Membership Inference Against DNA Methylation Databases

Inken Hagestedt*, Mathias Humbert†, Pascal Berrang*, Irina Lehmann‡,
Roland Eils§, Michael Backes*, Yang Zhang*
*CISPA Helmholtz Center for Information Security, {inken.hagestedt, pascal.berrang, backes, zhang}@cispa.saarland
†Cyber-Defence Campus, armasuisse S+T, mathias.humbert@armasuisse.ch
‡Helmholtz Centre for Environmental Research Leipzig, UFZ, irina.lehmann@ufz.de
§Berlin Institue of Health, r.eils@dkfz-heidelberg.de

***Abstract*—Biomedical data sharing is one of the key elements fostering the advancement of biomedical research but poses severe risks towards the privacy of individuals contributing their data, as already demonstrated for genomic data. In this paper, we study whether and to which extent DNA methylation data, one of the most important epigenetic elements regulating human health, is prone to membership inference attacks, a critical type of attack that reveals an individual's participation in a given database. We design and evaluate three different attacks exploiting published summary statistics, among which one is based on machine learning and another is exploiting the dependencies between genome and methylation data. Our extensive evaluation on six datasets containing a diverse set of tissues and diseases collected from more than 1,300 individuals in total shows that such membership inference attacks are effective, even when the target's methylation profile is not accessible. It further shows that the machine-learning approach outperforms the statistical attacks, and that learned models are transferable across different datasets.**

***Index Terms*—epigenetics, membership inference**

1. Introduction

With the rapidly decreasing costs of molecular profiling, the available biomedical data types are becoming increasingly diverse and go beyond the genomes of individuals. DNA methylation is one of the most important new types of biomedical data and is a key regulator of gene transcription. Abnormal methylation patterns can lead to severe diseases, such as cancer [11], [16], [56]. Moreover, DNA methylation is highly related to environmental cues, such as pollution, exposure to stress, or cigarette smoke [7], [51], [52], [54]. Despite being linked with such sensitive information, DNA methylation data are already available on various open research platforms, e.g., the Gene Expression Omnibus (GEO) [20].

Contrary to genomic data whose privacy has been extensively studied by the security research community [15], [33], [36], the privacy risks of epigenomic data have not yet been well investigated. One of the most critical attacks in the biomedical research setting is membership inference, popularized by Homer et al. [24]. Its idea is as follows: Given some raw data about a targeted individual, the attacker wants to know whether this individual is a member of a dataset (i.e., has contributed his data) by relying solely on aggregated statistics about this dataset. The attack on genomic data and its countermeasures have been investigated in numerous research papers over the last decade [26], [45], [53], [57], [59], [60].

In this paper, we aim at evaluating whether DNA methylation databases are also vulnerable to membership inference attacks. Because some regions of our methylation profiles are highly correlated with the genome, leakage of such data can indirectly expose family members' private data. Furthermore, it is uncertain how legal frameworks such as the US Genetic Information Nondiscrimination Act (GINA) apply to epigenomic data like DNA methylation [14], [43]. As a consequence, anticipating privacy risks and mitigating them with technical means is of utmost importance. On the other hand, it is unclear to what degree measurement noise in DNA methylation data and naturally occurring variability serve as privacy-protective noise. This is especially important when testing across tissue types since methylation varies across tissues as opposed to the genome that is the same in all kinds of body cells.

Contributions We present three different membership inference attacks against DNA methylation databases. These attacks differ in the attacker's knowledge (see Table 1 for a summary). For the first attacker, we assume she knows the methylation profile of the target and the means of the corresponding methylation positions from the targeted database. We rely on state-of-the-art statistical tests such as the log-likelihood ratio test to infer membership of the target in the database. For the second attacker, we assume individual methylation values of other patients are available, which enables her to learn meaningful features to improve her attack. Given that methylation values may vary across different tissues, we further study data transferability and show that a machine-learning (ML) model trained on one database can be used to perform membership inference on another database of different tissues or diseases. For the third attacker, we assume the methylation profile of the target not to be available and propose a new membership inference attack that only relies on genomic variants that are correlated with the methylation positions.

We conduct an extensive evaluation of our attacks on six diverse datasets containing the methylation profiles from different tissues, such as blood, brain, and breast cancer, of a total of 1,320 patients. Our results consistently demonstrate the success of membership inference attacks over different tissues and diseases. We further observe that

DOI 10.1109/EuroSP48549.2020.00039

Attacker's knowledge	Statistics about methylation database	Raw target data		External/auxiliary data	
		methylation	genome	methylation	genome + methylation
Statistical attack	✓	✓	-	-	-
Machine-learning attack	✓	✓	-	✓	-
Genome-based attack	✓	-	✓	-	✓

TABLE 1: Overview of our different attack settings. ✓means the attack needs this information, and - means it does not. When relying on methylation data, we also use training/test data from different tissues or with different diseases.

ML approaches outperform statistical attacks: While the best performing statistical test, the LLR test, exceeds 0.9 AUC (area under the ROC curve) for one dataset only, the machine-learning attack almost always reaches AUC of at least 0.9. Our empirical results also show that data is transferrable between different diseases and tissues: The model trained on a different type of dataset from the target dataset achieves similar performance to the model trained on the same type of dataset. Finally, the genome-based attack provides excellent performance with around 0.9 AUC, even though it performs slightly worse than the methylation-based attack,

To summarize, we make the following contributions:

- We study the feasibility of membership inference attacks against one of the most prevalent type of epigenomic data, DNA methylation.
- We propose new ML-based attacks that are able to learn relevant features even from DNA methylation data coming from other tissues.
- We propose an indirect membership inference attack that shows that DNA methylation databases are prone to membership inference attacks even without having access to the methylation data of the target but only to some of her genomic variants.

Organization In Section 2, we present the relevant biomedical background for our attacker models in Section 3 and the theoretical foundations of our attacks in Section 4. In Section 5, we detail the diverse datasets used for the evaluation of our attacks in Section 6. Finally, we summarize related work in Section 7 before concluding in Section 8.

2. Background

DNA methylation is one of the most important epigenetic modifications, affecting both the structure and activity of the DNA molecule [27], [46]. The methylation process consists in the addition of a molecule, namely, a methyl group, to the C (cytosine) nucleotide. Since DNA methylation may vary between copies of the DNA and across different cells, its value is quantified as the fraction of methylated nucleotides at a given genome position. Therefore, any DNA methylation position takes value in $\mathbb{R}_{[0,1]}$. With the current DNA methylation profiling technology, we can easily get access to several hundreds of thousands of DNA methylation positions in the human genome (e.g., the Illumina array provides 450k positions). We can get even more positions (up to tens of millions) by relying on more advanced technology such as whole-genome sequencing. Recent studies show that environmental factors such as exposure to stress or cigarette smoke, as well as the individual's age, correlate with changes in methylation values [7], [30], [31], [51], [54]. Moreover, aberrant DNA methylation patterns are often correlated with cancer stemming from the activation of genes such as oncogenes or the silencing of tumor suppressor genes [16].

	# positions	value range	time evolution
DNA methylation	$\sim 10^7$	$\mathbb{R}_{[0,1]}$	varying
SNPs	$\sim 10^8$	$\{0,1,2\}$	stable

TABLE 2: Key differences between DNA methylation and genomic variants (SNPs). Note that the number (#) of positions shows the total number of currently known positions but that this number can be orders of magnitude smaller in popular profiling technology such as the Illumina array (450k positions).

Besides environmental factors, methylation regions can also be influenced by genomic variants at some specific positions [18], [32], [49]. Single genomic positions that vary among individuals in a population are referred to as single nucleotide polymorphisms (SNPs). A SNP is determined by a pair of nucleotides (among {A,C,G,T}): one that is called major allele as it is most frequent in the population and the other that is called the minor allele. Therefore, a given SNP can take three values: two major alleles, typically encoded as 0, one major allele and one minor allele, encoded as 1, and two minor alleles, encoded as 2. We summarize the main differences between DNA methylation and SNPs in Table 2.

3. Threat Model

The adversary's objective is to determine whether an individual (referred to as a *target*) is a member of a group of study that we will refer to as a *pool*. By leveraging such an attack, the adversary can infer sensitive information about her target, as pools in medical studies can be associated with severe diseases.

To run her attack, the adversary gets access to aggregated methylation data that describe the statistical properties of the considered methylation data pool. While these aggregate data are usually published alongside biomedical case-control studies (see for example [28], [35], [48]), such aggregate data can nowadays also be queried from federated systems such as i2b2 [34], SHRINE [58] or MedCo [42]. In this work, we assume that the *mean* statistics about the pool are available to the adversary as it is the most common statistics currently available. Additionally, we assume the adversary has access to general methylation statistics of the reference population. Currently, these statistics have to be estimated by the adversary using a subset of the underlying population. However, we expect that population-wide statistics for

DNA methylation will become publicly available, as for genomic data. We will refer to the subset of the reference population as the *reference group* here.

In order to perform the attack, the adversary also needs access to some raw data of the target. For our methylation-based attack, we assume the adversary knows the target's DNA methylation at m positions encoded as $\overrightarrow{x} \in \mathbb{R}^m_{[0,1]}$, from similar or different tissue/disease type as the targeted database. Full individual DNA methylation profiles are increasingly available in public databases such as the Gene Expression Omnibus (GEO) [20] or ArrayExpress [4]. Moreover, with the increasing adoption in medical practice, DNA methylation data will also certainly be stored on hospital servers, potentially putting such profiles at risk. For instance, cyber-attacks against healthcare companies have increased by 72% from 2013 to 2014 [8].

As genomic data is currently more accessible than methylation data, we also propose and investigate a genome-based attack. We assume the adversary knows (part of) the genotype of the target instead of his methylation data. By now, more than 10 million individual genotypes have been sequenced through direct-to-consumer genetic testing [10], such as 23andMe [2] or AncestryDNA [3]. Those individuals can also share their sequenced genotypes online, on open platforms such as GEDmatch [19], OpenSNP [37], or the Personal Genome Project (PGP) [39], sometimes with their real identifiers. Therefore, even without considering the genomic databases at clinical premises, millions of genomic profiles are already freely available online.

4. Attacks

In this section, we present the analytical details of our membership inference attacks. We start with the methylation-based attack upon which we will build the genome-based membership inference attack.

4.1. Methylation-based Attack

Assuming the adversary has access to summary statistics of the pool, we analyze whether it is possible to infer whether the target is part of it by relying on statistical or machine-learning methods.

4.1.1. Difference of L_1 Distances. Homer et al. [24] have shown for genomic statistics that one can rely on the L_1 distance to infer membership in databases based on mean values only. We first evaluate how this method performs when applied to DNA methylation. The attack compares, for a methylation position j, the differences between the target's methylation value x^j and the mean statistics of the pool and reference group, and it determines which mean statistics is closest to x^j. Defining the mean values as μ_p^j for the pool and μ_r^j for the reference group, we have the following L_1 distances' difference:

$$D(x^j) = |x^j - \mu_r^j| - |x^j - \mu_p^j| \tag{1}$$

for the methylation position j. A value greater than 0 indicates that x^j is more likely to belong to the pool, while a value smaller than 0 indicates x^j is more likely to belong to the reference group. Intuitively, the L_1 test exploits the fact that the target's methylation value x^j influences the mean of its group. Therefore, the target's value x^j is expected to be closer to the mean value of the target's group than to the mean value of the other group.

Finally, we rely on the one-sided Student's t-test on the outcome of $D(x^j)$ for all methylation points j to test whether the target is part of the pool or reference group.

4.1.2. Log-Likelihood Ratio (LLR) Test. Additionally, we exploit the likelihood-ratio (LR) test, which has the notable advantage of reaching the maximum achievable power (true-positive rate) for a given false-positive level. This is explained theoretically by the Neyman-Pearson lemma, and its higher power compared to the L_1 test has been demonstrated empirically with genomic data [45].

However, the LR test poses assumptions on the data distribution. We rely on the normal distribution to model the distribution of methylation values, which is the continuous probability distribution that best fits the observed methylation data.[1] We evaluate in the next section whether this model is good enough to keep LR test's power high with actual methylation data.
The general formula for the LR test at position j is:

$$LR_j(x^j) = \frac{\sigma_r^j}{\sigma_p^j} e^{\frac{(x^j-\mu_r^j)^2}{2(\sigma_r^j)^2} - \frac{(x^j-\mu_p^j)^2}{2(\sigma_p^j)^2}} \tag{2}$$

where σ_r^j is the standard deviation of the reference group and σ_p^j the standard deviation of the pool at methylation position j. By taking the logarithm and summing over the m known methylation positions, we get the following log-likelihood ratio (LLR) formula:

$$LLR(\overrightarrow{x}) = \sum_{j=1}^{m} \frac{(x^j-\mu_r^j)^2}{2(\sigma_r^j)^2} - \frac{(x^j-\mu_p^j)^2}{2(\sigma_p^j)^2} + \log\frac{\sigma_r^j}{\sigma_p^j} \tag{3}$$

In this work, we assume the adversary gets access to the mean values of the pool but not to its standard deviations. A reasonable approximation of the standard deviation can be computed from the reference population under the assumption that the standard deviation is approximately the same for the pool.

Hence, we have $\sigma_p^j \approx \sigma_r^j := \sigma^j$, and the above expression simplifies to:

$$LLR(\overrightarrow{x}) = \sum_{j=1}^{m} \frac{(x^j-\mu_r^j)^2 - (x^j-\mu_p^j)^2}{2(\sigma^j)^2} \tag{4}$$

Note that, following an assumption made in previous works on membership privacy [6], [24], [45], we do not consider dependencies that may exist between different methylation points.

4.1.3. Machine-Learning Approach. The two previous statistical tests assume implicitly that the distance between mean and methylation value is equally informative for membership inference no matter the methylation position

1. We tested for equality to the normal distribution using the Kolmogorov-Shmirnov test and a p-value of 0.1 and observed $\frac{1}{3}$ to $\frac{2}{3}$ of the methylation regions being normally distributed, the value varying between datasets. We also tested other distributions, such as the beta distribution, but did not find anything fitting the methylation data better.

j. This assumption may not be true: There might be methylation positions that are sensitive to environmental or genetic variants, leading to a higher variance and thereby easier membership detection in the dataset.

To model a realistic attacker, we assume her to use the data itself to detect informative methylation regions and increase the success probability. We expect that an exceptionally high or low distance of the target to the pool means is more informative for membership inference. Similar to the statistical approaches, we rely on the L_1 and L_2 distances, both to pool and reference means. A division by the standard deviation additionally takes the data variability of the position into account, which simplifies comparison across multiple positions.

All of the aforementioned metrics have to be explored systematically, which we do by using machine learning. We fit a logistic regression classifier[2] that learns how to weight features obtained from different methylation regions. We explore the metrics using the following types of features:

1) L_1 distance to pool mean, formally: $|x^j - \mu_p^j|$ (referred to as L_1 distance feature)
2) Squared L_2 distance to pool mean, formally: $(x^j - \mu_p^j)^2$ (referred to as L_2 distance feature)
3) L_1 distance divided by the standard deviation, formally: $\frac{|x^j - \mu_p^j|}{\sigma^j}$ (referred to as scaled L_1 feature)
4) Squared L_2 distance divided by the variance, formally: $\frac{(x^j - \mu_p^j)^2}{(\sigma^j)^2}$ (referred to as scaled L_2 feature)
5) L_1 distance as used in the L_1 test, formally: $|x^j - \mu_r^j| - |x^j - \mu_p^j|$ (referred to as L_1 feature)
6) Log-likelihood ratio as used in the LLR test, formally: $\frac{(x^j - \mu_r^j)^2 - (x^j - \mu_p^j)^2}{2(\sigma^j)^2}$ (referred to as LLR feature)

To compute these features, we first obtain pool and reference means and approximate standard deviations as before for the LLR test. For each training value tr_j from a training patient, we compute the feature with the mean and standard deviation of the respective methylation position j. Features from different positions are combined into a feature vector. We then sort the features of each vector by increasing order of magnitude. This breaks the link between the learned weight and the methylation position j from which tr_j originated, but recall that our training objective is not which position j is more informative, but rather which distance is more informative for membership inference.

Subsampling: To increase the number of samples for learning while keeping the total amount of patients' data the attacker needs to know low, we generate more than one feature vector from each patient by randomly sampling s disjoint subsets of l methylation positions each. These multiple feature vectors are treated separately during training, but at test time they are combined with majority voting to eventually classify each patient into a single group. Details on the number of feature vectors per patient and length of the feature vectors are empirically evaluated in Section 6.

2. We opted for logistic regression due to its simplicity and the interpretability of the learned model.

We also apply subsampling to the L_1 and LLR tests to compare these directly with the ML approach, i.e, to tell apart the effect of the different settings for machine learning and the benefit of machine learning itself.

4.2. Genome-based Attack

In the following, we assume the attacker does not know the target's methylation values, but the target's genome instead. Genomic data is currently more available and easier to find online or via direct-to-consumer genetic testing services. The adversary can rely on correlations between the genome and methylation in specific regions. After inferring the methylation values, the attacker can mount the same attack as previously described, i.e., against a pool of methylation data. In the experimental evaluation, we will investigate if and to which extent the performance drops when genomic data is used instead of methylation data.

We still assume the attacker knows the mean methylation values of pool and reference group and estimates of the standard deviation from the reference group. Additionally, we assume the attacker has a set of paired methylation and genome data to identify the pairs of correlated methylation and genomic positions and to learn the conditional distribution of methylation values given the genomic values. This section shows how to extend our statistical tests and how to implement the necessary estimates to handle this attack scenario.

As demonstrated by Backes et al. [5], the conditional distribution of a methylation value x^j given a specific SNP g^i can be modeled with a normal distribution. Dropping the position index i of the SNP for simplicity, we define the probability distribution over the methylation values for a specific SNP value $g \in \{0, 1, 2\}$ as

$$f_g(x^j) = p(X_j = x^j \mid G = g) = \frac{1}{\sqrt{2\pi}\sigma_{j,g}} e^{-\frac{(x^j - \mu_{j,g})^2}{(\sigma_{j,g})^2}} \tag{5}$$

where $\mu_{j,g}$ and $\sigma_{j,g}$ denote the mean and standard deviation of $f_g(x^j)$, respectively.

Given this probability distribution, the following theorem shows that the expected log-likelihood ratio test for an individual carrying a given genotype boils down to using $\mu_{j,g}$ in place of the target's methylation value.

Theorem 1. *Assuming $\sigma_p^j \approx \sigma_r^j := \sigma^j$ for all methylation positions correlated with the genome, the LLR test based on the individual's genome is:*

$$LLR(g) = \sum_{j=1}^{m_c} \frac{(\mu_{j,g} - \mu_r^j)^2 - (\mu_{j,g} - \mu_p^j)^2}{2(\sigma^j)^2}, \tag{6}$$

where m_c represents the number of methylation positions correlated with the genome.

Proof. We derive hereafter the formula for the general case with different σ_p^j and σ_r^j. For a given methylation

point j, we need to integrate x^j over all its possible values given g:

$$LLR^j(g) = \frac{1}{2(\sigma_r^j)^2}\int_0^1 (x^j-\mu_r^j)^2 f_g(x^j)dx^j$$
$$-\frac{1}{2(\sigma_p^j)^2}\int_0^1 (x^j-\mu_p^j)^2 f_g(x^j)dx^j + \log\frac{\sigma_r^j}{\sigma_p^j}\int_0^1 f_g(x^j)dx^j$$

By setting $\Delta_j^c = \mu_{j,g} - \mu_p^j$ and $\Delta_j^r = \mu_{j,g} - \mu_r^j$:

$$LLR^j(g) = \frac{1}{2(\sigma_r^j)^2}\int_0^1 (x^j-\mu_{j,g}+\Delta_j^r)^2 f_g(x^j)dx^j$$
$$-\frac{1}{2(\sigma_p^j)^2}\int_0^1 (x^j-\mu_{j,g}+\Delta_j^c)^2 f_g(x^j)dx^j + \log\frac{\sigma_r^j}{\sigma_p^j}$$
$$=\frac{1}{2(\sigma_r^j)^2}\int_0^1 (x^j-\mu_{j,g})^2 f_g(x^j)dx^j$$
$$+\frac{\Delta_j^r}{(\sigma_r^j)^2}\int_0^1 (x^j-\mu_{j,g}) f_g(x^j)dx^j + \frac{(\Delta_j^r)^2}{2(\sigma_r^j)^2}\int_0^1 f_g(x^j)dx^j$$
$$-\frac{1}{2(\sigma_p^j)^2}\int_0^1 (x^j-\mu_{j,g})^2 f_g(x^j)dx^j$$
$$-\frac{\Delta_j^c}{(\sigma_p^j)^2}\int_0^1 (x^j-\mu_{j,g}) f_g(x^j)dx^j$$
$$-\frac{(\Delta_j^c)^2}{2(\sigma_p^j)^2}\int_0^1 f_g(x^j)dx^j + \log\frac{\sigma_r^j}{\sigma_p^j}$$

By using the central moments of the normal distribution, we eventually get:

$$LLR^j(g) = \frac{\sigma_{j,g}^2}{2(\sigma_r^j)^2} + \frac{(\Delta_j^r)^2}{2(\sigma_r^j)^2} - \frac{\sigma_{j,g}^2}{2(\sigma_p^j)^2} - \frac{(\Delta_j^c)^2}{2(\sigma_p^j)^2} + \log\frac{\sigma_r^j}{\sigma_p^j}$$

If $\sigma_p^j \approx \sigma_r^j := \sigma^j$, the above formula simplifies to

$$LLR^j(g) = \frac{\sigma_{j,g}^2}{2(\sigma_r^j)^2} + \frac{(\Delta_j^r)^2}{2(\sigma_r^j)^2} - \frac{\sigma_{j,g}^2}{2(\sigma_p^j)^2} - \frac{(\Delta_j^c)^2}{2(\sigma_p^j)^2}$$
$$= \frac{(\Delta_j^r)^2 - (\Delta_j^c)^2}{2(\sigma^j)^2} = \frac{(\mu_{j,g}-\mu_r^j)^2 - (\mu_{j,g}-\mu_p^j)^2}{2(\sigma^j)^2}$$

We obtain the final formula by summing over all methylation points m_c correlated with the genome. □

Similarly, for the L_1 test, we use the expected methylation value $\mu_{j,g}$ given the genotype g as the target's methylation value.

5. Datasets

For our evaluation, we rely on six datasets containing methylation profiles from diverse tissues of patients carrying different diseases. In total, we use the methylation profiles of 1,320 patients. Table 3 summarizes our datasets.

All but the last dataset were generated with the Illumina 450k array that determines the DNA methylation at 450,000 fixed positions. We refer to these datasets by the disease the respective patients carry. Our last dataset, the WGBS dataset, contains both the genome and the methylation of 75 patients

where the DNA methylation profiles have been generated by whole-genome bisulfite sequencing (WGBS). This results in a full view of DNA methylation patterns in the whole genome of blood cells.

Preprocessing Most of the datasets have missing methylation sites (positions) for specific patients or even missing methylation sites for all the patients sharing the same disease. We remove all methylation positions with missing data, which provides us with 299,998 different methylation positions for the combination of brain cancers and IBD, and about 360,000 different methylation positions for the breast cancer dataset.

For our WGBS dataset, we focus on highly correlated pairs of DNA methylation positions and SNPs. We follow the approach of Backes et al. [5] and only keep the pairs with a Spearman rank correlation coefficient larger than 0.49. This provides us with about 300 methylation positions and the single most correlated SNP position for each of those.

Human Subjects and Ethical Considerations The study on WGBS has received an approval from the responsible institutional ethics review board. All other datasets were publicly available in their anonymized form. All datasets have been stored and analyzed in anonymized form without access to non-anonymized data. Moreover, since we only randomly split the patients into pool and reference sets, the membership inference attacks do not reveal any more information than previously known by us. This way, we ensure that all participants were treated equally and with respect.

6. Attack Evaluation

We start by evaluating the statistical and machine-learning methylation-based attacks. Then, we present the results of our genome-based attack.

6.1. Methylation-based Attack Evaluation

Our evaluation studies the following research questions:

RQ 1	Does the LLR test outperform the L_1 test in the statistics attack setting?
RQ 2	What is the effect of our subsampling approach on the performance of the L_1 and LLR tests?
RQ 3	Which feature is best in the ML attack? Does the performance increase compared to the L_1 and LLR tests with subsampling?
RQ 4	Is it possible to train an attack model on a dataset of a different tissue or disease than the target dataset for the machine learning attack?
RQ 5	What is the influence of the dataset size on the performance of the membership inference attacks?

While RQ1 studies the statistical approach and verifies that the Neyman-Pearson Lemma applies to our data, RQ2 and RQ3 study the foundations of the ML approach. With RQ4 and RQ5 we explore how the ML case works in non-ideal situations, namely, different training and test data and larger dataset sizes.

RQ 1: Comparing the statistical L_1 and LLR test, does the LLR test outperform the L_1 test? To apply the L_1 and LLR tests, we first define pool and reference group.

Abbreviation	Description	Tissue Type	Number of Patients	GSE identifier	by
GBM	glioblastoma	brain cancer	136	GSE36278	[48]
PA	pilocytic astrocytoma	brain cancer	61	GSE44684	[29]
IBD CD	Crohn's disease	blood	77	GSE87640	[55]
IBD UC	ulcerative colitits	blood	79	GSE87640	[55]
BC	breast cancer	breast cancer	892	not publicly available	-
WGBS	genome and methylation data	blood	75	not publicly available	-

TABLE 3: Datasets used in our experiments. THE GSE identifier refers to the accession number in the Gene Expression Omnibus (GEO) database. The BC dataset was available on https://portal.gdc.cancer.gov in April 2017 but is not anymore.

We present a realistic attacker that cannot exploit any disease-specific differences between the databases. For each of our five first datasets, we first randomly sample 60 patients,[3] which are then randomly split into a pool of 30 patients and a reference group of 30 patients. We assume the attacker has means of these 30 patients available as μ_p^j and μ_r^j respectively. Further, we sample 15 patients from the pool and 15 from the reference group at random. The remaining 30 patients are not used in this setup, they serve as training set of the machine learning attack later in this section. We repeat the random splitting five times and present averaged results.

As discussed previously, we assume the attacker has access to the mean of the pool (μ_p^j) and reference group (μ_r^j) for each methylation position j. Moreover, we estimate σ^j by computing the standard deviation over the whole considered dataset.

We simulate membership inference attacks against each patient individually, i.e., all patients from the respective pool and reference group are attacked by applying the L_1 and LLR tests to each methylation position j and summarize across all methylation positions for the given patient as defined by the tests. Using multiple thresholds in the tests, we get a receiver operating characteristic (ROC) curve displaying the false-positive rate ($\frac{FP}{FP+TN}$) on the x-axis and the true-positive rate ($\frac{TP}{TP+FN}$) on the y-axis. The AUC is the area under this curve. An AUC of 0.5 indicates a performance similar to a random guess, whereas an AUC of 0.9 or above indicates an excellent performance. Finally, we average the results over the five random splits. The results are shown in the first five groups of bars in Figure 1.

We observe that the LLR test outperforms the L_1 test, complying with the Neyman-Pearson Lemma. The performance reaches > 0.7 AUC for all diseases when the LLR test is used, and even > 0.95 AUC for PA. Interestingly, the tissue type seems to have an influence on the attack performance, both IBD datasets are sampled from blood and are harder to attack compared to samples from brain cancer tissue for the diseases GBM and PA or breast cancer tissue for BC.

Finally, in order to evaluate a more realistic setting where the reference population is very large, we use our largest dataset on breast cancer (BC) patients. Instead of sampling 30 patients as the reference group, we use all remaining patients, i.e., 862 patients to compute μ_r^j. We observe that the AUC drops by only a few 0.1 compared to the case with a much smaller reference group and conclude that the privacy risks remain valid with a very large reference group.

3. Note that this is the maximum number we can consider if we want to compare the results across all datasets as PA contains 61 patients.

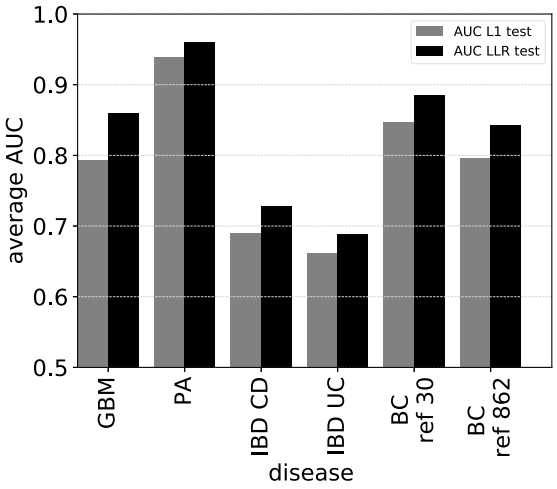


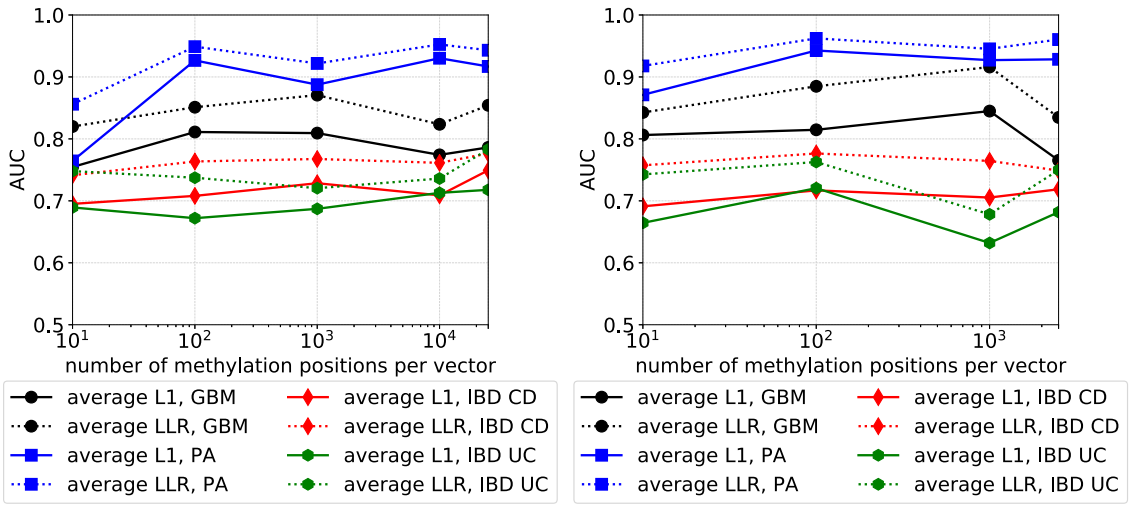
Figure 1: RQ1 (statistical attack setting): AUC of the L_1 and LLR tests applied to all methylation positions, averaged over five random splits of the data simulating attacks against each patient in both pool and reference groups.

Take-home message: the LLR test outperforms the L_1 test with DNA methylation data.

RQ 2: What is the effect of subsampling on the performance on L_1 and LLR test? Which values for the hyperparameters s and l are the best? We subsample each data vector before computing the L_1 test and the LLR test. For each patient, we randomly sample l methylation positions s times without replacement for various settings for s and l. At the end, we combine the inference labels of the s vectors of the same patient with majority voting to get a single outcome for each patient. As before, we first randomly sample 60 patients for each disease, which are then randomly split into 15 pool and 15 reference patients. Again, the remaining 30 patients are not used.

Figure 2a shows the performance of the L_1 (solid lines) and LLR tests (dotted lines) for four of our disease sets, with 10 repetitions of the sampling process. Observing no general trend of l increasing from 10^3 to 10^4, we drop $l = 10^4$ from the parameters which allows to increase s to 100, see Figure 2b.

Comparing the AUC with the standard setup before (Figure 1) shows that for most diseases, the performance of the L_1 test is similar or increased slightly, and the



(a) (b)

Figure 2: RQ2 (subsampling): Influence of the length l of the feature vectors on L_1 and LLR tests performance for four disease datasets when using (a) $s = 10$ vectors or (b) $s = 100$ vectors.

performance of the LLR test increased slightly in almost all cases, independent of how the parameters s and l are set. For example, in GBM the traditional LLR test performance is below 0.85 AUC and with s=100 and l = 10^3 raises to more than 0.9 AUC. The difference between the previous and the current setup is how the membership information from different methylation positions j are combined. Simply taking all of them into account performs worse than first combining a few of them into a binary answer and then taking the majority vote. In the latter case, some methylation positions will therefore not contribute to the answer. This experiment shows that not all methylation positions are informative.

Finally, we observe that a reasonable trade-off between l and s satisfying the constraint $l \cdot s \leq m$ is bounding l to 10^3 and setting $s = 100$.

Take-home message: Subsampling slightly increases performance, and the hyperparameters $l = 10^7$ and $s = 100$ represent a reasonable trade-off.

RQ 3: Which feature is best in the machine learning model, and can the performance be increased compared to the L_1 and LLR test with subsampling? For the machine-learning attack, we use the subsampling trade-off as found before and set $l = 10^3$ and $s = 100$. The remaining 15 pool and 15 reference patients are used as training set. After transforming each value in the training vectors into a feature using the formulas in Section 4.1.3, we sort the vectors' values in ascending order. Then, the vectors are fed into a logistic regression classifier: We rely on the Python library *sklearn* [38] and leave the regularization parameter C at its default 1.0. The classifier learns l coefficients that indicate importance of small, intermediate and large distances (as most of our features are distance-based).

Figure 3 shows the absolute value of the learned coefficients for IBD UC, plots for other diseases look similar. The higher the absolute value of the coefficient,

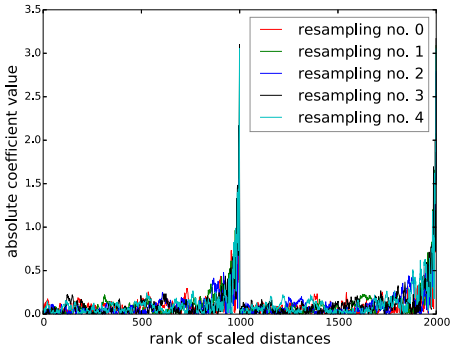
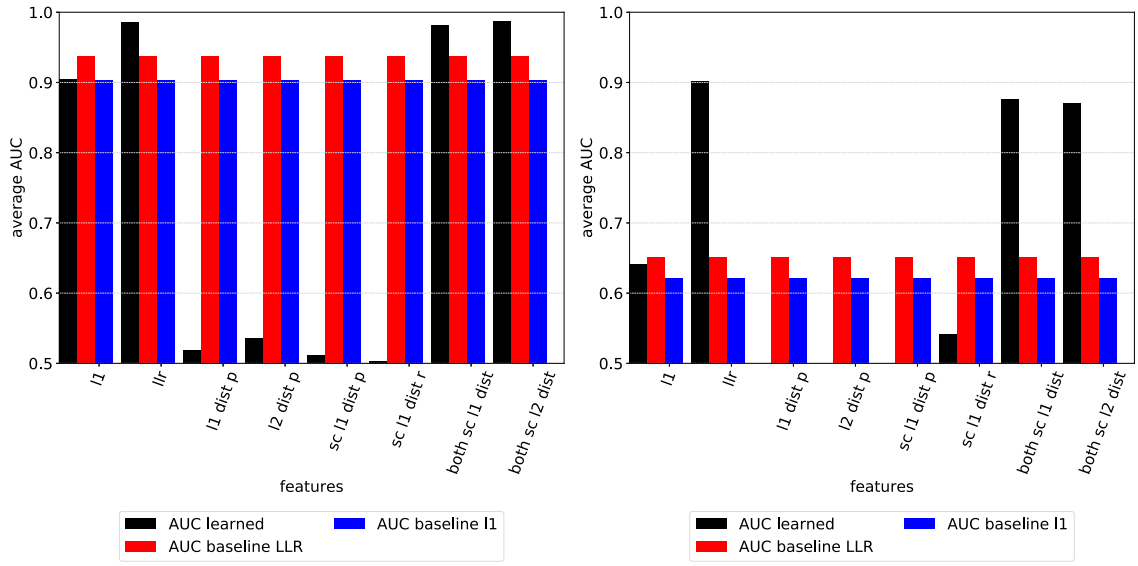


Figure 3: RQ3 (features): Absolute value of learned coefficients on IBD UC with both scaled L_1 features. We show the five repetitions (with different data sampling) of the experiment in different colors to check whether the coefficient values are consistent and not due to randomness. X-axis between 0 and 999 represent the coefficient values for scaled L_1 to the *pool* mean and x-axis between 1,000 and 1,999 represent the coefficient values for scaled L_1 to the *reference* mean.

the more informative the distance is for the classifier. The symmetric pattern arises due to the use of two features: the scaled L_1 distances to both pool and reference means. There is a tendency towards higher values on the right, indicating that higher distance values are more important for the attack. Nevertheless, the lower values do not get zero coefficients, which suggests they also contribute to the model. Additionally, we applied sklearn's recursive feature elimination [21], but the resulting classifiers performed worse in terms of AUC, supporting again the hypothesis that all distances are necessary.

We compare the performance of different features using the AUC of the learned model when applied to the



(a) (b)

Figure 4: RQ3 (features): Performance of different features evaluated on disease datasets (a) PA and (b) IBD UC.

test data. Figure 4 shows exemplarily the performances for PA and IBD UC, the "easiest" and one of the "hardest" disease datasets to attack. We test all the feature types introduced in Section 4.1.3. The distance-based features exist in two versions as distance to the *pool* and to the *reference* mean, respectively, indicated by "p" and "r" in the plot. We omitted the "r" version for some features which performed similarly to their "p" versions. Additionally, we trained on *both* versions of the distance features by concatenating the respective feature vectors. As a baseline, we rely on the L_1 and LLR tests with subsampling. We observe that some features work well, e.g., the LLR feature and using the L_1 or L_2 to both pool and reference group in their scaled form. Other features perform poorly and result in an AUC of around 0.5, e.g., the L_1 and L_2 both with and without scaling. This is why for those features, the black bar is barely visible in Figure 4. Nevertheless, the statistical tests L_1 and LLR are clearly outperformed, especially for the IBD UC and IBD CD datasets.

Take-home message: the performance can be increased by using a machine-learning approach with the LLR features or L_1 or L_2 to both pool and reference groups in their scaled form.

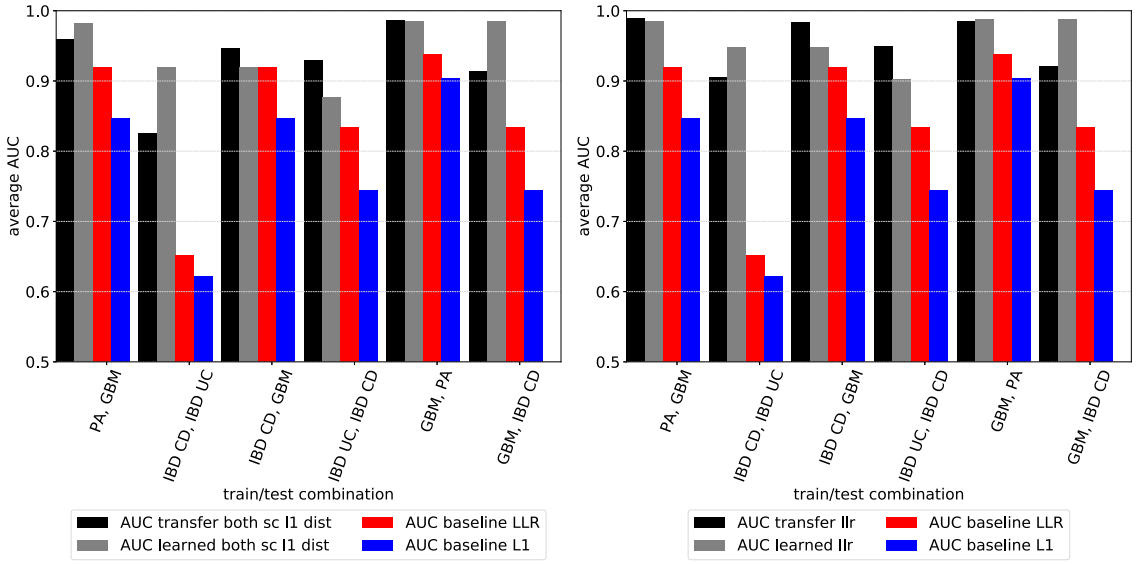
RQ 4: Is it possible to train an attack model on a dataset of a different tissue or disease than the target dataset for the machine learning attack? We study now another, more challenging attack scenario where the attacker trains her machine learning model with pool and reference groups extracted from one dataset and applies this model to pool and reference groups of another dataset. We keep the previous experimental setup, but test the learned model on dataset with different tissue or disease. This setup allows us to evaluate if our membership inference attack is prone to data transferability.

Figure 5 displays the resulting AUCs when learning on the first mentioned (in the x-axis labels) dataset and testing on the second one. Since the performance of the scaled L_1 and scaled L_2 distances are similar (see Figure 4), we show in Figure 5 only the scaled L_1 feature and the LLR feature due to space constraints. Comparing the black and gray bars, we can observe that most cases show a small loss of performance when the attacker learns on patients from a different disease or tissue compared to learning from the same one. However, the learned models still perform well on the different disease set and clearly outperform the statistical L_1 and LLR tests. Recall that the datasets GBM and PA are sampled from brain tumors, while both IBD datasets are from blood samples. According to biomedical research, part of the methylation patterns are tissue specific. However, our results show that our attack based on relative distance instead of methylation positions is prone to transferability even across different tissues.[4]

Take-home message: training and target datasets do not need to be the same for a successful attack.

RQ 5: What is the influence of a larger dataset on the performance of the machine learning model? Our larger dataset on breast cancer allows to study the impact of larger reference group and pool on the attack performance. First, we focus on the reference group size, increasing it from 30 to 800 patients, and keep the number of patients in the pool at 30. This allows us to evaluate whether a more realistic (i.e., larger) reference population has an impact on the attack performance. We evaluate the impact of increasing reference group size on machine learning classifiers trained on the LLR feature and both scaled L_1 features, and on the statistical LLR test using 30 patients for training and testing respectively. We observe in

4. Note that it is very unlikely that these results are due to the same patients being in the datasets because we obtained data from different studies and different diseases.



(a) (b)

Figure 5: RQ4 (transferability): Transferability of learned models based on (a) both scaled L_1 and (b) LLR features.

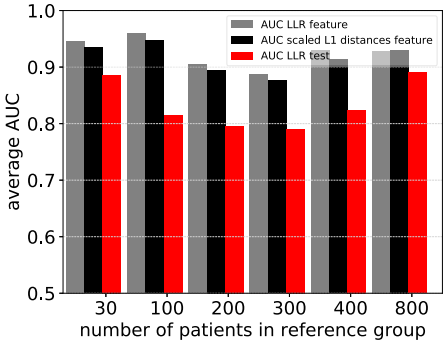


Figure 6: RQ5 (larger datasets): Performance with respect to an increasing number of patients in the reference group only.

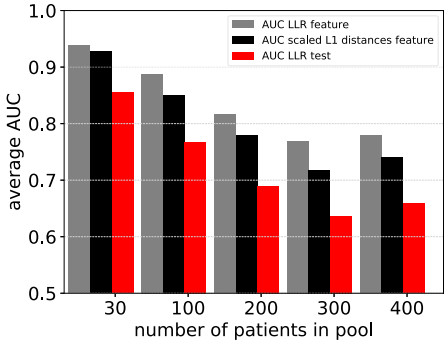


Figure 7: RQ5 (larger datasets): Performance with respect to an increasing number of patients in the pool and reference groups.

Figure 6 that both statistical and ML-based tests perform similarly under increased reference group sizes and that reference group size does not observe any clear influence on the attack performance. This demonstrates that privacy risks remain true with a large reference population, and allows us to extrapolate that membership inference would be possible in non-closed-world settings.

Second, we increase the dataset size from 30 patients in both pool and reference group to 100, 200, 300, and 400 patients. In all cases, we use disjoint training and test sets of the same size which contain the same number of pool and reference patients.

Figure 7 shows that the more patients there are in the pool, the worse the performance of the membership inference attack. As we see in Figure 6, reference group size does not influence the attack success. This confirms previous empirical results with genomic [45] and transcriptomic [6] data, as well as theoretical findings [13]. We further observe that the attack success decreases similarly for both the statistical attack and the ML attacks. We hypothesize that the performance decrease is due to the fact that the more patients are included in the pool, the less each patient contributes to its statistics, in our case, the means, which makes membership inference harder.

On the upside, we can foresee that with declining costs of molecular profiling, the size of epigenomic databases will rapidly grow. Nevertheless, we notice that the ML attack is quite robust to this increase, with still relatively good performance (AUC > 0.8) with 200 patients in the pool.

Take-home message: The attack performance is especially robust with respect to an increase in only the reference dataset size. However, when increasing both pool and reference groups, the attack performance decreases. We conclude that the privacy threat remains even with larger reference population, but also with pool sizes up to

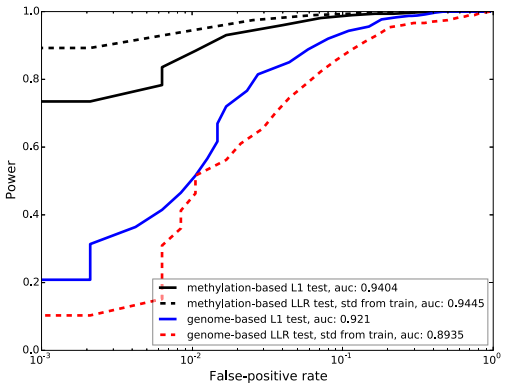


Figure 8: ROC curves of methylation-based L_1 and LLR tests and genome-based L_1 and LLR tests.

200 individuals.

6.2. Genome-based Attack Evaluation

Next, we evaluate the scenario in which the attacker has access to the target's genomic data instead of methylation data. We use the WGBS dataset, containing methylation and genome data of 75 patients. Notice that the data was generated with a different technique (WGBS), which targets different regions of the genome than the Illumina 450k array used for the previous datasets.

We randomly sample half of the patients as a training set (37 patients) which we use to estimate the relationship between genome and methylation data. The second half (38 patients) is used as a test: Half of the patients are chosen at random to be in the pool and the remaining half are in the reference group. The standard deviation is estimated from the training set, which we assume the attacker has full access to. Notice that we have only $m = 300$ methylation positions correlated with the genome, which is tremendously less than $m = 299,998$ used for the previous attacks. Therefore, we do not subsample from the patients. We repeat the experiment five times with different random splits into training and test sets. Moreover, for each of these five splits, we also repeat the splitting into pool and reference group five times, effectively yielding 25 randomly generated runs. As a baseline, we compute the L_1 and LLR tests under the previous assumption that the attacker knows the target's methylation values.

Figure 8 compares the performance of the attacks based on methylation and genomic data. The LLR test exploits the underlying normal distribution of methylation values given a specific genome value. The same technique is used for the L_1 test. We observe first that both L_1 and LLR tests perform worse with access to genome instead of methylation values, as expected. However, they still achieve high performance, which shows that an attacker with only access to the genome of the target can still successfully infer her membership in methylation databases.

Surprisingly, the performance decrease is higher in the case of the LLR test, where AUC drops from 0.94 to 0.89. For the L_1 test, the AUC decreases from 0.94 to 0.92 when we rely on the genomic data instead of the methylation data. One possible explanation for the LLR performance being lower than the L_1 performance in the genome attack is as follows. The estimation of the methylation values for the target given the genome is approximated and induces small errors. Such noisy values have a larger negative influence on the LLR test compared to the L_1 test.

We also applied our ML techniques to the WGBS set, but learning was not possible due to the relatively low number of methylation values. Nevertheless, we conclude that, despite the small drop in overall performance, membership inference is still possible with genomic data that is currently easier to obtain than methylation data.

Take-home message: We conclude that privacy is at risk even if the attacker has not access to the target's methylation data and must estimate them from their genome.

7. Related Work

In the following, we first present previous works related to membership inference attacks, then other attacks against DNA methylation data, and finally defense mechanisms.

Attacks Homer et al. were the first to present a membership inference attack by relying on summary statistics over genomic data and the L_1 distance between those and the target's data [24]. An extension to this attack was proposed by Wang et al. [57] using the intra-genome correlations which allowed to rely on only a few hundreds genomic positions. The theoretical complexity was further studied by Zhou et al. as well as recovery attacks based on summary statistics [60]. Moreover, Sankararaman et al. derived an upper bound on the power of membership inference with genomic data, and showed empirically that the likelihood-ratio (LR) test was more powerful than the L_1 distance attack [45].

Backes et al. [6] were the first to propose a membership inference attack against another type of biomedical data, namely transcriptomic data (microRNA expression). Despite the smaller dimensionality of the microRNA profiles (a few thousands points instead of millions with the genome), the attack based on L_1 distance and the likelihood-ratio test proved to be successful against disease-related databases.

Shokri et al. studied membership inference attacks against the training datasets of machine-learning models such as neural networks [47], while Hayes et al. studied the same attacks against generative models [23]. The authors showed that their attacks can be successfully performed against medical image datasets, further demonstrating the extent of the privacy threat. Moreover, Pyrgelis et al. [41] carried out a membership inference attack against location data. They used statistical features and fit several machine learning classifiers to infer membership. Additionally, they use various differential privacy mechanisms to protect the location data. Recently, Salem et al. showed that membership inference against machine-learning models was even possible with fewer assumptions on the adversarial power than in the first attack model [44]. They further proposed effective defense mechanisms against such attacks and showed that they could still provide a high level of utility for the ML model. Besides, there exist multiple other recent works in this field [9], [25].

Other than membership attacks, Philibert et al. showed that methylation data could be relied upon to infer part of the genotype and behavioral attributes such as alcohol consumption and smoking [40]. Besides also identifying methylation points correlated with genomic variants, Dyke et al. proposed high-level guidelines for methylation data disclosure that preserves privacy [14]. However, neither Philibert et al. nor Dyke et al. proposed concrete attacks and defenses on raw or aggregated methylation data. Backes et al. used the correlations between certain positions of the genome and methylation data in order to re-identify DNA methylation profiles by matching them to their corresponding genome [5]. As opposed to Backes et al. whose goal was to match a methylation profile to the genome of the same person, we use here the statistical relationships between genomic and methylation data in our second attack type to run a membership attack without access to the target's raw methylation data. Finally, Hagestedt et al. [22] designed an online service for finding relevant research datasets of methylation data similar to the Beacon Network [1]. For each methylation position, the service returns a binary answer whether data is available. Despite the coarse output format, they showed that membership inference attacks are feasible on unprotected methylation Beacons and propose a differentially private mechanism to mitigate the privacy threat. While the service itself can be designed in a privacy-preserving way, exact means and standard deviations are used to interpret the answers. Our work shows that these exact means alone pose a significant privacy threat.

Defenses Erlich and Narayanan [15] provided a general overview of the privacy threats to genomic data, which partially also apply to methylation data. They also presented an overview of the defense mechanisms.

How to apply differential privacy to genomic databases has been extensively studied. Johnson and Shmatikov have proposed algorithms that protect the output of data exploration (p-values and correlations, number and location of SNPs most likely associated with a disease) with differential privacy [26]. Uhler et al. have also proposed to release differentially-private summary statistics (allele frequencies, p-values, and χ^2 statistics) [53]. This was extended by Yu et al. to allow for arbitrary numbers of case and control samples [59].

Nevertheless, finding a reasonable trade-off between privacy and utility is not always feasible as Fredrikson et al. pointed out in their case study of warfarin dosing [17]. They notably showed that reasonable privacy risks cannot be attained without putting at risk the health of the patients taking warfarin. Differential privacy mechanisms that have been published to date for data of this type result in high utility loss for modest privacy protection. Therefore, Tramèr et al. [50] proposed a relaxation of differential privacy that assumes a weaker adversary in order to reach a better privacy-utility trade-off.

Finally, Backes et al. have applied differential privacy to microRNA expression's summary statistics for preventing membership inference attacks with such data [6]. Their results confirmed the difficulty of finding a reasonable privacy-utility trade-off, especially when the number of participants in the database is small.

8. Conclusion and Future Work

In this paper, we have thoroughly analyzed whether and to what extent DNA methylation databases are prone to membership inference attacks. In particular, we have considered two attacker models: one assuming the adversary to know her victim's methylation profile, and the second assuming the adversary to know only her victim's genotype. For both settings, we have studied traditional statistical attacks based on the L_1 distance and on the likelihood-ratio test. Additionally, we have proposed a new machine-learning attack that is able to exploit the fact that not all methylation data are equally informative for membership inference. In this setting, we have further studied data transferability, i.e., to which extent learning features from a dataset different than the targeted dataset influences the attack results. For the genome-based inference of membership, we have specifically designed the LLR attack to capture the probabilistic dependencies between the two types of data, and have identified a sufficient statistic for this attack.

We have evaluated our attacks on six different datasets, overall containing the DNA methylation profiles of 1,320 patients. Our empirical results consistently demonstrate the success of membership inference attacks over different tissues and diseases. Even though we were limited by the small number of patients in most of the datasets, the experiments with the larger breast cancer dataset suggested that our findings may scale. We concluded that the membership privacy of contributors to DNA methylation databases is put at risk even if the adversary does not directly get access to their methylation data but only their genomes.

Performing the membership inference attacks with DNA methylation data at different points in time is a future direction that is worth investigating. Moreover, designing attacks that exploit dependencies between methylation points is another interesting direction for future work.

Given the severe privacy risks that we uncover with our attacks, future work should study protection mechanisms. One direction would be to employ a privacy mechanism that allows the data publisher to balance the trade-off between the privacy loss to the individuals resulting from the data publication with the increased utility and benefit to society. Differential privacy [12] is a framework for creating and evaluating such mechanisms. The challenge is to find a mechanism applicable to this specific case that features both high data dimensionality and few individuals (currently) contributing their data. In line with other applications such as MBeacon [22], we believe that there is a clear benefit from sharing population- wide mean methylation values. Mean methylation values could become as relevant and well-studied as minor allele frequencies are today for the genome.

Acknowledgements

The authors from CISPA are partially supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0656). Mathias Humbert carried out most of this work while at the Swiss Data Science Center. He was supported by the

grant #2017-201 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain. The authors thank their shepherd for their guidance throughout the whole shepherding process.

References

[1] Beacon network. https://beacon-network.org. Accessed: 2019-14-05.

[2] 23andme. https://www.23andme.com. Accessed: 2019-18-11.

[3] AncestryDNA. https://www.ancestry.com/dna. Accessed: 2019-18-11.

[4] Arrayexpress. http://www.ebi.ac.uk/arrayexpress. Accessed: 2019-20-07.

[5] Michael Backes, Pascal Berrang, Matthias Bieg, Roland Eils, Carl Herrmann, Mathias Humbert, and Irina Lehmann. Identifying personal dna methylation profiles by genotype inference. In *2017 IEEE Symposium on Security and Privacy (S&P)*, pages 957–976. IEEE, 2017.

[6] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. Membership privacy in microrna-based studies. In *Proceedings of the 23rd ACM Conference on Computer and Communications Security (CCS)*, pages 319–330. ACM, 2016.

[7] T. Bauer, Saskia Trump, Naveed Ishaque, L. Thu rmann, L. Gu, Mario Bauer, Matthias Bieg, Zuguang Gu, Dieter Weichenhan, J.-P. Mallm, S. Ro der, G. Herberth, Eiko Takada, O. Mu cke, Marcus Winter, Kristin M Junge, K. Gru tzmann, U. Rolle-Kampczyk, Qi Wang, Christian Lawerenz, Michael Borte, Tobias Polte, Matthias Schlesner, Michaela Schanne, Stefan Wiemann, C. Geo rg, Hendrik G Stunnenberg, Christoph Plass, Karsten Rippe, Junichiro Mizuguchi, Carl Herrmann, Roland Eils, and Irina Lehmann. Environment-induced epigenetic reprogramming in genomic regulatory elements in smoking mothers and their children. *Molecular Systems Biology*, 12(3):861–861, mar 2016.

[8] 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: 2016-02-03.

[9] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against GANs. CoRR abs/1909.03935, 2019.

[10] 2017 was the year consumer dna testing blew up. https://www.technologyreview.com/s/610233/2017-was-the-year-consumer-dna-testing-blew-up/. Accessed: 2018-06-28.

[11] Partha M Das and Rakesh Singal. DNA methylation and cancer. *Journal of clinical oncology*, 22(22):4632–4642, 2004.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[13] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 650–669. IEEE, 2015.

[14] Stephanie OM Dyke, Warren A Cheung, Yann Joly, Ole Ammerpohl, Pavlo Lutsik, Mark A Rothstein, Maxime Caron, Stephan Busche, Guillaume Bourque, Lars Rönnblom, et al. Epigenome data release: a participant-centered approach to privacy protection. *Genome biology*, 16:1–12, 2015.

[15] Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy. *Nature reviews. Genetics*, 15(6):409, 2014.

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

[17] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *USENIX Security Symposium*, pages 17–32, 2014.

[18] Tom R Gaunt, Hashem A Shihab, Gibran Hemani, Josine L Min, Geoff Woodward, Oliver Lyttleton, Jie Zheng, Aparna Duggirala, Wendy L. McArdle, Karen Ho, Susan M Ring, David M Evans, George Davey Smith, and Caroline L Relton. Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, 17(1):61, 2016.

[19] GEDmatch. https://www.gedmatch.com. Accessed: 2019-18-11.

[20] Gene expression omnibus. https://www.ncbi.nlm.nih.gov/geo. Accessed: 2019-20-07.

[21] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.

[22] Inken Hagestedt, Yang Zhang, Mathias Humbert, Pascal Berrang, Haixu Tang, XiaoFeng Wang, and Michael Backes. Mbeacon: Privacy-preserving beacons for dna methylation data. In *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS). Internet Society*, 2019.

[23] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2019.

[24] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David 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.

[25] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 259–274. ACM, 2019.

[26] Aaron Johnson and Vitaly 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*, pages 1079–1087. ACM, 2013.

[27] Peter a Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics*, 13(7):484–92, jul 2012.

[28] Claudia L Kleinman, Noha Gerges, Simon Papillon-Cavanagh, Patrick Sin-Chan, Albena Pramatarova, Dong-Anh Khuong Quang, Véronique Adoue, Stephan Busche, Maxime Caron, Haig Djambazian, et al. Fusion of ttyh1 with the c19mc microrna cluster drives expression of a brain-specific dnmt3b isoform in the embryonal brain tumor etmr. *Nature genetics*, 46(1):39, 2014.

[29] Sally R Lambert, Hendrik Witt, Volker Hovestadt, Manuela Zucknick, Marcel Kool, Danita M Pearson, Andrey Korshunov, Marina Ryzhova, Koichi Ichimura, Nada Jabado, et al. Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. *Acta neuropathologica*, 126(2):291–301, 2013.

[30] Riccardo E Marioni, Sonia Shah, Allan F McRae, Brian H Chen, Elena Colicino, Sarah E Harris, Jude Gibson, Anjali K Henders, Paul Redmond, Simon R Cox, et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome biology*, 16(1):25, 2015.

[31] Riccardo E Marioni, Sonia Shah, Allan F McRae, Stuart J Ritchie, Graciela Muniz-Terrera, Sarah E Harris, Jude Gibson, Paul Redmond, Simon R Cox, Alison Pattie, et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *International journal of epidemiology*, 44(4):1388–1396, 2015.

[32] Joseph L. McClay, Andrey A. Shabalin, Mikhail G. Dozmorov, Daniel E. Adkins, Gaurav Kumar, Srilaxmi Nerella, Shaunna L. Clark, Sarah E. Bergen, Christina M. Hultman, Patrik K. E. Magnusson, Patrick F. Sullivan, Karolina A. Aberg, and Edwin J. C. G. van den Oord. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. *Genome Biology*, 16(1):291, 2015.

[33] Alexandros Mittos, Bradley Malin, and Emiliano De Cristofaro. Systematizing genomic privacy research–a critical analysis. *arXiv preprint arXiv:1712.02193*, 2017.

[34] Shawn N Murphy, Griffin Weber, Michael Mendis, Vivian Gainer, Henry C Chueh, Susanne Churchill, and Isaac Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, 17(2):124–130, 2010.

[35] Anna K Naumova, Abeer Al Tuwaijri, Andréanne Morin, Vanessa T Vaillancout, Anne-Marie Madore, Soizik Berlivet, Hamid-Reza Kohan-Ghadr, Sanny Moussette, and Catherine Laprise. Sex-and age-dependent dna methylation at the 17q12-q21 locus associated with childhood asthma. *Human genetics*, 132(7):811–822, 2013.

[36] Muhammad Naveed, Erman Ayday, Ellen W Clayton, Jacques Fellay, Carl A Gunter, Jean-Pierre Hubaux, Bradley A Malin, and XiaoFeng Wang. Privacy in the genomic era. *ACM Computing Surveys (CSUR), 2015*, 2015.

[37] Opensnp. https://opensnp.org. Accessed: 2019-18-11.

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

[39] Personal genome project. http://www.personalgenomes.org. Accessed: 2017-20-07.

[40] Robert A Philibert, Nicolas Terry, Cheryl Erwin, Winter J Philibert, Steven RH Beach, and Gene H Brody. Methylation array data can simultaneously identify individuals and convey protected health information: an unrecognized ethical concern. *Clinical epigenetics*, 6(1):28, 2014.

[41] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145*, 2017.

[42] Jean Louis Raisaro, Juan Ramón Troncoso-Pastoriza, Mickaël Misbach, E Sousa Gomes de Sá, Joao André, Sylvain Pradervand, Edoardo Missiaglia, Olivier Michielin, Bryan Alexander Ford, and Jean-Pierre Hubaux. Medco: Enabling privacy-conscious exploration of distributed clinical and genomic data. Technical report, 2017.

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

[44] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security Symposium (NDSS)*. Internet Society, 2019.

[45] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[46] Dirk Schübeler. Function and information content of DNA methylation. *Nature*, 517(7534):321–326, jan 2015.

[47] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 3–18. IEEE, 2017.

[48] Dominik Sturm, Hendrik Witt, Volker Hovestadt, Dong-Anh Khuong-Quang, David TW Jones, Carolin Konermann, Elke Pfaff, Martje Tönjes, Martin Sill, Sebastian Bender, et al. Hotspot mutations in h3f3a and idh1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer cell*, 22(4):425–437, 2012.

[49] Ai Ling Teh, Hong Pan, Li Chen, Mei Lyn Ong, Shaillay Dogra, Johnny Wong, Julia L. MacIsaac, Sarah M. Mah, Lisa M. McEwen, Seang Mei Saw, Keith M. Godfrey, Yap Seng Chong, Kenneth Kwek, Chee Keong Kwoh, Shu E. Soh, Mary F F Chong, Sheila Barton, Neerja Karnani, Clara Y. Cheong, Jan Paul Buschdorf, Walter Stunkel, Michael S. Kobor, Michael J. Meaney, Peter D. Gluckman, and Joanna D. Holbrook. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. *Genome Research*, 24(7):1064–1074, 2014.

[50] Florian Tramèr, Zhicong Huang, Jean-Pierre Hubaux, and Erman 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. ACM, 2015.

[51] Saskia Trump, Matthias Bieg, Zuguang Gu, Loreen Thürmann, Tobias Bauer, Mario Bauer, Naveed Ishaque, Stefan Röder, Lei Gu, Gunda Herberth, Christian Lawerenz, Michael Borte, Matthias Schlesner, Christoph Plass, Nicolle Diessl, Markus Eszlinger, Oliver Mücke, Horst-Dietrich Elvers, Dirk K. Wissenbach, Martin von Bergen, Carl Herrmann, Dieter Weichenhan, Rosalind J. Wright, Irina Lehmann, and Roland Eils. Prenatal maternal stress and wheeze in children: novel insights into epigenetic regulation. *Scientific Reports*, 6:28616, jun 2016.

[52] Loukia G. Tsaprouni, Tsun-Po Yang, Jordana Bell, Katherine J. Dick, Stavroula Kanoni, James Nisbet, Ana Viñuela, Elin Grundberg, Christopher P. Nelson, Eshwar Meduri, Alfonso Buil, Francois Cambien, Christian Hengstenberg, Jeanette Erdmann, Heribert Schunkert, Alison H. Goodall, Willem H. Ouwehand, Emmanouil Dermitzakis, Tim D. Spector, Nilesh J. Samani, and Panos Deloukas. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics*, (December):00–00, oct 2014.

[53] Caroline Uhler, Aleksandra Slavković, and Stephen E Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of privacy and confidentiality*, 5(1):137, 2013.

[54] Jenny van Dongen, Michel G. Nivard, Gonneke Willemsen, Jouke-Jan Hottenga, Quinta Helmer, Conor V. Dolan, Erik A. Ehli, Gareth E. Davies, Maarten van Iterson, Charles E. Breeze, Stephan Beck, Peter A.C.'t Hoen, René Pool, Marleen M.J. van Greevenbroek, Coen D.A. Stehouwer, Carla J.H. van der Kallen, Casper G. Schalkwijk, Cisca Wijmenga, Sasha Zhernakova, Ettje F. Tigchelaar, Marian Beekman, Joris Deelen, Diana van Heemst, Jan H. Veldink, Leonard H. van den Berg, Cornelia M. van Duijn, Bert A. Hofman, André G. Uitterlinden, P. Mila Jhamai, Michael Verbiest, Marijn Verkerk, Ruud van der Breggen, Jeroen van Rooij, Nico Lakenberg, Hailiang Mei, Jan Bot, Dasha V. Zhernakova, Peter van't Hof, Patrick Deelen, Irene Nooren, Matthijs Moed, Martijn Vermaat, René Luijk, Marc Jan Bonder, Freerk van Dijk, Michiel van Galen, Wibowo Arindrarto, Szymon M. Kielbasa, Morris A. Swertz, Erik W. van Zwet, Aaron Isaacs, Lude Franke, H. Eka Suchiman, Rick Jansen, Joyce B. van Meurs, Bastiaan T. Heijmans, P. Eline Slagboom, and Dorret I. Boomsma. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, 7:11115, 2016.

[55] NT Ventham, NA Kennedy, AT Adams, Rahul Kalla, Simon Heath, KR O'leary, H Drummond, DC Wilson, Ivo Glynne Gut, ER Nimmo, et al. Integrative epigenome-wide analysis demonstrates that dna methylation may mediate genetic risk in inflammatory bowel disease. *Nature communications*, 7:13507, 2016.

[56] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 2013.

[57] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong 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.

[58] Griffin M Weber, Shawn N Murphy, Andrew J McMurry, Douglas MacFadden, Daniel J Nigrin, Susanne Churchill, and Isaac S Kohane. The shared health research information network (shrine): a prototype federated query tool for clinical data repositories. *Journal of the American Medical Informatics Association*, 16(5):624–630, 2009.

[59] Fei Yu, Stephen E Fienberg, Aleksandra B Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of biomedical informatics*, 50:133–141, 2014.

[60] Xiaoyong Zhou, Bo Peng, Yong Fuga Li, Yangyi Chen, Haixu Tang, and XiaoFeng 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.