Membership Privacy in MicroRNA-based Studies

Michael Backes
CISPA, Saarland University &
MPI-SWS
Saarland Informatics Campus

Pascal Berrang
CISPA, Saarland University
Saarland Informatics Campus

Mathias Humbert
CISPA, Saarland University
Saarland Informatics Campus

Praveen Manoharan
CISPA, Saarland University
Saarland Informatics Campus

ABSTRACT

The continuous decrease in cost of molecular profiling tests is revolutionizing medical research and practice, but it also raises new privacy concerns. One of the first attacks against privacy of biological data, proposed by Homer et al. in 2008, showed that, by knowing parts of the genome of a given individual and summary statistics of a genome-based study, it is possible to detect if this individual participated in the study. Since then, a lot of work has been carried out to further study the theoretical limits and to counter the genome-based membership inference attack. However, genomic data are by no means the only or the most influential biological data threatening personal privacy. For instance, whereas the genome informs us about the risk of developing some diseases in the future, epigenetic biomarkers, such as microRNAs, are directly and deterministically affected by our health condition including most common severe diseases.

In this paper, we show that the membership inference attack also threatens the privacy of individuals contributing their microRNA expressions to scientific studies. Our results on real and 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: one relying on the L_1 distance and 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 both a differentially private mechanism and a hiding mechanism. We also consider two types of adversarial 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 to only release summary statistics of datasets containing at least a couple of hundred individuals.

CCS'16, October 24 - 28, 2016, Vienna, Austria

ISBN 978-1-4503-4139-4/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2976749.2978355

Keywords

Health privacy; Membership privacy; Differential privacy

1. INTRODUCTION

During the last decade, the cost of molecular profiling tests, such as DNA sequencing, has significantly dropped, enabling a new breakthrough in biomedical science and the subsequent advent of personalized medicine. 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 [19]. 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 [4]), 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's success more precisely and to propose defense mechanisms for countering it.

The genome is, however, not the only element correlated with human health that can have a detrimental effect on privacy. A variety of new biomarkers, such as epigenomic and transcriptomic data, are currently being studied by biomedical researchers towards a more precise and personalized medicine. One class of these biomarkers is the microRNA (miRNA). MiRNAs are small RNA molecules that regulate the majority of human genes. 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 [27, 35, 21, 29, 15]. Therefore, miRNA expression profiling promises to enable 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] However, despite the disease-leakage risk stemming from miRNAs, their growing importance and *pub-*

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

lic availability in biomedical databases,[2] privacy of miRNA data has been largely overlooked by the research community. Moreover, as miRNAs might not be strictly defined as genetic information, it is still unclear if the current genetic nondiscrimination laws, such as the US Genetic Information Nondiscrimination Act, would apply to them [30, 14].

Contributions.

In this paper, we first study whether, and to what extent, membership inference can be successfully carried out against miRNA expression datasets. Notable challenges we needed to overcome are that miRNA expressions are real-valued rather than discrete, but of several orders of magnitude lower dimension and more noisy 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 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 of 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 randomly generated datasets. This is essentially due to the fact that miRNA expressions are highly 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 to assume an attacker knowing 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 prohibitive loss for research and medical utility. One key observation is that the differentially private mechanism is able to reduce the attacks' 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.

[2]The current most prominent examples are the Gene Expression Omnibus (GEO) [3] and the ArrayExpress [1] databases.

Organization.

In Section 2, we introduce the required background and the adversarial model. In Section 3, we present analytical and experimental results of the membership inference attack against miRNA-based studies. In Section 4, we introduce defense mechanisms for countering the attack, and evaluate the privacy and utility they provide. We present the related work in Section 5 before concluding in Section 6.

2. PRELIMINARIES

In this section, we briefly present the required background on microRNA expressions, then describe our threat model, and finally review the basic definitions of differential privacy and membership privacy, which will be used in this work.

2.1 MicroRNA Expressions

MicroRNAs (abbreviated miRNAs) are small non-coding RNA molecules that regulate gene expression in plants and animals. These molecules notably regulate 60% of the genes coding human proteins [17]. Currently, there are more than 5,000 miRNAs known in human beings [8], and this number will certainly keep increasing [26].

A miRNA *expression* is a (positive) real value quantified in a two-step polymerase chain reaction (PCR) process that measures how much the miRNA is active in a given cell or tissue. A miRNA expression profile represents the set of miRNA expressions of an individual at one point in time.

Biomedical research is especially interested in discovering how miRNA expressions affect human pathologies. Recent studies of miRNA expression profiling have already demonstrated that dysregulation of miRNA is linked to neurodegenerative diseases (Alzheimer's and Parkinson's), heart diseases, diabetes, and the majority of cancers [27, 35, 21, 29, 15]. Hence, miRNA expression profiling promises to enable a 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 [22, 24]. In this work, we focus on membership privacy in the context of blood-based miRNAs. A summary of the relation between miRNA and human pathologies can be found in the Human miRNA Disease Database [6].

2.2 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 hundreds dollars (and the cost will certainly decrease over time). Full individual miRNA expression data are also increasingly available in public research databases, such as the Gene Expression Omnibus (GEO) [3] or ArrayExpress [1] 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 [7] such as the Anthem's breach, in which the medical records of around 80 million patients were leaked [5].

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, previous work has shown that miRNA expression profiles can be efficiently linked over time frames of up to one year [9].

Second, we assume the adversary has access to some summary statistics released for the pool. Formally, the pool is defined as a set $\mathcal{T} \in \mathbb{R}^{n \times m}$ 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 has also 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 for genomic data.

2.3 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 [10]. We review here the definitions and results on differential privacy and positive membership privacy relevant to this paper.

Definition 1 (Differential Privacy [10]). *A mechanism $\mathcal{A}$ provides ϵ-differential privacy if and only if for any two datasets T_1 and T_2 differing in one element, and any $S \subseteq \mathsf{range}(\mathcal{A})$, it holds that*

$$\Pr[\mathcal{A}(T_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(T_2) \in S]$$

In this paper, 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. [25], 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 [25]). *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 $t \in \mathcal{U}$, it holds that*

$$\Pr_{D,\mathcal{A}}[t \in T \mid \mathcal{A}(T) \in S] \leq \gamma \cdot \Pr_{D}[t \in T] \quad (1)$$

$$\Pr_{D,\mathcal{A}}[t \notin T \mid \mathcal{A}(T) \in S] \geq \frac{1}{\gamma} \cdot \Pr_{D}[t \notin T] \quad (2)$$

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_t of an entity t to a database is restricted to $p_t \in [a, b] \cup \{0, 1\}$, for $0 < a \leq b < 1$, then achieving weaker differential privacy is sufficient to achieve (positive) membership privacy, as shown by Tramèr et al. [32].

Theorem 1 (Tramèr et al. [32]). *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*

$$e^{\epsilon} = \begin{cases} \min(\frac{(1-a)\gamma}{1-a\gamma}, \frac{\gamma+b-1}{b}) & \text{if } a\gamma < 1, \\ \frac{\gamma+b-1}{b} & \text{otherwise.} \end{cases} \quad (3)$$

3. 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. [19] 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 [23] and compare their performance.

3.1 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 $x_j^i \in \mathbb{R}$. $\mathbf{x}^i \in \mathbb{R}^m$ is the vector of all miRNA expression values 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.

3.1.1 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 (i) x_j^v and μ_j, and (ii) x_j^v and $\hat{\mu}_j$. By computing the difference between these distances we obtain the following statistic:

$$D(x_j^v) = |x_j^v - \mu_j| - |x_j^v - \hat{\mu}_j| \quad (4)$$

Under the null hypothesis, if x_j^v is not part of the pool, $D(x_j^v)$ should approach zero. Under the alternative hypothesis, where x_j^v is member of the pool, it should be greater than zero because the victim's contribution x_j^v to $\hat{\mu}_j$ will shift $\hat{\mu}_j$ away from μ_j. When $D(x_j^v)$ is negative, x_j^v 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 $D(x_j^v)$ 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.

3.1.2 Likelihood-Ratio Test

Although the aforementioned test can be very accurate, there is no known theoretical guarantee on the power[3] 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 and thus, provides 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 [28]. 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 [31]. 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[4] 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:

$$f(x_j^v) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{x_j^v - \mu_j}{2\sigma_j^2}} \tag{5}$$

Similarly, under the alternative hypothesis, following a similar reasoning as in the theoretical analysis of [31], we consider the miRNA expressions of the victim to be drawn according to the probability distribution of the pool:

$$\hat{f}(x_j^v) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_j} e^{-\frac{x_j^v - \hat{\mu}_j}{2\hat{\sigma}_j^2}} \tag{6}$$

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

$$LR = \frac{\sigma}{\hat{\sigma}} e^{\frac{x_j^v - \mu_j}{2\sigma_j^2} - \frac{x_j^v - \hat{\mu}_j}{2\hat{\sigma}_j^2}} \tag{7}$$

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

$$LLR = \sum_{j=1}^{m} \frac{(x_j^v - \mu_j)^2}{2\sigma_j^2} - \frac{(x_j^v - \hat{\mu}_j)^2}{2\hat{\sigma}_j^2} + \log\frac{\sigma_j}{\hat{\sigma}_j} \tag{8}$$

If the adversary has access to the average values $\hat{\mu}_j$ of miRNA expressions in the pool, as assumed in this paper, 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 Subsection 3.2, 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*

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{2m}{n^2}}, \tag{9}$$

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

PROOF. First of all, we need to compute the statistics of the LLR defined in (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:

$$\mu_{j,0} := E[LLR_j \mid H_0] = \frac{1}{2\sigma_j^2}\int_{-\infty}^{\infty}(x_j - \mu_j)^2 f(x_j)dx_j \tag{10}$$

$$- \frac{1}{2\hat{\sigma}_j^2}\int_{-\infty}^{\infty}(x_j - \hat{\mu}_j)^2 f(x_j)dx_j + \log\frac{\sigma_j}{\hat{\sigma}_j}\int_{-\infty}^{\infty} f(x_j)dx_j \tag{11}$$

$$= \frac{1}{2} - \frac{1}{2\hat{\sigma}_j^2}\int_{-\infty}^{\infty}(x_j - \hat{\mu}_j)^2 f(x_j)dx_j + \log\frac{\sigma_j}{\hat{\sigma}_j} \tag{12}$$

$$= \frac{1}{2} - \frac{1}{2\hat{\sigma}_j^2}\int_{-\infty}^{\infty}(x_j - \mu_j - \frac{x_j - \mu_j}{n})^2 f(x_j)dx_j + \log\frac{\sigma_j}{\hat{\sigma}_j} \tag{13}$$

$$= \frac{1}{2} - \frac{\sigma_j^2}{2\hat{\sigma}_j^2} + \frac{\sigma_j^2}{n\hat{\sigma}_j^2} - \frac{\sigma_j^2}{2n^2\hat{\sigma}_j^2} + \log\frac{\sigma_j}{\hat{\sigma}_j}. \tag{14}$$

From (12) to (13), we assume that the pool is constituted of the victim v and $n-1$ individuals drawn as under the null, i.e., $\hat{\mu}_j = \frac{(n-1)\mu_j + x_j}{n}$. Using our assumption $\forall j : \hat{\sigma}_j = \sigma_j$, we obtain

$$\mu_{j,0} = \frac{1}{n} - \frac{1}{2n^2} = \frac{2n-1}{2n^2}. \tag{15}$$

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

$$\mu_{j,1} := E[LLR_j \mid H_1] = \frac{2n-1}{2(n-1)^2}, \forall j \tag{16}$$

[3]Power refers to the true-positive rate, also called sensitivity.
[4]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.

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

$$\sigma^2_{j,k} := E[LLR^2_j \mid H_k] - \mu^2_{j,k}, k \in \{0,1\} \tag{17}$$

$E[LLR^2_j \mid H_k]$ 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:

$$\sigma_{j,0} = \frac{2n-1}{\sqrt{2}n^2}, \tag{18}$$

$$\sigma_{j,1} = \frac{2n-1}{\sqrt{2}(n-1)^2} \tag{19}$$

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:

$$\sigma_{j,0}z_\alpha + \sigma_{j,0}z_{1-\beta} = \sqrt{m}(\mu_{j,1} - \mu_{j,0}) \tag{20}$$

$$\frac{1}{\sqrt{2}n^2}z_\alpha + \frac{1}{\sqrt{2}(n-1)^2}z_{1-\beta} = \sqrt{m}\left(\frac{1}{2(n-1)^2} - \frac{1}{2n^2}\right) \tag{21}$$

$$(n-1)^2 z_\alpha + n^2 z_{1-\beta} = \sqrt{2m}(n - \frac{1}{2}) \tag{22}$$

$$z_\alpha + z_{1-\beta} \approx \sqrt{\frac{2m}{n^2}} \tag{23}$$

□

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 [31]. 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 constantly keep discovering new miRNAs and, thereby, implicitly increase the number m of available statistics [8]. 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 highly affected by diseases.

3.2 Experimental Results

In this section, we evaluate the two aforementioned attacks and compare their respective performances. Before that, we provide details on the dataset we use for our evaluations (including those in Section 4.2).

3.2.1 Dataset Description

The dataset was first presented and used by Keller et al. in [23], and is publicly available in the gene expression omnibus (GEO) database under reference GSE61741. 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.

3.2.2 Results

We evaluate our attacks on the aforementioned dataset in two different settings: (i) we randomly pick a varying number n of individuals from the dataset to form a pool, and (ii) we consider every case group (carrying a disease) desribed above 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's 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.

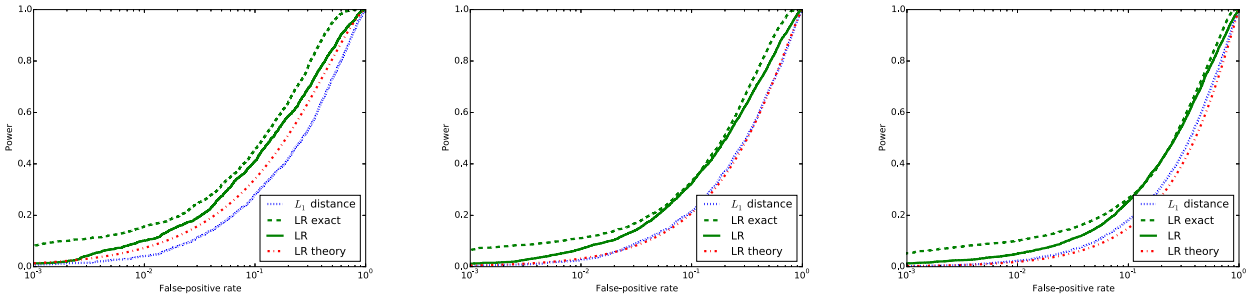
We evaluate each attack on aforementioned pools, using each of the 1,049 individuals as a potential victim. Given an attack and a pool, we obtain a test statistic T_v for every victim v. We then say v is more likely to be part of the pool than to be part of the reference population if the test statistic is greater than a given threshold t, i.e., $T_v > t$. 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 $T_v > t$), false-positive (v is not part of the pool and $T_v > t$), true-negative (v is not part of the pool and $T_v \leq t$) or false-negative (v is part of the pool and $T_v \leq t$). 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 (i) the attack based on the L_1 distances' difference, (ii) 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), (iii) the LR attack not knowing $\hat{\sigma}$, and approximating it as $\hat{\sigma} \approx \sigma$ (corresponding to our assumed threat model), and (iv) 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 Fig. 1, we depict three diagrams of randomly constructed pools for $n \in \{35, 65, 124\}$. We select these num-



(a) (b) (c)

Figure 1: ROC curves for pools of n randomly chosen individuals: (a) $n = 35$, (b) $n = 65$, (c) $n = 124$.

bers because they are representative for our dataset and also correspond to the numbers of cases of three disease-specific groups shown in Fig. 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 worst 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. This discrepancy is probably due to the fact that the reference population is supposed to be infinite, whereas in practice it is approximated by a finite group of samples.

Case Groups.

Fig. 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 Fig. 1 and on case groups in Fig. 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 (Fig. 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 (Fig. 1(c)).

Furthermore, as shown by our dataset, it often happens that the case group is very small. 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% (Fig. 2(a)). 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's success, it 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 [24].

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 to apply probabilistic sanitization before disclosing the summary statistics, or to drastically reduce the number of released means.

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 [31]. 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.

4. 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 our techniques on the privacy of pool's contributors and on the utility for research.

4.1 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.

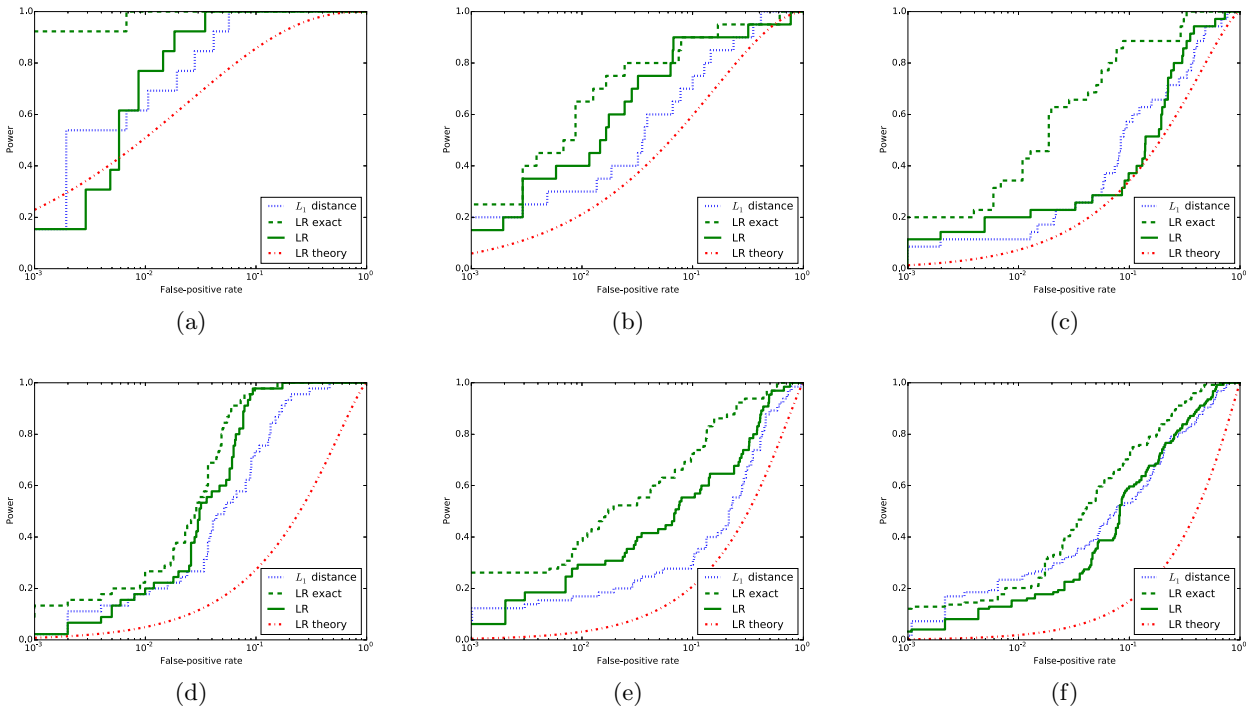


Figure 2: ROC curves for case groups of n individuals carrying: (a) stomach tumor ($n = 13$), (b) renal cancer ($n = 20$), (c) benign prostate hyperplasia ($n = 35$), (d) ductal adenocarcinoma ($n = 45$), (e) prostate cancer ($n = 65$), and (f) Wilms tumor ($n = 124$).

A standard method to achieve differential privacy for real-valued functions is to add Laplace noise: we replace the original mechanism $f_{\mathsf{avg}} : \mathcal{T} \to \mathbb{R}^m$ 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 $L(b)$. As shown by Dwork et al. [11], 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 $f_{\mathsf{avg}} : \mathcal{T} \to \mathbb{R}^m$ 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*

$$\Delta(f_{\mathsf{avg}}) = \max_{T_1, T_2 \in \mathcal{T}} \|f_{\mathsf{avg}}(T_1) - f_{\mathsf{avg}}(T_2)\|_1$$
$$= \max_{T_1, T_2 \in \mathcal{T}} \sum_{i}^{m} |f_{\mathsf{avg},i}(T_1) - f_{\mathsf{avg},i}(T_2)| = \sum_{i}^{m} \frac{\delta_i}{n},$$

where T_1 and T_2 are two datasets differing in one element.

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

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 y. Note that, as shown by Ghosh et al. [18], 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 $f : \mathcal{T} \to \mathbb{R}^m$ and let $f'_{\mathsf{avg}} = f_{\mathsf{avg}} + (Y_1, \ldots, Y_m)$, $Y_i \sim L(\frac{\Delta(f_{\mathsf{avg}})}{\epsilon})$. Then, $\forall y \geq 0$*

$$\Pr\left[|f_{\mathsf{avg},i}(T) - f'_{\mathsf{avg},i}(T)| \geq y\right] \leq e^{-\frac{\epsilon n y}{\sum_{k=1}^{m} \delta_k}}$$

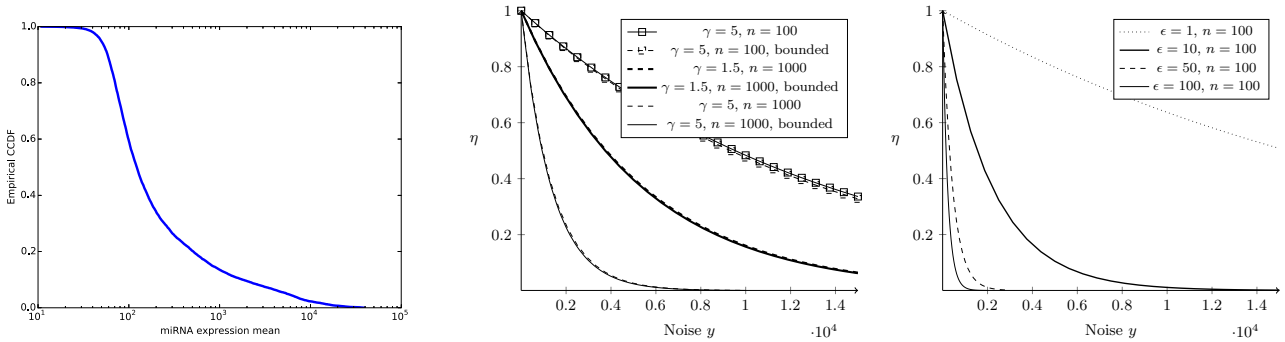
Proof. By Theorem 3.8 in [12], it holds that

$$\Pr\left[|f_i(T) - f'_i(T)| \geq ln\left(\frac{1}{\alpha}\right)\left(\frac{\Delta(f)}{\epsilon}\right)\right] \leq \alpha$$

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 anyone of the output values f_i.

By setting $y = ln\left(\frac{1}{\alpha}\right)\left(\frac{\Delta(f)}{\epsilon}\right)$, replacing $\Delta(f)$ by the formula derived in Definition 3, and solving 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. Fig. 3(a), which



(a) (b) (c)

Figure 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 $n = 100$.

represents the miRNA expression means' empirical complementary cumulative distribution function, 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's. As the sensitivity is, for every miRNA, by definition, the sum over all miRNAs' ranges, it affects the noise distribution added to every miRNA similarly. The probability bound on the maximum noise added to $f_{\mathsf{avg},i}$ is thus large unless the pool contains a large number n of samples, 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 [32], 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,[5] but rather the probability that an individual contributed his data to a pool, given that he is part of a given population, 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 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.

[5]Under this assumption, the probability of being in the case or control group is typically 0.5 [32].

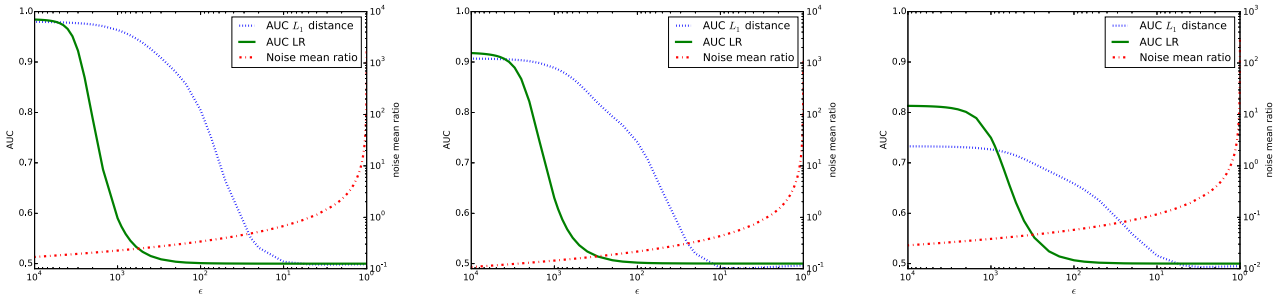
		ϵ		
D	a, b	$\gamma = 1.3$	$\gamma = 1.5$	$\gamma = 5$
D19	0.0003	0.2624	0.4056	1.6104
D17	0.0013	0.2627	0.4061	1.6145
D3	0.009	0.2651	0.41	1.6464

Table 1: Privacy parameters for the diseases D19 (stomach tumor), D17 (renal cancer) and D3 (prostate cancer – male only) achieving γ-membership privacy under prior probability determined from disease prevalence rate in the US (collected on [2]).

Fig. 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 1.

We make the following observations from Fig. 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 1 leads us to the same conclusion. Therefore, we suggest to make 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 in the next subsection.

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



(a) (b) (c)

Figure 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, (b) renal cancer, (c) prostate cancer.

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 both privacy, 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 bad unless we significantly increase the privacy parameter ϵ. Fig. 3(c) shows how our utility bound evolves depending on the parameter ϵ. By comparing the noise values y with the means' CCDF of Fig. 3(a), we clearly notice that the noise is too large with respect to most of the miRNAs' means with 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 3 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).

4.2 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 3.2, 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.

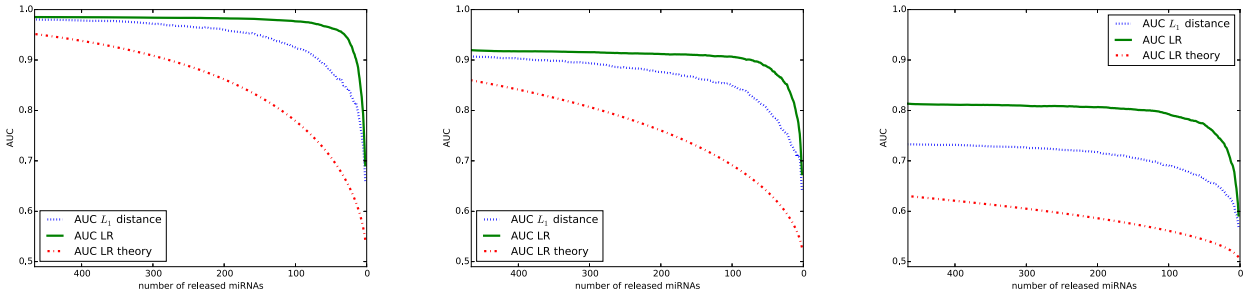
In this subsection, we focus on three case groups related to cancer that represent the groups for which the membership attack was most successful (see Fig. 2). Figures 4(a)–(c) show the AUC of the L_1 distance and LR attacks, and the noise-to-mean ratio $\frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i|}{\hat{\mu}_i}$ 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 pool's mean values. 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 bringing 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 (Fig. 4(a)) to around 10% added noise for prostate cancer (Fig. 4(c)). We clearly observe that the number n of participants 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



(a) (b) (c)

Figure 5: Membership inference attacks in the presence of a hiding mechanism. AUCs for three case groups: (a) stomach tumor, (b) renal cancer, (c) prostate cancer.

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 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(a)–(c) show the AUCs of the attacks presented in Section 3.

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, yielding the LR attack to always outperform the L_1 distance attack. We also observe that the theoretical LR's AUC slightly underestimates the success of the attack, as already noticed in Subsection 3.2, due to the disease-specific pool. Moreover, we notice that theoretical AUC curves are shaped like $\sqrt{m}$, as expected from relation (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 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. RELATED WORK

Here, we present the previous work on membership privacy in genome-wide association studies (GWAS) and how it relates to our work.

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 [19]. Wang et al. extend this attack by making use of the correlations among the different positions in the genome [34]. This improvement on the attack allows them to use the statistics related to only a few hundreds genetic variants. Zhou et al. further analyze the theoretical complexity of membership and recovery attacks based on summary statistics [38]. Sankararaman et al. show empirically that the likelihood-ratio test is more powerful than the L_1 distance attack proposed by Homer et al. [31]. Moreover, they derive a theoretical bound on the LR test that provides a very good approximation of the empirical LR test. Our work confirms that, for miRNA expression data, the empirical LR test is better than the L_1 distance attack. In contrast, 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: (i) the dimensions of both m and n are relatively small compared to those in the genomic setting considered in [31], typically an order of magnitude smaller for both, and (ii) 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).

On the defense side, various papers have studied how to properly apply noise on 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 [20]. Uhler et al. have also proposed to rely on differential privacy for sharing GWAS results privately. In [33], they present methods for privately disclosing allele frequencies, chi-square statistics, and p-values. In [36], 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. In [37], Yu et al. present a differentially private mechanism for logistic regression and show how it can be applied to the analysis of GWAS data. 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 [16]. 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. [32] 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. 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 can rather be used 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 [13]. Contrary to their work, we have 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 greater 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 result, 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.

6. CONCLUSION AND FUTURE WORK

This work sheds light on privacy risks stemming from miRNA expression data, showing that it is possible to detect membership in miRNA-based studies' datasets 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 contain individuals carrying specific diseases, their statistics are further away from healthy general population's statistics, which 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 attacks' 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 couple of hundreds 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 rather wants to detect membership between two pools.

7. ACKNOWLEDGEMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0656) and by the German Research Foundation (DFG) via the collaborative research center "Methods and Tools for Understanding and Controlling Privacy" (SFB 1223), project A5.

8. REFERENCES

[1] Arrayexpress. https://www.ebi.ac.uk/arrayexpress. Accessed: 2016-02-12.

[2] Cancer statistics. http://seer.cancer.gov/statfacts/html/all.html. Accessed: 2016-05-18.

[3] Gene expression omnibus. http://www.ncbi.nlm.nih.gov/geo. Accessed: 2016-02-12.

[4] Genome-wide association studies. https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/. Accessed: 2016-04-29.

[5] 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: 2016-02-03.

[6] Human miRNA disease database. http://www.cuilab.cn/hmdd. Accessed: 2016-05-15.

[7] Medical data - a new target for hackers. https://www.logpoint.com/se/about-us/blog/249-medical-data-a-new-target-for-hackers. Accessed: 2016-02-03.

[8] Number of microRNAs in human genome skyrockets. http://www.genengnews.com/gen-news-highlights/number-of-micrornas-in-human-genome-skyrockets/81250958/. Accessed: 2016-04-29.

[9] M. Backes, P. Berrang, A. Hecksteden, M. Humbert, A. Keller, and T. Meyer. Privacy in epigenetics: Temporal linkability of microRNA expression profiles. In *Proceedings of the 25th USENIX Security Symposium*, 2016.

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

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data

analysis. In *Proceedings of the Third Conference on Theory of Cryptography (TCC)*, pages 265–284, 2006.

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

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

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

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

[16] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Security Symposium*, pages 17–32, 2014.

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

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

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

[20] A. Johnson and V. Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1079–1087, 2013.

[21] P. A. Jones and S. B. Baylin. The epigenomics of cancer. *Cell*, 128:683–692, 2007.

[22] A. Keller, P. Leidinger, A. Bauer, A. ElSharawy, J. Haas, C. Backes, A. Wendschlag, N. Giese, C. Tjaden, K. Ott, et al. Toward the blood-borne mirnome of human diseases. *Nature methods*, 8:841–843, 2011.

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

[24] P. Leidinger, C. Backes, S. Deutscher, K. Schmitt, S. C. Mueller, K. Frese, J. Haas, K. Ruprecht, F. Paul, C. Stahler, et al. A blood based 12-mirna signature of alzheimer disease patients. *Genome Biol*, 14, 2013.

[25] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 889–900, 2013.

[26] E. Londin, P. Loher, A. G. Telonis, K. Quann, P. Clark, Y. Jing, E. Hatzimichael, Y. Kirino, S. Honda, M. Lally, 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):E1106–E1115, 2015.

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

[28] J. Neyman and E. S. Pearson. *On the problem of the most efficient tests of statistical hypotheses*. 1992.

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

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

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

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

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

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

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

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

[37] F. Yu, M. Rybar, C. Uhler, and S. E. Fienberg. Differentially-private logistic regression for detecting multiple-snp association in gwas databases. In *Privacy in Statistical Databases*, pages 170–184, 2014.

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