MBeacon: Privacy-Preserving Beacons for DNA Methylation Data

Inken Hagestedt*, Yang Zhang*§, Mathias Humbert†, Pascal Berrang*,
Haixu Tang‡, XiaoFeng Wang‡, Michael Backes*
*CISPA Helmholtz Center for Information Security,
{inken.hagestedt, yang.zhang, pascal.berrang, backes}@cispa.saarland
†Swiss Data Science Center, ETH Zurich and EPFL, mathias.humbert@epfl.ch
‡Indiana University Bloomington, {hatang, xw7}@indiana.edu

***Abstract*—The advancement of molecular profiling techniques fuels biomedical research with a deluge of data. To facilitate data sharing, the Global Alliance for Genomics and Health established the Beacon system, a search engine designed to help researchers find datasets of interest. While the current Beacon system only supports genomic data, other types of biomedical data, such as DNA methylation, are also essential for advancing our understanding in the field. In this paper, we propose the first Beacon system for DNA methylation data sharing: MBeacon. As the current genomic Beacon is vulnerable to privacy attacks, such as membership inference, and DNA methylation data is highly sensitive, we take a privacy-by-design approach to construct MBeacon.**

First, we demonstrate the privacy threat, by proposing a membership inference attack tailored specifically to unprotected methylation Beacons. Our experimental results show that 100 queries are sufficient to achieve a successful attack with AUC (area under the ROC curve) above 0.9. To remedy this situation, we propose a novel differential privacy mechanism, namely SVT^2, which is the core component of MBeacon. Extensive experiments over multiple datasets show that SVT^2 can successfully mitigate membership privacy risks without significantly harming utility. We further implement a fully functional prototype of MBeacon which we make available to the research community.[1]

I. Introduction

The advancement of molecular profiling technologies during the last decade has resulted in a deluge of biomedical data becoming available. The large quantity of data is considered the fuel for the next-generation bio-engineering industry. Leading researchers as well as practitioners have predicted the biotech era is coming.

Data sharing is essential for advancing biomedical research. However, large-scale data sharing has been limited, primarily due to privacy concerns [12], [22], [2], [24]. Homer et al. [16] have shown that an adversary can effectively predict the presence of an individual in a genomic dataset. This attack is known as membership inference attack [5], [35], [28], [31] and its implication is beyond membership status: For instance, if the dataset is collected from individuals carrying a certain disease, then the adversary can immediately infer this sensitive information about her target(s). A recent study [5] further shows that not only genomic data, but also other types of biomedical data, are vulnerable to membership inference attacks.

Aiming for a responsible and effective genomic data sharing solution, the Global Alliance for Genomics and Health (GA4GH)[2] established the Beacon system[3] in 2014. The Beacon system is essentially a search engine indexed over multiple *Beacons*. Each single Beacon is constructed by a partner institution of the Beacon system with its own database. Only one type of query is supported by a Beacon: whether its database contains any record with the specified nucleotide at a given position and chromosome, and the corresponding response is a binary "Yes" or "No". Upon a query from a researcher, the search engine, i.e., the Beacon system, will return the names of the partner institutions that answer "Yes", and the researcher can directly contact these institutions to obtain access to the data.

The current Beacon system only supports genomic data. However, other types of biomedical data, like epigenetic data, are also essential for biomedical research. In particular, DNA methylation, as one of the most important epigenetic elements, has been demonstrated to be very influential to human health. For instance, anomalous changes in the DNA methylation patterns are frequently observed in cancer [13]. Consequently, there exists a huge demand for methylation data sharing.

A. Contributions

We construct the first Beacon system for sharing DNA methylation data, namely, the MBeacon system. Similar to the current genomic Beacon system, the MBeacon system is also a search engine. Each institution taking part in the MBeacon system establishes its own *MBeacon* that implements the following query: "Are there any patients with a certain methylation value at a specific methylation position?", and provides a binary "Yes" or "No" response.

Despite the coarse-grained answer format, researchers have shown that the genomic Beacon is vulnerable to privacy

§Corresponding author

[1] In this version, we corrected a technical inconsistency in the SVT^2 algorithm.

[2] https://www.ga4gh.org/

[3] https://beacon-network.org/

Network and Distributed Systems Security (NDSS) Symposium 2019
24-27 February 2019, San Diego, CA, USA
ISBN 1-891562-55-X
https://dx.doi.org/10.14722/ndss.2019.23064
www.ndss-symposium.org

attacks, in particular membership inference attacks [36], [29], [1], [45]. In addition, previous works have demonstrated the serious privacy risks stemming from sharing DNA methylation data [3], [8]. Therefore, we follow a privacy-by-design approach to construct the MBeacon system.

Membership Inference Attack. The first step towards a privacy-preserving MBeacon is to evaluate the privacy threat of membership inference attacks against a plain (unprotected) methylation Beacon. Since existing attacks on the current Beacons are tailored to genomic data only, we design a membership inference attack suitable for DNA methylation data. Our membership inference attack relies on the likelihood-ratio test and uses as probability estimate a normal distribution calibrated to the mean and standard deviation of the general population's methylation values.

We empirically evaluate our attack on several unprotected methylation Beacons composed of various methylation datasets and show that the attack achieves a superior performance. For instance, the simulated attacker can achieve an AUC value (area under the ROC curve) of over 0.9 after submitting only 100 queries to the Beacon.

Defense Mechanism. The effectiveness of our membership inference attack demonstrates the privacy threat of the Beacon system for methylation data. To mitigate this threat, we propose a novel differential privacy mechanism, namely the double sparse vector technique ($\text{SVT}^{\mathbf{2}}$), which is the core component of MBeacon. We consider a MBeacon's query response to be highly privacy-sensitive if it differs from the expected response over the general population data. In fact, these differences are also the major reason why our membership inference attack is effective. A MBeacon is usually constructed over a database collected from people with a certain disease, and biomedical studies show that, for data of this kind, only a few methylation regions differ from the general population. As a consequence, only a few queries are highly privacy-sensitive. Therefore, we aim for a solution that scales noise to the sensitive responses in order to reduce the overall noise level of MBeacon, thus maintaining utility.

One possible solution for the problem is the sparse vector technique, a differential privacy mechanism that is designed to scale noise to a subset of highly privacy-sensitive responses. The sparse vector technique determines whether a response is sensitive by comparing it to a fixed threshold. However, it cannot be applied to MBeacon, as we need to check whether the MBeacon response and the expected response agree with each other. The novelty of our proposed $\text{SVT}^{\mathbf{2}}$ lies in checking this agreement through two comparisons to a fixed threshold: one for the MBeacon response, the other for the expected response. We prove that $\text{SVT}^{\mathbf{2}}$ guarantees differential privacy.

Utility Metrics. The goal of the MBeacon system is to facilitate DNA methylation data sharing. Therefore, the main users of the system are researchers who want to discover institutions that possess data of interest. In order to quantify the impact of $\text{SVT}^{\mathbf{2}}$ on the real-world utility of our MBeacon system, we introduce a new utility metric by simulating a legitimate researcher who tries to find other institutions that possess methylation data similar to her own data.

We evaluate the performance of our privacy-preserving MBeacon through extensive experiments (simulating 2,100 researchers). The results show that the privacy loss on membership inference attacks can be minimized while the researcher utility still remains high. For carefully chosen privacy parameters, it is possible to decrease the attacker's performance to random guessing (AUC < 0.6) while preserving a high utility for the researcher (AUC > 0.8). Furthermore, we conduct a large-scale evaluation of privacy parameters for $\text{SVT}^{\mathbf{2}}$ and provide the necessary tools for an institution to tune these parameters to their needs.

In addition, we have implemented a fully-functional prototype of the MBeacon system[4] which we make available to the research community.

B. Organization

The rest of the paper is organized as follows. We briefly introduce the current Beacon system and necessary biomedical background in Section II. MBeacon is formally defined in Section III. Section IV and V present our membership inference attack and its evaluation, respectively. In Section VI, we describe our defense mechanism $\text{SVT}^{\mathbf{2}}$. Section VII introduces the utility metric. The effectiveness of our defense is evaluated in Section VIII. The MBeacon prototype is introduced in Section IX. We summarize the related work in Section X, and then conclude in Section XI.

II. Background

In this section, we provide the necessary background on the current Beacon system as well as on DNA methylation.

A. Beacon System

Current biomedical data sharing has limited success due to its inherent privacy risks. To tackle this problem, GA4GH has established the Beacon system, also referred to as the Beacon network.

The Beacon system is a search engine that allows researchers to query whether any of the institutions taking part in the system possesses data of their interests. Each partner institution implements its own Beacon with its onsite data. These Beacons only support one simple type of query, i.e., the presence of a specified nucleotide (A, C, G, T) at a given position within a certain chromosome. The response is a binary "Yes" or "No". To give a concrete example, query "13 : 32936732 G $>$ C" stands for "Are there any patients that have allele C at position 32936732 (with reference allele G) on chromosome 13?". When the Beacon system receives such a query, it forwards the query to each of its partner institutions' Beacons. If an institution's dataset contains at least one record matching the query, then the Beacon answers "Yes". The names of all Beacons with "Yes" answers are sent back to the querier. In the end, the querier can contact the corresponding institutions for data access offline.

B. DNA Methylation

DNA methylation is one of the most important and best understood epigenetic elements. It consists of molecules, so-called methyl groups, added to the nucleotides at positions

[4] https://mbeacon-network.github.io/MBeacon-network/

TABLE I. NOTATIONS.

Notation	Description
v	A victim
$m(v)$	Methylation profile of v
$\mathbb{I}$	An institution's database
$B_{\mathbb{I}}$	A MBeacon built on $\mathbb{I}$
q	A query to a MBeacon
$\overrightarrow{Q}$	A vector of queries
K	An adversary's background knowledge
b	No. of bins for methylation values
$\mathbb{A}$	Membership inference attack
δ	Measurement error
SVT^2	The defense mechanism for MBeacon
α_i	No. of patients for q_i in MBeacon
β_i	Estimated No. of patients for q_i
P	Methylation of interest for researcher
D	Methylation of no-interest for researcher
$B_{P,D}$	MBeacon built with P and D
B_D	MBeacon built with D
T	MBeacon responses "Yes" if there are $p \geq T$ patients with the requested value

where a *C* nucleotide is followed by a *G* nucleotide (called CpG-dinucleotides). Usually, DNA methylation at a given CpG-dinucleotide is measured as a real value between 0 and 1. This value represents the fraction of methylated dinucleotides at this position. The whole DNA methylation profile of an individual can thus be represented as a vector of real values between 0 and 1. Intermediate values occur due to DNA methylation varying between copies of the DNA within the same cell, or due to mixtures of cells from different tissues being measured.

Whether the DNA is methylated at certain positions affects the DNA activity and structure [17], [33]. Some anomalous changes in methylation patterns are correlated with cancer [13], leading to activation of genes such as oncogenes, or the silencing of tumor suppressor genes. Meanwhile, environmental factors, such as pollution, smoking and stress, can cause the changes of methylation values [7], [40], [42], [41]. Therefore, an increasing number of studies concentrate on methylation, which require large amounts of DNA methylation data, and thus data sharing.

In this paper, we propose the first Beacon system for sharing DNA methylation data, namely the MBeacon system. Since an individual's methylation data may carry information about her current disease status and environmental factors influencing her health, methylation data is considered highly privacy-sensitive. Also, a recent study has shown that methylation data can be re-identified by inferring the corresponding genomes [3] given an individual's methylation profile. Therefore, our MBeacon system is built following a privacy-by-design approach.

III. MBEACON DESIGN

The MBeacon system is a search engine that indexes over multiple MBeacons. Each MBeacon is established by an institution with its own database, and this institution is referred to as a partner of the MBeacon system. We denote an institution by $\mathbb{I}$ and its MBeacon by $B_{\mathbb{I}}$. Without ambiguity, we also use $\mathbb{I}$ to represent the institution's database itself, which consists of multiple patients' methylation profiles. Moreover, we denote a patient by v, and her methylation profile, i.e., the sequenced methylation values, by a vector $m(v) \in \mathbb{R}^M_{[0,1]}$. The vector length M is equal to the total number of methylation positions considered, e.g., $M = 450,000$.

Similar to the genomic Beacon, our MBeacon supports one type of query, that is "*Are there any patients with this methylation value at a specific methylation position?*". Formally, we define a query q as a tuple (pos, val) where *pos* represents the queried position and *val* represents the queried value. A Beacon $B_{\mathbb{I}}$ is essentially a function,

$$B_{\mathbb{I}} : q \to \{0, 1\}, \tag{1}$$

where 0 represents "No" and 1 represents "Yes". It is worth noting that this general query format also allows researchers to infer answers to more complex queries, such as "*Are there any patients with methylation value above some threshold for a specific position?*". When a researcher issues a query to the MBeacon system, the system forwards this query to all the MBeacons, and returns the names of those MBeacons with "Yes" answers to the researcher.

For presentation purposes, we summarize the notations introduced here and in the following sections in Table I.

IV. MEMBERSHIP INFERENCE ATTACK

To demonstrate the privacy risks of unprotected methylation Beacons, we propose a membership inference attack against them. In this section, we first present the considered adversarial model, then the methodology of our attack.

A. Threat Model

In general, the goal of membership inference attacks is to predict whether the victim is a member of the database given certain knowledge about the victim. For instance, an attacker with access to the sequenced methylation values of her victim aims to infer whether the victim is in the database containing methylation data collected from some HIV carriers. By knowing who is member of the study, the attacker is able to infer the HIV status of her victim, even though (to the best of our knowledge) the HIV status is not directly detectable from the methylation values. This example demonstrates the severe consequence of membership inference. Moreover, all the existing attacks against genomic Beacons are membership inference attacks [36], [29], [1], [45].

We assume that the adversary has access to the victim's methylation data $m(v)$ and additional background knowledge K that we instantiate later. The adversary's goal is to perform an attack $\mathbb{A}$, to decide whether v is in the database of institution $\mathbb{I}$ by querying the MBeacon $B_{\mathbb{I}}$. Formally, the membership inference attack is defined as follows:

$$\mathbb{A} : (m(v), B_{\mathbb{I}}, K) \to \{0, 1\}, \tag{2}$$

where 1 means that the victim is in the MBeacon database and 0 that she is not. If v's methylation values are indeed part of the MBeacon's database ($m(v) \in \mathbb{I}$) and the attack output is 1, then the attack achieves a true positive for v. If the output is 0, then it is a false negative. However, if v's methylation values are not part of $B_{\mathbb{I}}$ (i.e., $m(v) \notin \mathbb{I}$) and the attack output is 0, this is a true negative, otherwise, if the output is 1, it is a false positive.

B. Attacking Methylation Beacons

We rely on the likelihood-ratio (LR) test to realize our membership inference attack for two main reasons. First, by the Neyman-Pearson Lemma [20], [37], the LR test achieves the highest power (true-positive rate) for a given false-positive rate in binary hypothesis testing if the theoretical preconditions are met. Second, the LR test has been successfully used by Shringarpure and Bustamante [36] and Raisaro et al. [29] for attacking genomic Beacons.

In general, the LR test formulates a null hypothesis H_0 and an alternative hypothesis H_1, and compares the quotient of the two hypotheses' likelihoods to a threshold. Our null hypothesis H_0 is defined as the queried victim v not being in the MBeacon ($m(v) \notin \mathbb{I}$), and the alternative hypothesis H_1 as the queried victim being in the MBeacon ($m(v) \in \mathbb{I}$).

The adversary submits a series $\overrightarrow{Q} = \langle q_1, \ldots, q_n \rangle$ ($n \leq M$) of queries to $B_{\mathbb{I}}$ with her victim's methylation values, i.e., $m(v)$, and get a list of responses, denoted by $B_{\mathbb{I}}(\overrightarrow{Q}) = \langle B_{\mathbb{I}}(q_1), \ldots, B_{\mathbb{I}}(q_n) \rangle$. Assuming that the different responses are independent,[5] the log-likelihood of the responses is

$$\begin{aligned} L(B_{\mathbb{I}}(\overrightarrow{Q})) = \sum_{i=1}^{n} & B_{\mathbb{I}}(q_i) \log(\Pr(B_{\mathbb{I}}(q_i) = 1)) + \\ & (1 - B_{\mathbb{I}}(q_i)) \log(\Pr(B_{\mathbb{I}}(q_i) = 0)). \end{aligned} \tag{3}$$

To implement the two hypotheses H_0 and H_1, we need to model $\Pr(B_{\mathbb{I}}(q) = 1)$ and $\Pr(B_{\mathbb{I}}(q) = 0)$. The approach in [36] cannot be directly applied as it is designed for genomic data, which is discrete. In contrast to that, methylation data is represented as a continuous value between 0 and 1. We propose to bin the methylation values into b equal-width bins that represent the range of values the querier might be interested in.[6] Here, b is a parameter of the MBeacon system, and we empirically study the influence of different values for b on the attack performance in Section V.

Thus, we represent a methylation Beacon as $B_{\mathbb{I}}^b$. The probability $\Pr(B_{\mathbb{I}}^b(q) = 0)$ to get a "No" answer, respectively $\Pr(B_{\mathbb{I}}^b(q) = 1)$ to get a "Yes" answer can be described in our case as:

$$\Pr(B_{\mathbb{I}}^b(q) = 0) = (1 - \tau^b(q))^N \tag{4}$$

$$\Pr(B_{\mathbb{I}}^b(q) = 1) = 1 - (1 - \tau^b(q))^N \tag{5}$$

Here, N is the number of patients in the Beacon. Following previous works on genomic Beacons [36], [29], we assume N to be publicly known and therefore being part of the attacker's background knowledge K. Meanwhile, τ^b is the probability of a patient having a methylation value in the interval determined by the respective bin. We can assume that the adversary has the exact probability as part of her background knowledge K. However, if the exact probability is not available and the adversary only knows the mean and standard deviation of people's methylation values at a certain position, she can approximate the probability with normal (Gaussian) distribution using μ_{pos} as the mean and σ_{pos} as the standard deviation of the queried position.[7] Concretely, $\tau^b(q)$ is estimated as:

$$\begin{aligned} & \widetilde{\tau^b}(q) = \widetilde{\tau^b}((pos, val)) = \\ & cdf(\mu_{pos}, \sigma_{pos}, b_r) - cdf(\mu_{pos}, \sigma_{pos}, b_l) \end{aligned} \tag{6}$$

where *cdf* is the cumulative distribution function of the normal distribution, and b_r (b_l) denotes the value of the corresponding bin's right (left) edge. Notice that, like in the genomic setting, the general probability of having a specific allele is required as well, and it is realized by assuming the population's allele frequencies are part of the attacker's background knowledge K.

By inserting the probabilities from Equations 4 and 5 into Equation 3, we get

$$\begin{aligned} L_{H_0}(B_{\mathbb{I}}^b(\overrightarrow{Q})) = \sum_{i=1}^{n} & B_{\mathbb{I}}^b(q_i) \log(1 - (1 - \tau^b(q_i))^N) + \\ & (1 - B_{\mathbb{I}}^b(q_i)) \log((1 - \tau^b(q_i))^N) \end{aligned} \tag{7}$$

$$\begin{aligned} L_{H_1}(B_{\mathbb{I}}^b(\overrightarrow{Q})) = \sum_{i=1}^{n} & B_{\mathbb{I}}^b(q_i) \log(1 - \delta(1 - \tau^b(q_i))^{N-1}) + \\ & (1 - B_{\mathbb{I}}^b(q_i)) \log(\delta(1 - \tau^b(q_i))^{N-1}). \end{aligned} \tag{8}$$

Notice that for the H_0 hypothesis, we consider all N patients in the database. However, for the H_1 hypothesis where we assume the target being part of the database, we consider only $N-1$ other patients that contribute to the answer in addition to the target. It might occur that two measurements of methylation data from the same patient and tissue type differ, either due to measurement errors or changes over time. Thus, the target may be part of the Beacon, but the attacker's data differs from the data entry in the Beacon. Similar to previous works, we denote this probability, i.e., measurement error, by δ and empirically evaluate its influence on our attack. We assume δ to be part of the attacker's background knowledge.

In the end, the log of the likelihood-ratio is given by:

$$\begin{aligned} \Lambda = & L_{H_0}(B_{\mathbb{I}}^b(\overrightarrow{Q})) - L_{H_1}(B_{\mathbb{I}}^b(\overrightarrow{Q})) \\ = & \sum_{i=1}^{n} (1 - B_{\mathbb{I}}^b(q_i)) \log\left(\frac{(1 - \tau^b(q_i))^N}{\delta(1 - \tau^b(q_i))^{N-1}}\right) + \\ & B_{\mathbb{I}}^b(q_i) \log\left(\frac{1 - (1 - \tau^b(q_i))^N}{1 - \delta(1 - \tau^b(q_i))^{N-1}}\right). \end{aligned} \tag{9}$$

If Λ is lower than some threshold t, we reject the null hypothesis and predict that the victim is in the MBeacon database. Otherwise, we conclude that the victim is not.

Finally, the choice of the set of queries $\langle q_1, \ldots, q_n \rangle$ influences the attack performance as well. We follow the same

[5] We assume the adversary does not submit a single query for multiple times, and we assume correlations between different methylation positions are not exploited, because they are not (yet) well studied. Note that the same independence assumption has been used in previous works on genomic Beacons [36], [29].

[6] There are two reasons why we only study equal-width bins: First, without further knowledge about the data distribution underlying the Beacon, it is hard to define a suitable bin width. Second, all Beacons should share the same interface to combine the answers in a well-defined way. This would not be possible if the bins vary across different Beacons based on the dataset they are composed of.

[7] We experimentally found that the normal distribution fits methylation data best, using the Kolmogorov-Smirnov test and a p-value of 0.1. Other ways to approximate the probability are left for future work.

TABLE II. DATASETS USED FOR OUR EXPERIMENTS.

Abbreviation	Description	number of patients	GSE identifier	by
Ependymoma	Ependymoma	48	GSE45353	[30]
GBM	glioblastoma	136	GSE36278	[38]
PA	pilocytic astrocytoma	61	GSE44684	[19]
ETMR-PNET	embryonal brain tumor and primitive neuroectodermal tumor	38	GSE52556	[18]
mHGA	4 different subtypes of pediatric glioblastomas	96	GSE55712	[14]
DIPG	diffuse intrinsic pontine glioma	28	GSE50022	[9]
IBD CD	Crohn's disease	77	GSE87640	[43]
IBD UC	ulcerative colitits	79	GSE87640	[43]

approach as Raisaro et al. [29] to rank all possible queries with their expected information gain: For each methylation position *pos*, the attacker computes the difference between the victim's methylation value $m(v)_{pos}$ and the general population's value $\widetilde{\tau^b}(pos, m(v)_{pos})$. The larger this difference, the higher the probability of getting a "Yes" answer if the target is part of the Beacon, and simultaneously, the higher the probability of getting a "No" answer if the target is not part of the Beacon. Therefore, we assume the attacker decides on the set of queries $\langle q_1, \ldots, q_n \rangle$ using this difference and querying the n most informative queries.

V. Attack Evaluation

In this section, we evaluate the performance of our membership inference attack against simulated methylation Beacons to demonstrate the privacy threat.

A. Datasets

For our experiments, we rely on eight diverse datasets containing methylation profiles of patients carrying specific diseases. In total, we use methylation profiles of 563 individuals. The datasets are available online in the Gene Expression Omnibus database (GEO),[8] and we summarize them in Table II. We use six brain tumor datasets, where the methylation data was sequenced from the respective brain tumor. Moreover, we also make use of an additional dataset with two types of inflammatory bowel disease, where the methylation data was sequenced from blood samples, reported in the last two lines of Table II. All of these data were generated with the Illumina 450k array, effectively determining the DNA methylation at 450,000 fixed positions.

Preprocessing. Most of the datasets have missing methylation sites for specific patients or even for all the patients sharing the same disease. We remove all methylation positions with missing data, which leaves us with 299,998 different methylation sites for the combination of all our eight datasets.

Human Subjects and Ethical Considerations. All datasets are publicly available in their anonymized form. Moreover, they have been stored and analyzed in anonymized form without having access to non-anonymized data. The membership inference we carry out does not reveal any more information than previously known by us.

B. Evaluation Results

We use our three largest[9] datasets, i.e., GBM, and both IBD datasets (referred to as IBD CD and IBD UC), to simulate three methylation Beacons, respectively. For each methylation Beacon, we randomly sample 60 patients to construct its Beacon database. We follow the approach of previous works on Beacons testing with uniform sets of patients [36], [29], [1], [45]. This ensures the attacker can only exploit individual variances and not disease-induced systematic differences, i.e., variances that are unavoidably in the data. Later in Section VIII, we explore another attack scenario on heterogeneous methylation sets.

We assume the adversary has access either to a randomly chosen sample from the methylation Beacon ("in" patient), or from the patients with the same disease who are not included in the methylation Beacon ("out" patient). For the "out" patients, we use the remainder of the patients that we do not sample into the methylation Beacon. For the "in" patients, we sample the same number of patients from the methylation Beacon to not introduce a bias between "in" and "out" test patients. To reduce the size bias between GBM and the two IBD sets, we sample at most 25 test patients. We repeat the random split of patients into methylation Beacon and testing set 10 times, which corresponds to a simulation of 500 attackers for GBM, 340 for IBD CD and 300 for IBD UC.

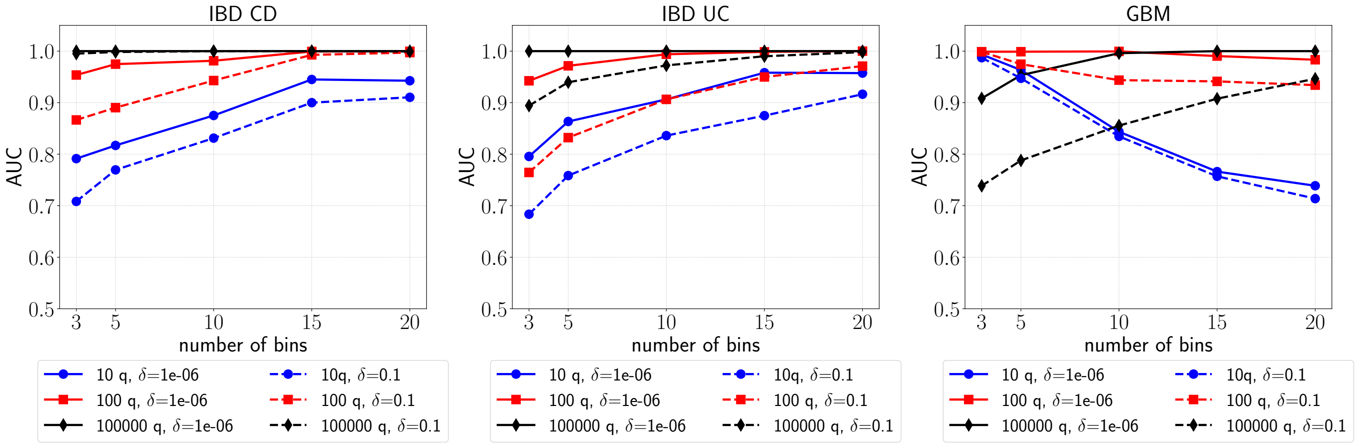
The attackers carry out the LR test as described previously in Section IV. We simulate attackers without access to the exact probability $\tau^b(q)$, because it is an unrealistic assumption that these are available. In fact, if such knowledge would be available, a lot of privacy would already be lost. Instead, we model attackers estimating the probabilities from a general background population. We combine the main datasets GBM, IBD UC, IBD CD with the other datasets (Ependymoma, mHGA, ETMR-PNET, PA and DIPG) as an estimate for the general population.[10] From this combined background data, we compute the attacker's background knowledge K as mean and standard deviation for each methylation position. Apart from being used in the LR test to estimate frequencies, the means are used to rank possible queries by their expected information gain, as discussed in Section IV.

We adopt the AUC, i.e., area under the ROC curve, as our evaluation metric since it does not involve picking a specific threshold for the LR test. The ROC curve is a 2D plot which reflects the relation between true positive rate and false positive rate over a series of thresholds for the LR test. The AUC summarizes the ROC curve as a single value. A ROC curve closer to top-left border of the plot, thus a larger AUC value, indicates a better prediction performance. Moreover,

[8] https://www.ncbi.nlm.nih.gov/geo/

[9] We exclude the mHGA dataset, since it is not uniform but a combination of 4 subtypes.

[10] Since general population statistics do not exist yet for methylation values, we had to estimate them. If the estimate was not accurate and a realistic attacker could get better estimates, the attack performance could increase.



(a)

(b)

(c)

Fig. 1. Influence of number of bins used and number of queries submitted on attacker's performance of the membership inference attack (a) on IBD CD, (b) IBD UC and (c) GBM.

there exists a conventional standard[11] to interpret AUC values: AUC $= 0.5$ is equivalent to random guessing, whereas an AUC greater than 0.9 shows the prediction is excellent. It is worth noting that AUC has been adopted by many recent works for assessing privacy attacks [5], [25], [15], [23], [6], [26], [28].

To get an overview of the attack and the influence of various parameters, we vary the number of bins b from 3 to 20, and let the attacker submit 10, 100, and 100,000 unique queries to the respective methylation Beacon. We vary δ between 0.1 and 10^{-6}.

Figure 1 shows the attacker's performance as a function of b. Different numbers of queries submitted are displayed in different colors, and line styles indicate two choices for δ. As expected, the number of bins influences the attacker's performance. The more bins, the fewer patients' values are expected in each of them, which makes the membership inference easier.

The attacker's performance is high as soon as the number of bins is reasonably large (larger than 3), no matter whether 100,000 or just 10 queries are submitted. This demonstrates the privacy risk of unprotected methylation Beacons. Nevertheless, the GBM curve for only 10 queries demonstrates that asking too few queries may just not be enough for a successful attack. The choice of δ has only little influence on the attack performance in case more than 100 queries are submitted.

We observe a different attack performance depending on the dataset, which is expected because we are testing different populations, diseases and tissues here. We note that both IBD datasets provide similar high AUCs, which can be explained by the fact that they are taken from the same tissue, namely blood cells.

As the increase in the attacker's performance is only slight for more than 10 bins, we fix the number of bins to 10 in the remainder of the experiments to reduce the number of parameters and simplify the presentation. Additionally, we fix δ to 10^{-6} to model the worst-case for privacy, even though the privacy risk differs not much for other choices of δ.

VI. Defense

The results in Section V demonstrate the privacy risks stemming from unprotected methylation Beacons. To mitigate this threat, we present our defense mechanism, the double sparse vector technique (SVT$^{\mathbf{2}}$). We first explain the intuition behind it and then the defense mechanism in detail. In the end, we prove that our defense mechanism is differentially private.

A. *Intuition*

Recall that we assume the background knowledge K contains the means and standard deviations of the general population at the methylation positions of interest. That means, if one judges by the background knowledge that there should (or should not) be an individual with some value in a MBeacon and the MBeacon output confirms this, then not much privacy is lost. Yet, if MBeacon's answer deviates from the background knowledge, one learns an additional piece of information about the real distribution in the MBeacon for the queried position. In consequence, the privacy of patients in the MBeacon is at risk. More formally, we consider a MBeacon response as highly privacy-sensitive if it deviates from the answer we expect from the general population.

A MBeacon is usually built with data collected from people with certain disease. According to biomedical research [39], [41], [43], for data of this kind, only a few methylation regions differ from the general population. This indicates that just a few query responses are expected to be privacy-sensitive. Therefore, we aim for a solution that calibrates the noise specifically to those few responses in order to reduce the overall noise level of MBeacon, thus maintaining utility.

B. *Background on SVT*

One possible solution in such a scenario is the sparse vector technique (SVT), a differential privacy mechanism which is designed to scale noise to a subset of sensitive responses.

[11] http://gim.unmc.edu/dxtests/roc3.htm

In SVT, whether a response is sensitive or not is determined by a threshold T defined by the data owner: A response $\alpha \geq T$ is considered as privacy-sensitive, and one assumes most responses will yield $\alpha < T$. SVT guarantees differential privacy while scaling noise only to the privacy-sensitive answers. To this end, SVT has an additional privacy parameter c which refers to as the maximal amount of answers $\alpha \geq T$ the mechanism can give over its whole lifetime. SVT adds noise to all queries (no matter whether they are privacy sensitive or not) before comparing to the threshold to ensure differential privacy. However, this noise is scaled to c instead of the much larger number of queries in total. For a detailed and formal description of SVT, we refer the reader to [11].

Algorithm 1: $\mathcal{A}$ outputs whether the database and prior agree on the number of patients in the queried position being above the threshold in a differentially private manner.

```
Input: base threshold $T$, privacy parameters $\epsilon_1, \epsilon_2$ and
       $c$, query sensitivity $\Delta$, query vector $\overrightarrow{Q}$, database
       $\mathbb{I}$ and prior frequency $\mathtt{P}$
Result: sanitized responses $R$ such that $r_i \in \{\bot, \top\}$ for
        each $i$
1  $z_1 = \text{LAP}(\frac{\Delta}{\epsilon_1})$;    $z_2 = \text{LAP}(\frac{\Delta}{\epsilon_1})$;
2  count = 0;
3  for each query $q_i$ in $\overrightarrow{Q}$ do
4      $y_i = \text{LAP}(\frac{2c\Delta}{\epsilon_2})$;    $y'_i = \text{LAP}(\frac{2c\Delta}{\epsilon_2})$;
5      get $\alpha_i$ from $\mathbb{I}$ and $\beta_i$ from $\mathtt{P}$;
6      if ($\alpha_i + y_i < T + z_1$ and $\beta_i + y_i < T + z_1$) or
         ($\alpha_i + y'_i \geq T + z_2$ and $\beta_i + y'_i \geq T + z_2$) then
7          $r_i$ = $\bot$ ;
8      else
9          $r_i$ = $\top$;
10         count = count + 1 ;
11     end
12     if count $\geq c$ then
13         Halt
14     end
15 end
```

Algorithm 2: $\mathcal{B}$ transforms the output of Algorithm 1 to the MBeacon output format.

```
Input: base threshold $T$, privacy parameters $\epsilon_1$, $\epsilon_2$ and
       $c$, query sensitivity $\Delta$, query vector $\overrightarrow{Q}$,
       database $\mathbb{I}$ and prior frequency $\mathtt{P}$
Result: sanitized MBeacon responses $B_{\mathbb{I}}(\overrightarrow{Q})$
1  $\overrightarrow{R} = \mathcal{A}(T, \epsilon_1, \epsilon_2, c, \Delta, \overrightarrow{Q}, \mathbb{I}, \mathtt{P})$ ;
2  for each query $q_i$ in $\overrightarrow{Q}$ do
3      get $r_i$ from $\overrightarrow{R}$;    get $\beta_i$ from $\mathtt{P}$;
4      if $r_i = \bot$ then
5          $B_{\mathbb{I}}(q_i) = \beta_i \geq T$;
6      else
7          $B_{\mathbb{I}}(q_i) = \neg(\beta_i \geq T)$;
8      end
9  end
```

C. SVT²

However, we cannot directly apply SVT to protect our methylation Beacon, as our privacy-sensitive responses depend on whether we expect a "No" or a "Yes" answer, thus cannot be judged by a simple, fixed threshold. Concretely, suppose that we expect β patients in the queried bin, then the true number of patients in the bin, i.e., α, is privacy-sensitive if β and α lie on opposite sides of a predefined threshold T and the Beacon gives another answer than the one we expected. This means we need two comparisons to determine whether the answer is privacy-sensitive. Therefore, we propose double sparse vector technique (SVT$^\mathbf{2}$) to protect MBeacon. Since SVT is not applicable, we cannot compare our new technique SVT$^\mathbf{2}$ to SVT.

Formally, the ith query is not privacy-sensitive if the following expectation is met:

$$\begin{aligned} &((\alpha_i + y_i < T + z_1) \wedge (\beta_i < T + z_1)) \\ \vee&((\alpha_i + y'_i \geq T + z_2) \wedge (\beta_i \geq T + z_2)) \end{aligned} \tag{10}$$

where α_i is the number of patients in the MBeacon that corresponds to the query q_i, β_i is the estimated number of patients given by the general population,[12] and T is the threshold determining whether the α_i and β_i agree with each other. This (dis-)agreement is used to check whether the current query is privacy-sensitive or not: Only Condition 10 being false implies the query is privacy-sensitive. Moreover, z_1, z_2 and y_i, y'_i are noise variables sampled independently from the Laplace distribution. The sampling procedure is explained in detail later in this section.

Similar to the sparse vector technique, SVT$^\mathbf{2}$ bounds the total number of highly privacy-sensitive queries by maintaining a counter. Each privacy-sensitive query increases the counter. If a predefined maximal budget c is exceeded, the algorithm stops answering. In practice, that would mean that the corresponding MBeacon goes offline. We study when this is the case and whether this negatively influences the MBeacon utility in Section VIII.

We disassemble our method SVT$^\mathbf{2}$ into Algorithms 1 and 2, also referred to as $\mathcal{A}$ and $\mathcal{B}$, for technical reasons of the differential privacy proof. Algorithm 1 answers whether the Beacon returns the requested answer in a differentially private way, Algorithm 2 then transforms this into the desired MBeacon answer format. Moreover, we formulate the expected answer as a query to a database to allow practitioners to instantiate it with the most suitable estimation for their purpose. In our evaluation, we use the normal distribution fitted to population-wide means and standard deviations, since the LR test also relies on their knowledge.

Algorithm 1 determines whether the prior and the MBeacon database agree on the answer. Condition 10 can be found in its generalized form in line 6 of Algorithm 1, where noise is added to the prior as well. This removes the assumption that β is publicly known from Algorithm 1. In the less privacy relevant case, answer can be directly given (line 7); in the more privacy relevant case, the privacy budget has to be decreased in addition to returning the answer. If the current privacy budget

[12] We assume the number of patients in the MBeacon database to be publicly known, so we can set $\beta_i = \tau^b(q_i)^N$.

count exceeds the maximal budget c, the algorithm has to stop answering (lines 12 and 13).

Algorithm 2 takes the output of Algorithm 1 and provides the differentially-private MBeacon answer by flipping the expected answer if necessary (line 7).

Notice that genomic Beacons usually set $T = 1$, but we generalize that setting by allowing other threshold values in a k-anonymity like fashion. For low values of T, the regions where the MBeacon answer differs from the expected answer grow, while for higher values they shrink. Furthermore, a user might not ask all queries at once, but in an adaptive manner. This is taken into consideration by SVT and consequently by SVT**²**, another important aspect in the on-line setting of MBeacon.

Repeated Queries. All differential privacy mechanisms, including our proposed mechanism, assume all queries are unique. Otherwise, the noise might eventually cancel out. A single person has no (legitimate) interest in asking the same query multiple times, but in an online Beacon setting, multiple users might ask the same question. However, the assumption is not a limitation: we maintain a database of responses and, if a question has been asked before, we answer the same way we did before. Initially, such a database can be empty and it gets filled with responses over time. Its size is in O(number of methylation regions×number of bins), but the total MBeacon database is O(number of methylation regions×number of patients) and we expect much more patients than bins, so the space overhead is acceptable.

D. Differential Privacy Proof

We first prove that Algorithm 1, i.e., $\mathcal{A}$, is differentially private. Then, we show that the transformation of its output to our desired MBeacon output using Algorithm 2, i.e., $\mathcal{B}$, is also differentially private. The combination of these arguments proves that SVT² is differentially private.

Theorem 1. *Algorithm 1 is* $2(\epsilon_1 + \epsilon_2)$*-differentially private.*

We present a proof sketch of Theorem 1 in the following, the full proof is presented in the appendix.

Proof sketch. Consider any output of $\mathcal{A}$ as a vector $\overrightarrow{R} \in \{\top, \bot\}^l$, we refer to its elements as $\overrightarrow{R} = \langle r_1, ..., r_l \rangle$. We define two sets $I_\top = \{i : r_i = \top\}$ and $I_\bot = \{i : r_i = \bot\}$ of indices for the different answers. For the analysis, let the noise values y_i, y'_i for all $i \in I_\top \cup I_\bot$ be arbitrary but fixed [11]. We concentrate on the probabilities over the randomness of z_1, z_2, i.e., the noise added to the threshold T. Moreover, let the two databases $\mathbb{I}$ and $\mathbb{I}'$ be arbitrary but fixed, such that $\mathbb{I}$ and $\mathbb{I}'$ are neighboring databases.

We begin by disassembling the probability of Algorithm 1 getting a specific answer $\overrightarrow{R}$ from $\mathbb{I}$ as follows.[13]

$$\Pr[\mathcal{A}(\mathbb{I}) = \overrightarrow{R}] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\rho_1 = z_1 \wedge \rho_2 = z_2] \\ f_{\mathbb{I}}(z_1, z_2) g_{\mathbb{I}}(z_1, z_2) dz_1 dz_2 \tag{11}$$

where

$$f_{\mathbb{I}}(z_1, z_2) = \Pr[\wedge_{i \in I_\bot} r_i = \bot | \rho_1 = z_1 \wedge \rho_2 = z_2] \tag{12}$$

$$g_{\mathbb{I}}(z_1, z_2) = \Pr[\wedge_{i \in I_\top} r_i = \top | \rho_1 = z_1 \wedge \rho_2 = z_2] \tag{13}$$

To prove the theorem, it is sufficient to show that, for sensitivity Δ, the following inequalities hold:

$$f_{\mathbb{I}}(z_1, z_2) \leq f_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) \tag{14}$$

$$g_{\mathbb{I}}(z_1, z_2) \leq e^{2\epsilon_2} g_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) \tag{15}$$

$$\Pr[\rho_1{=}z_1 \wedge \rho_2{=}z_2] {\leq} e^{2\epsilon_1} \Pr[\rho_1 = z_1 + \Delta \wedge \rho_2 = z_2 - \Delta] \tag{16}$$

which gives us the required connection between the two neighboring databases $\mathbb{I}$ and $\mathbb{I}'$.

To prove Inequality 14, we utilize only the sensitivity Δ, i.e., $|\alpha_i - \alpha'_i| \leq \Delta$ and $|\beta_i - \beta'_i| \leq \Delta$. For Inequality 15, as g argues about the negation of the query formulation, if we simply follow the proof for Inequality 14, we would get $g_{\mathbb{I}'}(z_1 - \Delta, z_2 + \Delta)$. Therefore, we rely on the fact that noise values y_i are Laplace distributed (formally, $\textsc{Lap}(\frac{2c\Delta}{\epsilon_2})$) and use Inequalities 17 and 18 to prove it.

$$\Pr[\rho = y_i] \leq e^{\frac{\epsilon_2}{c}} \Pr[\rho = v_i + 2\Delta] \tag{17}$$

$$\Pr[\rho = y_i] \leq e^{\frac{\epsilon_2}{c}} \Pr[\rho = v_i - 2\Delta] \tag{18}$$

To prove Inequality 16, we use the fact that z_1 and z_2 are sampled from $\textsc{Lap}(\frac{\Delta}{\epsilon_1})$.

In the end, by combining Inequalities 14, 15 and 16, we prove Theorem 1 as follows:

$$\begin{aligned} &\Pr[\mathcal{A}(\mathbb{I}) = \overrightarrow{R}] \\ =&\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\rho_1 = z_1 \wedge \rho_2 = z_2] \\ &f_{\mathbb{I}}(z_1, z_2) g_{\mathbb{I}}(z_1, z_2) dz_1 dz_2 \\ \leq& \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{2\epsilon_1} \Pr[\rho_1 = z_1 + \Delta \wedge \rho_2 = z_2 - \Delta] \\ &f_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) e^{2\epsilon_2} g_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) dz_1 dz_2 \\ =&e^{2\epsilon_1 + 2\epsilon_2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\rho_1 = z'_1 \wedge \rho_2 = z'_2] \\ &f_{\mathbb{I}'}(z'_1, z'_2) g_{\mathbb{I}'}(z'_1, z'_2) dz'_1 dz'_2 \\ =&e^{2(\epsilon_1 + \epsilon_2)} \Pr[\mathcal{A}(\mathbb{I}') = \overrightarrow{R}] \end{aligned}$$

■

The purpose of Algorithm 1 is to answer whether the database is approximated well by the background knowledge in a differentially private way. To output a Beacon answer of the format "Yes, such data is available" resp. "No, such data is not available", we need to remove the background knowledge from Algorithm 1's answer. This is performed by Algorithm 2, which preserves the differential privacy of the answer. Intuitively, the transformation maintains differential privacy due to the composition and post-processing theorems. However, these theorems are not directly applicable due to our database format. Therefore, we prove the following theorem.

Theorem 2. *Algorithm 2 is* $2(\epsilon_1 + \epsilon_2)$*-differentially private.*

[13] As the other inputs are fixed, we use $\mathcal{A}(\mathbb{I})$ to represent Algorithm 1 in the proof, omitting the other input parameters for better readability.

Proof. Once the prior frequency P is fixed, the output of Algorithm 2 only depends on the output of Algorithm 1, namely, whether the prior is correct or has to be flipped. Formally, we describe this as follows.

First, fixing any output $\overrightarrow{\mathcal{R}} \in \{$ "Yes", "No" $\}^l$ of Algorithm 2 on $Q = \langle q_1, ..., q_l \rangle$, we have:

$$\frac{\Pr[\mathcal{B}(T, \epsilon_1, \epsilon_2, c, Q, \mathbb{I}, \mathtt{P}) = \overrightarrow{\mathcal{R}}]}{\Pr[\mathcal{B}(T, \epsilon_1, \epsilon_2, c, Q, \mathbb{I}', \mathtt{P}) = \overrightarrow{\mathcal{R}}]} = *$$

As Algorithm 2 is deterministic, we have:

$$* = \frac{\Pr[\mathcal{A}(T, \epsilon_1, \epsilon_2, c, Q, \mathbb{I}, \mathtt{P}) = \overrightarrow{R}]}{\Pr[\mathcal{A}(T, \epsilon_1, \epsilon_2, c, Q, \mathbb{I}', \mathtt{P}) = \overrightarrow{R}]} = *$$

Algorithm 1 is $2(\epsilon_1 + \epsilon_2)$-differentially private, thus:

$$* \leq e^{2(\epsilon_1+\epsilon_2)}$$

■

Notice that, for technical reasons, we disassemble our proposed method into two stages. However, one can of course perform both stages at once and directly output the MBeacon response. Since we assume the prior frequency is publicly known, we do not have to add noise to its result, which yields Condition 10 above.

Setting the Parameters. We have shown that $\mathcal{A}$ is $2(\epsilon_1 + \epsilon_2)$-differentially private to make the connection between privacy-sensitive and less privacy-sensitive queries as well as the connection to the sparse vector technique visible. However, for tuning parameters, it is desirable to have only a single privacy parameter ϵ in addition to the budget c. Lyu et al. [21] showed that the ratio $\epsilon_1 : \epsilon_2 = 1 : (2c)^{\frac{2}{3}}$ maximizes utility, while preserving $\epsilon = \epsilon_1 + \epsilon_2$. We adopt Lyu's ratio between ϵ_1 and ϵ_2. The sensitivity Δ is 1 in our case, since removing a participant's entry from the database or changing it can affect the bin count by at most one. For a given privacy parameter and using $\Delta = 1$, we set:

$$\epsilon_1 = \frac{\frac{\epsilon}{2}}{(2c)^{\frac{2}{3}} + 1} \qquad \epsilon_2 = (2c)^{\frac{2}{3}} \epsilon_1$$

Application to other Domains. We emphasize that SVT$^\mathbf{2}$ is a general differential privacy mechanism, and can be applied in other cases beyond MBeacon: SVT$^\mathbf{2}$ is useful for comparing a database to a general belief in a differentially-private way. Moreover, comparing two databases is possible using Algorithm 1 since it applies noise to both databases α and β. In the future, we plan to apply SVT$^\mathbf{2}$ to other data domains, such as location data [27], [46], social network data [23], [47], and other types of biomedical data [4].

VII. Researcher Utility

The goal of the MBeacon system is to facilitate biomedical data sharing among the research community. Therefore, we quantify the utility of MBeacon as the ability of a legitimate researcher to find methylation data of interest.

Concretely, a researcher is interested in methylation profiles of people with a certain phenotype or disease. We use the set P to represent all these methylation profiles. Moreover, the researcher already has multiple profiles in P on her site, denoted by P' with $P' \subset P$. Then, her goal is to find those MBeacons with methylation profiles from $P \backslash P'$. A central assumption here is that methylation profiles in P are similar to each other.

As the MBeacon system only supports queries on single methylation positions, the researcher also relies on the LR test to find MBeacons that contain patients in P. Moreover, there often exist measurement errors when collecting methylation values. To increase the reliability of her LR test, the researcher further averages all the methylation profiles in P'.

Ideally, the researcher queries a MBeacon B_P only containing patients of interest. To simulate a more realistic case, we assume the existence of another population D the researcher is not interested in. Notice that D might also be a mixture of populations. The researcher tries to distinguish a MBeacon B_D containing no patients of interest from a MBeacon $B_{P,D}$ containing some patients of interest. In the worst case, there are only a few patients from P in $B_{P,D}$. In that case, the contribution of patients from P is small and may be hidden due to the SVT$^\mathbf{2}$ protection.

To get the lower bound of the MBeacon utility, we concentrate on this worst case scenario. Figure 2a depicts a graphical summary of the researcher setup. The researcher achieves a true positive if the MBeacon she selects contains some profiles in P. A false positive indicates that the MBeacon she finds does not have the data of her interest. True negative and false negative are defined accordingly. Given these numbers, in particular the true-positive and false-positive rates, we can derive the AUC as our core utility metric.

Attack Scenarios. In order to find a good trade-off between utility and privacy, we have to evaluate the attacker's success under the same scenario as the researcher. The attacker's goal is to detect with high probability whether a target is part of the MBeacon database or not. Of course, the attacker does not know whether she is querying a MBeacon of the form $B_{P,D}$ or B_D, similar to the researcher not knowing the distinction a priori. Moreover, the attacker's target might be a patient in D or in P. We refer to such an attacker as "full" attacker; a graphical overview is displayed in Figure 2b.

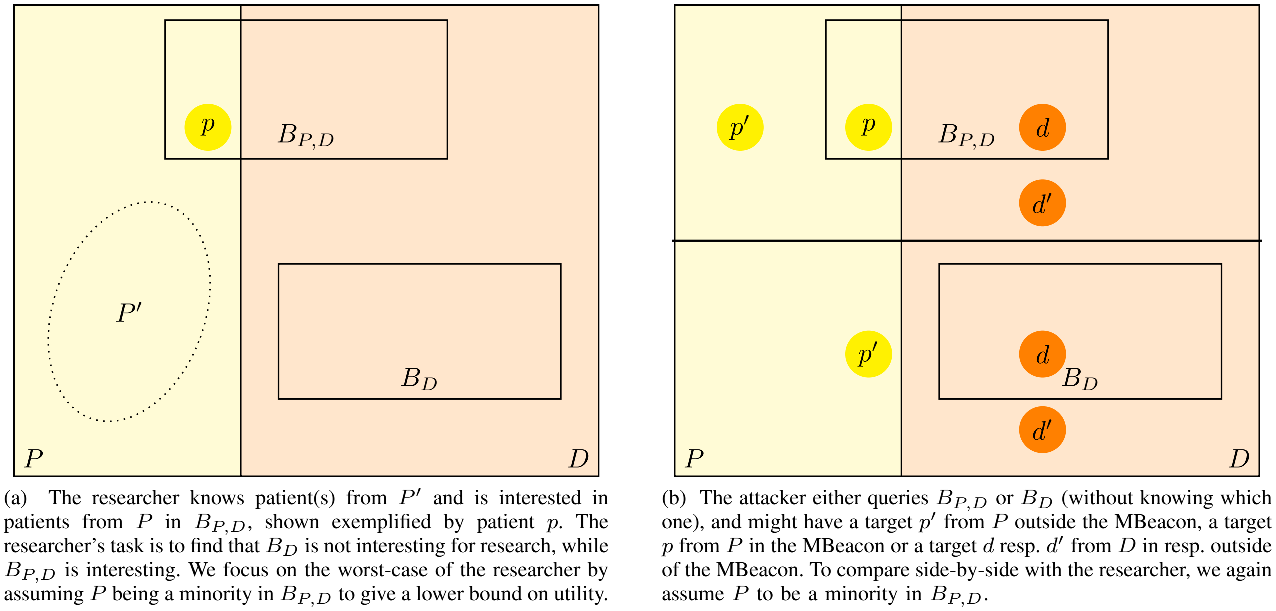
The evaluation of the "full" attacker is comparable to the researcher evaluation, but not to existing works [36], [29], [1], [45], where the MBeacon and the victim are from one uniform dataset. Therefore, we additionally model an attacker querying only B_D and trying to infer whether a victim in D is part of the MBeacon. We refer to this second attacker as the "standard" attacker, since it is the same as the one considered in Section V.

VIII. Defense Evaluation

We evaluate our defense mechanism SVT$^\mathbf{2}$ in this section with respect to the attack performance and utility as defined in Section VII.

A. Experimental Setup

For the set of researcher's interest, P, we use Ependymoma, which contains data from 48 patients. For the set D the researcher is not interested in, we use either GBM, IBD



(a) The researcher knows patient(s) from P' and is interested in patients from P in $B_{P,D}$, shown exemplified by patient p. The researcher's task is to find that B_D is not interesting for research, while $B_{P,D}$ is interesting. We focus on the worst-case of the researcher by assuming P being a minority in $B_{P,D}$ to give a lower bound on utility.

(b) The attacker either queries $B_{P,D}$ or B_D (without knowing which one), and might have a target p' from P outside the MBeacon, a target p from P in the MBeacon or a target d resp. d' from D in resp. outside of the MBeacon. To compare side-by-side with the researcher, we again assume P to be a minority in $B_{P,D}$.

Fig. 2. A graphical overview on the general utility setup for researcher (a) and the general utility setup for the attacker (b).

CD or IBD UC as before, forming three different types of MBeacons.

Each of these MBeacons consists of a certain number of patients in P, we test 7 different choices for this number including 1, 3, 5, 10, 13, 15 and 20. The remaining patients are randomly sampled from the respective D such that a total size of 60 is reached. Moreover, we sample randomly 60 patients from the respective D to construct B_D. We simulate 5 researchers querying each pair of corresponding MBeacons $B_{P,D}$ and B_D. The researcher possesses P' containing 5 randomly sampled patients in P that are not used in the MBeacon.[14] As mentioned in Section VII, the researcher averages these patients' profiles to reduce measurement errors. The whole sampling process is repeated 10 times to ensure the observations are not due to randomness.

For the attacker simulations, we re-use the MBeacons we constructed before for the researcher, but sample test patients differently. The "full" attacker has access to only a single patient. We randomly sample 12 patients from each of $B_{P,D}$ and B_D as the ones in the MBeacon. Accordingly, we sample 24 patients from $P \cup D$ as the patients that are outside the MBeacon. Since we assume throughout the experiments that patients in P are the minority, we use only up to a third of patients in P and the remainder in D. As before, we repeat random sampling 10 times. The "best" attacker does not have access to $B_{P,D}$ and, consequently, does not get test patients in P. Instead, we sample 24 test patients from B_D and 24 test patients from $D \setminus B_D$ for each of the B_D MBeacons.

We assume both researchers and attackers have access to the mean and standard deviation of the general population, that we estimate by a union of all our available datasets as before. These means and standard deviations are used to carry out LR tests and rank queries, up to 250,000 queries are allowed per researcher resp. attacker. Moreover, both researchers and attackers sort their queries based on expected information gain as explained in Section IV and used in the previous experiments in Section V.

To sum up, we test three different choices for D, and 7 different numbers of patients from P in $B_{P,D}$, simulate 5 researchers querying each of the MBeacons and re-sample the experiments 10 times, so simulate in total 2,100 researchers. Due to the attackers not averaging over multiple patients, we can simulate more membership inference attacks: 10,500 carried out by the "full" and the "standard" attacker each.

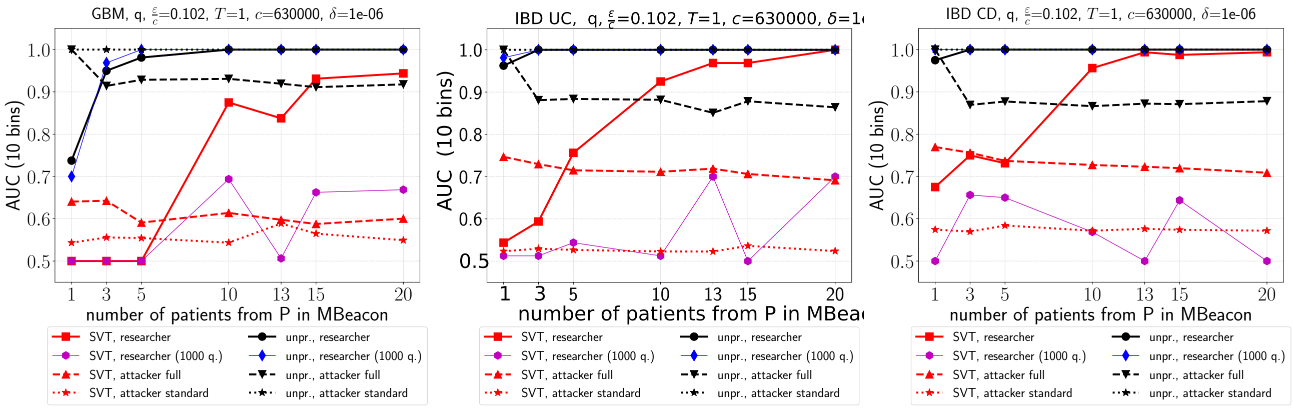
B. Evaluation of SVT^2

First, we evaluate the influence of the number of patients in P in the MBeacons of type $B_{P,D}$. We observe that if there are 5 or more patients of interest, the researcher's performance is maximized. The "full" attacker, however, suffers from more patients in P, probably due to the higher variance in the MBeacon.

Second, we focus on SVT^2. Our protection mechanism has three parameters: a threshold T determining how many patients have to be in the respective bin to answer "Yes", as well as the privacy parameter ϵ and the query budget c, which both calibrate the noise.

The Privacy Budget. We aim for parameters that drop the "standard" attackers' performance to about 0.5 AUC, equivalent to random guessing, while minimizing the noise. Moreover, exceeding the query budget is something MBeacon providers would want to avoid, because the MBeacon has to stop answering in that case. Therefore, we choose a budget that is never exceeded in our simulations. The researchers and the two different types of attackers ("standard" and "full") are all simulated separately, so our budget has to be sufficient for 50 attackers submitting 12,500,000 (50×250,000) queries in total.

[14] If the researcher averages fewer patients, the performance could decrease slightly since individual, non-disease related changes in the patients' methylation values become more pronounced in the search.

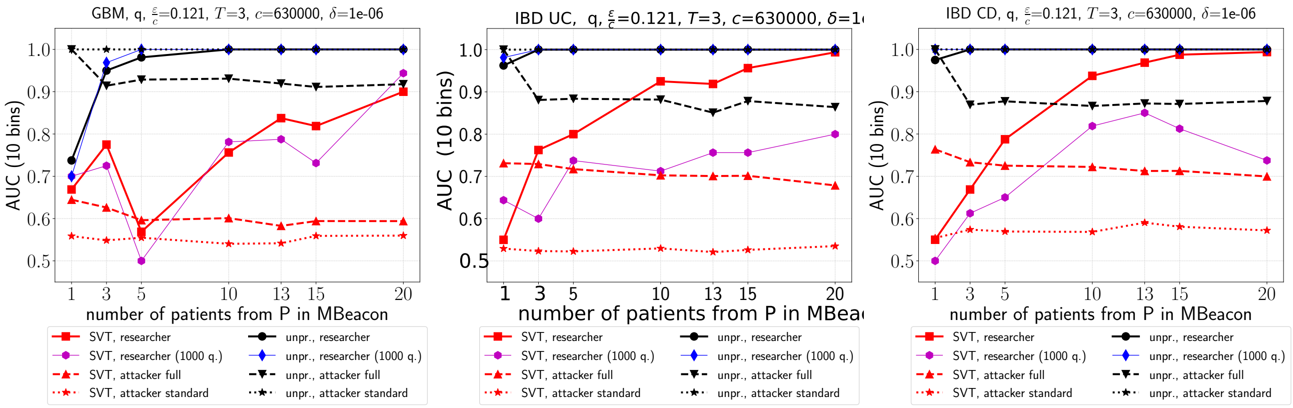


(a)

(b)

(c)

Fig. 3. Comparison of researchers' and attackers' performances in unprotected MBeacon (black, abbreviated as "unpr.") and protected MBeacon (red) using GBM (left), IBD UC (middle) and IBD CD (right) as D using up to 100,000 queries. Additionally, we plot the researchers' performances for 1,000 queries in blue (unprotected) and magenta (protected). AUCs with values smaller than 0.5 are displayed as 0.5.



(a)

(b)

(c)

Fig. 4. Comparison of researchers' and attackers' performances when setting $T = 3$ in unprotected MBeacon (black, abbreviated as "unpr.") and protected MBeacon (red) using GBM (left), IBD UC (middle) and IBD CD (right) as D using up to 100,000 queries. Additionally, we plot the researchers' performances for 1,000 queries in blue (unprotected) and magenta (protected). AUCs with values smaller than 0.5 are displayed as 0.5.

Notice that not all of those queries are expected to be unique and not all of them fall into the category of privacy-sensitive queries for which the budget must be reduced.

Threshold $T = 1$. We start with the default threshold $T = 1$, i.e., the MBeacon answers "Yes" if there is at least one patient's methylation value in the queried bin. A budget of $c = 630,000$ is sufficient for our simulations. This might seem large at first glance, but notice that, having 10 bins, there are 300,000×10 different queries that can be asked, so our c corresponds to about 21% of them. Due to space constraints, we report the privacy level that we found as a suitable trade-off between privacy and utility at $\frac{\epsilon}{c} = 0.102$. We report the privacy levels as in [34].

As shown in Figure 3, the privacy level is sufficient to drop the "standard" attackers' performance to less than 0.6 AUC which shows that the privacy threat can be mitigated successfully. In the more realistic "full" attacker scenario, however, the attacker's performance is higher, which is explained by the fact that membership attacks with patients from P against the B_D MBeacon are most successful. Nevertheless, we see a significant drop in performance due to the application of SVT^2.

The researcher's performance is still good with 0.8 AUC or more, depending on the number of patients from P in the MBeacon.

The impact of noise gets even more pronounced if we assume the researcher to submit only 1,000 queries. On the unprotected methylation Beacon, the AUC is about the same, however, the researcher cannot get good answers from an SVT^2-protected MBeacon. This shows the price of the SVT^2 protection: more queries have to be submitted.

Threshold $T = 3$. Next, we increase the threshold. We keep the same budget c since we just want to study the influence of the increased threshold, for which we also have to increase the privacy budget to $\frac{\epsilon}{c} = 0.121$. Figure 4 shows the result, we see a similar performance. This is the same for $T = 2$, which we do not show here for space constraints. A threshold $T > 3$ would probably not be accepted by researchers given this MBeacon sizes, therefore, we did not experiment with higher thresholds.

Setting the Parameters. The above results demonstrate that the threshold and other privacy parameters have to be chosen dependent on the use case to maximize utility and minimize the privacy loss. We believe that our general method of parameter tuning, namely, setting a budget c that is not exceeded, then changing values of ϵ based on attacker's and researcher's performance and increasing c if needed by a higher noise level (or reducing it if the consumed budget is much smaller), yields a good trade-off between utility and privacy for any dataset and MBeacon size.

IX. Prototype

We implement a fully functional prototype of our MBeacon system, which can be accessed at https://mbeacon-network.github.io/MBeacon-network/. Our prototype is built based on the same algorithms used in our experiments. All our implementation is done in Python, with packages including Pandas, Numpy, and Scipy. We rely on Flask[15] to build the web frontend. This allows us to seamlessly integrate our implementation into the backend, forming a centralized service as part of our prototype. The backend of our prototype is responsible to query the datasets and to return the MBeacon output after $\text{SVT}^{\mathbf{2}}$ has been applied. As an input, it takes a CpG identifier as well as the methylation value at this position to look for. Upon getting a query, our MBeacon system will return all the institutions that have the corresponding data.

In the future, we envision our prototype to be run in a decentralized manner, so that every data provider runs their own MBeacon service. In this scenario, the centralized service is only required to provide the frontend.

X. Related Work

Homer et al. [16] are among the first to perform a membership inference attack on genomic data. In their attack, summary statistics are used as the adversary's background knowledge and the L_1 distance to measure the similarity between summary and victim. Sankararaman et al. [32] further improved Homer's attack by incorporating the LR test in the algorithm. More recently, Backes et al. [5] have shown that membership inference attacks can be also successfully performed on epigenetic data, such as microRNA. Due to the threat demonstrated by the attacks, sharing biomedical data (or even summary statistics) has to take privacy into account which often prolongs the process for researchers to get data. In response, GA4GH established the Beacon system [10] to facilitate genomic data sharing.

Attacks on Genomic Beacons. Shringarpure and Bustamante [36] showed that even only given binary responses, it is possible to infer whether a patient is in a Beacon with the LR test. Moreover, their attack's probability estimation is not dependent on the allele frequencies, but the more stable allele distribution. While they studied the influence of several factors (population structure, Beacon size and others) on the attack's effectiveness, they did not propose any feasible solutions to establish a privacy-preserving genomic Beacon.

Raisaro et al. [29] extended the attack in [36] by adopting a sophisticated selection strategy. The attacker in this setting has direct access to allele frequencies and selects the most informative positions to query first. This setup serves as a blueprint for our attack against MBeacons.

The authors of [44] proposed an attack using the correlations between different single nucleotide polymorphisms (SNPs) to infer alleles that are missing or systematically hidden. This attack drops the number of queries necessary to infer membership with strong confidence, and renders privacy-preserving mechanisms based on hiding low-frequency SNPs useless. However, for DNA methylation, such correlations are not (yet) well studied. Therefore, we decide to postpone an in-depth study about the influence of correlations between methylation positions on the privacy risks to future work.

Privacy Protection for Beacons. Besides the attack, Raisaro et al. [29] proposed three protection mechanisms and experimentally showed their effectiveness even in their stronger attacker setting. However, they do not provide any formal guarantees on their protection mechanisms.

Wan et al. [45] further analyzed the protection mechanisms presented in [29], and additionally proposed a new one. They empirically evaluated utility, privacy and effectiveness of the protection methods under several settings with respect to the hyperparameters. Here, the corresponding utility, privacy and effectiveness measures were proposed in the iDASH challenge for genomic data.

Two additional privacy protection mechanisms are proposed by Al Aziz et al. [1], one of which, the biased randomized response, is proven to be differentially private. Apart from that, they analyzed both mathematically and experimentally how the decision boundary for membership relates to the number of queries and the number of patients in the Beacon.

To the best of our knowledge, the existing attacks are all conducted on genomic Beacons, and we propose the first membership inference attack on Beacons with DNA methylation data. Moreover, by simulating legitimate and adversarial behavior, we believe that our utility measures provide a more realistic picture. It is worth noting that the privacy and utility measures we propose in this paper are not limited to MBeacons, we leave their application on other types of biomedical data as a future work.

XI. Conclusion

In this paper, we propose the first Beacon system for sharing DNA methylation data, namely, the MBeacon system. Due to the severe privacy risks stemming from DNA methylation data, our construction of MBeacon follows a privacy-by-design approach.

We first illustrate the severe privacy risks by conducting a membership inference attack based on the LR test. Experimental results on multiple datasets show that with 100 queries, the adversary is able to achieve a superior performance. Then, we propose a defense mechanism, $\text{SVT}^{\mathbf{2}}$, to implement our privacy-preserving MBeacon. Our $\text{SVT}^{\mathbf{2}}$ is an advancement of the sparse vector technique, one type of differential privacy algorithms. We theoretically prove that $\text{SVT}^{\mathbf{2}}$ is differentially private. Since the goal of MBeacon is to facilitate biomedical data sharing, we propose a new metric for measuring researchers' utility considering a realistic scenario.

[15] http://flask.pocoo.org/

Extensive experiments demonstrate that, using carefully chosen parameters, MBeacon can degrade the performance of the membership inference attack significantly without substantially hurting the researchers' utility.

There are two directions we want to explore in the future. First, we plan to extend the Beacon-style system to other types of biomedical data, such as gene expression, microRNA or laboratory tests. In particular, this requires to adapt the estimate of the general population accordingly. Second, the current Beacon systems only support queries on a single position. We plan to extend the Beacon system to support multiple-position queries. On one hand, this new system should improve the utility for the researchers. On the other hand, it will raise new privacy challenges.

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). The author from the Swiss Data Science Center is supported by the grant #2017-201 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain. The authors from Indiana University are supported in part by National Science Foundation CNS-1408874 and National Health Institute U01EB023685. The authors would like to thank Dr. Rose Hoberman for her valuable comments on the submitted manuscript, and Dr. Ninghui Li for pointing out the technical inconsistency of the previous version.

References

[1] M. M. Al Aziz, R. Ghasemi, M. Waliullah, and N. Mohammed, "Aftermath of bustamante attack on genomic beacon service," *BMC medical genomics*, vol. 10, no. 2, p. 43, 2017.

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

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

[4] 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 (USENIX)*. USENIX Association, 2016, pp. 1223–1240.

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

[6] M. Backes, M. Humbert, J. Pang, and Y. Zhang, "walk2friends: Inferring Social Links from Mobility Profiles," in *Proceedings of the 24th ACM Conference on Computer and Communications Security (CCS)*. ACM, 2017, pp. 1943–1957.

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

[8] P. Berrang, M. Humbert, Y. Zhang, I. Lehmann, R. Eils, and M. Backes, "Dissecting privacy risks in biomedical data," in *Proceedings of the 3rd IEEE European Symposium on Security and Privacy (Euro S&P). IEEE*, 2018.

[9] P. Buczkowicz, C. Hoeman, P. Rakopoulos, S. Pajovic, L. Letourneau, M. Dzamba, A. Morrison, P. Lewis, E. Bouffet, U. Bartels *et al.*, "Genomic analysis of diffuse intrinsic pontine gliomas identifies three molecular subgroups and recurrent activating ACVR1 mutations," *Nature genetics*, vol. 46, no. 5, pp. 451–456, 2014.

[10] J. Burn, "A federated ecosystem for sharing genomic, clinical data," *Science*, vol. 352, pp. 1278–1280, 2016.

[11] C. Dwork, A. Roth *et al.*, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

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

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

[14] A. M. Fontebasso, S. Papillon-Cavanagh, J. Schwartzentruber, H. Nikbakht, N. Gerges *et al.*, "Recurrent somatic mutations in ACVR1 in pediatric midline high-grade astrocytoma," *Nature genetics*, vol. 46, no. 5, pp. 462–466, 2014.

[15] 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 (USENIX)*. USENIX Association, 2014, pp. 17–32.

[16] 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*, vol. 4, no. 8, p. e1000167, 2008.

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

[18] C. L. Kleinman, N. Gerges, S. Papillon-Cavanagh, P. Sin-Chan, A. Pramatarova, D.-A. K. Quang, V. Adoue, S. Busche, M. Caron, H. 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*, vol. 46, no. 1, pp. 39–44, 2014.

[19] S. R. Lambert, H. Witt, V. Hovestadt, M. Zucknick, M. Kool, D. M. Pearson, A. Korshunov, M. Ryzhova, K. Ichimura, N. Jabado *et al.*, "Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma," *Acta neuropathologica*, vol. 126, no. 2, pp. 291–301, 2013.

[20] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[21] M. Lyu, D. Su, and N. Li, "Understanding the Sparse Vector Technique for Differential Privacy," *Proceedings of the VLDB Endowment*, vol. 10, no. 6, pp. 637–648, 2017.

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

[23] M. Ni, Y. Zhang, W. Han, and J. Pang, "An Empirical Study on User Access Control in Online Social Networks," in *Proceedings of the 2016 ACM Symposium on Access Control Models and Technologies (SACMAT)*. ACM, 2016, pp. 12–23.

[24] B. Oprisanu and E. De Cristofaro, "Anonimme: Bringing anonymity to the matchmaker exchange platform for rare disease gene discovery," *bioRxiv*, p. 262295, 2018.

[25] J. Pang and Y. Zhang, "Location Prediction: Communities Speak Louder than Friends," in *Proceedings of the 2015 ACM Conference on Online Social Networks (COSN)*. ACM, 2015, pp. 161–171.

[26] J. Pang and Y. Zhang, "DeepCity: A Feature Learning Framework for Mining Location Check-Ins," in *Proceedings of the 2017 International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 2017, pp. 652–655.

[27] J. Pang and Y. Zhang, "Quantifying Location Sociality," in *Proceedings of the 2017 ACM Conference on Hypertext and Social Media (HT)*. ACM, 2017, pp. 145–154.

[28] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "Knock Knock, Who's There? Membership Inference on Aggregate Location Data,"

in *Proceedings of the 25th Network and Distributed System Security Symposium (NDSS)*, 2018.

[29] J. L. Raisaro, F. Tramèr, Z. Ji, D. Bu, Y. Zhao, K. Carey, D. Lloyd, H. Sofia, D. Baker, P. Flicek *et al.*, "Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks," *Journal of the American Medical Informatics Association*, p. ocw167, 2017.

[30] H. A. Rogers, J.-P. Kilday, C. Mayne, J. Ward, M. Adamowicz-Brice, E. C. Schwalbe, S. C. Clifford, B. Coyle, and R. G. Grundy, "Supratentorial and spinal pediatric ependymomas display a hypermethylated phenotype which includes the loss of tumor suppressor genes involved in the control of cell growth and death," *Acta neuropathologica*, vol. 123, no. 5, pp. 711–725, 2012.

[31] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

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

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

[34] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in *Proceedings of the 22nd ACM conference on computer and communications security (CCS)*. ACM, 2015, pp. 1310–1321.

[35] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks against Machine Learning Models," in *Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2017, pp. 3–18.

[36] S. S. Shringarpure and C. D. Bustamante, "Privacy Risks from Genomic Data-Sharing Beacons," *The American Journal of Human Genetics*, vol. 97, no. 5, pp. 631–646, 2015.

[37] T. F. M. Statistics, "A Decision Theoretic Approach," 1967.

[38] D. Sturm, H. Witt, V. Hovestadt, D.-A. Khuong-Quang, D. T. Jones, C. Konermann, E. Pfaff, M. Tönjes, M. Sill, S. Bender *et al.*, "Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma," *Cancer cell*, vol. 22, no. 4, pp. 425–437, 2012.

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

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

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

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

[43] N. Ventham, N. Kennedy, A. Adams, R. Kalla, S. Heath, K. O'leary, H. Drummond, D. Wilson, I. G. Gut, E. Nimmo *et al.*, "Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease," *Nature communications*, vol. 7, p. 13507, 2016.

[44] N. von Thenen, E. Ayday, and A. E. Cicek, "Re-identification of individuals in genomic data-sharing beacons via allele inference," *bioRxiv*, p. 200147, 2017.

[45] Z. Wan, Y. Vorobeychik, M. Kantarcioglu, and B. Malin, "Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services," *BMC medical genomics*, vol. 10, no. 2, p. 39, 2017.

[46] Y. Zhang, M. Humbert, T. Rahman, C.-T. Li, J. Pang, and M. Backes, "Tagvisor: A Privacy Advisor for Sharing Hashtags," in *Proceedings of the 2018 Web Conference (WWW)*. ACM, 2018, pp. 287–296.

[47] Y. Zhang, M. Humbert, B. Surma, P. Manoharan, J. Vreeken, and M. Backes, "CTRL+Z: Recovering Anonymized Social Graphs," *CoRR abs/1711.05441*, 2017.

XII. Appendix

A. Full Proof of Theorem 1

We begin by disassembling the probability of getting a specific answer $\overrightarrow{R}$ from a database $\mathbb{I}$ as in Equation 19.

$$\begin{aligned}\Pr[\mathcal{A}(\mathbb{I}) = \overrightarrow{R}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} & \Pr[\rho_1 = z_1 \wedge \rho_2 = z_2] \\ & f_{\mathbb{I}}(z_1, z_2) g_{\mathbb{I}}(z_1, z_2) dz_1 dz_2\end{aligned} \tag{19}$$

where

$$f_{\mathbb{I}}(z_1, z_2) = \Pr[\wedge_{i \in I_\perp} r_i = \perp | \rho_1 = z_1 \wedge \rho_2 = z_2] \tag{20}$$

$$g_{\mathbb{I}}(z_1, z_2) = \Pr[\wedge_{i \in I_\top} r_i = \top | \rho_1 = z_1 \wedge \rho_2 = z_2] \tag{21}$$

Intuitively, g deals with the positive answers indicating highly privacy-sensitive results and f deals with the negative answers. We will show that, for sensitivity Δ,

$$f_{\mathbb{I}}(z_1, z_2) \leq f_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) \tag{22}$$

$$g_{\mathbb{I}}(z_1, z_2) \leq e^{2\epsilon_2} g_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) \tag{23}$$

$$\Pr[\rho_1 {=} z_1 \wedge \rho_2 {=} z_2] \leq e^{2\epsilon_1} \Pr[\rho_1 = z_1 + \Delta \wedge \rho_2 = z_2 - \Delta] \tag{24}$$

which gives us the required connection between the two neighboring databases $\mathbb{I}$ and $\mathbb{I}'$.

Proof of Inequality 22. Due to the independence of the database entries Equation 20 is equivalent to

$$f_{\mathbb{I}}(z_1, z_2) = \prod_{i \in I_\perp} \Pr[r_i = \perp | \rho_1 = z_1 \wedge \rho_2 = z_2] = *$$

By plugging in our query formula, we have:

$$\begin{aligned} * = \prod_{i \in I_\perp} & \Pr[(\alpha_i + y_i < T + z_1 \wedge \beta_i + y_i < T + z_1) \\ & \vee (\alpha_i + y_i' \geq T + z_2 \wedge \beta_i + y_i' \geq T + z_2)] \\ = \prod_{i \in I_\perp} & \Pr[(y_i < T + z_1 - \alpha_i \wedge y_i < T + z_1 - \beta_i) \\ & \vee (y_i' \geq T + z_2 - \alpha_i \wedge y_i' \geq T + z_2 - \beta_i)] = * \end{aligned}$$

Next, we want to exploit the sensitivity to change to the other database. We know that $|\alpha_i - \alpha_i'| \leq \Delta$ leads to

$$\alpha_i \leq \alpha_i' + \Delta \quad \text{and} \quad \alpha_i \geq \alpha_i' - \Delta. \tag{a}$$

Similarly, $|\beta_i - \beta_i'| \leq \Delta$ indicates

$$\beta_i \leq \beta_i' + \Delta \quad \text{and} \quad \beta_i \geq \beta_i' - \Delta. \tag{b}$$

By using Equation (a) and (b), we have the following relation.

$$\begin{aligned} * \leq \prod_{i \in I_\perp} & \Pr[(y_i < T + z_1 - (\alpha_i' - \Delta) \wedge y_i < T + z_1 - (\beta_i' - \Delta)) \\ & \vee (y_i' \geq T + z_2 - (\alpha_i' + \Delta) \wedge y_i' \geq T + z_2 - (\beta_i' + \Delta))] \\ = \prod_{i \in I_\perp} & \Pr[(\alpha_i' + y_i < T + (z_1 + \Delta) \wedge \beta_i' + y_i < T + (z_1 + \Delta)) \\ & \vee (\alpha_i' + y_i' \geq T + (z_2 - \Delta) \wedge \beta_i' + y_i' \geq T + (z_2 - \Delta))] \\ = \; & f_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) \end{aligned}$$

Therefore, Inequality 22 if proven. Notice that the last step of simplification would not be possible if we had just used one noise variable $z = z_1 = z_2$. □

Proof of Inequality 23. Again, by independence of the database entries and the negation of our query formulation, we have:

$$g_{\mathbb{I}}(z_1, z_2) = \prod_{i \in I_\top} \Pr[\neg((\alpha_i + y_i < T + z_1 \wedge \beta_i + y_i < T + z_1) \\ \vee (\alpha_i + y_i' \geq T + z_2 \wedge \beta_i + y_i' \geq T + z_2))] = *$$

We push the negation inwards:

$$* = \prod_{i \in I_\top} \Pr[(\alpha_i + y_i \geq T + z_1 \vee \beta_i + y_i \geq T + z_1) \\ \wedge (\alpha_i + y_i' < T + z_2 \vee \beta_i + y_i' < T + z_2)] = *$$

The sensitivities $|\alpha_i - \alpha_i'| \leq \Delta$ and $|\beta_i - \beta_i'| \leq \Delta$ allow us to introduce the other database $\mathbb{I}'$ similar to before:

$$* \leq \prod_{i \in I_\top} \Pr[(y_i \geq T + z_1 - \alpha_i' - \Delta \vee y_i \geq T + z_1 - \beta_i' - \Delta) \\ \wedge (y_i' < T + z_2 - \alpha_i' + \Delta \vee y_i' < T + z_2 - \beta_i' + \Delta)] = *$$

We could go on as before with f, but it would not provide the desired bounds, as the signs of Δ would be flipped. Instead, we exploit that the noise values y_i are $\textsc{Lap}(\frac{2c\Delta}{\epsilon_2})$ distributed:

$$\Pr[\rho = y_i] \leq e^{\frac{\epsilon_2}{c}} \Pr[\rho = y_i + 2\Delta] \tag{c}$$

$$\Pr[\rho = y_i] \leq e^{\frac{\epsilon_2}{c}} \Pr[\rho = y_i - 2\Delta] \tag{d}$$

We cannot use that directly, as we have a logical formula in the probabilities. The outer conjunction can be rewritten to a multiplication due to independence of the noise variables v_i, v_i'. The inner disjunction is not problematic, as we show below. We prove it generally for any $x, x', Y_1, Y_2, Y_3, Y_4$ to increase readability. Later, we just need the following instantiations:

$$x = y_i \qquad x' = y_i'$$
$$Y_1 = T + z_1 - \alpha_i - \Delta \qquad Y_2 = T + z_1 - \beta_i - \Delta$$
$$Y_3 = T + z_2 - \alpha_i + \Delta \qquad Y_4 = T + z_2 - \beta_i + \Delta$$

We want to re-formulate $\Pr[x \geq Y_1 \vee x \geq Y_2]$ for some arbitrary, but fixed x, Y_1, Y_2. For probabilities, the following holds:

$$\Pr[x \geq Y_1 \vee x \geq Y_2] = \Pr[x \geq \min(Y_1, Y_2)]$$

Then, we apply (c):

$$\begin{aligned}
&\Pr[x \geq \min(Y_1, Y_2)] = \Pr[x \geq M] = \int_M^\infty \Pr[x = m] dm \\
&\leq e^{\frac{\epsilon_2}{c}} \int_M^\infty \Pr[x = m + 2\Delta] dm \text{ (substitute } t = \phi(m) = m + 2\Delta) \\
&= e^{\frac{\epsilon_2}{c}} \int_{\phi(M)}^{\phi(\infty)} \Pr[x = t] dt = e^{\frac{\epsilon_2}{c}} \Pr[x \geq \phi(M)] \\
&= e^{\frac{\epsilon_2}{c}} \Pr[x \geq \min(Y_1, Y_2) + 2\Delta] \\
&= e^{\frac{\epsilon_2}{c}} \Pr[x - 2\Delta \geq \min(Y_1, Y_2)] \\
&= e^{\frac{\epsilon_2}{c}} \Pr[x - 2\Delta \geq Y_1 \vee x - 2\Delta \geq Y_2]
\end{aligned} \tag{25}$$

Similarly, we re-formulate $\Pr[x' < Y_3 \vee x' < Y_4]$ for some arbitrary, but fixed x', Y_3, Y_4.

$$\Pr[x' < Y_3 \vee x' < Y_4] = \Pr[x' < \max(Y_3, Y_4)]$$

Now, we apply (d) as above:

$$\begin{aligned}
&\Pr[x' < \max(Y_3, Y_4)] \leq e^{\frac{\epsilon_2}{c}} \Pr[x' < \max(Y_3, Y_4) - 2\Delta] \\
&= e^{\frac{\epsilon_2}{c}} \Pr[x' + 2\Delta < \max(Y_3, Y_4)] = e^{\frac{\epsilon_2}{c}} \Pr[x' + 2\Delta < Y_3 \vee x' + 2\Delta < Y_4]
\end{aligned} \tag{26}$$

Now, we come back to the proof for Inequality 23. Since v_i and v_i' are independent, we have the following.

$$* = \prod_{i \in I_\top} \Pr[y_i \geq T + z_1 - \alpha_i' - \Delta \vee y_i \geq T + z_1 - \beta_i' - \Delta] \\ \Pr[y_i' < T + z_2 - \alpha_i' + \Delta \vee y_i' < T + z_2 - \beta_i' + \Delta] = *$$

Next, by utilizing Inequalities 25 and 26, we have:

$$\begin{aligned}
* &\leq \prod_{i \in I_\top} e^{\frac{\epsilon_2}{c}} \Pr[y_i \geq T + z_1 - \alpha_i' + \Delta \vee y_i \geq T + z_1 - \beta_i' + \Delta] \\
&\quad e^{\frac{\epsilon_2}{c}} \Pr[y_i' < T + z_2 - \alpha_i' - \Delta \vee y_i' < T + z_2 - \beta_i' - \Delta] \\
&= \prod_{i \in I_\top} e^{2\frac{\epsilon_2}{c}} \Pr[y_i \geq T + z_1 - \alpha_i' + \Delta \vee y_i \geq T + z_1 - \beta_i' + \Delta] \\
&\quad \Pr[y_i' < T + z_2 - \alpha_i' - \Delta \vee y_i' < T + z_2 - \beta_i' - \Delta] \\
&= e^{\frac{2\epsilon_2 |I_\top|}{c}} \prod_{i \in I_\top} \Pr[y_i \geq T + z_1 - \alpha_i' + \Delta \vee y_i \geq T + z_1 - \beta_i' + \Delta] \\
&\quad \Pr[y_i' < T + z_2 - \alpha_i' - \Delta \vee y_i' < T + z_2 - \beta_i' - \Delta] = *
\end{aligned}$$

As we have at most c answers for privacy-sensitive queries, i.e., $|I_\top| \leq c$, thus we have:

$$\begin{aligned}
* &\leq e^{2\epsilon_2} \prod_{i \in I_\top} \Pr[((y_i \geq T + z_1 - \alpha_i' + \Delta) \vee (y_i \geq T + z_1 - \beta_i' + \Delta)) \\
&\quad \wedge ((y_i' < T + z_2 - \alpha_i' - \Delta) \vee (y_i' < T + z_2 - \beta_i' - \Delta))] \\
&= e^{2\epsilon_2} g_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta)
\end{aligned}$$
□

Proof of Inequality 24. As ρ_1 and ρ_2 are sampled independently, $\Pr[\rho_1 = z_1 \wedge \rho_2 = z_2]$ equals to:

$$\Pr[\rho_1 = z_1 \wedge \rho_2 = z_2] = \Pr[\rho_1 = z_1] \Pr[\rho_2 = z_2] = *$$

Moreover, as ρ_1 and ρ_2 are sampled from $\textsc{Lap}(\frac{\Delta}{\epsilon_1})$, we have

$$\begin{aligned}
* &\leq e^{\epsilon_1} \Pr[\rho_1 = z_1 + \Delta] * e^{\epsilon_1} \Pr[\rho_2 = z_2 - \Delta] \\
&= e^{2\epsilon_1} \Pr[\rho_1 = z_1 + \Delta \wedge \rho_2 = z_2 - \Delta]
\end{aligned}$$
□

Let us wrap up using the above proofs on Inequalities 22 to 24 on 19.

$$\begin{aligned}
&\Pr[\mathcal{A}(\mathbb{I}) = \overrightarrow{R}] \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\rho_1 = z_1 \wedge \rho_2 = z_2] f_{\mathbb{I}}(z_1, z_2) g_{\mathbb{I}}(z_1, z_2) dz_1 dz_2 \\
&\leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{2\epsilon_1} \Pr[\rho_1 = z_1 + \Delta \wedge \rho_2 = z_2 - \Delta] \\
&\quad f_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) e^{2\epsilon_2} g_{\mathbb{I}'}(z_1 + \Delta, z_2 - \Delta) dz_1 dz_2 \\
&= e^{2\epsilon_1 + 2\epsilon_2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\rho_1 = z_1' \wedge \rho_2 = z_2'] \\
&\quad f_{\mathbb{I}'}(z_1', z_2') g_{\mathbb{I}'}(z_1', z_2') dz_1' dz_2' \\
&= e^{2(\epsilon_1 + \epsilon_2)} \Pr[\mathcal{A}(\mathbb{I}') = \overrightarrow{R}]
\end{aligned}$$

■

B. Fix of Technical Inconsistency in this Version

We fixed a technical inconsistency in the SVT^2 algorithm in this version. The prior version re-sampled the noise z_1, z_2 for the threshold T after line 10, i.e., if the answer is privacy-sensitive. That was, however, not consistent with the proof where we argue over all answers $I_\top, I_\bot$ with an arbitrary, but fixed noise z_1, z_2. Therefore, we removed the problematic line from the algorithm. Additionally, we re-evaluated our experiments and confirm that the observed trade-off between privacy and utility exists, even though with different privacy parameters. The Figures 3 and 4 as well as their description in Section VIII-B are updated accordingly.