a standard normal distribution. The y-axis displays the normalized quantiles of the sample distribution for each $G_j^i = g_j^i$. Given the minor discrepancies between the points and the diagonal, we can expect that the normal distribution will be a sufficiently good fit for the attack. Second, the part of the figure at the bottom right confirms that the normal distribution is indeed a good approximation for the conditional probability.

More importantly, it also shows that the overlap between the distributions conditioned on different genotype values is small, which can be used to recover the correct genotype, given the methylation level. This gives the intuition behind our re-identification attack.

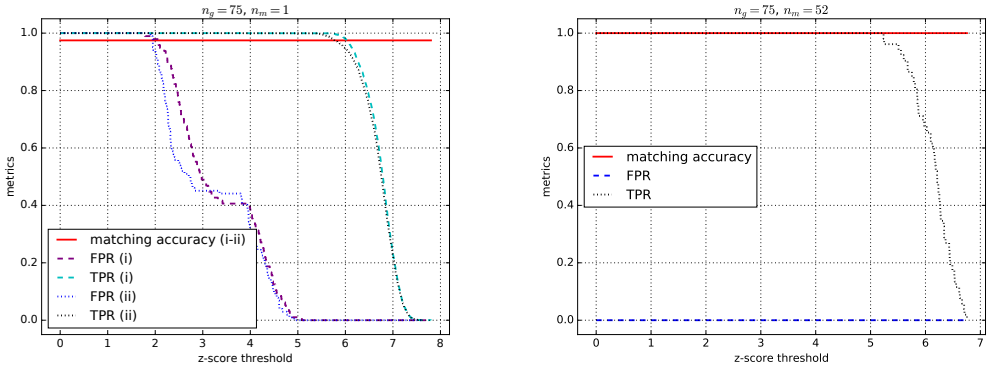
6.6.2 Experiment-specific Training and Testing Sets

In this second phase, we quantify the magnitude of the influence of each meQTL on its corresponding methylation region. From now on, in order to illustrate the performance of the attack under different scenarios, we build our training dataset from different subsets of the whole dataset described in Section 6.5. We consider three different training/testing experimental setups. In the first scenario, referred to as (i), we select one methylation profile per individual, i.e., 75 profiles, as follows: We pick the 67 profiles available at time t_0 and, in addition, the profiles of individuals not yet selected at t_0 (because of absence of data) at the smallest time point as possible: 1 at t_1, 1 at t_3, 3 at t_4, 2 at t_5, and 1 at t_6. We further select the 75 genotypes corresponding to these methylation profiles. Then, we randomly choose 37 pairs for the training set, and 38 for the testing set, or attack set. We repeat the random splitting 100 times.

In the second setup, (ii), we want to make sure that there are no individuals in the training and testing sets who have familial relationships, i.e., we want to avoid a child being in the training set, and his mother being in the test set, or the other way around. We also aim at 37 samples in the training set and 38 in the test set. Thus, we first randomly select from 2 to 18 mother-child pairs to be included in the training set, which leads us to 4 to 36 samples. Then, we randomly select the remaining samples among the isolated individuals (i.e., those who have no child or mother in our dataset) to attain 37 samples. We repeat this random selection 100 times and select the 38 remaining profiles to be part of the test set. This process ensures that there is no individual in the test set who is member of the same family as somebody in the training set.

The third experimental setup, (iii), is used for the scenarios where we want to map more than one methylation profile at a time with the genotypes. In both previous settings, we consider $n_m = 1$ and $n_g = 75$ (or more, as we will see later), but we repeat the attack with all 38 methylation profiles independently. Now, we want to match $n_m > 1$ methylation profiles to $n_g = 75$ genotypes. We then select our samples in order to maximize the number of methylation profiles in the test set, as follows. We select all individuals at time t_1 and at time points $t > t_1$ that do not have methylation profiles at t_0 and t_1. This gives us 16 methylation profiles at t_1 plus 7 at later time points, thus 23 methylation profiles for the training set. Then, for the test set, we select all methylation profiles at t_0 whose owners do not overlap with those in the training set. This leads to 52 methylation profiles for the test set.

Note that the requirement of having two samples per genotype value to learn the variance of the normal distribution reduces the number of meQTL-methylation pairs when we apply it to the training set and not the whole dataset. The total number of pairs ranges from 237 to 248 with a median value 240 in setup (i). It ranges from 208 to 236 with a median of 222.5 for (ii), and it is of 187 pairs for setup (iii) for which there is only one run and the number of samples in the training set is smaller (due to stronger constraints).



(a) (b)

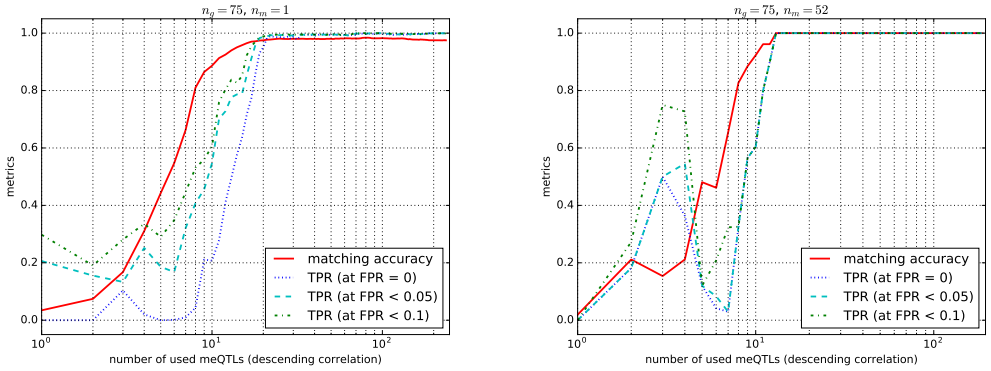
Figure 6.2: Matching of (a) one and (b) 52 methylation profiles among 75 genotypes: Average accuracy of the matched pairs, and true-positive, false-positive rates for a varying z-score threshold.

6.6.3 Results

We start by showing the performance of the attack with all available meQTL-methylation pairs (given the constraints above), $n_m = 1$, and $n_g = 75$. We include all 75 individual genotypes to be potentially matched to the methylation profiles as we assume that this can only make the attack harder for the adversary than considering just the 38 or 52 genotypes corresponding to the methylation profiles of the test set. Of course, we only select the 38 methylation profiles present in the test set to run our experiments. Therefore, we try to match one methylation profile with 75 genotypes, 38 times, over 100 runs, i.e., 3,800 times, and average the results.

Figure 6.2(a) shows: (1) the matching accuracy, i.e., the fraction of matched pairs containing genotypes and methylation profiles of the same individual, (2) the true-positive rate (TPR) after applying the z-score test, i.e., the number of correct matchings divided by the sum of the number of correct matching pairs and the number of matching pairs that are wrongly identified as non-matching, and (3) the false-positive rate (FPR) after applying the z-score test, i.e., the number of false mappings that are identified as correct divided by the sum of the latter value and the number of correct mappings identified as false. We could have also depicted other metrics, such as accuracy after z-score, but we consider the TPR and FPR as sufficient metrics to depict the success of the matching attack.

First, Figure 6.2(a) shows that, on average, the attack accurately matches the methylation profile to its corresponding genotypes around 97.5% of the time. Then, we notice that there exists a z-score for which we always reject all wrongly matched pairs (FPR = 0 for z-score approximately greater than 5), and never reject those that are correct (TPR = 1 for z-score approximately smaller than 5.5). This means that for the 2.5% of the pairs that are wrongly matched, we are able to identify that they are false positives. Finally, we notice that the matching accuracy is the same for both scenarios (i) and (ii) and that the FPR and TPR are also very similar.



(a) (b)

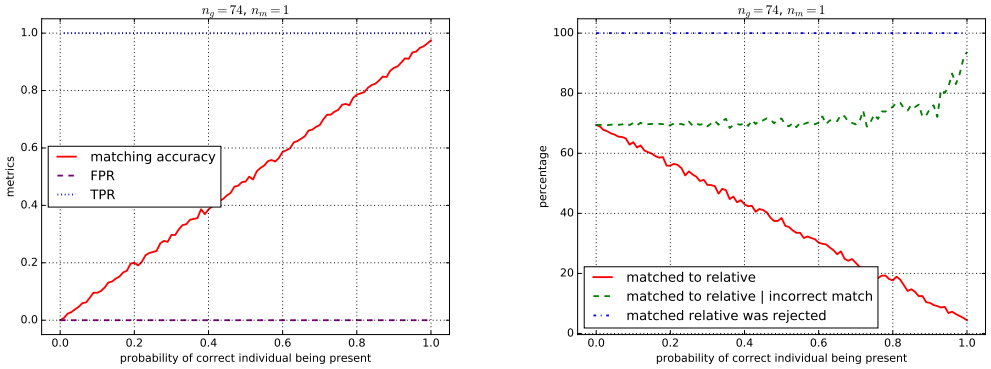
Figure 6.3: Matching of (a) one and (b) 52 methylation profiles among 75 genotypes with an increasing number of observed meQTLs/methylation regions (in descending levels of correlation): average accuracy of the matched pairs and true-positive rates at various false-positive levels.

Figure 6.2(b) shows the attack when there is more than one methylation profile to match to the genotypes. Given the experimental setup (iii), we have 52 methylation profiles that we match against the whole 75 genotypes. First of all, we notice that the matching accuracy is 100%, i.e., that the attack correctly matches the 52 pairs. Then, by looking at the z-score to validate the matched pairs, we note that it starts rejecting valid pairs from around 5.2. As we only have correctly matched pairs after the matching algorithm, there is no point in displaying the FPR, because there is no wrong pair to reject. We conclude from Figure 6.2(a) and 6.2(b) that the attack is more successful when matching more than one methylation profile to multiple genotypes.

Next, we evaluate the impact of reducing the number of use methylation-meQTL pairs on the attack success. In this endeavor, we gradually use an increasing number of observed methylation-meQTL pairs, from 1 to 237, in decreasing order of correlation. Figure 6.3(a) shows the evolution of the matching accuracy and of the TPR after applying the z-test, for three possible FPR values: 0, 0.05, and 0.1. First, we notice that we reach the maximum matching accuracy with only 20 methylation-meQTL pairs and almost 90% accuracy with 10 pairs. Second, we see that we attain a TPR of 0.6 at an FPR of 0.05 when we apply the z-test (at 10 pairs). Furthermore, we reach a 0.95 TPR at 0.05 FPR with 20 methylation-meQTL pairs, and 0.99 with 30 pairs.

When evaluating the same experiment with a fixed threshold of 5.5 (as found suitable in Figure 6.2(a)), we notice that 80 methylation-meQTL pairs are necessary to achieve a TPR of almost 0.9 and an FPR of 0. This arises from the fact that a larger number of methylation-meQTL pairs provides more information and thus gives a more accurate match score, which also allows for higher z-score thresholds to perform better.

Similarly, Figure 6.3(b) shows the evolution of the various metrics with respect to an increasing number of observed methylation-meQTL pairs, for $n_m = 52$. The less smooth behavior of the curves is due to the fact that we have one run here, compared to 100 runs in the case where $n_m = 1$. We notice that the matching accuracy and TPR



(a) (b)

Figure 6.4: (a) Identification of one methylation profiles among 74 genotypes with an increasing probability of the correct matching genotype being present in the dataset: Average accuracy of the matched pairs, true-positive and false-positive rates. (b) Analysis of the wrongly matched first-degree relatives in the aforementioned scenario.

reach highest values for a number of methylation-meQTL pairs that is lower than when $n_m = 1$. Precisely, the attack achieves full accuracy and TPR at 0 false-positives with only 13 pairs. Again, we see that matching more than one methylation profiles to their corresponding genotypes induces higher attack success.

We evaluate now how the attack performance evolves when the genotype corresponding to the targeted methylation profile is not present in the genotype dataset. We have $n_g = 74$ genotypes if the targeted genotype is not present and, for the sake of comparison, we keep the same number when it is present, by removing another of the 74 genotypes at random. Figure 6.4(a) shows the evolution of this performance with respect to an increasing probability that the targeted genotype is in the dataset, from 0 to 1, by intervals of 0.01. For each probability value x, we randomly sample a value v between 0 and 1, uniformly, and keep the targeted genotype in the dataset if and only if $v < x$. We repeat this sampling process 100 times, evaluate our attack, and average its outcomes. As expected, the matching accuracy increases with the probability that the correct genotype is present in the dataset, since the adversary cannot find the correct genotype if it is not there. The crucial point here is that the adversary can detect that the genotype is not present for any presence probability. Indeed, with the appropriate z-score (between 4.9 and 5.4), the adversary always rejects the wrongly matched genotypes (FPR=0) while accepting the correctly matched genotypes (TPR=1).

We also investigate the effect of a relative's genotype being in the genotype dataset, with a varying presence probability of the targeted genotype, as in Figure 6.4(a). The relative here is either the mother or the child. Figure 6.4(b) shows the percentage of times the relative's genotype is matched to the methylation profile, in absolute value, and relative to the condition that the matched pair was wrong. It also shows the percentage of times this wrongly matched pairs were rejected by the z-test. First, we observe a linear decrease of the probability of being matched to the relative with respect to the presence probability. We also see that this curve does not start at 1 but at

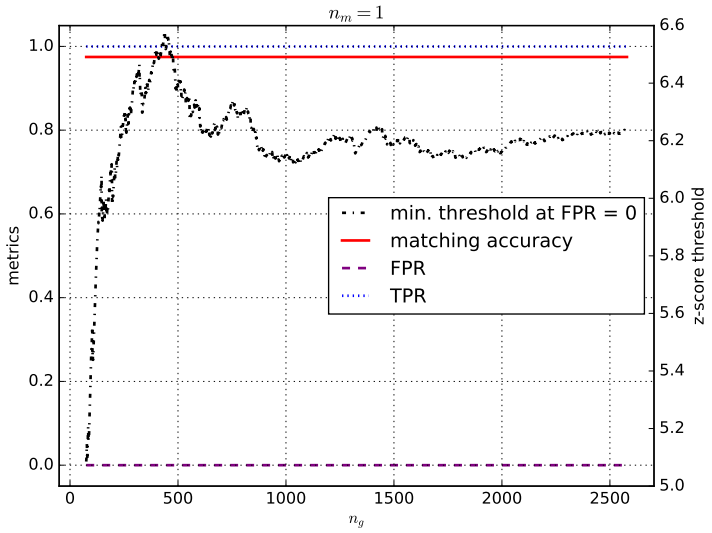


Figure 6.5: Identification of one methylation profiles among an increasing number of genotypes, from 75 to 2579: Average accuracy of the matched pairs, true-positive and false-positive rates and minimum z-score threshold for a null false-positive rate.

around 0.7. This means that, when the targeted genotype is not in the dataset, the wrongly matched genotype is in 70% of the cases the relative's genotype, and in the 30% remaining cases the one of an unrelated individual.

In order to better understand these proportions, we display the fraction of familial matches among all wrong matches (green dashed curve). We observe that this fraction increases with the presence probability. In order to understand this behavior, we must recall that the matching accuracy also increases with the presence probability. This means that the fewer wrong matched pairs there are, the more likely these are pairs containing the genotype of a relative and not of an unrelated individual. Also, it means that, when the chance that the targeted genotype is present in the dataset is high, the only genotype that can mislead the adversary's matching is the relative's genotype in the vast majority of cases.

Finally, we study the robustness of our attack for an increasing number of genotypes, from 75 to 2579, by including the 2504 genotypes of the 1000 Genomes Project (phase 3) [65]. Figure 6.5 shows the evolution of the matching accuracy, of the false-positive and true-positive rates after the z-test, and of the minimum z-score for reaching a null FPR. First, we notice that the matching accuracy remains constant, at 97.5%, for every genotype dataset size n_g. Moreover, there always exists a z-score that enables us to reject all wrongly matched pairs while keeping all correctly matched pairs. We notably notice that this z-score evolves quite a lot until around $n_g = 1000$ and that it tends to converge to a fixed value when n_g gets closer to 2579. We conclude from this figure that the attack is very robust to an increase in the number of genotypes we have to match the methylation profile to.

We also evaluated this experiment with fixed thresholds on the z-score. When less

than 100 genotypes are present, a threshold of 5.5 provides a TPR of 1 and FPRs below 0.05. When more than 100 genotypes are part of the test set, a threshold of 6 achieves the same effect. Since these observations conform with previous experiments, we believe that an adversary is able to determine a suitable threshold from her training data.

6.7 Private Classification with Random Forests

As we have shown, publicly releasing methylation profiles has a hugely detrimental effect on the patients' privacy, with a risk close to 100% to have one's methylation data re-identified. Therefore, we first strongly recommend reconsidering if the existing DNA methylation datasets should remain publicly available in online databases. Moreover, it is vital to understand the needs of the medical community for designing appropriate protection mechanisms that provide the doctors and patients with privacy guarantees and diagnosis utility. In this section, we propose a novel cryptographic scheme for privately classifying tumors based on random forests. We first describe the preliminaries on random forests and then present our private random forest-based classifier.

Random forests are a promising technique used in the medical community for classifying diseases [58]. This ensemble method bases its classification on a multitude of classification trees in order to prevent overfitting and to reduce the prediction variance [41]. For example, Danielsson et al. developed a random forest classifier tool enabling the identification of pediatric brain tumor subtypes with an accuracy of 98% [22].

In practice, when diagnosing a patient's disease, a sample is taken from the patient by a medical practitioner. Then, the sample needs to be analyzed either by the hospital or by a medical laboratory, resulting, e.g., in the DNA methylation profile of the patient. The actual classification based on these data can then be outsourced to a third-party company providing data-driven medicine, such as Sophia Genetics [114]. The DNA methylation profile is sent to the third party, which then provides the physician or hospital with the diagnosis. While the business model of this third party is inherently protected by keeping the classification model secret, the patient's privacy is clearly at risk, as his data are available to the third party.

Hence, when classifying a patient's disease, two privacy goals must be achieved: (1) protecting the company's classification model, and (2) protecting the patient's data from the third-party company. Note that, in order to construct its classifier, the company must have access to a training set of DNA methylation data in clear. Our scheme protects the data on which only classification has to be carried out (e.g., for diagnostic purposes). Finally, our scheme is flexible in the sense that it can release two outcomes: (1) only the class with the plurality vote (most frequently chosen by the random forest algorithm), or (2) the class of every tree in the random forest, which enables the medical practitioner to carry out a more fine-grained analysis of the distribution over the possible classes.

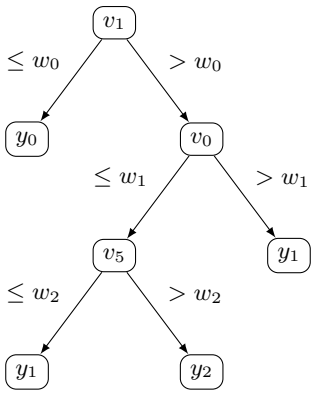


Figure 6.6: Example of a binary classification tree with classes $Y = \{y_0, y_1, y_2\}$, thresholds w_0, w_1, w_2 and an input vector $\mathbf{v}$.

6.7.1 Preliminaries

In this section, we briefly introduce the concepts of classification trees and random forests.

6.7.1.1 Classification Trees

Classification trees (or decision trees) are a predictive tool that is popular in machine learning. They are used to classify an input $\mathbf{v}$ into a set of different classes $Y = \{y_0, \ldots, y_k\}$. As the name suggests, a classification tree can be represented by a simple, usually binary, tree, in which each interior node corresponds to an input value v_i. The two edges of each interior node partition the node's input domain into two distinct sets. Each leaf node of the tree is labeled with a class y_j. It is worth noting that a single class may occur at more than one leaf.

In order to classify an input using a classification tree, one starts at the root node and walks down the tree until a leaf node is reached. At each interior node, the decision which edge to select is determined by the partition to which the corresponding input value belongs. Finally, the class label of the leaf node determines the result of the classification task. In the following, we will focus on the most common form of classification trees as implemented in many libraries: binary classification trees in which the partitioning at each interior node is given by a comparison of the input value with a threshold w_i. The model of such a classification tree is completely described by the structure of the tree, the input values v_i corresponding to each node, as well as the thresholds w_i applied at each node.

6.7.1.2 Random Forests

Classification trees usually suffer from a high prediction variance and can easily suffer from overfitting to their training set. In order to reduce the prediction variance, random forests put together multiple noisy, but approximately unbiased classification trees.

In general, a random forest consists of k classification trees, where the number k is subject to tuning. The training of a random forest is performed on a training dataset

$T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, consisting of n samples together with their corresponding class label. During the training, each tree is grown on n randomly chosen (with replacement) training samples, using only a randomly chosen set of input predictors (components of the training samples) $P \subseteq \{1, \ldots, len(\mathbf{x})\}$. It is this random subset of input predictors that distinguishes random forests from simple tree bagging and ensures the trees to be *de-correlated* so that the same input predictors are not used in all of the trees. This step is important to reduce the correlation of the trees, which then enables further reduction of the prediction variance [41].

Given a random forest model and an input $\mathbf{v}$, the classification algorithm evaluates each of the model's trees individually. Then, depending on the application, implementation or preference, the resulting class can be determined by plurality vote (or majority vote for binary classification), averaging class predictions, or providing class probabilities in terms of relative vote counts.

6.7.2 Private Classification with Random Forests

Next, we introduce our construction that allows for secure evaluation of random forests between a third party and a querier. More specifically, we do not want the querier (referred to as client) to learn the structure of the trees, nor should the third party (referred to as server) learn anything about the input sample or the result of the classification.

We build our construction on top of the work of Bost et al. [13] and extend it to work with random forests. In their work, they introduced three major classification protocols, namely for hyperplane decision, Naïve Bayes, and classification trees. They all satisfy the constraint to keep both the classifier model and the data confidential. Since classification trees are an essential component of random forests, we first discuss the details of the classification tree protocol, before extending it to random forests.

It is important to note that the classifier is trained upfront on data in the clear, whereas only the actual classification of new samples is performed securely on encrypted data.

6.7.2.1 Cryptosystem and Notation

In the following, we will rely on three different additively homomorphic public-key cryptosystems. An additively homomorphic public-key encryption scheme allows, given the two encrypted messages $Enc(a)$ and $Enc(b)$, to compute $Enc(a + b)$ using a public-key operation on the encrypted messages. Moreover, one of our cryptosystems is a leveled fully homomorphic encryption, which also allows performing a bounded number of multiplications in sequence, i.e., to compute $Enc(a \cdot b)$ on the encrypted messages. Bounded means that the cryptographic scheme allows to evaluate polynomials only up to a certain multiplicative depth L. Below, we list the cryptosystems we use and also mention the corresponding plaintext spaces M:

1. the QR (Quadratic Residuosity) cryptosystem of Goldwasser-Micali [50] ($M = \mathbb{F}_2$, bits),

2. the Paillier cryptosystem [96] ($M = \mathbb{Z}_N$ with N being the public modulus of Paillier),

3. a leveled fully homomorphic encryption (FHE) scheme based on the Brakerski-Gentry-Vaikuntanathan [14] scheme as implemented by HELib [55] ($M = \mathbb{F}_2$).

We denote the client in our protocols by C and the server by S. $[b]_A$ denotes a bit b encrypted by the QR scheme under party A's key (so only A can decrypt the message using her secret key). Similarly, $[\![m]\!]_A$ denotes an integer m encrypted by the Paillier scheme, and $[\![\![b]\!]\!]_A$ denotes a bit b encrypted by the leveled FHE scheme. SK_A^s is used for party A's secret key for the encryption scheme Paillier ($s = P$), QR ($s = QR$) or leveled FHE ($s = FHE$), and PK_A^s is the respective public key. For a distribution D, $a \leftarrow D$ means that we assign a a random sample from that distribution.

6.7.2.2 Cryptographic Assumptions and Adversarial Model

The security of our protocol relies on the semantic security [49] of the cryptosystems we use and, hence, also on the well-studied assumptions underlying those systems, namely the Quadratic Residuosity assumption, the Decisional Composite Residuosity assumption, and the Ring Learning With Errors (RLWE) assumption.

We prove our protocol to be secure in the two-party computation framework for passive adversaries (or honest-but-curious [49]), by relying on modular sequential composition of smaller protocols as described below.

6.7.2.3 Building Blocks

Specifically, we will reuse existing building blocks from the work of Bost et al. and also design a new one that is needed for our protocol: changing encryption ownership. Their work already introduced several smaller building blocks, such as different comparison protocols on encrypted data, or a protocol to evaluate the arg max function on encrypted data. Those building blocks that are necessary for our own construction are briefly reviewed hereunder before we introduce our own building blocks as well as the full construction.

Comparison Protocols. Bost et al. introduce five slightly different comparison protocols, two of which we will need in our construction. Let A, B be two parties. A has $\mathsf{PK}_B^P, \mathsf{PK}_B^{QR}$ and B has the corresponding secret keys $\mathsf{SK}_B^P, \mathsf{SK}_B^{QR}$.

The first comparison protocol (referred to as (i) later) assumes that A has two values $[\![a]\!]_B, [\![b]\!]_B$. This protocol then allows comparing a and b, so that A learns $[a \leq b]_B$, and B learns nothing about the comparison.

The second comparison protocol, (ii), works the same way, the only difference being that B also learns $a \leq b$.

More details, as well as the other comparison protocols, can be found in [13].

arg max on Encrypted Data. Based on their comparison protocol (ii), Bost et al. develop a protocol to compute the arg max on encrypted data. Let A, B be two parties.

A has k encrypted values $([\![a_1]\!]_B, \ldots, [\![a_k]\!]_B)$ (where k is also known to B) and wants to know the arg max over unencrypted values (i.e., the index i of the largest value a_i), but neither party should learn anything else.

Hence, this protocol allows to compute $\arg\max_{1 \leq i \leq k} a_i$, given only the values encrypted under B's key. In particular, during the computation, B should neither learn the values a_i, nor should B learn the order relations between the a_i's. The full details of this protocol are described in [13].

Changing the Encryption Scheme. In order to convert ciphertexts from one of the cryptosystems to another, Bost et al. rely on a simple protocol to change the encryption scheme. Since this protocol is crucial for essential parts of our construction, we will provide a more detailed description of the protocol.

First, we consider the case, for which $M_{s_1} = M_{s_2} = \mathbb{F}_2$, i.e., the two cryptosystems have the same message space: Let A, B be two parties, A having $\mathsf{PK}_B^{s_1}, \mathsf{PK}_B^{s_2}$ and a ciphertext $c = Enc_{s_1}(x)$. B has the corresponding secret keys $\mathsf{SK}_B^{s_1}, \mathsf{SK}_B^{s_2}$. The goal is to re-encrypt x using the cryptosystem s_2, without B learning x.

Intuitively, the protocol works as follows: First, A uniformly picks a random noise $r \leftarrow M_{s_1}$, encrypts it using $\mathsf{PK}_B^{s_1}$ and adds it to the ciphertext c, before sending the result to B. B then decrypts the ciphertext to $x + r \in M_{s_1}$, re-encrypts it using $\mathsf{SK}_B^{s_2}$ and sends $Enc_{s_2}(x + r)$ to A, who can strip off r using the homomorphic property of s_2. B only obtains $x + r$, which hides x information-theoretically (this can be seen as a one-time pad).

In our construction, for $M_{s_1} \neq M_{s_2}$, we only require the transformation from $M_{s_1} = \mathbb{F}_2$ to $M_{s_2} = \mathbb{Z}_N$, i.e., from FHE to Paillier. Hence, we will only discuss this case in more detail. In this case, the beginning of the protocol remains the same, and A obtains $[\![x \oplus r]\!]_B$ with $x, r \in \mathbb{F}_2$. The important difference compared to the previous case now arises when A wants to strip off $r \in M_{s_1} = \mathbb{F}_2$ from the encryption. Since the additive operation on $\mathbb{F}_2$ is $\oplus$ and on $\mathbb{Z}_N$ is $+$, we have to emulate $\oplus$ in Paillier's message space. This can be easily done by computing:

$$[\![x]\!]_B = \begin{cases} [\![x \oplus r]\!]_B & \text{if } r = 0 \\ g([\![x \oplus r]\!]_B^{-1}) \mod N^2 & \text{if } r = 1 \end{cases}$$

Before giving the result to an adversary, who knows $[\![x \oplus r]\!]_B$, but not SK_B^P, the obtained result has to be refreshed to preserve semantic security. A pseudocode implementation as well as the security and correctness proofs of this protocol can be found in [13].

Private Evaluation of Classification Trees. The most useful protocol to us is the one for privately evaluating a classification tree. Here, the main idea is to represent the classification tree as a polynomial P, the output of which is the result of the classification.

Let b_i be the boolean outcome of a comparison between the input value v_j of the ith node and the corresponding threshold w_i, i.e., $w_i < v_j$. Then, given the class labels $Y = \{y_0, \ldots, y_k\}$, one can express a classification tree by a polynomial.

The polynomial is constructed recursively by a procedure $\mathcal{F}(T)$. If T is a leaf node, $\mathcal{F}(T) = y$, where y is the class label at the leaf T. If T is an internal node, and T_1 is the child tree in case the corresponding b is true, and T_2 is the child tree in case b is false, then $\mathcal{F}(T) = b\mathcal{F}(T_1) + (1-b)\mathcal{F}(T_2)$ is the polynomial that evaluates T_1 if b and T_2 otherwise. For the example in Figure 6.6, this polynomial would be $P(b_0, b_1, b_2, y_0, y_1, y_2) = b_0(b_1 \cdot y_1 + (1-b_1)(b_2 \cdot y_2 + (1-b_2)y_1)) + (1-b_0)y_0$.

Using this polynomial, Bost et al. then introduce a protocol to evaluate the tree, while revealing only the outcome and the number of comparisons. Let S and C denote the server and client respectively. First, S and C make use of the comparison protocol (i), so that S learns the bits $[b_i]_C$ for every node. Then, they interact in the protocol to change the encryption scheme from QR to FHE, thus obtaining $[\![b_i]\!]_C$.

The server S can then evaluate the polynomial P using the homomorphic properties of the FHE scheme. However, since the plaintext space is only $\mathbb{F}_2$ and the class labels potentially take more than one bit, we would have to evaluate the polynomial for each bit individually. Fortunately, the so-called SIMD slots of the FHE scheme (described in detail in [112]) allow the scheme to encrypt a vector of bits in one ciphertext and evaluate the polynomial on the whole vector at once, in parallel. Hence, for each class label y_i, the server encrypts its bit representation $y_{i0}, \ldots, y_{il}$ using these SIMD slots to $[\![y_{i0}, \ldots, y_{il}]\!]_C$ and can evaluate the polynomial for each bit in parallel.

The client can later decrypt the resulting class label and convert it back to the normal integer representation. A more detailed explanation as well as proofs of correctness can be found in [13].

Changing Encryption Owner. Next, we will introduce our own protocol to change the ownership of an encryption, which we will need in order to apply the arg max protocol in a way that only the client learns the result of the plurality vote.

Given two parties A and B, out of which A holds the encrypted message $[\![x]\!]_B$, we want B to hold the same encrypted message, but this time under A's key. However, neither A nor B should learn the message x itself. In the following, we design a protocol to meet this goal and give the security proof in Section 6.9.

Let A have $\mathsf{PK}_B^P, \mathsf{SK}_A^P, [\![x]\!]_B$ and B have $\mathsf{SK}_B^P, \mathsf{PK}_A^P$. Then A first blinds the encrypted message by uniformly sampling a random noise r from the plaintext space, encrypting it and adding it to the ciphertext. Then, A also encrypts r using her own secret key and sends both $[\![x+r]\!]_B$ and $[\![r]\!]_A$ to B. B then decrypts the first ciphertext to $x+r$, which hides x in an information-theoretic way and encrypts it again, using PK_A^P. Then B strips off r using the sent encryptions without learning r itself and obtains $[\![x]\!]_A$.

The complete protocol is shown in Algorithm 6.1.

Theorem 4. *The protocol in Algorithm 6.1 is secure in the honest-but-curious model.*

The proof of the theorem is given in Section 6.9, since we introduce additional notation not required to understand the general idea. Thus, we decided to postpone the proof and first focus on our protocol instead.

Algorithm 6.1 Changing Encryption Owner

Input: $A : (\llbracket x \rrbracket_B, \mathsf{SK}_A^P, \mathsf{PK}_B^P)$, $B : (\mathsf{PK}_A^P, \mathsf{SK}_B^P)$
Output: $B : \llbracket x \rrbracket_A$

1: A: uniformly pick a random noise $r \leftarrow M_P = \mathbb{Z}_N$ (Paillier's message space), encrypt it using PK_B^P and compute $\llbracket x + r \rrbracket_B$
2: A: encrypt r using SK_A^P to $\llbracket r \rrbracket_A$
3: A: send $(\llbracket x + r \rrbracket_B, \llbracket r \rrbracket_A)$ to B
4: B: decrypt $\llbracket x + r \rrbracket_B$ to get $x + r$ and encrypt it using PK_A^P to $\llbracket x + r \rrbracket_A$
5: B: compute $\llbracket x \rrbracket_A = \llbracket x + r \rrbracket_A \cdot \llbracket r \rrbracket_A^{-1}$ using the homomorphic property

6.7.2.4 Private Random Forests

Now that we introduced all building blocks necessary to privately evaluate a random forest, we first give an intuition of our protocol before presenting its pseudocode in Algorithm 6.2.

Naively, one could just evaluate each tree of a random forest individually, given the protocol introduced by Bost et al., and return the outcomes to the client. The client is then able to compute the plurality vote or any metric she is interested in. This, however, will not only leak the number of trees but most likely also the number of nodes within each tree to the client. Indeed, the scheme of Bost et al. reveals the number of comparisons, thus the number of inner nodes to the client. We modify this idea to only leak the total number of trees and the total number of nodes. Moreover, we extend it by giving the option to only reveal the plurality-vote class to the client. Thus, we do not evaluate one tree after another, but we perform the evaluations of all trees in a batch, e.g., running the comparison protocol for the b_i's of all trees in a row. This way, the client cannot distinguish between different trees during the evaluation.

In order to allow the protocol to only reveal the plurality-vote class, we have to modify the protocol further. Intuitively, for the server S to determine the plurality-vote class, S needs to be able to count the votes for each class without learning the actual outcomes of the trees. We can achieve this by slightly changing the way the class labels are encoded into the SIMD slots: Instead of encoding each integer class label as its binary representation, we encode a class label y_i by only setting the ith bit to 1. While encoding k labels into a binary representation needs only $\lceil log_2(k) \rceil + 1$ bits, our method will take exactly k bits. However, if enough SIMD slots, compared to the number of classes, are available, this should not have a substantial effect on the protocol performance. More specifically, a class label y_i is now encoded as $(y_{i1}, \ldots, y_{ik})$ with $y_{ij} = 1$ if $i = j$ and 0 otherwise.

After obtaining the outcomes of all trees, the server and client interact to change the encryption schemes from FHE to Paillier, resulting in Paillier ciphertexts for each outcome and class label $\llbracket y_{ij} \rrbracket_C$ for $i \in \{1, \ldots, n\}, j \in \{1, \ldots, k\}$, where $y_{ij} = 1$ if the outcome of the ith tree was class j and $y_{ij} = 0$ otherwise. This encoding allows to sum up all votes for each class (or vote count), so that the server obtains $\llbracket \sum_{i=1}^{n} y_{ij} \rrbracket_C$ using Paillier's homomorphic property.

However, we cannot directly apply the arg max protocol as this would reveal the

Algorithm 6.2 Evaluate a Random Forest

Input: Client C : $(\mathsf{SK}_C^P, \mathsf{SK}_C^{QR}, \mathsf{SK}_C^{FHE}, \mathsf{PK}_S^P, \mathbf{v})$, Server S : $(\mathsf{PK}_C^P, \mathsf{PK}_C^{QR}, \mathsf{PK}_C^{FHE}, \mathsf{SK}_S^P, \mathcal{F} = \{t_1, \ldots, t_n\})$

Output: Client C : the outcome of evaluating $\mathcal{F}$ on $\mathbf{v}$ in terms of a plurality vote or the individual votes

1: S: produces the polynomials $P_1, \ldots, P_n$ for each tree in $\{t_i\}_{i=1}^n$
2: C: sends the encrypted query $[\![v_0]\!]_C, \ldots, [\![v_m]\!]_C$ to S
3: S and C perform the comparison protocol (i) on a shuffled order of the nodes, so that S obtains $[b_i]_C$ for every node in the trees
4: S: changes the encryption obtaining $[\![b_i]\!]_C$
5: S: computes each class label y_i by setting only the ith bit to 1 and encrypts the class labels using FHE and SIMD slots to $[\![y_{i1}, \ldots, y_{ik}]\!]_C$ with $y_{ij} = 1$ if $i = j$ and 0 otherwise
6: S: evaluates the polynomials using the fully homomorphic encryption, obtaining the encrypted outcomes $\{[\![y_{j1}, \ldots, y_{jk}]\!]_C\}_{j=1}^n$ for each tree
7: **if** C is allowed to get all individual outcomes **then**
8: S: rerandomizes the encrypted outcomes, shuffles their order and sends them to C, who can decrypt them
9: **else**
10: S: rerandomizes the encrypted outcomes and changes their encryption scheme to Paillier, resulting in $[\![y_{ij}]\!]_C$ for $i \in \{1, \ldots, n\}, j \in \{1, \ldots, k\}$
11: S: sums the bits for each class separately, obtaining $[\![\sum_{i=1}^n y_{ij}]\!]_C = \sum_{i=1}^n [\![y_{ij}]\!]_C$ for every $j \in \{1, \ldots, k\}$, effectively computing the vote counts of each class
12: S and C change the ownership of the vote counts, so that C obtains $[\![\sum_{i=1}^n y_{ij}]\!]_S$ using our protocol
13: C and S perform the arg max protocol, so that C learns only the outcome of the plurality-vote class
14: **end if**

classification result to the party holding the ciphertexts, i.e., the server. Hence, we leverage our encryption ownership protocol to transfer the vote counts to the client under the server's key. The client thus has $[\![\sum_{i=1}^n y_{ij}]\!]_S$, which allows him to determine the plurality-vote class by applying the arg max protocol.

The complete protocol is provided in Algorithm 6.2.

Theorem 5. *The protocol presented in Algorithm 6.2 is secure in the honest-but-curious model.*

We refer again to Section 6.9 for the proof.

6.8 Evaluation of the Private Classifier

Now that we have introduced our protocol for private classification on random forests, we will evaluate its performance on a dataset and classifier used in practice. More specifically, we base our performance evaluation on MethPed [22, 1], a random forest classifier for

the identification of pediatric brain tumor subtypes based on DNA methylation data, which is available as an R package. From this package, we extract their random forest model and feed it into our protocol implementation for the performance evaluation.[1]

MethPed, in its standard configuration, trains a random forest model of 1000 trees based on its original training data, consisting of 472 clinically diagnosed brain tumor cases after data cleaning and k-nearest neighbor imputation of missing values [1]. The DNA methylation samples have been collected from several datasets, all of which are publicly available on the GEO database (GEO accession numbers GSE50022, GSE55712, GSE36278, GSE52556, GSE54880, GSE45353 and GSE44684). The random forest is then trained on a total of 900 methylation sites that were shown to yield the highest predictive power in a large number of regression analyses.

Our protocol implementation is based on the original implementation of the work of Bost et al.[2]. We extended it by implementing the protocol for changing the encryption scheme from FHE to Paillier, as well as by adding our own protocol for changing the ownership of the encryption. Moreover, we fully implemented the random forest classification protocol (Algorithm 6.2) and tested its correctness on sample inputs. Then, we ported the MethPed classifier into our implementation and included two methylation samples to evaluate the classifier on. The implementation of our private random forest classifier is written in C++ using GMP [52, 48], Boost [12], Google's Protocol Buffers [51], and HELib [55]. The source code of our implementation can be found at `https://github.com/paberr/ciphermed-forests`.

In order to represent the methylation levels as integers in our protocol, we multiply them by 10^8 and store the result as an integer. Since the data we used is available at a precision of eight digits after the decimal point and methylation values are bounded by the range $[0, 1]$, we do not lose any precision.

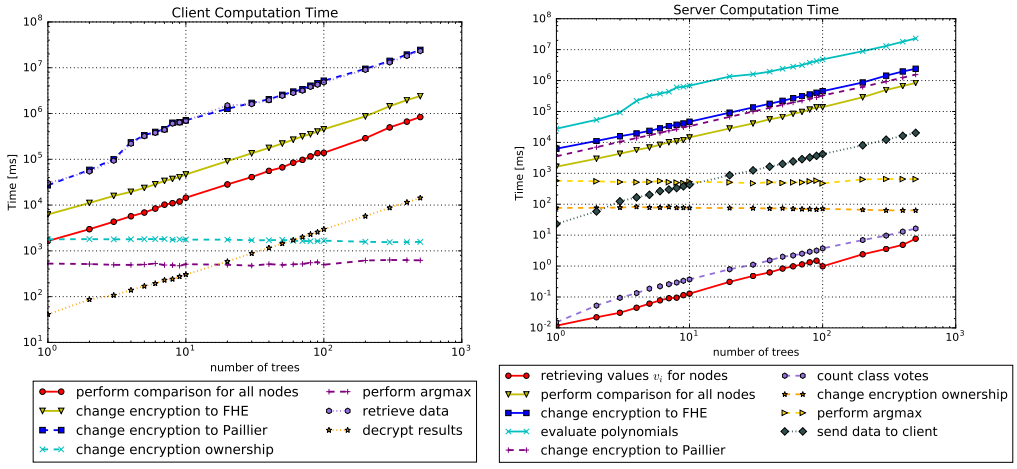
6.8.1 Evaluation Setup

To evaluate the performance of our protocol, we ran the client and server of the classification task on different machines, both, on the same network and on different networks. One client was run on a local computing server with approximately 775 GB RAM and four Intel Xeon E5-4650L processors, providing 64 cores (with hyperthreading enabled) running at 2.60 GHz. Another client was run on an Amazon AWS instance of the type `r4.2xlarge` with 61 GB RAM and 8 Intel Xeon E5-2686 v4 vCPUs and a network bandwidth up to 10 gigabits located in Frankfurt, Germany. The server was run on a local computing server with approximately 1.55 TB RAM and four Intel Xeon E7-8867 processors, providing 128 cores (with hyperthreading enabled) running at 2.50 GHz. Since our current implementation does not make use of any multithreading technique, we used the large number of cores to run multiple experiments, i.e., classification tasks, at once.

Similar to Bost et al., we also used 1024-bit cryptographic keys and chose the statistical security parameter λ to be 100. HELib was configured to use 80 bits of

[1] The R implementation and the used methylation sites are available at `http://bioconductor.org/packages/devel/bioc/html/MethPed.html`.

[2] Available at `https://github.com/rbost/ciphermed`.



(a) (b)

Figure 6.7: Duration of different protocol steps on the (a) client and (b) server side for varying number of trees and both protocol variations.

security, roughly corresponding to a 1024-bit asymmetric key [13].

6.8.2 Performance Evaluation

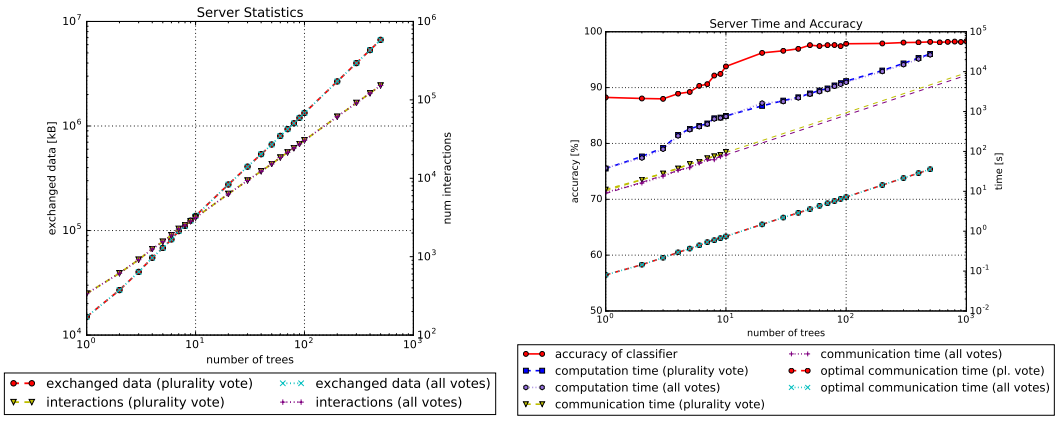
We evaluate our protocol for a varying number of trees

$$n \in \{1, 2, \ldots, 9, 10, 20, \ldots, 90, 100, 200, \ldots, 400, 500\}$$

and two independent classification queries provided in the MethPed R package [1]. We restricted the number of trees to a maximum of 500 in order to keep the computational costs low. We still can estimate the cost of running our protocol with 1000 trees by the general trend as seen in the following. Moreover, we evaluate both versions of our protocol, the first revealing only the plurality-vote class to the client, and the second revealing one outcome per tree to the client. For $n \leq 100$, we classify each of the samples five times, resulting in a total of 10 executions for each of our protocol instantiations. For $n > 100$, we classify each of the samples only once, due to the increased computational costs. The trees used for the classification consist of between 16 and 37 inner nodes, with an average of around 25 inner nodes.

In the following figures, a solid line is used for operations common to both of our protocol instantiations, a dashed line is used for the instantiation returning the plurality-vote class, and a dotted line is used for the one outputting the outcome for each tree. The performance evaluation of common operations groups together the results of both instantiations, yielding 20 executions if $n \leq 100$, and 4 executions if $n > 100$.

Figure 6.7(a) depicts the performance evaluation on the client side, both axes scaled logarithmically. Generally, the computational costs of most of our protocol steps scale approximately linearly in the number of trees. Only changing the ownership of the encryption and performing the arg max seem to have a constant execution time. These



(a) (b)

Figure 6.8: (a) Data exchange and number of interactions for varying number of trees and both protocol variations. (b) Total duration of a classification task and accuracy of the random forest for varying number of trees and both protocol variations.

two blocks scale linearly with the number of class labels, which are fixed (to the 9 types of brain tumors) in our experiments.

Next, we compare the execution time of both protocol instantiations. We see that changing the encryption scheme of the outcomes from FHE to Paillier and retrieving all the outcomes in the FHE cryptosystem take almost the same amount of time, since essentially the same operations are required. Performing the plurality vote protocol then only adds a constant computational burden on the client's side, thereby increasing the total computation time only to a negligible degree.

In Figure 6.7(b), we analyze the same scenarios on the server side. Unsurprisingly, the relationships between the number of trees in the random forest and the computational costs are the same as for the client. It is worth noting that the computationally most expensive operation is by far the FHE evaluation of the polynomials. Evaluating the polynomials takes almost an order of magnitude more time than the second most expensive protocol step. Thus, minimizing the number of trees and potentially also the number of inner nodes is a major concern when applying our protocol. Moreover, parallelizing the evaluation of the polynomials is a possible improvement, which we did not explore in our implementation.

In terms of the amount of exchanged data and the number of interactions, both protocol instantiations seem to be more or less equivalent as shown in Figure 6.8(a). Revealing the individual outcomes to the client is not noticeably different from performing the plurality vote protocol. While time is certainly the primary concern when running a classification task, the amount of data exchanged over the network should not be underestimated. For example, evaluating 50 trees involves transferring around 0.67 GB of data over the network. Increasing the number of trees to 100 involves around 1.33 GB of data exchange.

Finally, in Figure 6.8(b), we study the total time to run the protocol on the server side (excluding the time for sending packets over the network) in comparison with the

accuracy of the random forest built on the given number of trees. The accuracy was determined based on the out-of-bag samples during the training phase and averaged over 10 different runs. Since our private classification uses the same precision for the methylation values as the R implementation and builds on exactly the same trees, the accuracy provided by our private classification technique is also the same. While the computational costs clearly increase approximately linearly in the number of trees, the accuracy does not. While 1000 trees provide an accuracy of 98.3%, 50 trees are already sufficient to provide an accuracy of 97.6% at only an estimated 5% of the computational cost. We also depict the communication time between our Amazon AWS instance and the local computing server for a smaller range of number of trees. Evaluating 50 trees takes in total less than an hour, even when including the time for sending and receiving packets over the internet. We also evaluated the timing on the client's side, which exhibits the same behaviour as on the server's side.

We emphasize that our current implementation does neither aim at minimizing the number of interactions, nor does it make use of pipelining of interactions. Based on the measured throughput between the Amazon AWS instance and our computing server, we additionally depict the estimated optimal communication time over the network in Figure 6.8(b). Improving the transmission of data can potentially decrease the communication time for 500 trees down to 50 seconds.

Since, in the current medical scenario, it usually takes at least one day for a laboratory to analyze a sample, we assume a similar computational limit on the classification. Given such a limit, we conclude that a laboratory offering the privacy-preserving analysis using our protocol would be able to provide a good trade-off between computational costs and accuracy. Moreover, the structure of random forests offers a great potential to parallelize some of the operations (e.g., the polynomial evaluation), which we leave to future research.

We note that both protocol instantiations take approximately the same time to run. While returning the selected class for a number of 50 trees is about 2 minutes faster than returning the plurality vote, this difference only accounts to about 6 minutes for 100 trees and to about 23 minutes for 500 trees. Hence, we suggest selecting the instantiation based on the output the client needs and the information the server agrees to reveal. If the client wants a fine-grained output to analyze the distribution of the different classes, then he may request to get access not to the plurality-vote class, but to the selected class of each tree. However, this will leak more information about the underlying random forest model than only disclosing the plurality-vote class.

6.9 Proofs

This section is devoted to proving the security of our scheme in the presence of honest-but-curious participants. Although we assume the same security model as in the work by Bost et al. [13], we commence discussing the necessary concepts.

6.9.1 Secure Two-party Computation Framework

Both, our protocol to change the ownership of an encryption and the protocol to privately evaluate a random forest model are two-party protocols. Let the two parties be denoted by A and B. In order to show that all computations are done privately, we assume the honest-but-curious (semi-honest) model as described in [49].

Let $f = (f_A, f_B)$ be a (probabilistic) polynomial function and Π be a protocol computing f. Using A's input a and B's input b, the two parties want to compute $f(a, b)$ by applying the protocol Π with the security parameter λ.

We denote the view of a party $P \in \{A, B\}$ during the execution of Π by the tuple $V_P(\lambda, a, b) = (1^\lambda; a; r^P; m_1^P, \ldots, m_t^P)$ where r is P's random tape and $m_1^P, \ldots, m_t^P$ are the messages received by P. We define the outputs of parties A and B for the execution of Π as $\mathsf{Out}_A^\Pi(\lambda, a, b)$ and $\mathsf{Out}_B^\Pi(\lambda, a, b)$. The global output is defined as the tuple $\mathsf{Out}^\Pi(\lambda, a, b) = (\mathsf{Out}_A^\Pi(\lambda, a, b), \mathsf{Out}_B^\Pi(\lambda, a, b))$.

To ensure the private, secure computation, we require that whatever A can compute from its interactions with B can be computed from its input and output, yielding the following security definition.

Definition 4. *A two-party protocol Π securely computes the function f if there exist two probabilistic polynomial time algorithms S_A and S_B (also called simulators) such that for every possible input a, b of f,*

$$\{S_A(1^\lambda, a, f_A(a, b)), f(a, b)\} \equiv_c \{V_A(\lambda, a, b), \mathsf{Out}^\Pi(\lambda, a, b)\}$$

and

$$\{S_B(1^\lambda, b, f_B(a, b)), f(a, b)\} \equiv_c \{V_B(\lambda, a, b), \mathsf{Out}^\Pi(\lambda, a, b)\}.$$

$\equiv_c$ means computational indistinguishability against probabilistic polynomial time adversaries with negligible advantage in the security parameter λ.

6.9.2 Cryptographic Assumptions

In this section, we briefly review the cryptographic assumptions underlying the cryptosystems we use.

Assumption 1 (Quadratic Residuosity Assumption [50])**.** *Let $N = p \times q$ be the product of two distinct odd primes p and q. Let $\mathbb{QR}_N$ be the set of quadratic residues modulo N and $\mathbb{QNR}_N = \{x \in \mathbb{Z}_N^* \mid x$ is not a quadratic residue modulo N, but $\mathcal{J}_N(x) = +1\}$ be the set of quadratic non residues, where $\mathcal{J}_N(x)$ is the Jacobi symbol.*

$\{(N, \mathbb{QR}_N) \mid |N| = \lambda\}$ and $\{(N, \mathbb{QNR}_N) \mid |N| = \lambda\}$ are computationally indistinguishable with respect to probabilistic polynomial time algorithms.

Assumption 2 (Decisional Composite Residuosity Assumption [96])**.** *Let $N = p \times q$ with $|N| = \lambda$ be the product of two distinct odd primes p and q. We call z a Nth residue modulo N^2 if there exists $y \in \mathbb{Z}_{N^2}$ such that $z = y^N \mod N^2$. Nth residues and non Nth residues are computationally indistinguishable with respect to probabilistic polynomial time algorithms.*

Assumption 3 (RLWE [14])**.** *Let $f(x) = x^d + 1$ where $d = d(\lambda)$ is a power of 2. Let $q = q(\lambda) \geq 2$ be an integer. Let $R = Z[x]/(f(x))$ and let $R_q = R/qR$. Let $\chi = \chi(\lambda)$ be a distribution over R. The $RLWE_{d,q,\chi}$ problem is to distinguish between two distributions: In the first distribution, one samples (a_i, b_i) uniformly from R_q^2. In the second distribution, one first draws $s \leftarrow R_q$ uniformly and then samples $(a_i, b_i) \in R_q^2$ by sampling $a_i \leftarrow R_q$ uniformly, $e_i \leftarrow \chi$, and setting $b_i = a_i \cdot s + e_i$. The $RLWE_{d,q,\chi}$ assumption is that the $RLWE_{d,q,\chi}$ problem is infeasible.*

6.9.3 Modular Sequential Composition

In order to ease the security proof of our construction, we rely on sequential modular composition as defined in [18]. The idea is that two parties run a protocol Π and use calls to an ideal functionality f while running Π. This can be imagined as A and B privately computing f by sending their inputs to a trusted third party T and receiving the results from it. If we can now show that Π respects security and privacy in the honest-but-curious model and if we have a protocol ρ that securely and privately computes f in the same model, we can replace f by executions of ρ in Π. The resulting protocol Π^ρ is then still secure in the aforementioned model.

We call $(f_1, \ldots, f_m)$-*hybrid model* the semi-honest model augmented with an incorruptible trusted party T for evaluating the functionalities. The parties A and B run a protocol Π that contains calls to T for these functionalities. For each call, the parties send their input to T and wait until they receive the respective results. It is crucial that both parties must not communicate until receiving the result since we only consider sequential composition here. T does not keep state between different calls to the functionalities. Therefore the protocol may contain multiple calls even for the same function, which are all independent.

Let Π be a two-party protocol in the $(f_1, \ldots, f_m)$-hybrid model and $\rho_1, \ldots, \rho_m$ be secure protocols in the semi-honest model computing $f_1, \ldots, f_m$. We define $\Pi^{\{\rho_1,\ldots,\rho_m\}}$ as the protocol where all ideal calls of Π have been replaced by executions of the corresponding protocol: if party P_j needs to compute f_i with input x_j, it halts, starts an execution of ρ_i with the other party, gets the result β_j from ρ_i and continues as if β_j was received from T.

Theorem 6 (Modular Sequential Composition Theorem [18, 81])**.** *Let $f_1, \ldots, f_m$ be two-party probabilistic polynomial time functionalities and $\rho_1, \ldots, \rho_m$ be protocols that compute respectively $f_1, \ldots, f_m$ in the presence of semi-honest adversaries.*

Let g be a two-party probabilistic polynomial time functionality and Π a protocol that securely computes g in the $(f_1, \ldots, f_m)$-hybrid model in the presence of semi-honest adversaries.

Then $\Pi^{\rho_1,\ldots,\rho_m}$ securely computes g in the presence of semi-honest adversaries.

6.9.4 Proof of Changing Encryption Owner Protocol

Proof of Theorem 4. The function f this protocol computes is:

$$f((\llbracket x \rrbracket_B, \mathsf{SK}_A, \mathsf{PK}_B), (\mathsf{PK}_A, \mathsf{SK}_B)) = (\emptyset, \llbracket x \rrbracket_A)$$

For the sake of simplicity, we do not take into account the randomness used for the encryptions of r for A and $c' = x + r$ for B. The distribution of these coins for one party is completely independent of the other elements taken into account in the simulations, so we omit them in our security proof.

A's view is $V_A = (\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B; r; \emptyset)$. A does not output anything. The simulator $S_A(\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B)$ runs as follows:

1. Picks uniformly at random $\tilde{r} \leftarrow M_P$.
2. Outputs $(\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B; \tilde{r}; \emptyset)$

Since r and $\tilde{r}$ are sampled from the same distribution, independent from any other parameter,

$$\{(\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B; \tilde{r}; \emptyset), f([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\} = \{(\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B; r; \emptyset), f([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\}.$$

Moreover, it holds that

$$\{(\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B; r; \emptyset), f([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\} = \{(\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B; r; \emptyset), (\emptyset, [\![x]\!]_A)\}$$

and we can conclude

$$\{S_A(\mathsf{SK}_A, \mathsf{PK}_B, [\![x]\!]_B), f([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\} \equiv_c \{V_A([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B), \mathsf{Out}([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\}.$$

B's view is $V_B = (\mathsf{PK}_A, \mathsf{SK}_B; [\![x + r]\!]_B, [\![r]\!]_A)$. B outputs $[\![x]\!]_A$. We build a simulator $S_B(\mathsf{PK}_A, \mathsf{SK}_B)$ as follows:

1. Pick uniformly at random $\tilde{r} \leftarrow M_P$ and $\tilde{c} \leftarrow M_P$.
2. Generate the encryptions $[\![\tilde{r}]\!]_A$ and $[\![\tilde{c}]\!]$ using PK_A.
3. Output $(\mathsf{PK}_A, \mathsf{SK}_B; [\![\tilde{c}]\!]_B, [\![\tilde{r}]\!]_A)$

By semantic security of the encryption scheme (in our concrete case the Paillier cryptosystem), it holds that (proof see below)

$$\{(\mathsf{PK}_A, \mathsf{SK}_B; [\![\tilde{c}]\!]_B, [\![\tilde{r}]\!]_A), f([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\} \equiv_c \tag{6.3}$$

$$\{(\mathsf{PK}_A, \mathsf{SK}_B; [\![x + r]\!]_B, [\![r]\!]_A), f([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\} \tag{6.4}$$

and hence (using also the correctness of the scheme)

$$\{S_B(\mathsf{PK}_A, \mathsf{SK}_B), f([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\} \equiv_c \{V_B([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B), \mathsf{Out}([\![x]\!]_B, \mathsf{SK}_A, \mathsf{PK}_B, \mathsf{PK}_A, \mathsf{SK}_B)\}.$$

We will prove the computational indistinguishability of (4) and (5) in more detail by giving a reduction to the semantic security. To this end, we assume that we have a distinguisher $\mathcal{D}$ that can distinguish (4) and (5). In particular, given

$$\{(\mathsf{PK}, \mathsf{SK}', [\![y]\!]_{\mathsf{SK}'}, [\![r]\!]_{\mathsf{SK}}), [\![x]\!]_{\mathsf{SK}}\}$$

$\mathcal{D}$ outputs 1 if y, r and x are independent uniformly random values and 0 if $r = y - r'$ for a random r' and $x = y - r = r'$. Then, we construct a reduction $\mathcal{R}$ as follows:

1. On input PK, generate a new key pair $(\mathsf{SK}', \mathsf{PK}') \leftarrow \mathsf{KeyGen}(1^\lambda)$.
2. Pick uniformly at random $y, \tilde{r} \leftarrow M$.
3. Choose challenger messages $m_0 = y - \tilde{r}$, $m_1 = \tilde{r}$ and give them to the semantic security challenger.
4. Receive c from the challenger, compute $[\![\tilde{r}]\!]_{\mathsf{PK}}$ and query the distinguisher

   $$\mathcal{D}(\{(\mathsf{PK}, \mathsf{SK}', [\![y]\!]_{\mathsf{SK}'}, c), [\![\tilde{r}]\!]_{\mathsf{PK}}\}),$$

   which returns b.
5. Return b to the challenger.

Since we simulate both cases ((4) and (5)) perfectly to the distinguisher, its success probability in distinguishing (4) and (5) transfers exactly to our reduction in the semantic security game. Since Paillier encryption is shown to be semantically secure under the Decisional Composite Residuosity Assumption, the distinguisher must have not more than negligible success probability. And hence our scheme is secure. □

6.9.5 Proof of Private Random Forest Evaluation Scheme

The correctness of our protocol follows from the correctness of the private classification tree protocol in [13]. Moreover, we will provide a security proof for the protocol revealing only the plurality-vote class. Since our second protocol instantiation – revealing all trees' outcomes – is essentially only a shorter version of the main protocol, we do not provide a separate security proof for this protocol.

Proof of Theorem 5. Let A be the server S and B be the client C. We prove the security of our protocol (see Algorithm 6.2) in the hybrid model using the following 5 ideal functionalities, which we let execute by a trusted third party:

- the comparison protocol in step 3:
  $f_1([\![x]\!]_B, [\![y]\!]_B, l, \mathsf{SK}_B^{QR}, \mathsf{PK}_B^{QR}, \mathsf{SK}_B^P, \mathsf{PK}_B^P) = ([x \leq y]_B, \emptyset)$
- the protocol to change the encryption scheme in step 4:
  $f_2([b]_B, \mathsf{SK}_B^{QR}, \mathsf{PK}_B^{QR}, \mathsf{SK}_B^{FHE}, \mathsf{PK}_B^{FHE}) = ([\![\![b]\!]\!]_B, \emptyset)$
- the protocol to change the encryption scheme in step 10:
  $f_3([\![\![y_1, \ldots, y_k]\!]\!]_B, \mathsf{SK}_B^{FHE}, \mathsf{PK}_B^{FHE}, \mathsf{SK}_B^P, \mathsf{PK}_B^P) = (\{[\![y_1]\!]_B, \ldots, [\![y_k]\!]_C\}_{i=1}^k, \emptyset)$

- the protocol to change the ownership of the encryption in step 12:
  $f_4([\![x]\!]_B, \mathsf{SK}_A^P, \mathsf{PK}_B^P, \mathsf{PK}_A^P, \mathsf{SK}_B^P) = (\emptyset, [\![x]\!]_A)$
- the arg max protocol in step 13:
  $f_5(\{[\![a_i]\!]_A\}_{i=1}^k, l, \mathsf{SK}_A^P, \mathsf{PK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_A^{QR}) = (\emptyset, \arg\max_i\{a_i\}_{i=1}^k)$

We will conclude using Theorem 6, our own security proofs for those steps, as well as the proofs in [13].

The whole protocol computes the function:

$$\begin{aligned} f(&\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l,\\ &\mathsf{SK}_A^P, \mathsf{PK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_A^{QR},\\ &\mathsf{SK}_B^P, \mathsf{PK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{PK}_B^{QR},\\ &\mathsf{SK}_B^{FHE}, \mathsf{PK}_B^{FHE}) \end{aligned}$$

where $\{P_i\}_{i=1}^n$ are the polynomials, $\{w_h\}_h$ are the thresholds for each inner node, g is the number of features of the client's sample, $\{[\![v_i]\!]_B\}_{i=1}^g$ is the input by the client. f_A returns nothing, while f_B returns the plurality-vote class of the random forest evaluation.

A's view is now:

$$\begin{aligned} V_A = (&\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l,\\ &\mathsf{SK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_B^P, \mathsf{PK}_B^{QR}, \mathsf{PK}_B^{FHE};\\ &\mathsf{coins};\\ &\{[b_h]_B\}_h, \{[\![b_h, \ldots, b_h]\!]_B\}_h,\\ &\{[\![y_{ij}]\!]_B\}_{i\in\{1,\ldots,n\}, j\in\{1,\ldots,k\}}) \end{aligned}$$

where coins is the random tape for encryptions and $\{[b_h]_B\}_h$ the comparison result for each node. We simulate A's real view with the following simulator S_A:

1. Generate a random bit $\tilde{b}_h$ for each inner node in the random forest.
2. Generate random bits y_{ij} for $i \in \{1, \ldots, k\}, j \in \{1, \ldots, n\}$.
3. Generate a random tape $\widetilde{\mathsf{coins}}$ of the required length. The length can be determined based mainly on the polynomials, which encode the number of trees, number of classes and the number of nodes in the tree.
4. Output

$$\begin{aligned} H_0 = (&\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l,\\ &\mathsf{SK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_B^P, \mathsf{PK}_B^{QR}, \mathsf{PK}_B^{FHE};\\ &\widetilde{\mathsf{coins}};\\ &\{[\tilde{b}_h]_B\}_h, \{[\![\tilde{b}_h, \ldots, \tilde{b}_h]\!]_B\}_h,\\ &\{[\![\tilde{y}_{ij}]\!]_B\}_{i\in\{1,\ldots,n\}, j\in\{1,\ldots,k\}}) \end{aligned}$$

Since $\widetilde{\mathsf{coins}}$ and coins come from the same distribution, H_0 is indistinguishable from:

$$\begin{aligned} H_1 = (&\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l, \\ &\mathsf{SK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_B^P, \mathsf{PK}_B^{QR}, \mathsf{PK}_B^{FHE}; \\ &\mathsf{coins}; \\ &\{[\tilde{b}_h]_B\}_h, \{[\![\tilde{b}_h, \ldots, \tilde{b}_h]\!]_B\}_h, \\ &\{[\![\tilde{y}_{ij}]\!]_B\}_{i\in\{1,\ldots,n\},j\in\{1,\ldots,k\}}) \end{aligned}$$

Moreover, by the semantic security of QR and FHE (we abstain from the trivial reduction proof here), we can deduce that H_1 is computationally indistinguishable from:

$$\begin{aligned} H_2 = (&\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l, \\ &\mathsf{SK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_B^P, \mathsf{PK}_B^{QR}, \mathsf{PK}_B^{FHE}; \\ &\mathsf{coins}; \\ &\{[b_h]_B\}_h, \{[\![b_h, \ldots, b_h]\!]_B\}_h, \\ &\{[\![\tilde{y}_{ij}]\!]_B\}_{i\in\{1,\ldots,n\},j\in\{1,\ldots,k\}}) \end{aligned}$$

And by the semantic security of Paillier, we get that H_2 is computationally indistinguishable from:

$$\begin{aligned} H_3 = (&\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l, \\ &\mathsf{SK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_B^P, \mathsf{PK}_B^{QR}, \mathsf{PK}_B^{FHE}; \\ &\mathsf{coins}; \\ &\{[b_h]_B\}_h, \{[\![b_h, \ldots, b_h]\!]_B\}_h, \\ &\{[\![y_{ij}]\!]_B\}_{i\in\{1,\ldots,n\},j\in\{1,\ldots,k\}}) \end{aligned}$$

Hence, we showed that

$$\begin{aligned} &V_A(\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l, \\ &\quad \mathsf{SK}_A^P, \mathsf{PK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_A^{QR}, \\ &\quad \mathsf{SK}_B^P, \mathsf{PK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{PK}_B^{QR}, \\ &\quad \mathsf{SK}_B^{FHE}, \mathsf{PK}_B^{FHE}) \\ &\equiv_c S_A(\{P_i\}_{i=1}^n, \{w_h\}_h, \{[\![v_i]\!]_B\}_{i=1}^g, l, \\ &\quad \mathsf{SK}_A^P, \mathsf{SK}_A^{QR}, \mathsf{PK}_B^P, \mathsf{PK}_B^{QR}, \mathsf{PK}_B^{FHE}) \end{aligned}$$

B's view is

$$\begin{aligned} V_B = (&\{v_i\}_{i=1}^g, l, c, n, k \\ &\mathsf{PK}_A^P, \mathsf{PK}_A^{QR}, \mathsf{SK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{SK}_B^{FHE}; \\ &\mathsf{coins}; \\ &\{[\![\sum_{i=1}^n y_{ij}]\!]_A\}_{j=1}^n, \arg\max_j \{\sum_{i=1}^n y_{ij}\}_{j=1}^n) \end{aligned}$$

where c is the number of inner nodes over all trees, n is the number of trees, k is the number of classes, $[\![\sum_{i=1}^{n} y_{ij}]\!]_A$ is the encrypted vote count per class and $\arg\max_j\{\sum_{i=1}^{n} y_{ij}\}$ is the result of the arg max protocol and hence the output of B.

We simulate B by the simulator S_B as follows:

1. Generate n random Paillier encryptions $\{[\![\tilde{y}_j]\!]_A\}_{j=1}^{n}$.
2. Generate a random value between $v \leftarrow \{1, \ldots, n\}$.
3. Generate a random tape $\widetilde{\mathsf{coins}}$ of the required length, which can be determined by c, n and k.
4. Output

$$\begin{aligned} H_0' = (&\{v_i\}_{i=1}^{g}, l, c, n, k \\ &\mathsf{PK}_A^P, \mathsf{PK}_A^{QR}, \mathsf{SK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{SK}_B^{FHE}; \\ &\widetilde{\mathsf{coins}}; \\ &\{[\![\tilde{y}_j]\!]_A\}_{j=1}^{n}, v) \end{aligned}$$

Given that $\widetilde{\mathsf{coins}}$ and coins both are sampled from the same distribution with the same length, we can conclude that $H_0' \equiv_c H_1'$, with H_1' below:

$$\begin{aligned} H_1' = (&\{v_i\}_{i=1}^{g}, l, c, n, k \\ &\mathsf{PK}_A^P, \mathsf{PK}_A^{QR}, \mathsf{SK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{SK}_B^{FHE}; \\ &\mathsf{coins}; \\ &\{[\![\tilde{y}_j]\!]_A\}_{j=1}^{n}, v) \end{aligned}$$

Next, we show the indistinguishability of H_1' and V_B by giving a reduction to the semantic security of Paillier. To this end, we assume that we have a distinguisher $\mathcal{D}$ that can distinguish H' and V_B. In particular, given

$$\begin{aligned} &(\{v_i\}_{i=1}^{g}, l, c, n, k \\ &\mathsf{PK}_A^P, \mathsf{PK}_A^{QR}, \mathsf{SK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{SK}_B^{FHE}; \\ &\mathsf{coins}; \\ &\{[\![y_j]\!]_A\}_{j=1}^{n}, v) \end{aligned}$$

$\mathcal{D}$ outputs 1 if $v = \arg\max_j\{y_j\}_{j=1}^{n}$ and 0 otherwise. Then, we construct a reduction $\mathcal{R}$ as follows:

1. On input PK, pick uniformly at random $x, y, z \leftarrow M$, such that $x \neq y \neq z$.
2. Order the chosen values (w.l.o.g., we from here on assume $x < y < z$).
3. Generate new keys $\mathsf{PK}_A^{QR}, \mathsf{SK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{SK}_B^{FHE}$.
4. Choose challenger messages $m_0 = x$, $m_1 = z$ and give them to the semantic security challenger.

5. Receive c from the challenger and query the distinguisher

$$\mathcal{D}(\emptyset, 0, 0, 2, 0, \mathsf{PK}, \mathsf{PK}_A^{QR}, \mathsf{SK}_B^P, \mathsf{SK}_B^{QR}, \mathsf{SK}_B^{FHE}; \emptyset; \{[\![y]\!]_{\mathsf{PK}}, c\}, 2),$$

   which returns b.

6. Return b to the challenger.

Since we simulate both cases perfectly to the distinguisher, its success probability transfers exactly to our reduction in the semantic security game. Since Paillier encryption is shown to be semantically secure under the Decisional Composite Residuosity Assumption, the distinguisher must have at most negligible success probability.

Given the correctness of the protocol as well as the computational indistinguishability of both, simulators and views, we can apply Theorem 6. We replace the ideal calls by our provable secure building blocks. Theorem 6 then gives us the security of our scheme in the semi-honest model. □

6.10 Conclusion

In this chapter, we have first demonstrated that DNA methylation datasets can be re-identified by having access to an auxiliary database of genotypes. Following a Bayesian approach, we have shown that we could reach an accuracy of 97.5% to 100% depending on the attack scenario, with a few hundred methylation regions and genotype positions. Then, by using a statistical test upon our matching outcomes, we have empirically demonstrated that the very few wrongly matched pairs could be correctly identified and rejected, yielding a false-positive rate of 0 and true-positive rate of 1 for appropriate statistical thresholds. We have further shown that our identification attack was very robust to a decrease of methylation-meQTL pairs. When matching 52 methylation profiles with 75 genotypes, we could reach a full accuracy with only 13 meQTLs and methylation regions. We have also observed that the very few wrongly matched pairs often contain the genotype of the relative (in more than 90% of the cases). Finally, we have shown that our attack was robust to an increase of the database size to more than 2500 genotypes.

Facing this severe threat to epigenetic privacy, we have proposed a novel cryptographic scheme for privately classifying tumors based on methylation data. Our protocol relies on random forests and homomorphic encryption, and it is proven secure in the honest-but-curious adversarial model. We have implemented our private classifier in C++ and evaluated its performance on real data. We have shown that it can accurately classify brain tumors in nine classes of tumor subtypes based on 900 methylation levels in less than an hour. This constitutes an acceptable computational overhead in the considered clinical setting at hand. As a meta-consequence, we highly recommend removing DNA methylation profiles from public databases as these are extremely prone to re-identification, especially given that genotypes are also increasingly available online, sometimes with their owners' identifiers [95].

For future research, we plan to study if the attack is as successful when meQTL-methylation pairs are determined from a different tissue's data. At the defense side, we

would like to study other machine-learning algorithms and to propose private schemes for those that are efficient in classification with methylation data. Differentially private approaches could also be studied, although differential privacy may degrade utility too much for typical medical needs [40].

7

Privacy Risks in Interdependent Biomedical Data

Towards a Comprehensive Methodology for Quantifying Privacy Risks

7.1 Motivation

In the same vein as the previous chapter, we investigate interdependencies between different types of biomedical data. As pointed out, the main negative aspect of 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. 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. 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 [P1].

The goal of this work is to tackle the significant privacy concern of biomedical data from a more holistic perspective by encompassing different kinds of biomedical data and the statistical dependencies between them in the same framework. Furthermore, beyond genomic data, interdependent privacy risks between individuals from the same family have not been studied. This work aims at filling this gap by proposing a generic framework for dissecting and quantifying privacy risks in biomedical data on a large scale.

7.2 Contributions

In particular, 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 inheritance 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 learn the structure of the Bayesian network automatically. 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 that are 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 adversary's average 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 high 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, which we presented in Chapter 4.

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 [107, P1]). 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.

From these results, it becomes apparent that interdependencies across different biological layers, along with the temporal dimension, and in-between family members can pose a severe privacy threat towards health data privacy if an adversary is able to collect and leverage multiple pieces of evidence.

7.3 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 [113, 36, 23, 124]. Moreover, this kind of information and also the kinship between individuals can be further matched to side channels such as surname-genome associations databases [53] or online social networks [64].

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 (1) any background knowledge from domain experts, (2) any knowledge the adversary can construct based on auxiliary datasets, and (3) any data the adversary observes during his inference attack.

7.4 The Bayesian Network Model

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 main steps in Bayesian network inference: (1) learning the structure of the Bayesian network, (2) learning the necessary parameters of the network, and (3) 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.

7.4.1 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 $X_1, \ldots, X_l \in V$ in the graph corresponds to a random variable. An edge $X_i \to X_j \in E$ 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 7.4.3 to describe our structure learning algorithm.

A structure $X \to Z \leftarrow Y$ 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 5 (Active Trail [74])**.** *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 $X_{i-1} \to X_i \leftarrow X_{i+1}$, 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 6 (d-separation and Independencies [74])**.** *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 $\mathsf{d}\text{-}\mathsf{sep}_G(\mathbf{X}, \mathbf{Y} \mid \mathbf{Z})$, 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:

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

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:

$$\Pr[X_1, X_2, \ldots, X_n] = \prod_{i=1}^{n} \Pr[X_i \mid Parents(X_i)].$$

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*.

7.4.2 Notation and Networks

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

Let $\mathcal{V}$ 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 $a \in \mathcal{V}$. Similarly, $m_{a,t}^r \in [0, 1]$ denotes the average methylation within region $r \in \mathcal{R}$ for an individual $a \in \mathcal{V}$ at time point $t \in \mathcal{T}$. Let $\mathcal{M}$ denote the set of mothers, each member of which has a corresponding child in $\mathcal{V}$. Also, let $\mathcal{C}$ be the set to represent children who have their corresponding mothers in $\mathcal{V}$. For simplicity reasons, we assume that $\mathcal{M} \cap \mathcal{C} = \emptyset$. Also, note that $\mathcal{M} \cup \mathcal{C} \subseteq \mathcal{V}$.

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 work, 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 [106, 63]. This is a key element in simplifying the network structure as it allows us to build $|\mathcal{Q}|$ independent Bayesian networks. In Section 7.5, 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 $\mathcal{Q}$.

7.4.3 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 [61, 63], we have to learn most of the edges (dependencies) between the nodes in the Bayesian networks.

In literature, there exist general algorithms (as listed in [74]) 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 work, we introduce the novel notion of a minimal I-map given external knowledge.

Definition 7 (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 8 (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 $\mathcal{I}$.*

We propose an algorithm that achieves this definition in Algorithm 7.1: 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.

Algorithm 7.1 Build a minimal I-map given external knowledge

Input: External knowledge κ, 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 κ.

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 $G^* = \mathsf{None}$.
3: **for** $G \in \mathbf{G}$ **do**
4: Calculate $\mathcal{I}(G) = \{(X \perp Y \mid \mathbf{Z}) \mid \mathsf{d}\text{-}\mathsf{sep}_G(X, Y \mid \mathbf{Z})\}$.
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^*

Theorem 7 (Correctness of Algorithm 7.1)**.** *Algorithm 7.1 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 7 states the correctness of our algorithm, and we prove its validity in Section 7.8.

Scalability. The pseudocode given in Algorithm 7.1 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 Section 7.8. 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 $\mathcal{I}(G) \not\subseteq \mathcal{I}$, 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 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.

7.4.3.1 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:

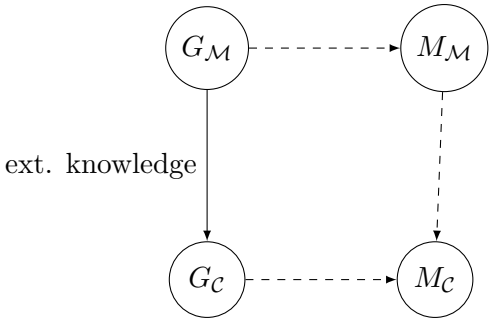


Figure 7.1: The graphical model for mother-child dependencies. The full edge represents external expert knowledge that is given, and dashed edges represent dependencies that need to be learned: if they exist (structure learning) and, if so, what is the magnitude of the dependency (parameter learning).

- $G_{\mathcal{M}} \to G_{\mathcal{C}} \in E$, 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 7.1. 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 $\mathcal{I}(P)$.

7.4.3.2 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,

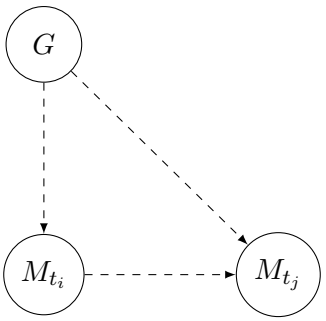


Figure 7.2: The graphical model for temporal inference of DNA methylation.

- $M_{t_i} \to M_{t_j} \in E \Rightarrow i < j$, 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 7.2, 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 $\mathcal{I}(P)$.

7.4.4 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 work, 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 $\Pr\left[G^i\right]$ for any Bayesian network.[1] More precisely, population statistics give us the minor allele frequency MAF_i (cf. Chapter 2) for each SNP. Let $p_k = \Pr\left[G^i = k\right]$, 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 of DNA methylation data, however, we need 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 $\Pr\left[M^r\right]$ can be considered to be generally multimodal. Therefore, following the general methodology in biomedical applications [132], 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 $\forall i, j \in \{1, \ldots, l\}$: $B_i \cap B_j = \emptyset \Leftrightarrow i \neq j$.

[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 $\Pr\left[G^i\right]$ 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 7.1: The probability distribution $\Pr[G_{\mathcal{C}} \mid G_{\mathcal{M}}]$ based on the laws of Mendelian inheritance, given external knowledge of $p_g = \Pr[G = g]$.

As stated before, we rely on MLE to learn the remaining distributions in our networks. Let $A \subseteq \mathcal{V}$ 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:

$$\Pr[M \in B_j \mid \mathbf{Z} = \mathbf{z}] = \frac{|\{m_a \mid a \in A \wedge m_a \in B_j \wedge \mathbf{z}_a = \mathbf{z}\}|}{|\{m_a \mid a \in A \wedge \mathbf{z}_a = \mathbf{z}\}|}. \tag{7.1}$$

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 7.6.2.

7.4.4.1 Mother-Child Networks

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

$$\Pr\left[G^i_{\mathcal{C}} = g^i_{\mathcal{C}} \mid G^i_{\mathcal{M}} = g^i_{\mathcal{M}}\right] = \sum_{g^i_{\mathcal{P}} \in \{0,1,2\}} \Pr\left[G^i_{\mathcal{P}} = g^i_{\mathcal{P}}\right] \Pr\left[G^i_{\mathcal{M}} = g^i_{\mathcal{M}}\right] \cdot \Pr\left[G^i_{\mathcal{C}} = g^i_{\mathcal{C}} \mid G^i_{\mathcal{M}} = g^i_{\mathcal{M}}, G^i_{\mathcal{P}} = g^i_{\mathcal{P}}\right], \tag{7.2}$$

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 sex or subgroup the individual is in and write $\Pr[G^i]$ instead of $\Pr[G^i_{\mathcal{M}}]$, $\Pr[G^i_{\mathcal{C}}]$ and $\Pr[G^i_{\mathcal{P}}]$. While $\Pr[G^i]$ – as stated before – is calculated from population statistics, $\Pr[G^i_{\mathcal{C}} \mid G^i_{\mathcal{M}}, G^i_{\mathcal{P}}]$ is exactly specified by the laws of Mendelian inheritance. Combining these finally results in the probability distribution as shown in Table 7.1.

7.4.4.2 Temporal Inference

Except for $\Pr[G]$, the parameters of the temporal inference network are learned by applying MLE, similarly to the mother-child network.

7.4.5 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 [74]. 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 [67] or (loopy) belief propagation algorithms [97], 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 $\Pr[X_i \mid \mathbf{E} = \mathbf{e}]$, 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.

7.4.6 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* [61, 63].

Expected estimation error has already been introduced in the context of genomic data by Humbert et al. [61]. 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 $\Pr[\hat{x} \mid \mathbf{Z} = \mathbf{z}]$, 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:

$$E_x(X \mid \mathbf{Z} = \mathbf{z}) = \sum_{\hat{x} \in \mathcal{X}} \Pr[X = \hat{x} \mid \mathbf{Z} = \mathbf{z}] \, ||\hat{x} - x||, \tag{7.3}$$

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 $\Pr\left[\hat{B} \mid \mathbf{Z} = \mathbf{z}\right]$ and the true methylation value m being part of

a bin B, the estimation error is calculated as follows:

$$E_B(M^r \mid \mathbf{Z} = \mathbf{z}) = \sum_{\hat{B} \in B^r} \Pr\left[M^r \in \hat{B} \mid \mathbf{Z} = \mathbf{z}\right] ||\mu(\hat{B}) - \mu(B)||. \tag{7.4}$$

The second metric (i.e., entropy) quantifies the *uncertainty* of the adversary [26, 109] and is defined as:

$$H_x(X \mid \mathbf{Z} = \mathbf{z}) = -\sum_{\hat{x} \in \mathcal{X}} \Pr\left[X = \hat{x} \mid \mathbf{Z} = \mathbf{z}\right] \log \Pr\left[X = \hat{x} \mid \mathbf{Z} = \mathbf{z}\right]. \tag{7.5}$$

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

For the binned methylation values, the definition of entropy easily translates to:

$$H_B(M^r \mid \mathbf{Z} = \mathbf{z}) = \\ -\sum_{\hat{B} \in B^r} \Pr\left[M^r \in \hat{B} \mid \mathbf{Z} = \mathbf{z}\right] \log \Pr\left[M^r \in \hat{B} \mid \mathbf{Z} = \mathbf{z}\right]. \tag{7.6}$$

7.5 Dataset Description

The dataset we use is the same as previously relied upon in Chapter 6. However, we apply a slightly different methodology here and hence end up with a different number of pairs. To recall the properties of the dataset, it 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 [111, 24] (version 141). To accomplish the task of determining the genotype from WGBS data, the Bis-SNP tool was used [82].

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 [115] 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 [10].

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 [104]. 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.

7.6 Evaluation

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

7.6.1 Structure Learning

Given the set $\mathcal{Q} \subseteq \mathcal{S} \times \mathcal{R}$ of SNP-methylation pairs as determined in Section 7.5, for each pair, we apply the algorithm presented in Section 7.4.3 for both settings we are interested in. Independence is tested using the χ^2-test at a significance level of $\alpha = 0.05$.

Figure 7.3(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 7.5).

Figure 7.3(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.

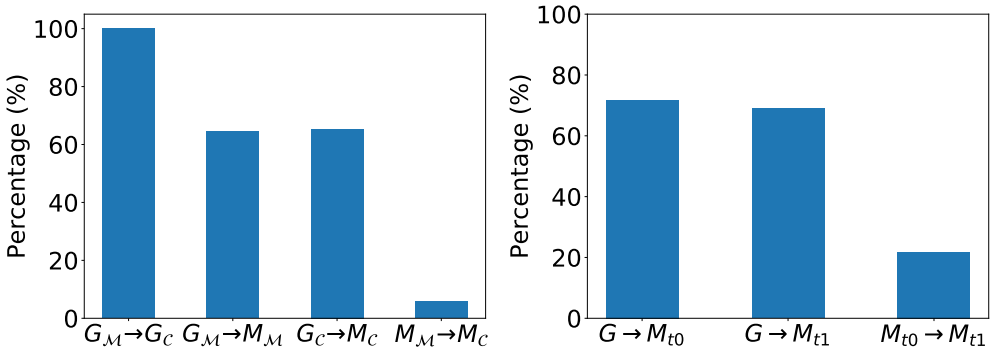


Figure 7.3: Distribution of edges after structure learning: (a) in the mother-child setting, (b) for the temporal inference of methylation data.

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.

7.6.2 Parameter Learning

We obtain the parameters of the Bayesian networks by relying on: (1) external knowledge (population statistics) and (2) maximum likelihood estimation on a training set.

We build the population statistics using Kaviar [47], a compilation of 162 million positions in the human genome. Kaviar contains data from 77,781 individuals. Using Kaviar, we estimate the prior probability of an individual carrying a specific variant $\Pr\left[G^i\right]$, and also calculate $\Pr\left[G^i_{\mathcal{C}} \mid G^i_{\mathcal{M}}\right]$, 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 7.6.3. 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 $B^r = \{B_1, \ldots, B_5\}$.

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 [85], which mitigates the problem of assigning 0 probabilities to rare methylation values by artificially adjusting the probability. More

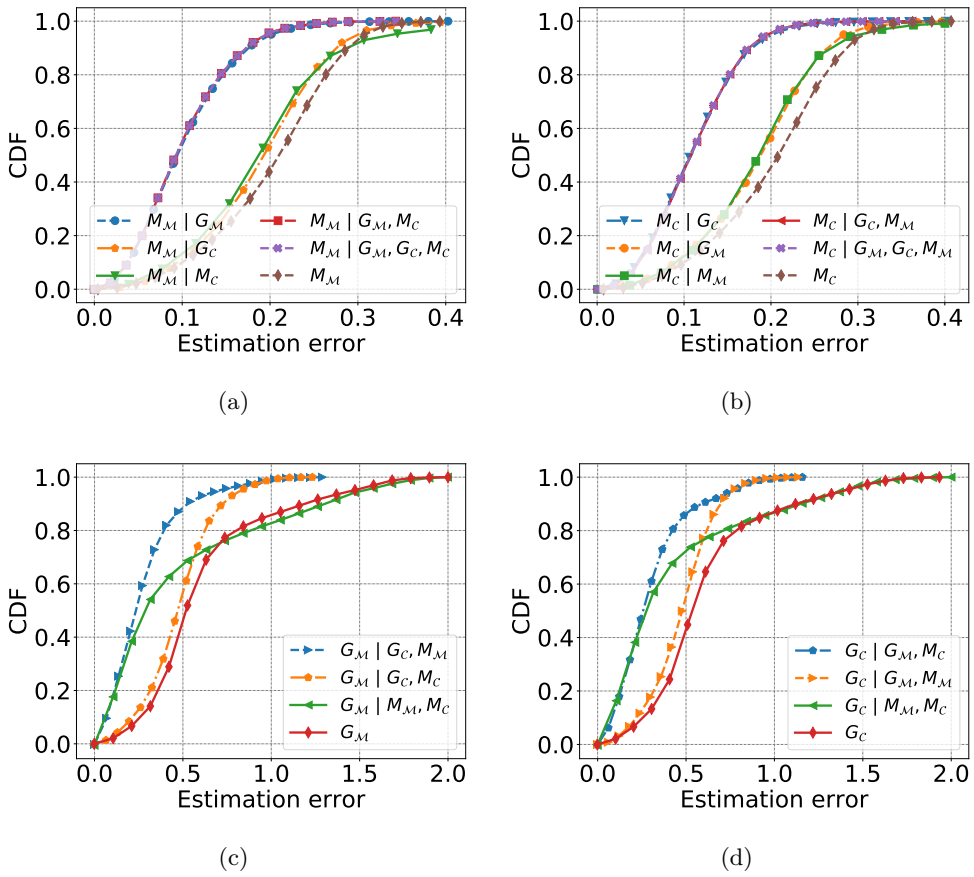


Figure 7.4: Estimation error when inferring the methylation of (a) mother and (b) child; the genome of (c) mother and (d) child.

precisely, Laplace smoothing gives us the following probability estimate:

$$\widehat{\Pr}\left[M \in B_j \mid \mathbf{Z} = \mathbf{z}\right] = \frac{|\{m_a \mid a \in A \wedge m_a \in B_j \wedge \mathbf{z}_a = \mathbf{z}\}| + \gamma}{|\{m_a \mid a \in A \wedge \mathbf{z}_a = \mathbf{z}\}| + \gamma |B^r|}. \tag{7.7}$$

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

7.6.3 Variable Prediction

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. For completeness, Figure 7.7 and Figure 7.8 contain the remaining experiments from our evaluation section. These figures are to be found at the end of our evaluation section and closely reflect the conclusions drawn below.

We begin with an analysis of the estimation error. Figure 7.4(a) and Figure 7.4(b) depict the CDFs of the privacy metrics for the methylation inference of the mother and the child, respectively. Analogously, Figure 7.4(c) and Figure 7.4(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., $\Pr[M_{\mathcal{M}} \mid G_{\mathcal{M}}, G_{\mathcal{C}}, M_{\mathcal{C}}]$ or $\Pr[M_{\mathcal{M}} \mid G_{\mathcal{M}}]$ results in less than 0.1 estimation error for almost 60% of the variables, while the same estimation error for the prior ($\Pr[M_{\mathcal{M}}]$) is only achieved in 10% of the networks (Figure 7.4(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, Figure 7.5 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 inference 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 $\Pr[M_{\mathcal{C}} \mid G_{\mathcal{C}}, M_{\mathcal{M}}]$ and $\Pr[M_{\mathcal{C}} \mid G_{\mathcal{M}}, G_{\mathcal{C}}, M_{\mathcal{M}}]$ 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 $\Pr[M_{\mathcal{C}} \mid G_{\mathcal{C}}]$, an observation which does not hold

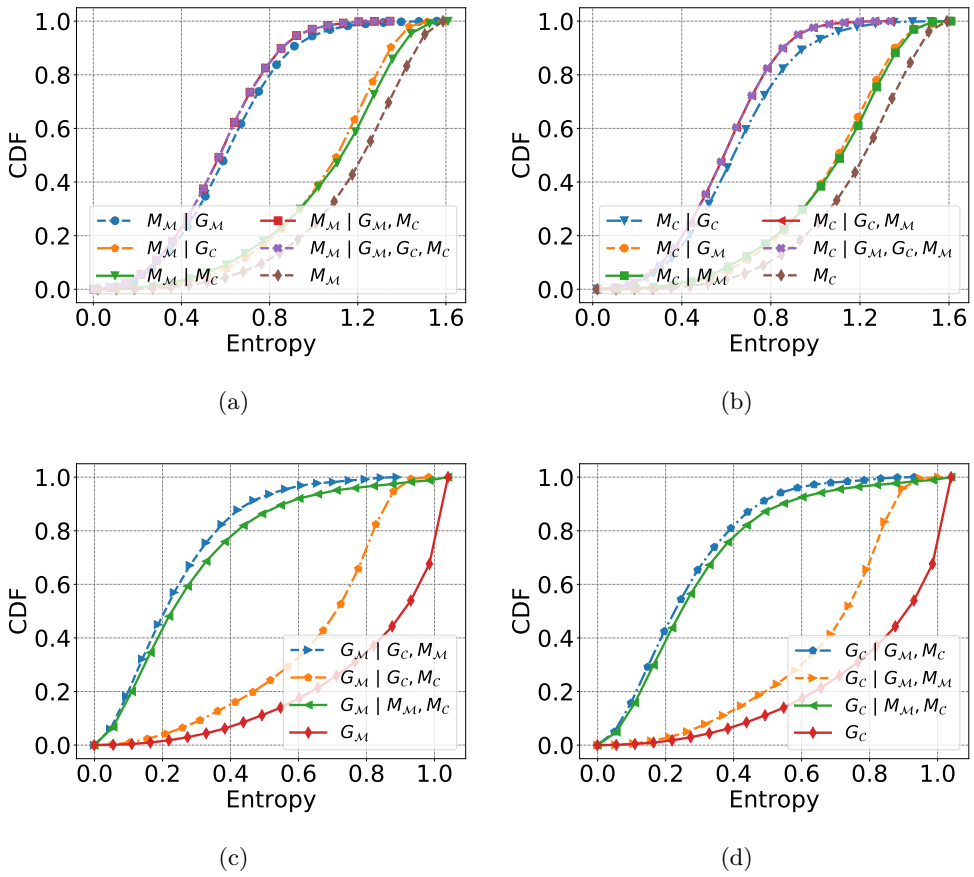


Figure 7.5: Entropy when inferring the methylation of (a) mother and (b) child; the genome of (c) mother and (d) child.

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, 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

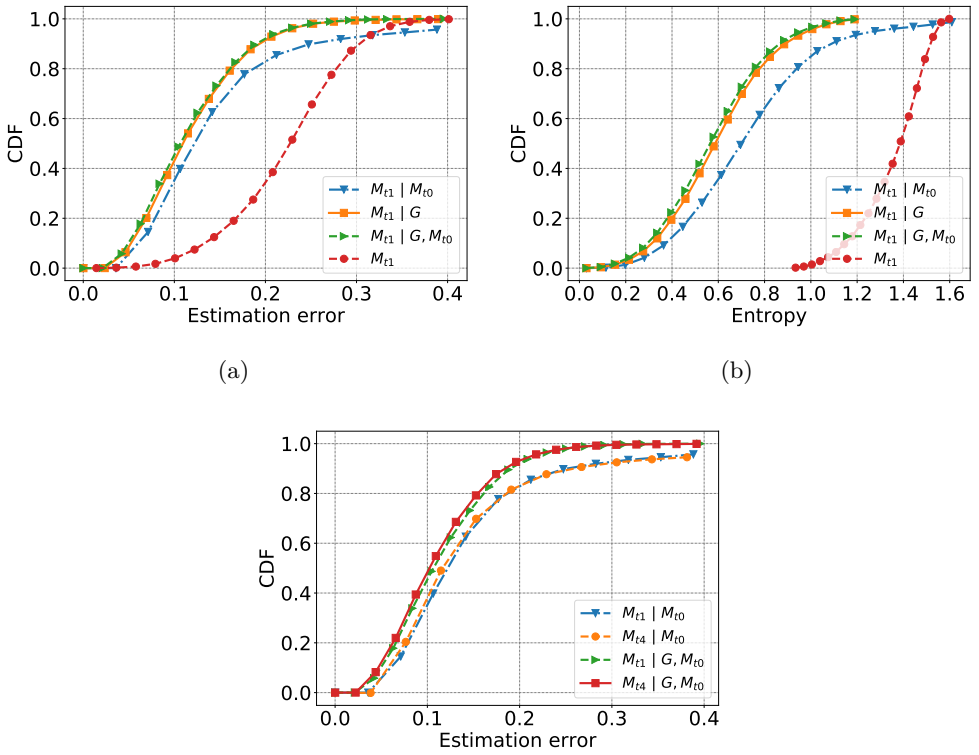


Figure 7.6: (a) Estimation error and (b) entropy when inferring the methylation at M_{t_1}; (c) comparison of estimation error between inferring M_{t_1} and M_{t_4}.

applying our framework for inferring $\Pr[G_{\mathcal{M}} \mid M_{\mathcal{M}}]$, we obtain an estimation error of 0.025 at this SNP, while the estimation error of the prior $\Pr[G_{\mathcal{M}}]$ 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.

Figure 7.6(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 (Figure 7.6(b)).

However, the genome is not the only strong evidence for the methylation. The

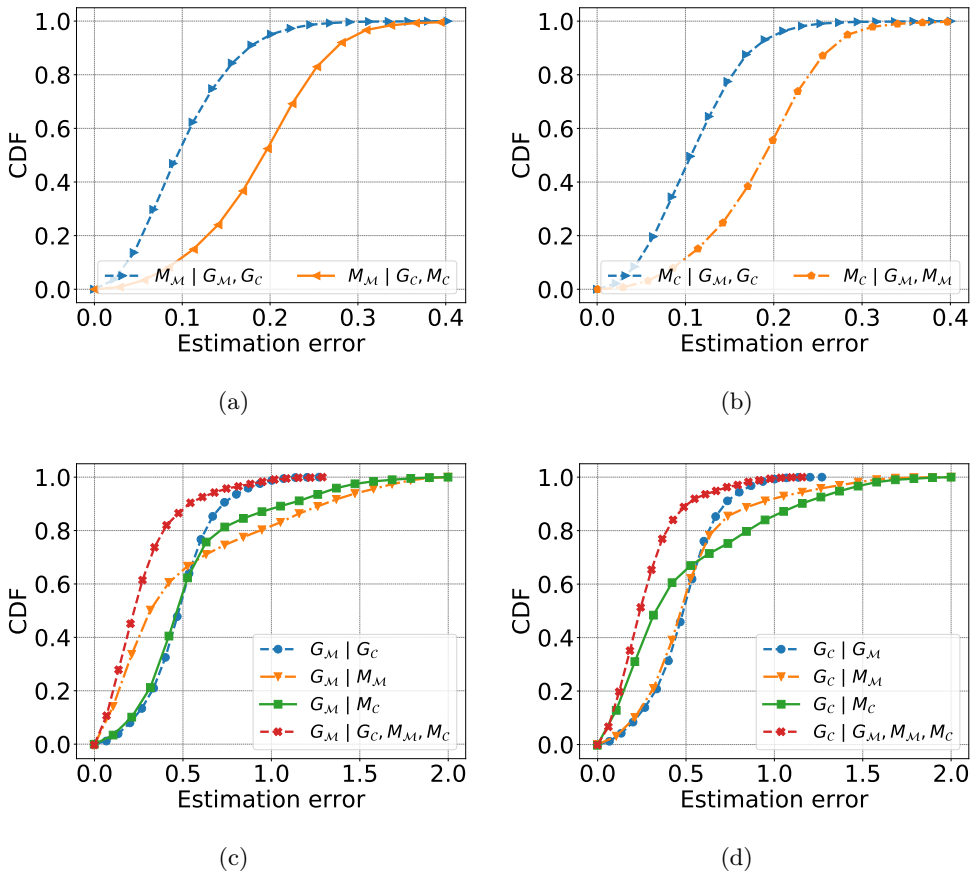


Figure 7.7: Additional graphs for estimation error when inferring the methylation of (a) mother and (b) child; the genome of (c) mother and (d) child.

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). Figure 7.6(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.

Considering all of the results in this section, we have clearly demonstrated severe privacy risks inherent to epigenetic and genetic data. Especially when observing multiple pieces of evidence, it becomes clear that an adversary – exploiting the interdependencies of these types of data – will successfully breach the privacy of many individuals.

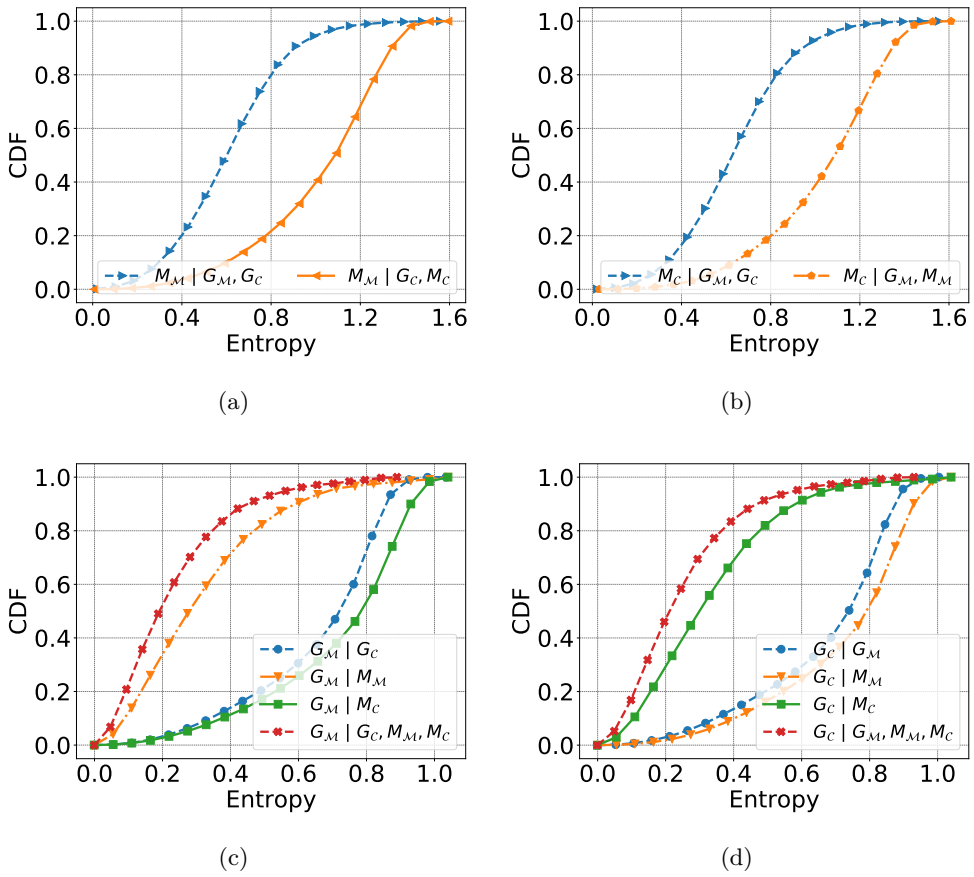


Figure 7.8: Additional graphs for entropy when inferring the methylation of (a) mother and (b) child; the genome of (c) mother and (d) child.

7.7 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 7.6.3, 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., $\Pr\left[M^r_{\mathcal{C}} = m^r_{v_t} \mid M^r_{\mathcal{M}} = m^r_{v_o}\right]$ 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.

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 $w_{r,a} = \Pr\left[M^r_{\mathcal{C}} = m^r_a \mid M^r_{\mathcal{M}} = m^r_{v_o}\right]$. 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:

$$\hat{v}_t = \arg\max_{a \in \mathcal{D}} \sum_{(\cdot,r) \in \mathcal{Q}'} w_{r,a} \tag{7.8}$$

$$= \arg\max_{a \in \mathcal{D}} \sum_{(\cdot,r) \in \mathcal{Q}'} \Pr\left[M^r_{\mathcal{C}} = m^r_a \mid M^r_{\mathcal{M}} = m^r_{v_o}\right] \tag{7.9}$$

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 7.6.1 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 $\mathcal{Q}$.

Interestingly, this application also has parallels to the previous chapter. It can be seen as a generalized version of a matching attack, as it is based on a generic model. In contrast to our previous work, however, we do not take into account the possibility that the corresponding mother or child is not present.

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

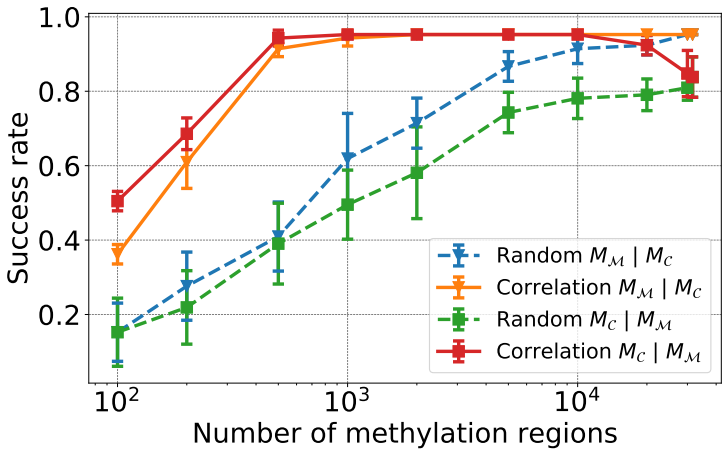


Figure 7.9: Success rate for discovering mother/child of each observed victim.

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. Figure 7.9 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.

Given these results, we have demonstrated the usefulness of both our Bayesian network and our heuristic for this kind of application. Indeed, the Bayesian network can serve as a fundamental building block to cope with more complex application scenarios. Moreover, this again demonstrates the severe privacy threat stemming from epigenetic data such as DNA methylation profiles.

7.8 Proof of Correctness

In this section, we prove the correctness of Algorithm 7.1. We begin recalling the definition of a Markov blanket in a Bayesian network, as stated by Koller and Friedman [74].

Proposition 1 (Markov Blanket)**.** *Given a node X in a Bayesian network G, by $MB_G(X)$ we denote the smallest set of nodes $\mathbf{U}$ that are needed 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.*

Removing 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 7.4.3. 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 independency 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 $\mathcal{I}(G)$.

This concludes both cases and proves the original statement. □

Main Proof. Leveraging the proposition from above, we are now able to prove the correctness of our structure learning algorithm as depicted in Algorithm 7.1.

Proof of Theorem 7. 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 $\mathcal{I}(G') \not\subseteq \mathcal{I}$.

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.

In the final step, we have to 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 that 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.

Combining these three proof steps proves the correctness of the algorithm and thus Theorem 7. □

7.9 Conclusion

In this chapter, 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 and effectiveness 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.

8
Conclusion

In this dissertation, we presented a line of work aiming at quantifying and protecting the privacy of individuals' biomedical data. We specifically focused on epigenetic data, a widely used type of biomedical data. Although our epigenome is closely linked with our health status, its privacy implications have received little to no attention so far, and epigenetic data have often been released (without identifiers) on open online platforms with nonrestricted access.

In this thesis, we thoroughly studied the privacy of individuals in the presence of multiple realistic attack scenarios, in which an adversary obtains parts of the individuals' biomedical data. Namely, we consider (1) temporal linkability, an attack scenario in which the adversary wishes to link epigenetic profiles of the same type of biomedical data taken at different points in time, (2) linkability between different types of biomedical data, (3) membership privacy, an attack scenario in which the adversary wishes to learn whether an individual participated in a study, and (4) inference, an attack scenario in which the adversary wishes to infer previously unknown data about an individual, given some observations. Besides quantifying privacy in the presence of such adversaries, we also presented and evaluated solutions to preserve the privacy of individuals. Our mitigation techniques stretch from the differentially private release of epigenetic data – and the quantification of utility in such cases – up to cryptographic constructions to securely and privately evaluate a random forest on a patient's data. We instantiated our use cases with two of the most important types of epigenetic data: miRNA expression profiles, and DNA methylation in combination with genomic data.

The content of this dissertation stretches across multiple peer-reviewed publications [P1, P2, P3, P4], covering both, privacy quantifications and mitigation techniques.

Our study on temporal linkability of miRNA expression profiles (Chapter 4) strikingly demonstrated that personal miRNA expression profiles can be successfully linked over time. We also proposed and evaluated two defense mechanisms when releasing datasets of epigenetic data: hiding a subset of the expression data, and adding noise to the released expression profiles in a fully distributed manner. We observed that, in most cases, the noise mechanism provides a better privacy-utility trade-off than the hiding method.

In our work on membership privacy of miRNA-expression-based studies (Chapter 5), we showed that it is possible for an adversary to detect membership in a miRNA-based study by relying on only the published mean statistics and the victim's profile. Moreover, we further studied two defense mechanisms for releasing the mean statistics of a dataset, following the example of our previous work: hiding a subset of the expression data, and adding noise to the released mean values in a differentially private manner. We observed that the noising mechanism is able to protect the privacy. However, the amount of noise might render the released statistics useless, in particular for small datasets. Hence, we recommend having a large number of participants, at least a few hundreds.

We then examined how correlations between the genome and DNA methylation can be exploited in order to infer an individual's genotype (Chapter 6). Achieving an accuracy of more than 97.5% with only a few hundred methylation regions and genotype positions, we demonstrated the disastrous impact of such an attack on privacy. We also presented a statistical test upon our matching outcome that is able to identify and reject the very few wrongly matched pairs and thus further degrades privacy. In

contrast to our first work, this constitutes a linkability attack between different types of biomedical data and shows the crucial effect of correlations between such types of data. On the mitigation side, we proposed a novel cryptographic scheme for privately classifying tumors based on methylation data. The protocol relies on random forests and homomorphic encryption, and it is proven secure in the honest-but-curious adversarial model.

Finally, leveraging the knowledge from our previous studies, we proposed a generic framework for quantifying privacy risks (Chapter 7), capturing and generalizing most of the attack scenarios previously presented. Our framework relies on a Bayesian network that allows us to quantify privacy implications due to various correlations between different types of biomedical data, along the temporal dimension, and between related individuals. We also proposed a general algorithm to learn the structure of the underlying Bayesian networks by combining data with external knowledge. Our extensive set of experiments demonstrated that our Bayesian network model is able to achieve a strong prediction performance. The results reinforce that the privacy risks inherent to interdependent biomedical data – and epigenetic data specifically – have to be taken seriously.

Future Research Directions. As this thesis shows, it is essential to provide the means for a quantitative assessment of the privacy risks induced by sharing or leaking medical data. Potential privacy risks include linkage, identification and inference attacks against the patients' data, carried out by a multitude of possible adversaries. Furthermore, after identifying and quantifying these risks, mitigation measures have to be designed. On the one hand, it is crucial to adjust those measures with the close collaboration of biomedical experts to fit their application scenarios and needs. On the other hand, the mitigation measures have to provide a sufficient amount of privacy, giving the patient herself the control over her data. Hence, a major challenge while designing those measures is to strike a balance between privacy risks, the utility of the resulting data, and ease of use. Not respecting any combination of these will result in the mitigation measure not being adopted for real use, or in the worst case even causing harm to a patient by providing inaccurate statements in return for a higher privacy.

The author of this dissertation envisions to come up with solutions to quantify and mitigate the privacy risks for various kinds of patients' data, encompassing the whole pipeline from measuring data and storing data, up to analyzing data and providing a diagnosis. To achieve this goal, a first step is to find solutions to concrete application scenarios, raise the awareness for health privacy, and then to generalize the specific solutions where appropriate to build up a generic framework for securing the privacy of biomedical data. In the same vein as this thesis, progress has to be made in four areas, three of which we explored in our work: (1) assessing the privacy risks for biomedical data, (2) providing mitigation measures by perturbing the data, (3) providing mitigation measures by relying on cryptographic constructions, and (4) leveraging secure hardware for trusted computing. While the second research direction is more relevant for the release of public medical datasets and statistics, the third research direction allows for the secure storage of medical data and the further analysis thereof without losing utility.

When it comes to assessing the privacy risks, more potential sources of data have

to be considered and also previously disregarded dimensions have to be incorporated. Building on our generic framework from Chapter 7, we could include various other types of biomarkers and characteristics. Moreover, other potential background knowledge should be an integral component of new models as well.

Mitigations of privacy risks involving perturbing the biomedical data have to be adapted to specific scenarios. This might include considering a more realistic adversary and weakening assumptions in exchange for achieving a sufficient level of privacy and still a high utility for the original purpose of the data. Side knowledge like the sex and parts of the genome need to be considered as well. Overall, these adaptations or relaxations require close collaboration with biomedical researchers and have to be tuned to fit the scenario at hand.

The third direction aims at securing biomedical data by cryptographic means. This first requires a deeper understanding of how the data is processed and what the data is used for, an understanding that has to be acquired in close collaboration with biomedical experts for a particular application. Then, concepts like homomorphic encryption or multiparty computation have to be adapted and combined into cryptographic protocols fulfilling the requirements. Due to the long-lasting impact (also on relatives) of certain medical data, a particular emphasis has to be put on the resistance of the underlying cryptographic building blocks against future attackers. Moreover, potential adaptations of cryptographic mechanisms have to be made to allow for efficient processing of the data.

Finally, as a fourth, newly emerging field, the application of secure hardware for trusted computing on biomedical data is also a promising research direction that will become more important in the future.

Datasets

Chapter	Type	Origin	Healthy	Diseased	Time points	Pub.	GEO accession number
4	miRNA expression	blood & plasma	29	0	2	[57]	-
4	miRNA expression	plasma	0	26 (lung cancer)	8	[78]	GSE68951
4 & 5	miRNA expression	blood	94	955 (various)	1	[73]	GSE61741
6	DNA methylation	brain tissue	0	472 (brain tumor)	1	[1]	GSE50022, GSE55712, GSE36278, GSE52556, GSE54880, GSE45353, GSE44684
6 & 7	DNA methylation & genome	blood	75 (including 21 mother/child pairs)		9	-	-

Overview over all datasets used in this thesis.

Bibliography

Author's Papers for this Thesis

[P1] Backes, M., Berrang, P., Hecksteden, A., Humbert, M., Keller, A., and Meyer, T. Privacy in epigenetics: temporal linkability of microRNA expression profiles. In: *Proceedings of the 25th USENIX Security Symposium (Security)*. USENIX Association, 2016, 1223–1240.

[P2] Backes, M., Berrang, P., Humbert, M., and Manoharan, P. Membership privacy in microRNA-based studies. In: *Proceedings of the 23rd ACM Conference on Computer and Communication Security (CCS)*. ACM, 2016, 319–330.

[P3] Backes, M., Berrang, P., Bieg, M., Eils, R., Herrmann, C., Humbert, M., and Lehmann, I. Identifying personal DNA methylation profiles by genotype inference. In: *Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2017, 957–976.

[P4] Berrang, P., Humbert, M., Zhang, Y., Lehmann, I., Eils, R., and Backes, M. Dissecting privacy risks in biomedical data. In: *Proceedings of the 2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2018.

Other Papers of the Author

[S1] Backes, M., Berrang, P., Goga, O., Gummadi, K., and Manoharan, P. Profile linkability despite anonymity in social media systems. In: *Proceedings of the 15th ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2016.

[S2] Backes, M., Berrang, P., Hecksteden, A., Humbert, M., Keller, A., and Meyer, T. On epigenomic privacy: tracking personal microRNA expression profiles over time. In: *Workshop on Understanding and Enhancing Online Privacy (UEOP), affiliated with NDSS*. 2016.

[S3] Backes, M., Berrang, P., Humbert, M., Shen, X., and Wolf, V. Simulating the large-scale erosion of genomic privacy over time. In: *3rd International Workshop on Genome Privacy and Security (GenoPri), Selected for publication in IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2016.

[S4] Backes, M., Berrang, P., and Manoharan, P. From zoos to safaris – from closed-world enforcement to open-world assessment of privacy. In: *Foundations of Security Analysis and Design VIII*. Springer-Verlag, 2016, 87–138.

Other references

[1] Ahamed, M. T., Danielsson, A., Nemes, S., and Carén, H. MethPed: an R package for the identification of pediatric brain tumor subtypes. *BMC Bioinformatics* 17, 1 (2016), 262.

[2] Andrés, M. E., Bordenabe, N. E., Chatzikokolakis, K., and Palamidessi, C. Geo-indistinguishability: differential privacy for location-based systems. In: *Proceedings of the 20th ACM Conference on Computer and Communication Security (CCS)*. ACM, 2013, 901–914.

[3] *ArrayExpress Archive of Functional Genomics Data*. `https://www.ebi.ac.uk/arrayexpress`. Accessed: 03/11/2017.

[4] Ayday, E., De Cristofaro, E., Hubaux, J.-P., and Tsudik, G. Whole genome sequencing: revolutionary medicine or privacy nightmare? *Computer* 48, 2 (2015), 58–66.

[5] Ayday, E., Raisaro, J. L., Hubaux, J.-P., and Rougemont, J. Protecting and evaluating genomic privacy in medical tests and personalized medicine. In: *Proceedings of the 12th ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2013, 95–106.

[6] Ayday, E., Raisaro, J. L., McLaren, P. J., Fellay, J., and Hubaux, J.-P. Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data. In: *USENIX Workshop on Health Information Technologies*. USENIX Association, 2013.

[7] Backes, C., Leidinger, P., Keller, A., Hart, M., Meyer, T., Meese, E., and Hecksteden, A. Blood born miRNAs signatures that can serve as disease specific biomarkers are not significantly affected by overall fitness and exercise. *PloS one* 9, 7 (2014), e102183.

[8] Baldi, P., Baronio, R., De Cristofaro, E., Gasti, P., and Tsudik, G. Countering GATTACA: efficient and secure testing of fully-sequenced human genomes. In: *Proceedings of the 18th ACM Conference on Computer and Communication Security (CCS)*. ACM. 2011, 691–702.

[9] Bauer, T., Trump, S., Ishaque, N., Thu rmann, L., Gu, L., Bauer, M., Bieg, M., Gu, Z., Weichenhan, D., Mallm, J.-P., Ro der, S., Herberth, G., Takada, E., Mu cke, O., Winter, M., Junge, K. M., Gru tzmann, K., Rolle-Kampczyk, U., Wang, Q., Lawerenz, C., Borte, M., Polte, T., Schlesner, M., Schanne, M., Wiemann, S., Geo rg, C., Stunnenberg, H. G., Plass, C., Rippe, K., Mizuguchi, J., Herrmann, C., Eils, R., and Lehmann, I. Environment-induced epigenetic reprogramming in genomic regulatory elements in smoking mothers and their children. *Molecular Systems Biology* 12, 3 (2016), 861–861.

[10] Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[11] *Bilans de santé en balade sur le net*. `http://www.lematin.ch/suisse/bilans-sante-balade-net/story/21621328`. Accessed: 03/11/2017.

[12] *Boost C++ libraries*. http://www.boost.org/. Accessed: 03/11/2017.

[13] Bost, R., Popa, R. A., Tu, S., and Goldwasser, S. Machine learning classification over encrypted data. In: *Proceedings of the 22nd Annual Network and Distributed System Security Symposium (NDSS)*. The Internet Society, 2015.

[14] Brakerski, Z., Gentry, C., and Vaikuntanathan, V. Fully homomorphic encryption without bootstrapping. *Cryptology ePrint Archive, Report 2011/277* (2011).

[15] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, 108–122.

[16] Cachin, C. Entropy measures and unconditional security in cryptography. PhD thesis. SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH, 1997.

[17] *Cancer Statistics*. https://seer.cancer.gov/statistics/summaries.html. Accessed: 03/11/2017.

[18] Canetti, R. Security and composition of multiparty cryptographic protocols. *Journal of Cryptology* 13, 1 (2000), 143–202.

[19] Chatzikokolakis, K., Andrés, M. E., Bordenabe, N. E., and Palamidessi, C. Broadening the scope of differential privacy using metrics. In: *Proceedings of the 13th International Privacy Enhancing Technologies Symposium (PETS)*. Springer-Verlag, 2013, 82–102.

[20] Cloud, J. Why your DNA isn't your destiny. *Time Magazine* 6 (2010).

[21] Danezis, G. and De Cristofaro, E. Fast and private genomic testing for disease susceptibility. In: *Proceedings of the 13th ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2014, 31–34.

[22] Danielsson, A., Nemes, S., Tisell, M., Lannering, B., Nordborg, C., Sabel, M., and Carén, H. MethPed: a DNA methylation classifier tool for the identification of pediatric brain tumor subtypes. *Clinical Epigenetics* 7, 1 (2015), 1.

[23] Das, P. M. and Singal, R. DNA methylation and cancer. *Journal of clinical oncology* 22, 22 (2004), 4632–4642.

[24] *dbSNP*. https://www.ncbi.nlm.nih.gov/SNP/. Accessed: 03/11/2017.

[25] De Cristofaro, E., Liang, K., and Zhang, Y. Privacy-preserving genetic relatedness test. In: *3rd International Workshop on Genome Privacy and Security (GenoPri)*. 2016.

[26] Diaz, C., Seys, S., Claessens, J., and Preneel, B. Towards measuring anonymity. In: *Proceedings of the 2nd International Workshop on Privacy Enhancing Technologies*. Springer-Verlag. 2002, 54–68.

[27] Dongen, J. van, Nivard, M. G., Willemsen, G., Hottenga, J.-J., Helmer, Q., Dolan, C. V., Ehli, E. A., Davies, G. E., Iterson, M. van, Breeze, C. E., Beck, S., Hoen, P. A., Pool, R., Greevenbroek, M. M. van, Stehouwer, C. D., Kallen, C. J. van der, Schalkwijk, C. G., Wijmenga, C., Zhernakova, S., Tigchelaar, E. F., Beekman, M., Deelen, J., Heemst, D. van, Veldink, J. H., Berg, L. H. van den, Duijn, C. M. van, Hofman, B. A., Uitterlinden, A. G., Jhamai, P. M., Verbiest, M., Verkerk, M., Breggen, R. van der, Rooij, J. van, Lakenberg, N., Mei, H., Bot, J., Zhernakova, D. V., Hof, P. van't, Deelen, P., Nooren, I., Moed, M., Vermaat, M., Luijk, R., Bonder, M. J., Dijk, F. van, Galen, M. van, Arindrarto, W., Kielbasa, S. M., Swertz, M. A., Zwet, E. W. van, Isaacs, A., Franke, L., Suchiman, H. E., Jansen, R., Meurs, J. B. van, Heijmans, B. T., Slagboom, P. E., and Boomsma, D. I. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications* 7 (2016), 11115.

[28] Duverle, D. A., Kawasaki, S., Yamada, Y., Sakuma, J., and Tsuda, K. Privacy-preserving statistical analysis by exact logistic regression. In: *Proceedings of the 2015 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2015, 7–16.

[29] Dwork, C. Differential privacy: a survey of results. In: *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation (TAMC)*. Springer-Verlag, 2008, 1–19.

[30] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In: *Proceedings of the 3rd Conference on Theory of Cryptography (TCC)*. Springer-Verlag, 2006, 265–284.

[31] Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407.

[32] Dwork, C., Smith, A., Steinke, T., Ullman, J., and Vadhan, S. Robust traceability from trace amounts. In: *Proceedings of the 56th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society, 2015, 650–669.

[33] Dyke, S. O., Cheung, W. A., Joly, Y., Ammerpohl, O., Lutsik, P., Rothstein, M. A., Caron, M., Busche, S., Bourque, G., Rönnblom, L., et al. Epigenome data release: a participant-centered approach to privacy protection. *Genome Biology* 16, 1 (2015), 1–12.

[34] Edmonds, J. Paths, trees, and flowers. *Canadian Journal of Mathematics* 17, 3 (1965), 449–467.

[35] Erlich, Y. and Narayanan, A. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics* 15, 6 (2014), 409–421.

[36] Esteller, M. and Herman, J. G. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. *The Journal of Pathology* 196, 1 (2002), 1–7.

[37] Feinberg, A. P. and Fallin, M. D. Epigenetics at the crossroads of genes and the environment. *JAMA* 314, 11 (2015), 1129–1130.

[38] Fierer, N., Lauber, C. L., Zhou, N., McDonald, D., Costello, E. K., and Knight, R. Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences* 107, 14 (2010), 6477–6481.

[39] Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannan, B. J., and Huttenhower, C. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences* 112, 22 (2015), E2930–E2938.

[40] Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., and Ristenpart, T. 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, 17–32.

[41] Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*. Vol. 1. Springer-Verlag, 2001.

[42] Friedman, R. C., Farh, K. K.-H., Burge, C. B., and Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* 19, 1 (2009), 92–105.

[43] Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W. L., Ho, K., Ring, S. M., Evans, D. M., Davey Smith, G., and Relton, C. L. Systematic identification of genetic influences on methylation across the human life course. *Genome Biology* 17, 1 (2016), 61.

[44] *Gene Expression Omnibus*. `http://www.ncbi.nlm.nih.gov/geo`. Accessed: 03/11/2017.

[45] *Genome-Wide Association Studies Fact Sheet*. `https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/`. Accessed: 03/11/2017.

[46] Ghosh, A., Roughgarden, T., and Sundararajan, M. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing* 41, 6 (2012), 1673–1693.

[47] Glusman, G., Caballero, J., Mauldin, D. E., Hood, L., and Roach, J. C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* 27, 22 (2011), 3216–3217.

[48] *GNU MP*. `https://gmplib.org`. Accessed: 03/11/2017.

[49] Goldreich, O. *Foundations of Cryptography - Volume 2 (Basic Applications)*. 2004.

[50] Goldwasser, S. and Micali, S. Probabilistic encryption & how to play mental poker keeping secret all partial information. In: *Proceedings of the 14th Annual ACM Symposium on Theory of Computing (STOC)*. ACM, 1982, 365–377.

[51] *Google's Protocol Buffers*. `https://github.com/google/protobuf`. Accessed: 03/11/2017.

[52] Granlund, T. and Team, G. D. *GNU MP 6.0 Multiple Precision Arithmetic Library*. Samurai Media Limited, 2015.

[53] Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., and Erlich, Y. Identifying personal genomes by surname inference. *Science* 339, 6117 (2013), 321–324.

[54] Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. 2008, 11–15.

[55] Halevi, S. and Shoup, V. *HElib – An Implementation of homomorphic encryption*. `https://github.com/shaih/HElib`. 2014.

[56] *Health insurer Anthem discloses customer and employee data breach*. `http://www.computerworld.com/article/2879649/health-insurer-anthem-discloses-customer-and-employee-data-breach.html`. Accessed: 03/11/2017.

[57] Hecksteden, A., Leidinger, P., Backes, C., Rheinheimer, S., Pfeiffer, M., Ferrauti, A., Kellmann, M., Sedaghat, F., Meder, B., Meese, E., et al. MiRNAs and sports: tracking training status and potentially confounding diagnoses. *Journal of Translational Medicine* 14, 1 (2016), 219.

[58] Ho, T. K. Random decision forests. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. 1995, 278–282.

[59] Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics* 4, 8 (2008), e1000167.

[60] *Human miRNA Disease Database*. `http://www.cuilab.cn/hmdd`. Accessed: 03/11/2017.

[61] Humbert, M., Ayday, E., Hubaux, J.-P., and Telenti, A. Addressing the concerns of the Lacks family: quantification of kin genomic privacy. In: *Proceedings of the 20th ACM Conference on Computer and Communication Security (CCS)*. ACM, 2013, 1141–1152.

[62] Humbert, M., Ayday, E., Hubaux, J.-P., and Telenti, A. Reconciling utility with privacy in genomics. In: *Proceedings of the 13th ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM, 2014, 11–20.

[63] Humbert, M., Ayday, E., Hubaux, J.-P., and Telenti, A. Quantifying interdependent risks in genomic privacy. *ACM Transactions on Privacy and Security (TOPS)* 20, 1 (2017), 3.

[64] Humbert, M., Huguenin, K., Hugonot, J., Ayday, E., and Hubaux, J.-P. Denanonymizing genomic databases using phenotypic traits. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2015, 2 (2015), 99–114.

[65] *IGSR: The International Genome Sample Resource (1000 Genomes Project)*. `http://www.internationalgenome.org/data`. Accessed: 03/11/2017.

[66] Im, H. K., Gamazon, E. R., Nicolae, D. L., and Cox, N. J. On sharing quantitative trait GWAS results in an era of multiple-omics data and the limits of genomic privacy. *The American Journal of Human Genetics* 90, 4 (2012), 591–598.

[67] Jensen, F. V. and Jensen, F. Optimal junction trees. In: *Proceedings of the 10th International Conference on Uncertainty in Artificial Intelligence (UAI)*. Morgan Kaufmann Publishers Inc. 1994, 360–366.

[68] Johnson, A. and Shmatikov, V. 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, 1079–1087.

[69] Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* 13, 7 (2012), 484–92.

[70] Jones, P. A. and Baylin, S. B. The epigenomics of cancer. *Cell* 128, 4 (2007), 683–692.

[71] Karvelas, N., Peter, A., Katzenbeisser, S., Tews, E., and Hamacher, K. Privacy-preserving whole genome sequence processing through proxy-aided ORAM. In: *Proceedings of the 13th ACM Workshop on Privacy in the Electronic Society (WPES)*. ACM. 2014, 1–10.

[72] Keller, A., Leidinger, P., Bauer, A., ElSharawy, A., Haas, J., Backes, C., Wendschlag, A., Giese, N., Tjaden, C., Ott, K., et al. Toward the blood-borne miRNome of human diseases. *Nature Methods* 8, 10 (2011), 841–843.

[73] Keller, A., Leidinger, P., Vogel, B., Backes, C., ElSharawy, A., Galata, V., Mueller, S. C., Marquart, S., Schrauder, M. G., Strick, R., et al. MiRNAs can be generally associated with human pathologies as exemplified for miR-144*. *BMC medicine* 12, 1 (2014), 224.

[74] Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[75] Koufogiannis, F., Han, S., and Pappas, G. J. Optimality of the Laplace mechanism in differential privacy. *arXiv preprint arXiv:1504.00065* (2015).

[76] Kuhn et al., M. *caret: Classification and Regression Training*. `http://caret.r-forge.r-project.org`. Accessed: 03/11/2017. 2017.

[77] Leidinger, P., Backes, C., Deutscher, S., Schmitt, K., Mueller, S. C., Frese, K., Haas, J., Ruprecht, K., Paul, F., Stähler, C., et al. A blood based 12-miRNA signature of alzheimer disease patients. *Genome Biology* 14, 7 (2013), R78.

[78] Leidinger, P., Galata, V., Backes, C., Stähler, C., Rheinheimer, S., Huwer, H., Meese, E., and Keller, A. Longitudinal study on circulating miRNAs in patients after lung cancer resection. *Oncotarget* 6, 10 (2015), 16674.

[79] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., et al. The diploid genome sequence of an individual human. *PLoS Biology* 5, 10 (2007), e254.

[80] Li, N., Qardaji, W., Su, D., Wu, Y., and Yang, W. Membership privacy: a unifying framework for privacy definitions. In: *Proceedings of the 20th ACM Conference on Computer and Communication Security (CCS)*. ACM, 2013, 889–900.

[81] Lindell, Y. and Pinkas, B. Secure multiparty computation for privacy-preserving data mining. *Journal of Privacy and Confidentiality* 1, 1 (2009), 5.

[82] Liu, Y., Siegmund, K. D., Laird, P. W., and Berman, B. P. Bis-SNP: combined DNA methylation and SNP calling for bisulfite-seq data. *Genome Biology* 13, 7 (2012), R61.

[83] Londin, E., Loher, P., Telonis, A. G., Quann, K., Clark, P., Jing, Y., Hatzimichael, E., Kirino, Y., Honda, S., Lally, M., et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate-and tissue-specific microRNAs. *Proceedings of the National Academy of Sciences* 112, 10 (2015), E1106–E1115.

[84] Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B. L., Mak, R. H., Ferrando, A. A., et al. MicroRNA expression profiles classify human cancers. *Nature* 435, 7043 (2005), 834–838.

[85] Manning, C. D., Raghavan, P., Schütze, H., et al. *Introduction to information retrieval*. Vol. 1. 1. Cambridge university press Cambridge, 2008.

[86] Massey, J. L. Guessing and entropy. In: *Proceedings of the 1994 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 1994, 204.

[87] McClay, J. L., Shabalin, A. A., Dozmorov, M. G., Adkins, D. E., Kumar, G., Nerella, S., Clark, S. L., Bergen, S. E., Hultman, C. M., Magnusson, P. K. E., Sullivan, P. F., Aberg, K. A., and Oord, E. J. C. G. van den. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biology* 16, 1 (2015), 291.

[88] McLaren, P. J., Raisaro, J. L., Aouri, M., Rotger, M., Ayday, E., Bartha, I., Delgado, M. B., Vallet, Y., Günthard, H. F., Cavassini, M., et al. Privacy-preserving genomic testing in the clinic: a model using HIV treatment. *Genetics in Medicine* 18, 8 (2016).

[89] *Medical Data – A New Target for Hackers*. `https://www.logpoint.com/se/about-us/blog/249-medical-data-a-new-target-for-hackers`. Accessed: 03/11/2017.

[90] *MicroRNAs: Definition and Overview*. `https://www.thermofisher.com/de/de/home/references/ambion-tech-support/microrna-studies/tech-notes/micrornas-definition-and-overview.html`. Accessed: 03/11/2017.

[91] Naveed, M., Ayday, E., Clayton, E. W., Fellay, J., Gunter, C. A., Hubaux, J.-P., Malin, B. A., and Wang, X. Privacy in the genomic era. *ACM Computing Surveys (CSUR)* 48, 1 (2015), 6.

[92] Neyman, J. and Pearson, E. S. *On the problem of the most efficient tests of statistical hypotheses*. Springer-Verlag, 1992, 73–108.

[93] Ngun et al., T. Abstract: a novel predictive model of sexual orientation using epigenetic markers. In: *American Society of Human Genetics 2015 Annual Meeting*. 2015.

[94] *Number of microRNAs in Human Genome Skyrockets*. http://www.genengnews.com/gen-news-highlights/number-of-micrornas-in-human-genome-skyrockets/81250958/. Accessed: 03/11/2017.

[95] *openSNP*. https://opensnp.org. Accessed: 03/11/2017.

[96] Paillier, P. Public-key cryptosystems based on composite degree residuosity classes. In: *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Springer-Verlag. 1999, 223–238.

[97] Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

[98] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research (JMLR)* 12 (2011), 2825–2830.

[99] Philibert, R. A., Terry, N., Erwin, C., Philibert, W. J., Beach, S. R., and Brody, G. H. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. *Clinical epigenetics* 6, 1 (2014), 28.

[100] *Premera, Anthem data breaches linked by similar hacking tactics*. http://www.computerworld.com/article/2898419/data-breach/premera-anthem-data-breaches-linked-by-similar-hacking-tactics.html. Accessed: 03/11/2017.

[101] *President Obama's Precision Medicine Initiative*. https://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative. Accessed: 03/11/2017.

[102] Qureshi, I. A. and Mehler, M. F. Advances in epigenetics and epigenomics for neurodegenerative diseases. *Current neurology and neuroscience reports* 11, 5 (2011), 464–473.

[103] R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/. Accessed: 03/11/2017. R Foundation for Statistical Computing, 2016.

[104] Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., et al. Linkage disequilibrium in the human genome. *Nature* 411, 6834 (2001), 199–204.

[105] Rothstein, M. A., Cai, Y., and Marchant, G. E. The ghost in our genes: legal and ethical implications of epigenetics. *Health matrix (Cleveland, Ohio: 1991)* 19, 1 (2009), 1.

[106] Sankararaman, S., Obozinski, G., Jordan, M. I., and Halperin, E. Genomic privacy and limits of individual detection in a pool. *Nature genetics* 41, 9 (2009), 965–967.

[107] Schadt, E. E., Woo, S., and Hao, K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nature genetics* 44, 5 (2012), 603–608.

[108] Schübeler, D. Function and information content of DNA methylation. *Nature* 517, 7534 (2015), 321–326.

[109] Serjantov, A. and Danezis, G. Towards an information theoretic metric for anonymity. In: *Proceedings of the 2nd International Workshop on Privacy Enhancing Technologies*. Springer-Verlag. 2002, 41–53.

[110] Sheaffer, K. L., Kim, R., Aoki, R., Elliott, E. N., Schug, J., Burger, L., Schubeler, D., and Kaestner, K. H. DNA methylation is required for the control of stem cell differentiation in the small intestine. *Genes & Development* 28, 6 (2014), 652–664.

[111] Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29, 1 (2001), 308–311.

[112] Smart, N. P. and Vercauteren, F. Fully homomorphic SIMD operations. *Designs, codes and cryptography* 71, 1 (2014), 57–81.

[113] *SNPedia*. `https://www.snpedia.com/index.php/SNPedia`. Accessed: 03/11/2017.

[114] *Sophia Genetics*. `http://www.sophiagenetics.com/`. Accessed: 03/11/2017.

[115] Spearman, C. The proof and measurement of association between two things. *The American journal of psychology* 15, 1 (1904), 72–101.

[116] Teh, A. L., Pan, H., Chen, L., Ong, M. L., Dogra, S., Wong, J., MacIsaac, J. L., Mah, S. M., McEwen, L. M., Saw, S. M., Godfrey, K. M., Chong, Y. S., Kwek, K., Kwoh, C. K., Soh, S. E., Chong, M. F. F., Barton, S., Karnani, N., Cheong, C. Y., Buschdorf, J. P., Stunkel, W., Kobor, M. S., Meaney, M. J., Gluckman, P. D., and Holbrook, J. D. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Research* 24, 7 (2014), 1064–1074.

[117] *The Black Market For Stolen Health Care Data*. `http://www.npr.org/sections/alltechconsidered/2015/02/13/385901377/the-black-market-for-stolen-health-care-data`. Accessed: 03/11/2017.

[118] Tipping, M. E. and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61, 3 (1999), 611–622.

[119] Tramèr, F., Huang, Z., Hubaux, J.-P., and Ayday, E. Differential privacy with bounded priors: reconciling utility and privacy in genome-wide association studies. In: *Proceedings of the 22nd ACM Conference on Computer and Communication Security (CCS)*. ACM, 2015, 1286–1297.

[120] Trump, S., Bieg, M., Gu, Z., Thürmann, L., Bauer, T., Bauer, M., Ishaque, N., Röder, S., Gu, L., Herberth, G., Lawerenz, C., Borte, M., Schlesner, M., Plass, C., Diessl, N., Eszlinger, M., Mücke, O., Elvers, H.-D., Wissenbach, D. K., Bergen, M. von, Herrmann, C., Weichenhan, D., Wright, R. J., Lehmann, I., and Eils, R. Prenatal maternal stress and wheeze in children: novel insights into epigenetic regulation. *Scientific Reports* 6 (2016), 28616.

[121] Tsaprouni, L. G., Yang, T.-P., Bell, J., Dick, K. J., Kanoni, S., Nisbet, J., Viñuela, A., Grundberg, E., Nelson, C. P., Meduri, E., Buil, A., Cambien, F., Hengstenberg, C., Erdmann, J., Schunkert, H., Goodall, A. H., Ouwehand, W. H., Dermitzakis, E., Spector, T. D., Samani, N. J., and Deloukas, P. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* 9, 10 (2014), 1382–1396.

[122] Uhler, C., Slavković, A., and Fienberg, S. E. Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality* 5, 1 (2013), 137.

[123] *Urgent probe as Michael Schumacher's medical records stolen and put on sale for £40k.* `http://www.express.co.uk/news/world/484495/Investigation-underway-after-Michael-Schumacher-s-medical-records-stolen`. Accessed: 03/11/2017.

[124] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. Cancer genome landscapes. *Science* 339, 6127 (2013), 1546–1558.

[125] Wang, R., Li, Y. F., Wang, X., Tang, H., and Zhou, X. Learning your identity and disease from research papers: information leaks in genome wide association study. In: *Proceedings of the 16th ACM Conference on Computer and Communication Security (CCS)*. ACM, 2009, 534–544.

[126] Wang, X. S., Huang, Y., Zhao, Y., Tang, H., Wang, X., and Bu, D. Efficient genome-wide, privacy-preserving similar patient query based on private edit distance. In: *Proceedings of the 22nd ACM Conference on Computer and Communication Security (CCS)*. ACM. 2015, 492–503.

[127] Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., et al. The genomic landscapes of human breast and colorectal cancers. *Science* 318, 5853 (2007), 1108–1113.

[128] Wu, D. J., Feng, T., Naehrig, M., and Lauter, K. Privately evaluating decision trees and random forests. *Proceedings on Privacy Enhancing Technologies (PoPETs)* 2016, 4 (2016), 335–355.

[129] *Your private medical data is for sale -– and it's driving a business worth billions.* `https://www.theguardian.com/technology/2017/jan/10/medical-data-multibillion-dollar-business-report-warns`. Accessed: 03/11/2017.

[130] Yu, F., Fienberg, S. E., Slavković, A. B., and Uhler, C. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics* 50 (2014), 133–141.

[131] Yu, F., Rybar, M., Uhler, C., and Fienberg, S. E. Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In: *Privacy in Statistical Databases*. Springer-Verlag, 2014, 170–184.

[132] Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T., and Engelhardt, B. E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome biology* 16, 1 (2015), 14.

[133] Zhou, X., Peng, B., Li, Y. F., Chen, Y., Tang, H., and Wang, X. To release or not to release: evaluating information leaks in aggregate human-genome data. In: *Proceedings of the 16th European Symposium on Research in Computer Security (ESORICS)*. Springer-Verlag, 2011, 607–627.