Mechanisms for Hiding Sensitive Genotypes With Information-Theoretic Privacy

Fangwei Ye, *Member, IEEE*, Hyunghoon Cho, *Member, IEEE*, and Salim El Rouayheb, *Senior Member, IEEE*

***Abstract*— Motivated by the growing availability of personal genomics services, we study an information-theoretic privacy problem that arises when sharing genomic data: a user wants to share his or her genome sequence while keeping the genotypes at certain positions hidden, which could otherwise reveal critical health-related information. A straightforward solution of erasing (masking) the chosen genotypes does not ensure privacy, because the correlation between nearby positions can leak the masked genotypes. We introduce an erasure-based privacy mechanism with perfect information-theoretic privacy, whereby the released sequence is statistically independent of the sensitive genotypes. Our mechanism can be interpreted as a locally-optimal greedy algorithm for a given processing order of sequence positions, where utility is measured by the number of positions released without erasure. We show that finding an optimal order is NP-hard in general and provide an upper bound on the optimal utility. For sequences from hidden Markov models, a standard modeling approach in genetics, we propose an efficient algorithmic implementation of our mechanism with complexity polynomial in sequence length. Moreover, we illustrate the robustness of the mechanism by bounding the privacy leakage from erroneous prior distributions. Our work is a step towards more rigorous control of privacy in genomic data sharing.**

***Index Terms*— Information-theoretic privacy, genomic privacy, genomic data sharing, data sanitization, hidden Markov models.**

I. Introduction

A. Motivation

THE rise of personal genomics, whereby private individuals are exposed to an increasing range of direct-to-consumer services for sequencing, sharing, or analyzing their genomes, is leading to growing concerns for genomic privacy [1]–[3]. A personal genome is a rich trove of information about the underlying individual, including predictors for disease risks and other health-related traits, which holds great potential for improving one's health, yet may cause harm if used against the individual. Unlike other types of personal data like passwords, one's genetic data cannot be replaced once leaked, and a data breach may even affect the relatives of the individual whose genome is leaked. In order to facilitate the sharing of genomes to improve public health and advance science, we need principled strategies for controlling the privacy risks associated with genomic data sharing.

A key need in this regard is to selectively limit the leakage of information about biological or health-related traits of an individual that can be inferred from the shared genetic data. For example, one may wish to hide certain *genotypes* (an individual's genetic information at specific genomic positions) with well-established disease association before sharing his or her data with others (*e.g.,* analytic service providers or researchers). Such a capability would give the individuals more fine-grained control over their genomic privacy.

A simple approach to privacy protection, whereby specific positions in the genome deemed sensitive by the individual are masked before sharing the data, does not provide sufficient privacy protection. This is because the correlation structure among nearby genomic positions induced by the biological processes of genetic inheritance can be used to reconstruct the masked data as demonstrated in a number of studies [4], [5]. To prevent such an attack, one could alternatively erase all positions that are highly correlated with the sensitive sites [6], which may be achieved by masking the data within a large window around each sensitive position. Unfortunately, depending upon the chosen size of window, these approaches either provide incomplete privacy protection or require an excessive amount of data to be erased in order to achieve strong privacy (as we demonstrate in our results), thus limiting the usefulness of the shared data. Here, we aim to design a principled and effective mechanism for sharing a personal genome that provably hides sensitive positions, while introducing a small amount of erasure. Our techniques build upon the recent work on ON-OFF privacy [31], [32] while extending the theory to general data distributions beyond Markov chains addressed in the previous work.

It is worth noting that information-theoretic approaches are being increasingly explored for a diverse range of applications in genomics, including sequencing [7], genome-wide

Manuscript received December 5, 2020; revised September 12, 2021; accepted February 24, 2022. Date of publication March 3, 2022; date of current version May 20, 2022. The work of Fangwei Ye was supported in part by NSF under Grant CCF 1817635, in part by the National Institutes of Health (NIH) under Grant DP5 OD029574-01, and in part by the Eric and Wendy Schmidt through the Schmidt Fellows Program at Broad Institute. The work of Hyunghoon Cho was supported in part by NIH under Grant DP5 OD029574-01 and in part by the Eric and Wendy Schmidt through the Schmidt Fellows Program at Broad Institute. The work of Salim El Rouayheb was supported in part by NSF under Grant CCF 1817635. An earlier version of this paper was presented at the 2020 IEEE International Symposium on Information Theory [DOI: 10.1109/ISIT44484.2020.9174492]. *(Corresponding author: Fangwei Ye.)*

Fangwei Ye was with the Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854 USA. He is now with the Broad Institute of MIT and Harvard, Cambridge, MA 02142 USA (e-mail: fye@broadinstitute.org).

Hyunghoon Cho is with the Broad Institute of MIT and Harvard, Cambridge, MA 02142 USA (e-mail: hhcho@broadinstitute.org).

Salim El Rouayheb is with the Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ 08854 USA (e-mail: salim.elrouayheb@rutgers.edu).

Communicated by A. Beimel, Associate Editor for Sequences and Cryptography.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3156276.

Digital Object Identifier 10.1109/TIT.2022.3156276

association study (GWAS) [8], [9], genome assembly [10], [11], regulatory network of gene interactions (RNGI) [12], and DNA-based information storage [13]. There are also recent works addressing the issue of genomic privacy, including a solution for private shotgun sequencing [14] based on the intensively researched private information retrieval (PIR) problems [15]–[20] and differential privacy mechanisms for sharing aggregate genomic data [21]–[23]. Broadly, our work can be viewed as a continuation of these efforts to develop effective genomic data processing tools from an information-theoretic perspective, yet for a novel problem that we introduce, *i.e.,* the design of mechanisms for selectively hiding sensitive positions in genetic sequences.

B. Genetics Background

An individual's genome consists of a pair of sequences, one from each parent, each consisting of around 3 billion nucleotides (A, C, G, and T). Each sequence is referred to as a *haplotype*. Since most of the genome sequence is identical between different individuals, a common way to compactly represent a personal genome is as a list of positions of variation, paired with the observed nucleotide(s) in the given individual (referred to as a *genotype*). In this work, we consider the problem of sharing a list of genotypes corresponding to a *single* haplotype of an individual. Although standard sequencing or genotyping pipelines produce a genotype at each position that convolves the two haplotypes, well-established methods exist [24], [25] for resolving this ambiguity in order to separate the two haplotypes (a process called *phasing*), after which each haplotype could be individually considered.

In the setting of our work, we consider an adversary whose goal is to infer the target individual's genotypes at specific positions in the genome, given a partially masked genetic sequence of the individual. In principle, this reconstruction task is equivalent to an extensively studied problem in bioinformatics known as *genotype imputation*, originally developed for coping with the presence of missing data in the existing experimental pipelines for characterizing personal genomes. If one were to mask only the sensitive positions before sharing the data, existing imputation algorithms are expected to be effective at revealing the hidden genotypes using other genotypes in their respective neighborhoods.

A state-of-the-art algorithm for genotype imputation, Minimac [26], is based on a classical model of genetic sequences introduced by Li and Stephens [27]. In this model, a person's genetic sequence is modelled as a mosaic of a large group of reference sequences from other individuals. This model intuitively captures the underlying biological process of *recombination*, which describes the interleaving of two haplotypes of each parent when their genetic material is passed onto the child. Formally, these models are expressed as hidden Markov models (HMMs), where a sequence of genotypes of an individual is generated from a sequence of hidden states indicating which reference haplotype to copy the genotype from, for each corresponding position. The parameters of these models are typically inferred from a large reference panel including tens of thousands of sequenced human genomes [28]. Although alternative approaches to imputation (e.g. based on matrix factorization [29]) exist, in our work we are especially interested in HMMs as the primary means to model the distribution of genotypes, considering the wide adoption of HMMs in genetics not only for imputation, but also for other standard tasks like phasing [24] and simulation [30]. Further details of this model is provided in Section VII-A.

C. Setup and Contributions

In this paper, we formulate the *genotype hiding* problem: We consider a user who wishes to share a partially erased version of their genetic sequence while protecting a list of sensitive positions. Privacy is measured by the mutual information between the sensitive positions and the released sequence, and we adopt a stringent privacy requirement that enforces zero mutual information (i.e., perfect privacy). The goal of the problem is to design a privacy mechanism that satisfies this requirement, while minimizing the number of erasures introduced so as to maximize the utility of the data.

We present such a mechanism with perfect privacy and provide a range of theoretical insights into its performance with respect to its utility, measured by the erasure rate. The proposed mechanism sequentially processes the positions in the sequence in a given ordering and determines a suitable erasure rate at each position based on the previously released positions and the data generating distribution. We prove that our mechanism can be viewed as a locally-optimal, greedy solution for minimizing the erasure rate at each position. Furthermore, we give a lower bound on the number of erasures required for any mechanism satisfying the privacy constraint, and show that our privacy mechanism is in fact (globally) optimal for a class of data generative distributions defined by Markov chains. We also show that finding the optimal ordering for the sequential mechanism is generally intractable (NP-hard), illustrating the limits of current techniques. Lastly, we derive an upper bound on potential privacy leakage due to inaccuracies in the estimation of the data generative model, suggesting that our mechanism is relatively robust to a small amount of noise in the data distribution.

For practical applications, we are particularly interested in data generating distributions induced by hidden Markov models (HMMs), which are broadly adopted in genetics as described in Section VII-A. To this end, we also present a computationally-efficient algorithm to implement the proposed privacy mechanism based on HMMs, and provide an empirical evaluation of its performance on simulated datasets.

The rest of this paper is organized as follows. In Section II, we formalize the genotype-hiding problem. Performance bounds are summarized in Section III. In Section IV, we introduce our privacy mechanism for hiding sensitive genotypes. In Section V, we describe its interpretation as a locally-optimal solution in detail and demonstrate the NP-hardness of finding the optimal ordering in general. The robustness of our privacy mechanism to model mismatch is discussed in Section VI. In Section VII, we propose an efficient implementation of the privacy mechanism for hidden Markov models.

Simulation experiments are presented in Section VIII. Finally in Section IX, we conclude the paper and discuss future directions.

II. The Genotype-Hiding Problem

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be the user's personal genome sequence of length n, and each X_i takes values in the alphabet $\mathcal{X}$. The user wishes to share $\mathbf{X}$ with others, but is concerned about revealing information about certain positions of $\mathbf{X}$. To hide the values at these sensitive positions, the user generates a masked version of the data $\mathbf{Y} = (Y_1, \ldots, Y_n)$, which only partially reveals $\mathbf{X}$.

The desired properties of $\mathbf{Y}$ are given as follows. First, since we expect substitution errors to be considerably more undesirable than erasures in genetic analyses, we impose a constraint that Y_i can be either X_i or the erasure symbol $*$. We refer to this property as the *faithfulness condition*, i.e.,

$$Y_i = X_i \text{ or } *. \qquad (\textbf{Faithfulness}) \tag{1}$$

Note that the alphabet of Y_i is $\mathcal{X} \cup \{*\}$.

Next, let $\mathcal{K} \subset [n] := \{1, \ldots, n\}$ be the user-provided set of indices of $\mathbf{X}$ containing sensitive information. We assume that $\mathcal{K}$ is chosen irrespective of the sequence (*i.e.,* independently from $\mathbf{X}$) based on information such as family history or curated disease associations. We use $X_{\mathcal{K}}$ to denote a collection of random variables, i.e., $X_{\mathcal{K}} := \{X_i : i \in \mathcal{K}\}$. We require that no information about $X_{\mathcal{K}}$ is revealed when $\mathbf{Y}$ is shared. In other words, we require that

$$I(X_{\mathcal{K}}; \mathbf{Y}) = 0, \qquad (\textbf{Privacy}) \tag{2}$$

where $I(\cdot)$ denotes the mutual information. We refer to this requirement as the *privacy condition*. Note that our notion of privacy is stronger than alternatives such as local differential privacy [33], which allows a small amount of leakage. Our work focuses on maximizing the utility over all mechanisms satisfying the perfect privacy condition.

We aim to design a *privacy mechanism* $w(\mathbf{y}|\mathbf{x})$ to generate $\mathbf{Y}$ from given $\mathbf{X}$ and $\mathcal{K}$ such that both the faithfulness and privacy conditions are satisfied. Here, we consider the ideal scenario where the data generating distribution $p(\mathbf{x})$ is known to the mechanism. We discuss the impact of having an inaccurate $p(\mathbf{x})$ in Section VI; even under this challenging scenario, we show that the potential privacy leakage is bounded by the divergence between the given $p(\mathbf{x})$ and the true distribution. Note that we use uppercase symbols to represent random variables and lowercase symbols to denote their realizations.

While satisfying the above two conditions, we wish to share as much of $\mathbf{X}$ as possible. More precisely, let $e(\mathbf{Y})$ be the number of erasure symbols in $\mathbf{Y}$. Our goal is to minimize the expected number of erasures $\mathbb{E}[e(\mathbf{Y})]$, or equivalently the *erasure rate* $\frac{1}{n}\mathbb{E}[e(\mathbf{Y})]$, where

$$\mathbb{E}[e(\mathbf{Y})] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}\{Y_i = *\}] = \sum_{i=1}^{n} p(y_i = *), \tag{3}$$

and $\mathbb{1}\{\cdot\}$ denotes the indicator function.

A formal description of the genotype-hiding problem is given below. We start by defining the privacy mechanism for the genotype-hiding problem as follows.

Definition 1: An $(n, \mathcal{K})$ privacy mechanism for a given data generative distribution $p(\mathbf{x})$ with input alphabet $\mathcal{X}^n$ and output alphabet $\mathcal{Y}^n$ is defined by a probabilistic encoding function

$$\mathsf{Enc} : \mathcal{X}^n \to \mathcal{Y}^n,$$

where Enc satisfies both the faithfulness condition ($Y_i \in \{X_i, *\}, \forall i$) and the privacy condition ($I(X_{\mathcal{K}}; \mathbf{Y}) = \mathbf{0}$).

The performance of the privacy mechanism is measured by the expected number of erasures per symbol in an output sequence $\mathbf{y}$. This measure captures the distortion between the input and output sequences induced by a set of single-letter erasures. Following the convention, we define the *rate* of a privacy mechanism as the fraction of positions that are not erased in the output:

Definition 2: The rate of an $(n, \mathcal{K})$ privacy mechanism for a given data generative distribution $p(\mathbf{x})$ is defined by $1 - \frac{1}{n}\mathbb{E}[e(\mathsf{Enc}(\mathbf{X}))]$ per symbol.

Definition 3: For any given data distribution $p(\mathbf{x})$, a rate R is achievable if there exists an $(n, \mathcal{K})$ privacy mechanism such that

$$1 - \frac{1}{n}\mathbb{E}[e(\mathbf{Y})] \geq R, \tag{4}$$

where $\mathbf{Y} = \mathsf{Enc}(\mathbf{X})$.

Clearly, if R is achievable then $R - \epsilon$ for any $\epsilon > 0$ is also achievable by the definition, so we are interested in finding the maximum achievable rate.

It is worth noting that the encoder $\mathsf{Enc}(\cdot)$ can be potentially stochastic, so we may use conditional probabilities $w(\mathbf{y}|\mathbf{x})$ to represent the encoding function. If we treat conditional probabilities $w(\mathbf{y}|\mathbf{x})$ where $\mathbf{x} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{Y}^n$ as decision variables, the genotype-hiding problem can be defined as the following optimization problem:

$$\begin{aligned} \underset{w(\mathbf{y}|\mathbf{x})}{\text{maximize}} \quad & 1 - \frac{1}{n}\sum_{i=1}^{n} p(y_i = *) & \\ \text{subject to} \quad & I(X_{\mathcal{K}}; \mathbf{Y}) = 0 & \text{(Privacy)} \\ & Y_i \in \{X_i, *\}, \forall i & \text{(Faithfulness)} \end{aligned} \tag{5}$$

Note that this problem maximizes the information rate (utility) under the stringent privacy constraint such that no information about the sensitive positions is leaked.

If we express the objective and the constraints explicitly in terms of the conditional probabilities $w(\mathbf{y}|\mathbf{x})$, the optimization problem (5) can be viewed as an instance of linear programming (LP). However, the scale of the problem is intractable in practice, given the exponential blowup in the number of variables and constraints as the length of the sequence n grows; the number of decision variables is $|\mathcal{X}|^n|\mathcal{Y}|^n$, and the number of constraints is in the order of $|\mathcal{X}|^{|\mathcal{K}|}|\mathcal{Y}|^n + n|\mathcal{X}||\mathcal{Y}|$.

Therefore, the ultimate goal of this paper is to identify a solution to the genotype-hiding problem in a tractable and computationally-efficient manner. To this end, we first present an achievable privacy mechanism as well as an upper bound on the maximum achievable rate. Then we show that the proposed

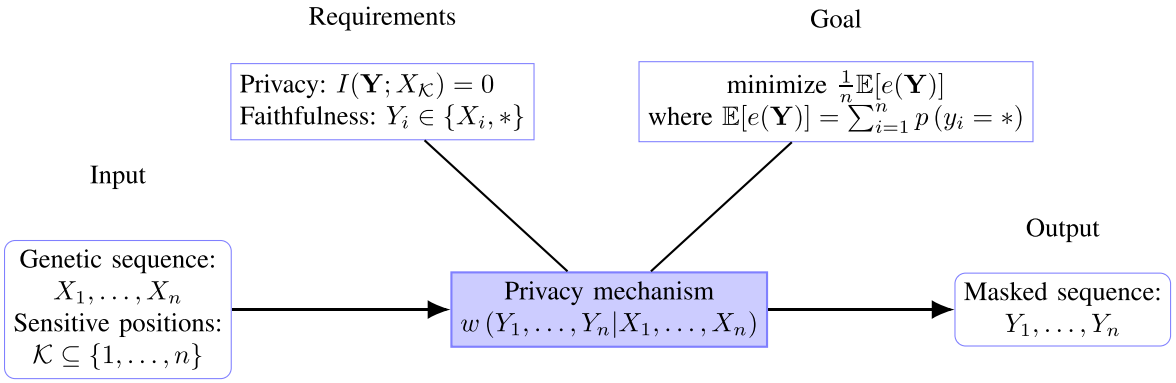

Fig. 1. An illustration of $(n, \mathcal{K})$ genotype-hiding privacy mechanism. The mechanism takes as input a genetic sequence along with a set of sensitive positions and outputs a masked sequence with erasures. We require the faithfulness and privacy conditions to be satisfied, and the goal is to minimize the expected number of erasures in the output.

privacy mechanism is computationally efficient for a particular data generative distribution, namely hidden Markov models, which is of broad interest in our motivating application in genomics.

III. Performance Bounds

In this section, we state the performance bounds on the achievable rate in the following theorems.

Theorem 1: For a given data distribution $p(\mathbf{x})$, a rate R is achievable if

$$R \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{x_i \in \mathcal{X}} \mathbb{E}_{Y_{[i-1]}} \left[\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i | x_{\mathcal{K}} = u, Y_{[i-1]}\right) \right]. \quad (6)$$

A detailed description of the achievable scheme will be presented in Section IV. The right-hand side of (6) may appear unconventional, given that conditioning on $Y_{[i-1]}$ for each i makes the probability term generally hard to compute as the sequence length n grows. However, this expression corresponds to a sequential mechanism where the encoder generates $Y_1, \ldots, Y_n$ one position at a time, and an efficient update exists for incrementally expanding the conditioning set. As an example, in Section VII, we present a concrete implementation of the privacy mechanism for data distributions governed by hidden Markov models, which indeed allows the right-hand side of (6) to be efficiently computed.

Theorem 2: For a given data distribution $p(\mathbf{x})$, any achievable rate R must satisfy

$$R \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p(x_i | x_{\mathcal{K}} = u). \quad (7)$$

It is worth noting that, given a data distribution $p(\mathbf{x})$, each summand in the right-hand side of (7) represents the conditional probability of the observation x_i at coordinate i when the sensitive positions $x_{\mathcal{K}}$ take on the *least*-likely values, which can be determined from the given $p(\mathbf{x})$.

Proof: From (3), we know that to establish (7), it is sufficient to show

$$p(y_i \neq *) \leq \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p(x_i | x_{\mathcal{K}} = u) \quad (8)$$

for any mechanism satisfying the privacy and faithfulness conditions. Consider

$$\begin{aligned} p(y_i \neq *) &= \sum_{y_i \in \mathcal{X}} p(y_i) \\ &\overset{(a)}{=} \sum_{y_i \in \mathcal{X}} \min_{u} p(y_i | x_{\mathcal{K}} = u) \\ &\overset{(b)}{=} \sum_{y_i \in \mathcal{X}} \min_{u} p(y_i = x_i | x_{\mathcal{K}} = u) \\ &= \sum_{x_i \in \mathcal{X}} \min_{u} p(x_i | x_{\mathcal{K}} = u)\, p(y_i = x_i | x_i, x_{\mathcal{K}} = u) \\ &\overset{(c)}{\leq} \sum_{x_i \in \mathcal{X}} \min_{u} p(x_i | x_{\mathcal{K}} = u), \end{aligned} \quad (9)$$

where (a) is due to the fact that Y_i is independent of $X_{\mathcal{K}}$ (privacy condition); (b) follows from the faithfulness condition $Y_i \in \{X_i, *\}$; and (c) follows from the fact that probabilities are bounded above by 1. □

Although not true in general, the upper bounds in (6) and (7) match under special circumstances, implying the optimality of an achievable mechanism. That is,

$$\begin{aligned} &\sum_{x_i \in \mathcal{X}} \sum_{y_{[i-1]}} p\left(y_{[i-1]}\right) \min_{x_{\mathcal{K}}} p\left(x_i | x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\qquad = \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p(x_i | x_{\mathcal{K}} = u). \end{aligned} \quad (10)$$

We observe that a *sufficient condition* for this equality is given by the following: for any x_i, if

$$u^* \in \arg\min_{u} p(x_i | x_{\mathcal{K}} = u), \quad (11)$$

then

$$u^* \in \arg\min_{u} p\left(x_i | x_{\mathcal{K}} = u, y_{[i-1]}\right) \quad (12)$$

for all possible $y_{[i-1]}$. Intuitively, this means that for any given position x_i, the least-likely values of the (unobserved) sensitive positions $x_{\mathcal{K}}$ remains the same regardless of the positions that have been previously released in the output $y_{[i-1]}$ during the course of the mechanism.

A special case that satisfies this optimality condition is when random variables $X_1, \ldots, X_n$ form a Markov chain (*i.e.,* $p(\mathbf{x})$

is induced by a Markov chain), with a single sensitive position. Without loss of generality, we assume $\mathcal{K} = \{1\}$.

Corollary 1 (Markov Chain): If $X_1, \ldots, X_n$ forms a Markov chain and the sensitive position is $\mathcal{K} = \{1\}$, then a rate R is achievable if and only if

$$R \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i | x_{\mathcal{K}} = u\right). \tag{13}$$

It is sufficient to justify the corollary by showing that the aforementioned sufficient condition holds. The proof is included in Appendix A.

IV. Privacy Mechanism

In this section, we present a privacy mechanism for generating $\mathbf{Y}$ based on a given $p(\mathbf{x})$, whose performance matches the bound given in (6), while satisfying both faithfulness and privacy conditions.

Let us first recall the genotype-hiding problem introduced in (5), i.e.,

$$\begin{aligned} \underset{w(\mathbf{y}|\mathbf{x})}{\text{maximize}} \quad & 1 - \frac{1}{n} \sum_{i=1}^{n} p\left(y_i = *\right) \\ \text{subject to} \quad & I(X_{\mathcal{K}}; \mathbf{Y}) = 0 \qquad \text{(Privacy)} \\ & Y_i \in \{X_i, *\}, \forall i. \qquad \text{(Faithfulness)} \end{aligned} \tag{14}$$

This problem is difficult to solve in its general form given the exponentially growing number of decision variables in $w(\mathbf{y}|\mathbf{x})$ as the sequence length n grows. Instead, we adopt a greedy optimization approach, whereby the erasure probability of y_i is locally minimized, one position at a time, from 1 to n. In other words, for each $i = 1, \ldots, n$, we solve

$$\begin{aligned} \underset{w\left(y_i | \mathbf{x}, y_{[i-1]}\right)}{\text{minimize}} \quad & p\left(y_i = * | y_{[i-1]}\right) \\ \text{subject to} \quad & I(X_{\mathcal{K}}; Y_i | Y_{[i-1]}) = 0 \\ & Y_i \in \{X_i, *\}, \end{aligned} \tag{15}$$

for any given $y_{[i-1]}$. Note that

$$I(X_{\mathcal{K}}; \mathbf{Y}) = \sum_{i=1}^{n} I(X_{\mathcal{K}}; Y_i | Y_{[i-1]}) = 0, \tag{16}$$

by the chain rule, so if the first constraint of (15) is satisfied for all i, then the solution preserves the required privacy constraint $I(X_{\mathcal{K}}; \mathbf{Y}) = 0$ as defined in (2). The second constraint is inherited directly from the faithfulness condition. In other words, any solution satisfying the constraints of (15) for all i will naturally be a feasible solution to the genotype-hiding problem in (5).

We observe that solving the local optimization problem (15) gives rise to a sequential mechanism for generating $\mathbf{Y}$. That is, we generate $\mathbf{Y}$ one position at a time, where the conditional distribution for Y_i may depend on the values of $Y_1, \ldots, Y_{i-1}$ that have been previously generated. The following defines our chosen privacy mechanism for any given position i, which is in fact an optimal solution to the local optimization problem (15). A detailed proof of the local optimality of this scheme is deferred to Section V.

A. Privacy Mechanism

Generate each Y_i according to the following conditional distribution

$$\begin{aligned} & w\left(y_i | x_i, x_{\mathcal{K}}, y_{[i-1]}\right) \\ & \quad = \begin{cases} \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i | x_{\mathcal{K}} = u, y_{[i-1]}\right)}{p\left(x_i | x_{\mathcal{K}}, y_{[i-1]}\right)}, & \text{if } y_i = x_i, \\ 1 - \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i | x_{\mathcal{K}} = u, y_{[i-1]}\right)}{p\left(x_i | x_{\mathcal{K}}, y_{[i-1]}\right)}, & \text{if } y_i = *, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{17}$$

for any x_i, $x_{\mathcal{K}}$ and $y_{[i-1]}$, where $[i-1] := \{1, \ldots, i-1\}$.

The expression for the erasure probability in the above mechanism can be intuitively understood as follows. We first identify the values of the sensitive positions with the smallest likelihood of generating the observed symbol x_i at the i-th position (as indicated by the numerator in the fractional term), conditioned on the previously released positions $y_{[i-1]}$. Note that u is an auxiliary variable denoting the possible values in the alphabet $\mathcal{X}^{|\mathcal{K}|}$, whereas $x_{\mathcal{K}}$ denotes the observed values at the sensitive positions. We then choose the erasure probability such that, the probability of releasing the original symbol (without erasure) becomes identical among different hypothetical values of $x_{\mathcal{K}}$, thus ensuring privacy.

It is worth noting that our privacy mechanism satisfies the faithfulness condition (*i.e.,* $y_i \in \{x_i, *\}$) by design, so we only need to verify that it satisfies the privacy constraint (2). Before verifying the privacy constraint, we note the following properties of the mechanism.

(1) If $i \in \mathcal{K}$, then

$$\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i | x_{\mathcal{K}} = u, y_{[i-1]}\right) = 0, \tag{18}$$

which yields

$$w\left(y_i = * | x_i, x_{\mathcal{K}}, y_{[i-1]}\right) = 1. \tag{19}$$

This implies that X_i is *always erased* if it corresponds to one of the sensitive positions in $\mathcal{K}$.

(2) We notice from (17) that X_i is *not* erased with some nonzero probability, so this mechanism is strictly better than the naïve approach of always erasing any position that have a nonzero correlation with the sensitive positions.

Proof of Privacy: To show that the proposed mechanism in (17) satisfies the privacy condition (2), it is sufficient to show

$$I(Y_i; X_{\mathcal{K}} | Y_1, \ldots, Y_{i-1}) = 0, \tag{20}$$

for all $i = 1, \ldots, n$, since this implies

$$I(X_{\mathcal{K}}; \mathbf{Y}) = \sum_{i=1}^{n} I(X_{\mathcal{K}}; Y_i | Y_{[i-1]}) = 0 \tag{21}$$

by the chain rule. To establish (20), we will equivalently prove that

$$p\left(y_i | x_{\mathcal{K}}, y_{[i-1]}\right) = p\left(y_i | y_{[i-1]}\right) \tag{22}$$

for any $x_{\mathcal{K}}$, $y_{[i-1]}$ and y_i. Since

$$\begin{aligned} & p\left(y_i | x_{\mathcal{K}}, y_{[i-1]}\right) \\ & \quad = \sum_{x_i \in \mathcal{X}} p\left(x_i | x_{\mathcal{K}}, y_{[i-1]}\right) w\left(y_i | x_i, x_{\mathcal{K}}, y_{[i-1]}\right), \end{aligned} \tag{23}$$

by substituting (17), we have

$$\begin{aligned} &p\left(y_i = *|x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\quad = \sum_{x_i} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right) w\left(y_i = *|x_i, x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\quad = 1 - \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right). \end{aligned} \tag{24}$$

Similarly, for $y_i \in \mathcal{X}$, we have

$$\begin{aligned} &p\left(y_i|x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\quad = \sum_{x_i \in \mathcal{X}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right) w\left(y_i = x_i|x_i, x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\quad = \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right). \end{aligned} \tag{25}$$

We can observe that the right-hand sides of both (24) and (25) are independent of $x_{\mathcal{K}}$, and hence by combining (24) and (25), we have

$$p\left(y_i|x_{\mathcal{K}}, y_{[i-1]}\right) = p\left(y_i|y_{[i-1]}\right), \tag{26}$$

for any $x_{\mathcal{K}}$, $y_{[i-1]}$ and y_i, which finishes the proof of (22). □

Finally, we can easily verify that our sequential privacy mechanism (17) achieves the rate

$$\begin{aligned} &1 - \frac{1}{n}\sum_{i=1}^{n} p\left(y_i = *\right) \\ &= 1 - \frac{1}{n}\sum_{i=1}^{n} \sum_{y_{[i-1]}} p\left(y_i = *|y_{[i-1]}\right) p\left(y_{[i-1]}\right) \\ &\overset{\text{(a)}}{=} 1 - \frac{1}{n}\sum_{i=1}^{n} \sum_{y_{[i-1]}} p\left(y_{[i-1]}\right) \\ &\qquad\qquad \left(1 - \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \sum_{y_{[i-1]}} p\left(y_{[i-1]}\right) \sum_{x_i \in \mathcal{X}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \sum_{x_i \in \mathcal{X}} \sum_{y_{[i-1]}} p\left(y_{[i-1]}\right) \min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right), \end{aligned} \tag{27}$$

where (a) follows by (24) and (26). The final expression is identical to the right-hand side of (6) as desired.

Example: We present an example to illustrate the operations of the proposed privacy mechanism in a simplified setting. Let us consider a data distribution $p\left(\mathbf{x}\right)$ where $X_1, \ldots, X_n$ form a Markov chain, as in Corollary 1, and a single sensitive position $\mathcal{K} = \{1\}$.

By inspecting the privacy mechanism in (17), we know that if $y_{i-1} \neq *$ for some $i > 1$, then

$$\begin{aligned} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right) &= p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}, x_{i-1} = y_{i-1}\right) \\ &= p\left(x_i|x_{i-1} = y_{i-1}\right), \end{aligned} \tag{28}$$

for any x_i and $y_{[i-1]}$ by the Markov property and the fact that $\mathcal{K} = \{1\}$. This implies that

$$\begin{aligned} w\left(y_i = x_i|x_i, x_{\mathcal{K}}, y_{[i-1]}\right) &= \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right)}{p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right)} \\ &= \frac{p\left(x_i|x_{i-1} = y_{i-1}\right)}{p\left(x_i|x_{i-1} = y_{i-1}\right)} \\ &= 1, \end{aligned} \tag{29}$$

which means that if $y_{i-1} \neq *$ then $y_i \neq *$ with probability one.

Thus, when $p\left(\mathbf{x}\right)$ is specified by a Markov chain, we see that the privacy mechanism erases all positions within a window from the sensitive position and releases the rest without erasure, and the size of the window is stochastically chosen. This observation suggests that, in contrast to the heuristic approach of deterministically choosing a window for erasure, our mechanism introduces additional uncertainty about sensitive data (in fact achieving perfect privacy) by randomizing the choice of the window. Later in Section VIII, we present a simulation experiment comparing our mechanism with the deterministic window-based erasure approach with respect to the privacy-utility trade-off, based on a more realistic data distribution defined by hidden Markov models.

V. Local Optimality

In the previous section, we proposed a privacy mechanism for the genotype-hiding problem satisfying both privacy and faithfulness conditions. Here, we provide further insights into the optimality of the proposed mechanism. We first prove that the mechanism is indeed an optimal solution to the local optimization problem in (15) as claimed, and thus can be viewed as a greedy solution to the general genotype-hiding problem in (5) given a fixed variable ordering (*i.e.,* the order in which Y_i's are sampled). We then present a negative result to inform future investigation, showing that finding an optimal variable ordering for the mechanism is intractable (NP-hard) in general, thus illustrating the limits of current techniques in achieving global optimality.

A. Optimality With Respect to the Local Optimization Problem

Let us first recall the local optimization problem (15), *i.e.,*

$$\begin{aligned} \underset{w\left(y_i|\mathbf{x}, y_{[i-1]}\right)}{\text{minimize}} \quad & p\left(y_i = *|y_{[i-1]}\right) \\ \text{subject to} \quad & I(X_{\mathcal{K}}; Y_i|Y_{[i-1]}) = 0 \\ & Y_i \in \{X_i, *\}. \end{aligned} \tag{30}$$

As we have shown,

$$I(X_{\mathcal{K}}; \mathbf{Y}) = \sum_{i=1}^{n} I(X_{\mathcal{K}}; Y_i|Y_{[i-1]}) = 0, \tag{31}$$

by the chain rule, so any solution satisfying the constraints of (15) for all i is a feasible solution to the general genotype-hiding problem in (5).

We now show that the privacy mechanism in (17) is optimal with respect to the above optimization problem. First, for any given $y_{[i-1]}$, note that

$$\begin{aligned} &p\left(y_i = *|y_{[i-1]}\right) \\ &\quad = 1 - \sum_{y_i \in \mathcal{X}} p\left(y_i|y_{[i-1]}\right) \\ &\quad \overset{(a)}{=} 1 - \sum_{y_i \in \mathcal{X}} \min_{x_{\mathcal{K}}} p\left(y_i|x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\quad \overset{(b)}{=} 1 - \sum_{y_i \in \mathcal{X}} \min_{x_{\mathcal{K}}} p\left(y_i = x_i|x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\quad = 1 - \sum_{x_i \in \mathcal{X}} \min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\qquad w\left(y_i = x_i|x_i, x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\quad \overset{(c)}{\geq} 1 - \sum_{x_i \in \mathcal{X}} \min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right), \end{aligned} \tag{32}$$

where (a) follows from the privacy condition, (b) follows from the faithfulness condition, and (c) holds because probability values are at most 1. This implies that any feasible solution to the local optimization problem (15) has to satisfy

$$p\left(y_i = *|y_{[i-1]}\right) \geq 1 - \sum_{x_i \in \mathcal{X}} \min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right), \tag{33}$$

and that it is optimal if the last step holds with equality, *i.e.,*

$$\begin{aligned} &\min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right) w\left(y_i = x_i|x_i, x_{\mathcal{K}}, y_{[i-1]}\right) \\ &\qquad = \min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right), \end{aligned} \tag{34}$$

for any x_i and $y_{[i-1]}$.

By plugging in the proposed mechanism in (17), we have

$$\begin{aligned} &\min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right) w\left(y_i = x_i|x_i, x_{\mathcal{K}}, y_{[i-1]}\right) \\ &= \min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right) \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right)}{p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right)} \\ &= \min_{x_{\mathcal{K}}} \min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right) \\ &= \min_{x_{\mathcal{K}}} p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right), \end{aligned} \tag{35}$$

where the last step follows because the two minimizations, both over the alphabet of $X_{\mathcal{K}}$, are equivalent and can be merged. This implies that the mechanism (17) attains the minimum probability of erasing Y_i and thus is an optimal solution to the local optimization problem (15). Therefore, our sequential privacy mechanism can be viewed as a locally-optimal algorithm for solving the general genotype-hiding problem (5), given a fixed variable ordering.

B. NP-Hardness of Finding an Optimal Variable Ordering

So far, we considered the privacy mechanism that generates a masked sequence $Y_1, \ldots, Y_n$ in a linear order from 1 to n. A natural question is then whether this linear ordering is optimal in terms of the erasure rate that the locally-optimal mechanism achieves. Here, we illustrate the difficulty of determining the optimal variable ordering for the mechanism from a complexity theory perspective, by proving that it is NP-hard in general. This suggests that devising an efficient mechanism with better optimality guarantees in the general setting requires additional assumptions or techniques to circumvent this impossibility result, which is an interesting direction for further research.

To formalize the problem, let $(o_1, \ldots, o_n)$ be any permutation of $(1, \ldots, n)$. We consider generating $\mathbf{Y}$ in the order of $o_1, \ldots, o_n$ instead. In this setting, the privacy mechanism (17) is defined by the conditional distribution

$$\begin{aligned} &w\left(y_{o_i}|x_{o_i}, x_{\mathcal{K}}, y_{o_{[i-1]}}\right) \\ &\quad = \begin{cases} \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_{o_i}|x_{\mathcal{K}} = u, y_{o_{[i-1]}}\right)}{p\left(x_{o_i}|x_{\mathcal{K}}, y_{o_{[i-1]}}\right)}, & \text{if } y_{o_i} = x_{o_i}, \\ 1 - \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_{o_i}|x_{\mathcal{K}} = u, y_{o_{[i-1]}}\right)}{p\left(x_{o_i}|x_{\mathcal{K}}, y_{o_{[i-1]}}\right)}, & \text{if } y_{o_i} = *, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{36}$$

for any x_{o_i}, $x_{\mathcal{K}}$ and $y_{o_{[i-1]}}$, where $o_{[i-1]} := \{o_1, \ldots, o_{i-1}\}$. It is easy to see that the faithfulness and privacy conditions are still satisfied regardless of the ordering.

In the following, we show that finding the best ordering $(o_1, \ldots, o_n)$ that minimizes the erasure rate of the mechanism is NP-hard by constructing a polynomial-time reduction of the well-known *hitting set* problem [34] to our problem. More specifically, given an arbitrary instance of a hitting set problem, we construct an instance of the genotype-hiding problem for which finding the optimal ordering for the privacy mechanism is equivalent to solving the original hitting set problem.

At the core of this reduction is a bipartite graph, illustrated in Figure 2, which we use to represent both an instance of the hitting set problem and to construct a corresponding instance of the genotype-hiding problem, as we explain in detail below. To clarify the dimensions of the problems upfront, note that we represent a hitting set problem for k sets over m elements using a bipartite graph with m left nodes and k right nodes, and the resulting genotype-hiding problem is over a sequence of length $n = m + k$ with k sensitive positions ($|\mathcal{K}| = k$) and a specially constructed $p(\mathbf{x})$.

We first review the hitting set problem. Consider a universe $U = \{v_1, \ldots, v_m\}$ and a collection of non-empty subsets $\mathcal{S} = \{S_1, \ldots, S_k\}$ such that $S_j \subseteq U$ for all $j \in [k]$. Without loss of generality, assume that $U = \bigcup_{j=1}^{k} S_j$, and $U = [m]$. A universe U and sets $\{S_1, \ldots, S_k\}$ can be represented by a bipartite graph, as depicted in Fig. 2. The goal of the hitting set problem is to find the minimum cardinality h^* of a set $V \subseteq U$ that satisfies $V \cap S_i \neq \emptyset$ for all i, that is

$$h^* = \min_{V \subseteq U : V \cap S_j \neq \emptyset, \forall j \in [k]} |V|. \tag{37}$$

Next, we construct the corresponding genotype-hiding problem from the given hitting set problem instance $(U, \mathcal{S})$. For any $i \in [m]$, $j \in [k]$ such that $i \in S_j$, let $b_{i,j}$ be a random variable which is independently and uniformly drawn from $\{0, 1\}$. In other words, each edge in the bipartite graph is associated with a random bit $b_{i,j}$ (see Fig. 2). Then, we define $\mathbf{X}$ to be a sequence of length $n = m + k$ as follows. Let X_i

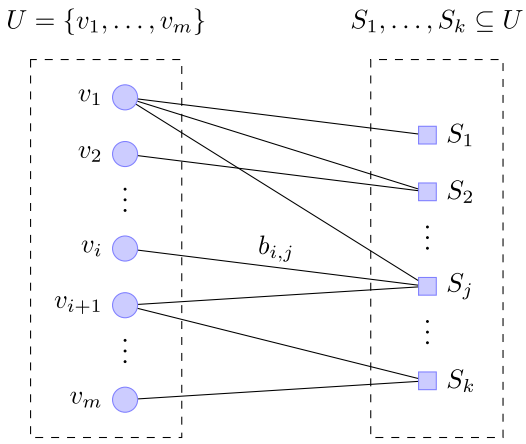

Fig. 2. A graphical illustration of the bipartite graph used in our NP-hardness proof, representing an instance of the hitting set problem. The universe U is represented by vertices on the left, sets are represented by vertices on the right, and the edges represent the inclusion of elements in each set. To facilitate reduction to the genotype-hiding problem, we associate each edge with an independent and uniformly random bit $b_{i,j}$.

for $i \in [m]$ be a tuple of random bits associated with edges connected to node v_i, *i.e.,*

$$X_i = (b_{i,j_1}, \ldots, b_{i,j_r}), \tag{38}$$

where $\{j_1, \ldots, j_r\} = \{j : i \in S_j\}$. Next, let X_{m+j} for $j \in [k]$ be

$$X_{m+j} = \bigoplus_{i \in S_j} b_{i,j}, \tag{39}$$

which can be viewed as a parity check bit over the edges connected to node S_j. In other words, the first m positions of the sequence are uniform and independently distributed symbols (a tuple of random bits), whereas the remaining k positions are parity check bits defined over the first m positions.

Note that the joint distribution $p(\mathbf{x}) = p(x_1, \ldots, x_{m+k})$ is succinctly characterized by the random bits $b_{i,j}$'s and the associated bipartite graph, and thus the description of the genotype-hiding problem can be generated in polynomial time with respect to m and k. In the following, we refer to the above data generating distribution as $p(\mathbf{x}; U, \mathcal{S})$, with respect to which the corresponding genotype-hiding problem is defined.

Theorem 3: Given a data generating distribution $p(\mathbf{x}; U, S)$ for a sequence of length $n = m+k$ and sensitive positions $\mathcal{K} = \{m+1, \ldots, m+k\}$, finding the best ordering $(o_1, \ldots, o_{m+k})$ that minimizes the erasure rate of our mechanism (36) is NP-hard.

We provide a sketch of the proof here and defer the details to Appendix C. First, we note the key property of $p(\mathbf{x}; U, S)$ that whether or not our mechanism erases the o_i-th position is deterministic given the variable ordering, as stated in the following lemma.

Lemma 2: Given a data generating distribution $p(\mathbf{x}; U, S)$ for a sequence of length $n = m + k$ and sensitive positions $\mathcal{K} = \{m+1, \ldots, m+k\}$, the conditional sampling distribution of our privacy mechanism satisfies

$$w\left(y_{o_i} = *|x_{o_i}, x_{\mathcal{K}}, y_{o_{[i-1]}}\right) \in \{0, 1\} \tag{40}$$

for all i, given any ordering $\pi = (o_1, \ldots, o_{m+k})$.

Proof: See Appendix B. □

As a result of Lemma 2, the overall erasure rate of the privacy mechanism can be calculated simply by counting the number of erased positions. Note that, if $o_i \in \mathcal{K}$, then

$$w\left(y_{o_i} = *|x_{o_i}, x_{\mathcal{K}}, y_{o_{[i-1]}}\right) = 1, \tag{41}$$

regardless of the ordering as we have previously shown. Thus, we need to compare only the erased indices in $[m] = [m+k]\backslash\mathcal{K}$ for finding the best ordering.

Let E_π be the set of erased indices in $[m]$ for a given ordering $\pi = (o_1, \ldots, o_n)$, *i.e.,*

$$E_\pi = \{i : y_i = *, i \in [m]\}, \tag{42}$$

where the distribution over $\mathbf{Y}$ is determined by the privacy mechanism. Then, finding the best ordering corresponds to finding π that leads to the minimum cardinality e^* of the corresponding E_π:

$$e^* = \min_\pi |E_\pi|. \tag{43}$$

Intuitively, whether a particular index $i \in [m]$ is included in E_π can be easily determined based on the bipartite graph representation of the underlying hitting set problem (see Fig. 2) as follows. The ordering $\pi = (o_1, \ldots, o_{m+k})$ specifies the order in which the m nodes on the left-hand side of the graph, each with a corresponding X_i, is visited by the mechanism (disregarding the sensitive indices $o_i \notin [m]$, which are always erased). As we show in the proof of Lemma 2, when we visit the node $o_i \in [m]$, X_{o_i} is erased if and only if there exists a node $j \in [k]$ on the right-hand side of the graph that is connected to o_i and only to other nodes (if any) that are previously visited *and* not erased. The presence of such a node j indicates that the sensitive variable X_{m+j} is directly revealed by X_{o_i} (since the rest of random bits contributing to X_{m+j} are already released in $\mathbf{Y}$ without erasure), while the absence of such j indicates the existence of other positions that are erased or have not been released that fully mask the correlation between X_{o_i} and the sensitive positions.

Finally, we complete the reduction by showing that solving (43) also produces a solution for the hitting set problem (37), *i.e.,* $e^* = h^*$. This is achieved by showing both that the set of erased indices E_π is in fact a valid hitting set ($e^* \geq h^*$), and that there exists an ordering π satisfying $|E_\pi| \leq |V|$ for any given hitting set V ($e^* \leq h^*$). A detailed proof is included in Appendix C.

Since the hitting set problem is equivalent to the set cover problem and is well-known to be NP-hard, our reduction proves that finding the best ordering π for our privacy mechanism given any $p(\mathbf{x})$ and $\mathcal{K}$ is also NP-hard. We note that this result does not preclude the possibility that for a restricted class of genotype-hiding problems (*e.g.,* with a structured $p(\mathbf{x})$ defined by HMMs), one could still find an efficient polynomial-time algorithm for determining the optimal variable ordering, which remains an interesting open question.

VI. Robustness

In this section, we discuss the robustness of our mechanism with respect to the underlying data distribution. In our

formulation of the privacy mechanism, the distribution (or the data generative model) $p(\mathbf{x})$, from which the input genome sequence originated, is assumed to be known. In practice, one can only empirically estimate this distribution based on existing data resources, *e.g.,* by obtaining maximum likelihood estimates of the model parameters based on a large collection of reference genomes in public data repositories. Consequently, the generative model used by the mechanism is bound to have deviations from the true generative process, both in terms of the limitations of the model as well as the noisy estimation of the parameters. These discrepancies can potentially lead to privacy leakage if the adversary has access to a more accurate distribution for the underlying input. Here, we study the potential privacy leakage under the worst-case scenario, where the adversary has access to the true underlying distribution. We bound the potential leakage as a function of the distance between the data distribution used by the mechanism and the true underlying distribution, suggesting that our mechanism is robust to small deviations in the noisy data distribution we expect to encounter in real-world use cases.

We denote the noisy data distribution used by the mechanism by $q(\mathbf{x})$ and the true distribution by $p(\mathbf{x})$. The privacy mechanism constructs the sampling distribution $w(\mathbf{y}|\mathbf{x})$ based on the available $q(\mathbf{x})$ such that the output $\mathbf{Y}$ is independent of sensitive genotypes $X_{\mathcal{K}}$ with respect to the joint distribution $q(\mathbf{x},\mathbf{y})$ induced by $q(\mathbf{x})$ and the mechanism $w(\mathbf{y}|\mathbf{x})$, *i.e.,*

$$\begin{aligned} q(x_{\mathcal{K}},\mathbf{y}) &= \sum_{x_{[n]\setminus\mathcal{K}}} q(\mathbf{x},\mathbf{y}) \\ &= \sum_{x_{[n]\setminus\mathcal{K}}} w(\mathbf{y}|\mathbf{x}) q(\mathbf{x}) = q(x_{\mathcal{K}})\, q(\mathbf{y}). \end{aligned} \tag{44}$$

Since $\mathbf{X}$ is actually generated from $p(\mathbf{x})$ not $q(\mathbf{x})$, we also define the true joint distribution $p(\mathbf{x},\mathbf{y})$ induced by $p(\mathbf{x})$ and the mechanism $w(\mathbf{y}|\mathbf{x})$; note that the mechanism is still based on $q(\mathbf{x})$.

Then, we can measure the unforeseen privacy leakage due to the mismatch in data distribution by the mutual information $I(p(x_{\mathcal{K}});p(\mathbf{y}))$ between the sensitive genotypes and the output sequence with respect to $p(\mathbf{x},\mathbf{y})$, as follows:

$$\begin{aligned} &I(p(x_{\mathcal{K}});p(\mathbf{y})) \\ &= \sum_{x_{\mathcal{K}},\mathbf{y}} p(x_{\mathcal{K}},\mathbf{y}) \log \frac{p(x_{\mathcal{K}},\mathbf{y})}{p(\mathbf{y})\, p(x_{\mathcal{K}})} \\ &= \sum_{x_{\mathcal{K}},\mathbf{y}} p(x_{\mathcal{K}},\mathbf{y}) \log \frac{p(x_{\mathcal{K}},\mathbf{y})\, q(x_{\mathcal{K}},\mathbf{y})}{p(\mathbf{y})\, p(x_{\mathcal{K}})\, q(x_{\mathcal{K}},\mathbf{y})} \\ &\overset{(a)}{=} \sum_{x_{\mathcal{K}},\mathbf{y}} p(x_{\mathcal{K}},\mathbf{y}) \log \frac{p(x_{\mathcal{K}},\mathbf{y})\, q(x_{\mathcal{K}})\, q(\mathbf{y})}{p(\mathbf{y})\, p(x_{\mathcal{K}})\, q(x_{\mathcal{K}},\mathbf{y})} \\ &= D(p(x_{\mathcal{K}},\mathbf{y})\,||\,q(x_{\mathcal{K}},\mathbf{y})) - D(p(x_{\mathcal{K}})\,||\,q(x_{\mathcal{K}})) \\ &\qquad - D(p(\mathbf{y})\,||\,q(\mathbf{y})), \end{aligned} \tag{45}$$

where $D(\cdot||\cdot)$ denotes relative entropy or equivalently Kullback-Leibler (KL) divergence, and (a) follows from (44). This leads to the following theorem.

Theorem 4: $I(p(x_{\mathcal{K}});p(\mathbf{y})) \leq D(p(\mathbf{x})\,||\,q(\mathbf{x}))$.

Proof: See Appendix D. □

This result implies that the amount of privacy leakage due to the potential mismatch between the data distribution used by the mechanism and the true underlying generative process gracefully scales with the extent to which the two distributions diverge.

VII. Privacy Mechanism for Hidden Markov Models

Thus far, we considered the data generative model $p(\mathbf{x})$ of the privacy mechanism to be an arbitrary distribution. Here, we address a particular form of $p(\mathbf{x})$ of great interest for our application setting in genomics, namely the Li and Stephens model [27], which is based on a hidden Markov model. This model is widely adopted in genetics for a wide range of tasks that require a probabilistic model of the genome [35]. For this class of $p(\mathbf{x})$, we propose an efficient algorithm to implement the privacy mechanism introduced in Section IV.

A. Review of Hidden Markov Models for Genomes

The classical hidden Markov model (HMM) describing the distribution of personal genomes [27] is as follows. First, let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ represent an individual's (haplotype) genetic sequence of length n. Following standard practice in genetics, we adopt a binary alphabet $\mathcal{X} = \{0,1\}$ for each element X_i, representing whether the observed nucleotide is identical to the one in the reference human genome (called *reference allele*) or not (*alternative allele*). In addition, we are given a reference dataset of m personal genome sequences $\mathcal{H} = \{\mathbf{h}_j : j = 1,\ldots,m\}$, where each sequence $\mathbf{h}_j$ is of length n. The i-th coordinate of $\mathbf{h}_j$ is denoted by $h_{i,j}$, which also takes a value in $\mathcal{X}$.

In this model, $\mathbf{X}$ is viewed as a "mosaic" of reference sequences in $\mathcal{H}$ with potential substitution errors arising from mutations or experimental noise in sequencing. Formally, $\mathbf{X}$ depends on a sequence of hidden states $\{S_i\}_{i=1}^n$ forming a Markov chain, where each S_i takes an integer in the range $\{1,\ldots,m\}$, representing an index into $\mathcal{H}$. Without loss of generality, we assume that the initial state S_1 is uniformly distributed over $\{1,\ldots,m\}$. The transition probability $\pi_{i,j}$ from state i to j is set to $\frac{\epsilon}{m-1}$ and $1-\epsilon$ for $i \neq j$ and $i = j$, respectively. The parameter ϵ is often called the recombination probability; in the following we also use the term *crossover probability* to refer to this quantity.

Next, each X_i is sampled based on the hidden state S_i by copying the corresponding symbol in the selected reference sequence with a small probability of error. In other words, X_i is equal to the symbol in the i-th position of $\mathbf{h}_{S_i}$ with *error probability* θ. The overall data distribution $p(\mathbf{x})$ is fully specified by the tuple $(\mathcal{H},\epsilon,\theta)$. We provide a graphical illustration of $p(\mathbf{x})$ in Fig. 3. In our work, we treat the parameters of the above model as given. In practice, these parameters are estimated from a large collection of reference genomes, *e.g.,* including hundreds of thousands of individuals, which are available in public data repositories such as the UK Biobank [36].

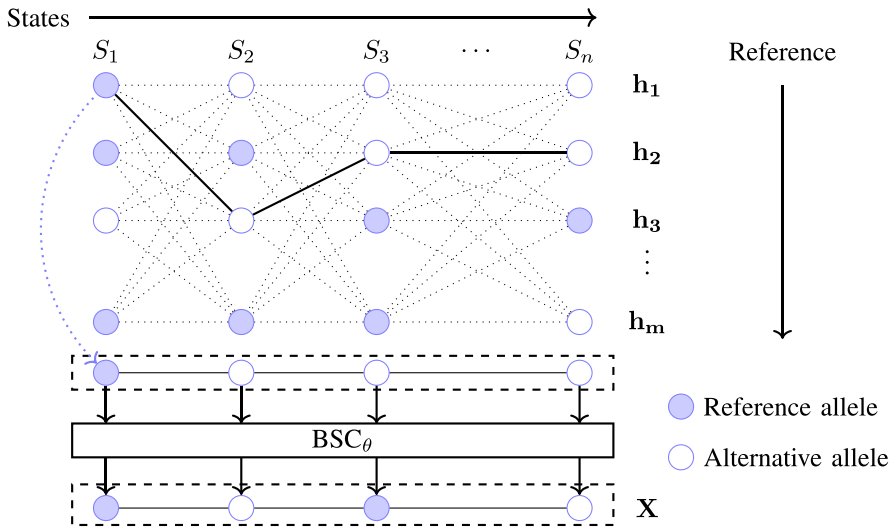

Fig. 3. A graphical illustration of HMM for genomes. The state space of the hidden states is $\{1, \ldots, m\}$, where each element corresponds to an index into the reference dataset $\{\mathbf{h}_1, \ldots, \mathbf{h}_m\}$ (each of length n). A Markov process $\{S_i\}_{i=1}^n$ indicates which reference sequence the user reads the data from at the i-th position. For each i, X_i differs from the i-th position of $\mathbf{h}_{S_i}$ with probability θ, representing noise in the data. BSC_θ: Binary symmetric channel with crossover probability θ.

B. An Efficient Algorithm for HMMs

In this section, we propose an efficient algorithm to implement the privacy mechanism introduced in Section IV for $p(\mathbf{x})$ based on a hidden Markov model $(\mathcal{H}, \epsilon, \theta)$ described in the previous section. The outline of our algorithm is provided in Algorithm 1.

As seen in (17), the privacy mechanism determines the probability of erasing x_i mainly based on the probability $p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right)$. By employing a belief propagation approach akin to the well-known forward-backward algorithm [37], we track the computation of $p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right)$ for all $u \in \mathcal{X}^{|\mathcal{K}|}$ efficiently. The novelty of our algorithm is that it incorporates the stochasticity of the privacy mechanism in addition to that of the HMM.

First, note that it is sufficient to describe how to compute $p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right)$ for all $u \in \mathcal{X}^{|\mathcal{K}|}$ and $i \in [n]$, which fully determines the distribution of $y_1, \ldots, y_n$ specified by our privacy mechanism, *i.e.,*

$$p\left(y_i|x_i, x_{\mathcal{K}}, y_{[i-1]}\right) = \begin{cases} \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}}=u, y_{[i-1]}\right)}{p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right)}, & \text{if } y_i = x_i, \\ 1 - \frac{\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i|x_{\mathcal{K}}=u, y_{[i-1]}\right)}{p\left(x_i|x_{\mathcal{K}}, y_{[i-1]}\right)}, & \text{if } y_i = *, \\ 0, & \text{otherwise.} \end{cases} \tag{46}$$

We begin by expressing $p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right)$ as

$$\begin{aligned} &p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right) \\ &= \sum_{s_i} p\left(s_i|x_{\mathcal{K}} = u, y_{[i-1]}\right) p\left(x_i|s_i, x_{\mathcal{K}} = u, y_{[i-1]}\right) \\ &= \sum_{s_i, s_{i-1}} p\left(s_{i-1}|x_{\mathcal{K}}=u, y_{[i-1]}\right) p\left(s_i|s_{i-1}, x_{\mathcal{K}}=u\right) p\left(x_i|s_i\right). \end{aligned} \tag{47}$$

Note that

$$\begin{aligned} &p\left(s_i|s_{i-1}, x_{\mathcal{K}} = u\right) \\ &\quad = \frac{p\left(s_i, s_{i-1}, x_{\mathcal{K}} = u\right)}{p\left(s_{i-1}, x_{\mathcal{K}} = u\right)} \\ &\quad = \frac{p\left(s_{i-1}|x_{\mathcal{K}_{i-}} = u_-\right) p\left(s_i|s_{i-1}\right) p\left(x_{\mathcal{K}_{i+}} = u_+|s_i\right)}{p\left(s_{i-1}|x_{\mathcal{K}_{i-}} = u_-\right) p\left(x_{\mathcal{K}_{i+}} = u_+|s_{i-1}\right)} \\ &\quad = \frac{p\left(s_i|s_{i-1}\right) p\left(x_{\mathcal{K}_{i+}} = u_+|s_i\right)}{p\left(x_{\mathcal{K}_{i+}} = u_+|s_{i-1}\right)} \\ &\quad = \frac{p\left(s_i|s_{i-1}\right) p\left(x_{\mathcal{K}_{i+}} = u_+|s_i\right)}{\sum_{s_i} p\left(s_i|s_{i-1}\right) p\left(x_{\mathcal{K}_{i+}} = u_+|s_i\right)}, \end{aligned} \tag{48}$$

where $\mathcal{K}_{i-} := \mathcal{K} \cap \{1, \ldots, i-1\}$, $\mathcal{K}_{i+} := \mathcal{K} \cap \{i, \ldots, n\}$, u_- and u_+ are corresponding values of $x_{\mathcal{K}_{i-}}$ and $x_{\mathcal{K}_{i+}}$ specified by u.

As $p\left(x_i|s_i\right)$ and $p\left(s_i|s_{i-1}\right)$ are directly given by the HMM, we need only to consider how to compute the two terms $p\left(s_{i-1}|x_{\mathcal{K}} = u, y_{[i-1]}\right)$ and $p\left(x_{\mathcal{K}_{i+}} = u_+|s_i\right)$. To simplify our notation, we introduce the following variables to represent these terms:

$$\begin{aligned} \psi^{(i)}(u, s_i) &:= p\left(s_i|x_{\mathcal{K}} = u, y_1, \ldots, y_i\right), \\ \gamma^{(i)}(u, s_i) &:= p\left(x_{\mathcal{K}_{i+}} = u_+|s_i\right). \end{aligned}$$

With $\psi^{(i)}(u, s_i)$ and $\gamma^{(i)}(u, s_i)$ for a given position i, we can calculate (47) as

$$\begin{aligned} &p\left(x_i|x_{\mathcal{K}} = u, y_{[i-1]}\right) \\ &= \sum_{s_{i-1}} \frac{\sum_{s_i} \psi^{(i-1)}(u, s_{i-1}) p\left(s_i|s_{i-1}\right) \gamma^{(i)}(u, s_i) p\left(x_i|s_i\right)}{\sum_{s_i} p\left(s_i|s_{i-1}\right) \gamma^{(i)}(u, s_i)}. \end{aligned} \tag{49}$$

First, note that $\gamma^{(i)}(u, s_i)$ can be recursively computed in the same manner as calculating the backward probabilities in the forward-backward algorithm, as described below:

1) Initialization: We initialize $\gamma^{(n)}(u, s_n)$ by

$$\gamma^{(n)}(u, s_n) = \begin{cases} p\left(x_n = u_n | s_n\right), & n \in \mathcal{K}, \\ 1, & n \notin \mathcal{K}. \end{cases} \tag{50}$$

2) Iterations: For $i = n-1, \ldots, 1$, we compute $\gamma^{(i)}(u, s_i)$ as

$$\gamma^{(i)}(u, s_i) = \begin{cases} \sum_{s_{i+1}} p\left(x_i = u_i | s_i\right) p\left(s_{i+1} | s_i\right) \gamma^{(i+1)}(u, s_{i+1}), & i \in \mathcal{K} \\ \sum_{s_{i+1}} p\left(s_{i+1} | s_i\right) \gamma^{(i+1)}(u, s_{i+1}), & i \notin \mathcal{K}. \end{cases} \tag{51}$$

Next, to efficiently compute $\psi^{(i)}(u, s_i)$ for $i \in [n]$, we analogously adopt the following iterative steps.

3) Initialization: $\psi^{(1)}(u, s_1)$ is initialized by

$$\psi^{(1)}(u, s_1) \propto p\left(s_1 | x_{\mathcal{K}} = u\right) p\left(y_1 | s_1, x_{\mathcal{K}} = u\right), \tag{52}$$

where $p\left(s_1 | x_{\mathcal{K}} = u\right)$ can be calculated by (48) given $\gamma^{(1)}(u, s_1)$, and $p\left(y_1 | s_1, x_{\mathcal{K}} = u\right)$ is given by our mechanism as shown in (46).

4) Iterations: Using Bayes' rule, we can express $\psi^{(i)}(u, s_i)$ as

$$\begin{aligned} \psi^{(i)}(u, s_i) &= p\left(s_i | x_{\mathcal{K}} = u, y_{[i]}\right) \\ &\propto p\left(s_i | x_{\mathcal{K}} = u, y_{[i-1]}\right) p\left(y_i | s_i, x_{\mathcal{K}} = u, y_{[i-1]}\right), \end{aligned} \tag{53}$$

where

$$\begin{aligned} &p\left(s_i | x_{\mathcal{K}} = u, y_{[i-1]}\right) \\ &\quad = \sum_{s_{i-1}} \psi^{(i-1)}(u, s_{i-1}) p\left(s_i | s_{i-1}, x_{\mathcal{K}} = u\right), \end{aligned} \tag{54}$$

and

$$\begin{aligned} &p\left(y_i | s_i, x_{\mathcal{K}} = u, y_{[i-1]}\right) \\ &\quad = \sum_{x_i} p\left(x_i | s_i\right) p\left(y_i | x_i, x_{\mathcal{K}} = u, y_{[i-1]}\right). \end{aligned} \tag{55}$$

Therefore, $\psi^{(i)}(u, s_i)$ can be computed based on $\psi^{(i-1)}(u, s_{i-1})$. We note that the probability $p\left(s_i | s_{i-1}, x_{\mathcal{K}} = u\right)$ can be calculated using $\gamma^{(i)}(u, s_i)$ as shown in (48), and $p\left(y_i | x_i, x_{\mathcal{K}} = u, y_{[i-1]}\right)$ is given by our mechanism as shown in (46). Using this recurrence relation, $\psi^{(i)}(u, s_i)$ for all $i \in [n]$ can be computed.

Analogous to the forward-backward algorithm, our algorithm has polynomial computational complexity of $\mathcal{O}(nm^2)$ for a fixed u, with respect to the sequence length n and the number of reference sequences m, for a given u. Clearly, $\min_{u \in \mathcal{X}^{|\mathcal{K}|}} p\left(x_i | x_{\mathcal{K}} = u, y_{[i-1]}\right)$ can be easily obtained once $p\left(x_i | x_{\mathcal{K}} = u, y_{[i-1]}\right)$ for all u have been computed. This overhead involves a factor of $2^{|\mathcal{K}|}$ in the computational complexity, but we expect $|\mathcal{K}|$ to be a small constant in practice (*e.g.,* less than 10); since genotype correlation is predominantly local, the user may apply our mechanism to local regions of the genome of a permissive length, each of which including only a few sensitive positions.

Algorithm 1 Mechanism for Hiding Sensitive Genotypes in $\mathbf{X}$

Input: Genome sequence $\mathbf{X} = (X_1, \ldots, X_n)$ from an HMM with parameters $(\mathcal{H}, \epsilon, \theta)$, and indices of sensitive positions $\mathcal{K} \subset [n]$

Output: Masked genome sequence $\mathbf{Y} = (Y_1, \ldots, Y_n)$, such that $I(X_{\mathcal{K}}; \mathbf{Y}) = 0$ and $Y_i \in \{X_i, *\}$ for all $i \in [n]$

```
Initialize γ^(n)(u, s_n) according to (50)
for i = n − 1, ..., 1 do
    for u ∈ X^|K| do
        Compute γ^(i)(u, s_i) according to (51)
        Compute p(s_i | s_{i−1}, x_K = u) according to (48)
    end for
end for
Initialize ψ^(1)(u, s_1) according to (52)
for i = 2, ..., n do
    Calculate the erasure probability for Y_i using (46)
    Generate Y_i ∈ {X_i, *} according to the erasure probability
    for u ∈ X^|K| do
        Compute ψ^(i)(u, s_i) according to (53)
    end for
end for
```

VIII. Simulations

In this section, we provide insights into the empirical performance of our privacy mechanism for hidden Markov models (HMMs) on simulated datasets. We randomly generated 100 haplotype sequences of length 100, which together with the choices of error probability θ and crossover probability ϵ induce $p(\mathbf{x})$, as described in Section VII-A. For simplicity, we suppose the sensitive position $\mathcal{K} = \{1\}$.

We first illustrate the privacy-utility trade-off of the heuristic window-based erasure approach described in the Introduction. In particular, this approach erases the first ω positions of the sequence to hide information about the sensitive position (the first position). The results are shown in Figure 4. The erasure rate is defined by the size of the erased window over the sequence length, *i.e.,* ω/n (note $n = 100$). The privacy leakage is measured by the mutual information between the released positions and the sensitive position X_1, normalized by the entropy of X_1, *i.e.,* $I(X_1; X_{[n]\setminus[\omega]})/H(X_1)$. We also show the expected erasure rate of our proposed privacy mechanism for comparison, whose privacy leakage is strictly zero by design. We observe that the window-erasure approach requires a high erasure rate (around 0.3) to keep the privacy leakage close to zero, whereas our mechanism achieves a considerably smaller erasure rate (around 0.12) while providing perfect privacy. On the other hand, choosing a window size for the baseline approach to match the erasure rate of our mechanism leads to a considerable privacy leakage.

We next evaluate our privacy mechanism over a range of different parameter settings. We consider $\theta \in \{0.01, 0.05\}$ and vary ϵ from 0.01 to 0.5, both of which reflect reasonable ranges of the parameters for the scale of the dataset we simulated. We provide each instance of $p(\mathbf{x})$ to our privacy mechanism with $\mathcal{K} = \{1\}$ to calculate its achievable rate R (*i.e.,* one minus

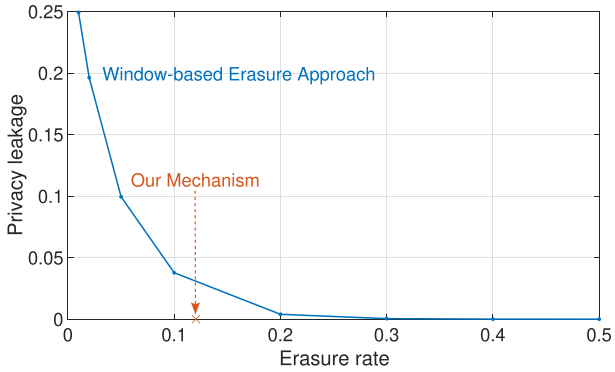

Fig. 4. Privacy-utility trade-off of the window-based erasure approach on simulated HMM data with $m = 100$, $n = 100$, $\mathcal{K} = \{1\}$, crossover probability $\epsilon = 0.1$ and error probability $\theta = 0.01$. Erasure rate denotes the size of window that is erased normalized by the sequence length n. Privacy leakage denotes the mutual information between the released data and the sensitive symbol normalized by the entropy of the sensitive symbol.

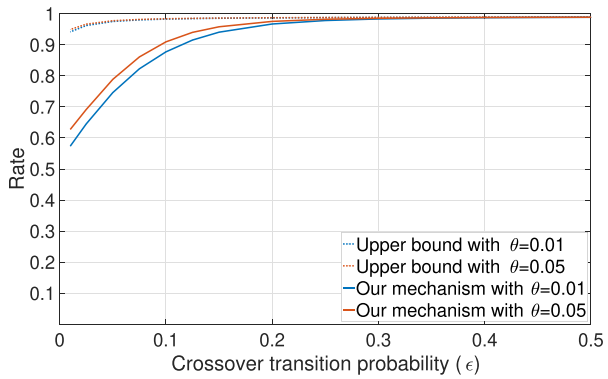

Fig. 5. Comparison of our mechanism and the upper bound on simulated HMM data with $m = 100$, $n = 100$, $\mathcal{K} = \{1\}$ and different choices of crossover probability ϵ and error probability θ.

the expected erasure rate). Figure 5 shows the comparison between the rate of our mechanism and the upper bound we derived in Section III. The results suggest that the performance of our mechanism shows varying degrees of closeness to the theoretical upper bound depending on the characteristics of the underlying data distribution. In particular, for higher values of ϵ, representing the regime where the hidden Markov model mixes faster and thus the correlation with the sensitive position decays more quickly, the rate of our mechanism is nearly identical to the upper bound. On the other hand, for lower values of ϵ, which lead to stronger correlations in the sequence, we observed that the gap between our mechanism and the upper bound can grow considerable large. Note that this does not necessarily imply that our mechanism achieves a significantly suboptimal performance, given that the upper bound we considered is not tight in general. We also note that the rate of our mechanism is generally higher when the error probability is larger ($\theta = 0.05$ vs 0.01), which agrees with the intuition that higher levels of noise in the data distribution lower the requirement for hiding sensitive information, thus leading to lower erasure probabilities and higher rates as a result.

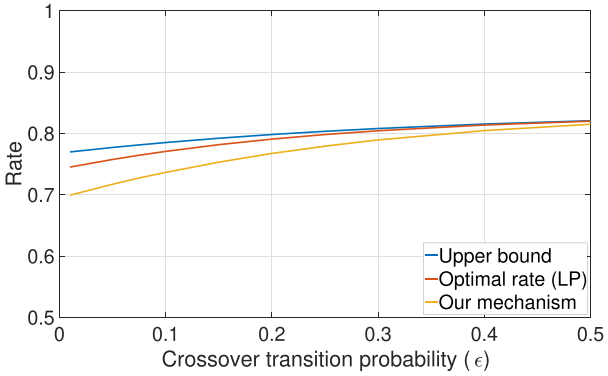

Fig. 6. Comparison of our mechanism, the upper bound and the optimal rate based on a linear programming (LP) solution on simulated HMM data with $m = 100$, $n = 6$, $\mathcal{K} = \{1\}$ based on a truncated version of the dataset used in Fig. 5.

To gain further insights into the noticeable gap between the upper bound and our mechanism in the small ϵ regime, we additionally implemented a linear programming (LP) approach for directly obtaining the optimal mechanism $w(\mathbf{y}|\mathbf{x})$. However, since the size of LP grows exponentially with the length of the sequence n, we could only evaluate this approach for small problem instances due to numerical instability. We took the same simulated data as before and truncated each reference haplotype down to the first six positions to obtain a tractable LP instance for this experiment ($n = 6$).

The rate comparisons of our privacy mechanism, LP-based optimal mechanism, and the upper bound in this setting are shown in Fig. 6. As expected, we observed that the optimal rate lies between the upper bound and the rate of our mechanism, demonstrating that the gap between the optimal rate and the rate of our mechanism is indeed smaller than the ostensible gap suggested by the upper bound.

Taken together, these results suggest that, although the performance of our mechanism is often quite close to the upper bound, the difference between the maximum achievable rate and the rate of our mechanism can vary based on the properties of the data distribution. We note that it is yet unknown whether there exists a privacy mechanism that can be as efficiently constructed as our mechanism while achieving performance that is closer to the optimal rate. Closing this performance gap both by devising enhanced privacy mechanisms that achieve higher rates and by developing tighter upper bounds are important directions for future work.

IX. Conclusion and Future Work

In this paper, we introduced the genotype hiding problem and proposed an information-theoretic privacy mechanism as a solution. We analyzed the theoretical properties of the mechanism, and proposed an efficient algorithmic implementation of the mechanism for hidden Markov models, a main model of interest for our application in genomics.

It is worth noting that our mechanism does not rule out the possibilities of genotype reconstruction attacks that leverage

(i) alternative genetic sequence models and imputation strategies or (ii) a larger set of reference dataset using which HMM parameters could be more accurately estimated. However, our model based on HMMs is consistent with the state-of-the-art techniques for genotype imputation, which is a relatively mature field. In addition, given the high cost of amassing large-scale genomic data, it would be a significant challenge for an attacker to gain access to a larger dataset than those in the public realm. As such, our mechanism could be thought of as providing privacy protection according to the best knowledge of the field. Our results in Section VI show that any unforeseen privacy leakage arising from the discrepancies in the data distribution scales gracefully with the relative entropy between the true distribution and the one used by the mechanism.

There are several key directions for future work. Our work focused on hiding the content of the sensitive positions, yet a potential concern remains regarding information revealed by the choice of sensitive positions $\mathcal{K}$. Any approach relying on erasures for privacy protection may inevitably leak information about $\mathcal{K}$, since preventing such leakage would generally require erasures to be consistently applied throughout the sequence, which is highly costly in terms of utility if only a small fraction is considered sensitive. An interesting extension of our work is then to relax the faithfulness condition when hiding the positions is deemed important. A promising approach is to re-sample the erased positions from the data distribution as a post-processing step to the mechanism presented in this paper. That said, we note that in our application setting, $\mathcal{K}$ is neither necessarily or nor solely decided by the sequence, as it may be determined based on family history of diseases or curated disease associations in public repositories. Thus, we believe the mechanism presented in this work is directly applicable in many practical scenarios.

Next, although we focused on achieving perfect privacy (with respect to the given data distribution), it may be useful in practice to consider a relaxed notion such as local differential privacy [38]. This may give the user the ability to determine a more desirable trade-off between the level of privacy and the amount of data to be erased. From an analytical standpoint, this direction would also lead to useful insights about the achievable points along the privacy-utility trade-off curve defined by the genotype-hiding problem with a relaxed notion of privacy, to complement the results in this work.

Furthermore, it would be interesting to explore the generalization of our efficient implementation strategies to a broader class of data generative models beyond HMMs, which may allow similar mechanisms to be employed to protect sensitive data in other domains.

Lastly, we plan to study the performance of our privacy mechanism on real genetic datasets and release the software implementation of our mechanism for the genetics community in the near future.

Growing threats to genetic privacy are necessitating principled strategies for protecting the privacy of individuals while maintaining the utility of data sharing. Our work illustrates how such a strategy could be designed from an information-theoretic perspective to enable selective disclosure of personal genomic data. Our methodology is broadly applicable to other data sharing scenarios involving sensitive data with complex correlation structure. We hope that our work will help spur the development of a wide range of information-theoretic tools for modelling and preserving private genomic information.

APPENDIX A
PROOF OF COROLLARY 1

We prove the sufficient condition of the optimality holds for the Markov chain case. We give an inductive proof for the sufficient condition by showing that, for a given x_i,

$$u^* \in \arg\min_u p\left(x_i|x_{\mathcal{K}} = u, y_{[j-1]}\right) \tag{56}$$

implies

$$u^* \in \arg\min_u p\left(x_i|x_{\mathcal{K}} = u, y_{[j]}\right) \tag{57}$$

for $j = 1, \ldots, i-1$. For each j, we consider the following two cases ($y_j \neq *$ and $y_j = *$):

(1) If $y_j \neq *$, then we have

$$\begin{aligned} p\left(x_i|x_{\mathcal{K}}, y_{[j]}\right) &= \sum_{x_j} p\left(x_j|x_{\mathcal{K}}, y_{[j]}\right) p\left(x_i|x_{\mathcal{K}}, y_{[j]}, x_j\right) \\ &\overset{(a)}{=} \mathbb{1}\{x_j = y_j\} p\left(x_i|x_{\mathcal{K}}, y_{[j]}, x_j\right) \\ &\overset{(b)}{=} \mathbb{1}\{x_j = y_j\} p\left(x_i|x_j\right), \end{aligned} \tag{58}$$

where (a) follows because Y_j can either be X_j or $*$, and (b) follows from Markovity. In this case, $\arg\min_u p\left(x_i|x_{\mathcal{K}} = u, y_{[j]}\right)$ is indeed independent of u, which means

$$\arg\min_u p\left(x_i|x_{\mathcal{K}} = u, y_{[j]}\right) = |\mathcal{X}|, \tag{59}$$

so the statement is trivially true.

(2) If $y_j = *$, then we have

$$\begin{aligned} &p\left(x_i|x_{\mathcal{K}}, y_{[j]}\right) \\ &= \sum_{x_j} p\left(x_j|x_{\mathcal{K}}, y_{[j]}\right) p\left(x_i|x_{\mathcal{K}}, y_{[j]}, x_j\right) \\ &\overset{(a)}{=} \sum_{x_j} p\left(x_j|x_{\mathcal{K}}, y_{[j]}\right) p\left(x_i|x_j\right) \\ &\overset{(b)}{\propto} \sum_{x_j} \left\{ p\left(x_j|x_{\mathcal{K}}, y_{[j-1]}\right) - \min_u p\left(x_j|x_{\mathcal{K}} = u, y_{[j-1]}\right) \right\} \\ &\qquad p\left(x_i|x_j\right) \\ &= \sum_{x_j} p\left(x_i|x_j\right) p\left(x_j|x_{\mathcal{K}}, y_{[j-1]}\right) \\ &\qquad - \sum_{x_j} p\left(x_i|x_j\right) \min_u p\left(x_j|x_{\mathcal{K}} = u, y_{[j-1]}\right) \\ &= p\left(x_i|x_{\mathcal{K}}, y_{[j-1]}\right) \\ &\qquad - \sum_{x_j} p\left(x_i|x_j\right) \min_u p\left(x_j|x_{\mathcal{K}} = u, y_{[j-1]}\right), \end{aligned} \tag{60}$$

where (a) follows from Markovity, and (b) follows from Bayes's rule and our privacy mechanism (17). Since the second term of the right-hand side in (60) is independent of $x_{\mathcal{K}}$, we obtain

$$\arg\min_{u} p\left(x_i|x_{\mathcal{K}}=u, y_{[j]}\right) = \arg\min_{u} p\left(x_i|x_{\mathcal{K}}=u, y_{[j-1]}\right). \tag{61}$$

For both cases, we have verified that the sufficient condition holds, which completes the proof.

Appendix B
Proof of Lemma 2

We will prove (40) by induction. First, consider the base case:

$$w\left(y_{o_1}=*|x_{o_1}, x_{\mathcal{K}}\right) = 1 - \frac{\min_{x_{\mathcal{K}}} p\left(x_{o_1}|x_{\mathcal{K}}\right)}{p\left(x_{o_1}|x_{\mathcal{K}}\right)}. \tag{62}$$

From the previous discussion, we know that if $o_i \in \mathcal{K}$, then

$$w\left(y_{o_i}=*|x_{o_i}, x_{\mathcal{K}}, y_{o_{[i-1]}}\right) = 1, \tag{63}$$

so without loss of generality, we assume that

$$o_1 \notin \mathcal{K} = \{m+1, \ldots, m+k\}. \tag{64}$$

Since

$$x_{\mathcal{K}} = \left\{ \sum_{i:i\in S_j} b_{i,j} : j \in [k] \right\}, \tag{65}$$

and

$$x_{o_1} = \{b_{o_1,j} : o_1 \in S_j\} \tag{66}$$

by definition, we can see that if there exists some j such that $S_j = \{o_1\}$, then $b_{o_1,j} \in x_{\mathcal{K}}$ and $b_{o_1,j} \in x_{o_1}$. In this case, we can always find some assignments such that

$$\min_{x_{\mathcal{K}}} p\left(x_{o_1}|x_{\mathcal{K}}\right) = 0, \tag{67}$$

implying that

$$w\left(y_{o_1}=*|x_{o_1}, x_{\mathcal{K}}\right) = 1. \tag{68}$$

If there is no j such that $S_j = \{o_i\}$, each $\sum_{i:i\in S_j} b_{i,j}$ constituting $x_{\mathcal{K}}$ is a binary summation of some $b_{o_1,j}$ and (independent) random bits $b_{i,j}$ such that $i \neq o_1$, where the latter render the result uniformly random. This means that $X_{\mathcal{K}}$ is independent of X_{o_1}, and thus we have

$$w\left(y_{o_1}=*|x_{o_1}, x_{\mathcal{K}}\right) = 1 - \frac{\min_{x_{\mathcal{K}}} p\left(x_{o_1}|x_{\mathcal{K}}\right)}{p\left(x_{o_1}|x_{\mathcal{K}}\right)} = 1 - \frac{\min_{x_{\mathcal{K}}} p\left(x_{o_1}\right)}{p\left(x_{o_1}\right)} = 0, \tag{69}$$

for all x_{o_1} and $x_{\mathcal{K}}$.

Assume the statement is true for $o_1, \ldots, o_{i-1}$. Then for o_i, note that

$$p\left(x_{o_i}|x_{\mathcal{K}}, y_{o_{[i-1]}}\right) = \frac{p\left(x_{o_i}|x_{\mathcal{K}}\right) p\left(y_{o_{[i-1]}}|x_{o_i}, x_{\mathcal{K}}\right)}{p\left(y_{o_{[i-1]}}|x_{\mathcal{K}}\right)}. \tag{70}$$

By letting

$$\tilde{\mathcal{E}}_i = \left\{o_j : y_{o_j} \neq *, j \leq i-1\right\}, \tag{71}$$

(70) can be written as

$$p\left(x_{o_i}|x_{\mathcal{K}}, y_{o_{[i-1]}}\right) = \frac{p\left(x_{o_i}|x_{\mathcal{K}}\right) p\left(x_{\tilde{\mathcal{E}}_i}|x_{o_i}, x_{\mathcal{K}}\right)}{p\left(x_{\tilde{\mathcal{E}}_i}|x_{\mathcal{K}}\right)} = p\left(x_{o_i}|x_{\tilde{\mathcal{E}}_i}, x_{\mathcal{K}}\right), \tag{72}$$

because of the inductive assumption that the decisions whether to erase $y_{o_1}, \ldots, y_{o_{i-1}}$ are deterministic.

Hence, we have

$$w\left(y_{o_i}=*|x_{o_i}, x_{\mathcal{K}}, y_{o_{[i-1]}}\right) = 1 - \frac{\min_{x_{\mathcal{K}}} p\left(x_{o_i}|x_{\mathcal{K}}, y_{o_{[i-1]}}\right)}{p\left(x_{o_i}|x_{\mathcal{K}}, y_{o_{[i-1]}}\right)} = 1 - \frac{\min_{x_{\mathcal{K}}} p\left(x_{o_i}|x_{\tilde{\mathcal{E}}_i}, x_{\mathcal{K}}\right)}{p\left(x_{o_i}|x_{\tilde{\mathcal{E}}_i}, x_{\mathcal{K}}\right)}. \tag{73}$$

Analogous to our argument for the base case, if there exists some j such that $S_j \subseteq \tilde{\mathcal{E}}_i \cup \{o_i\}$, then one can determine $b_{o_i,j} \in x_{o_1}$ from $x_{\tilde{\mathcal{E}}_i}, x_{\mathcal{K}}$, and thus

$$\min_{x_{\mathcal{K}}} p\left(x_{o_i}|x_{\tilde{\mathcal{E}}_i}, x_{\mathcal{K}}\right) = 0, \tag{74}$$

implying that

$$w\left(y_{o_i}=*|x_{o_i}, x_{\mathcal{K}}, y_{o_{[i-1]}}\right) = 1. \tag{75}$$

If there is no such j, each x_j for $j \in \mathcal{K}$ is the binary summation of some $b_{o_i,j} \in x_{o_i}$ and some independent random bits $b_{i',j}$ such that $i' \neq o_i$, which again guarantees that $X_{\mathcal{K}}$ is independent of X_{o_i} conditioning on $X_{\tilde{\mathcal{E}}_i}$. Thus, we have

$$w\left(y_{o_i}=*|x_{o_i}, x_{\mathcal{K}}, y_{o_{[i-1]}}\right) = 1 - \frac{\min_{x_{\mathcal{K}}} p\left(x_{o_i}|x_{\tilde{\mathcal{E}}_i}\right)}{p\left(x_{o_i}|x_{\tilde{\mathcal{E}}_i}\right)} = 0, \tag{76}$$

for all x_{o_i}, $x_{\mathcal{K}}$ and $y_{o_{[i-1]}}$, which completes the inductive proof.

Appendix C
Proof of Theorem 3

First, let us show that $e^* \geq h^*$ by showing that E_π is a hitting set for any order π, *i.e.,* $E_\pi \cap S_j \neq \emptyset$ for all $j \in [k]$. We prove it by contradiction. Suppose that there exists some S_j such that $E_\pi \cap S_j = \emptyset$, which implies that $S_j \subseteq [m]\backslash E_\pi$ for some j. Assume that $S_j = \{i_1, \ldots, i_t\}$, and i_t is the last index visited that specified by the given order π. Then, when we run our mechanism for i_t, since $i_1, \ldots, i_{t-1}$ are all visited and not erased, by recalling the proof of Lemma 2, we know that $\tilde{\mathcal{E}}_{i_t} \supseteq \{i_1, \ldots, i_{t-1}\}$, so we have $S_j \subseteq \tilde{\mathcal{E}}_{i_t} \cup \{i_t\}$. It means that y_{i_t} is erased or $i_t \in E_\pi$, which contradicts with our assumption $E_\pi \cap S_j = \emptyset$.

Next, we show that $e^* \leq h^*$ by showing that for any given hitting set V, there exists an order π such that $|E_\pi| \leq |V|$. Suppose V is a hitting set and $|V| = h$, *i.e.,* $V \cap S_j \neq \emptyset$ for all $j \in [k]$. Consider an order π such that $o_i \notin V \cup [m+1 : m+k]$ for $i \leq m-h$ and $o_i \in V$ for $i \in [m-h+1 : m]$, *i.e.,* visiting indices in the complementary of T before attaining V. When we visit o_i such that $i \leq m-h$ (or $o_i \in [m]\backslash V$), by the assumption that $V \cap S_j \neq \emptyset$ for all j, we know that there exists some index $t_j \in S_j \cap V$ for each j. By recalling

the definition (71), we know that $\tilde{\mathcal{E}}_i \supseteq [m]\backslash V$, so $t_j \notin \tilde{\mathcal{E}}_i$. Note that $t_j \in V$ while $o_i \notin V$, so $t_j \notin \tilde{\mathcal{E}}_i \cup \{o_i\}$. Hence, we know that y_{o_i} is not erased, or $o_i \notin E_\pi$ from the proof of Lemma 2. Since $o_i \notin E_\pi$ for $i \leq m-h$ given this particular order π, we have $|E_\pi| \leq h = |V|$, which completes the proof.

APPENDIX D
PROOF OF THEOREM 4

From (45), we have

$$I(p(x_{\mathcal{K}}); p(\mathbf{y})) = D(p(x_{\mathcal{K}}, \mathbf{y}) || q(x_{\mathcal{K}}, \mathbf{y})) - D(p(x_{\mathcal{K}}) || q(x_{\mathcal{K}})) - D(p(\mathbf{y}) || q(\mathbf{y})), \quad (77)$$

and it remains to show that the right-hand side is bounded above by $D(p(\mathbf{x}) || q(\mathbf{x}))$.

By applying the chain rule for relative entropy, we have

$$D(p(\mathbf{x}, \mathbf{y}) || q(\mathbf{x}, \mathbf{y})) = D(p(\mathbf{x}) || q(\mathbf{x})) + D(p(\mathbf{y}|\mathbf{x}) || q(\mathbf{y}|\mathbf{x})), \quad (78)$$

and

$$D(p(\mathbf{x}, \mathbf{y}) || q(\mathbf{x}, \mathbf{y})) = D(p(x_{\mathcal{K}}, \mathbf{y}) || q(x_{\mathcal{K}}, \mathbf{y})) + D(p(x_{[n]\backslash\mathcal{K}}|x_{\mathcal{K}}, \mathbf{y}) || q(x_{[n]\backslash\mathcal{K}}|x_{\mathcal{K}}, \mathbf{y})). \quad (79)$$

The definition of conditional relative entropy and the proof of the chain rule for relative entropy can be found in [40, p. 24]. From these equations, we obtain

$$D(p(x_{\mathcal{K}}, \mathbf{y}) || q(x_{\mathcal{K}}, \mathbf{y})) = D(p(\mathbf{x}) || q(\mathbf{x})) + D(p(\mathbf{y}|\mathbf{x}) || q(\mathbf{y}|\mathbf{x})) - D(p(x_{[n]\backslash\mathcal{K}}|x_{\mathcal{K}}, \mathbf{y}) || q(x_{[n]\backslash\mathcal{K}}|x_{\mathcal{K}}, \mathbf{y})). \quad (80)$$

By substituting (80) in (77), we have

$$\begin{aligned} &I(p(x_{\mathcal{K}}); p(\mathbf{y})) \\ &= D(p(\mathbf{x}) || q(\mathbf{x})) + D(p(\mathbf{y}|\mathbf{x}) || q(\mathbf{y}|\mathbf{x})) - D(p(x_{\mathcal{K}}) || q(x_{\mathcal{K}})) \\ &\quad - D(p(x_{[n]\backslash\mathcal{K}}|x_{\mathcal{K}}, \mathbf{y}) || q(x_{[n]\backslash\mathcal{K}}|x_{\mathcal{K}}, \mathbf{y})) - D(p(\mathbf{y}) || q(\mathbf{y})) \\ &\overset{(a)}{\leq} D(p(\mathbf{x}) || q(\mathbf{x})) + D(p(\mathbf{y}|\mathbf{x}) || q(\mathbf{y}|\mathbf{x})) \\ &\overset{(b)}{=} D(p(\mathbf{x}) || q(\mathbf{x})), \end{aligned} \quad (81)$$

where (a) follows from the non-negativity of relative entropy, (b) follows from the assumption $q(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}) = w(\mathbf{y}|\mathbf{x})$.

REFERENCES

[1] J.-P. Hubaux, S. Katzenbeisser, and B. Malin, "Genomic data privacy and security: Where we stand and where we are heading," *IEEE Secur. Privacy*, vol. 15, no. 5, pp. 10–12, Oct. 2017.

[2] D. Grishin, K. Obbad, and G. M. Church, "Data privacy in the age of personal genomics," *Nature Biotechnol.*, vol. 37, no. 10, pp. 1115–1117, Oct. 2019.

[3] B. Berger and H. Cho, "Emerging technologies towards enhancing privacy in genomic data sharing," *Genome Biol.*, vol. 20, no. 1, pp. 1–3, 2019.

[4] D. R. Nyholt, C.-E. Yu, and P. M. Visscher, "On jim Watson's APOE status: Genetic information is hard to hide," *Eur. J. Human Genet.*, vol. 17, no. 2, pp. 147–149, Feb. 2009.

[5] N. von Thenen, E. Ayday, and A. E. Cicek, "Re-identification of individuals in genomic data-sharing beacons via allele inference," *Bioinformatics*, vol. 35, no. 3, pp. 365–371, Feb. 2019.

[6] G. Gürsoy *et al.*, "Data sanitization to reduce private information leakage from functional genomics," *Cell*, vol. 183, no. 4, pp. 905.e16–917.e16, Nov. 2020.

[7] A. S. Motahari, G. Bresler, and D. N. C. Tse, "Information theory of DNA shotgun sequencing," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6273–6289, Oct. 2013.

[8] H. Cho, D. J. Wu, and B. Berger, "Secure genome-wide association analysis using multiparty computation," *Nature Biotechnol.*, vol. 36, no. 6, pp. 547–551, Jul. 2018.

[9] B. Tahmasebi, M. A. Maddah-Ali, and S. A. Motahari, "Information theory of mixed population genome-wide association studies," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[10] I. Shomorony, S. H. Kim, T. A. Courtade, and D. N. C. Tse, "Information-optimal genome assembly via sparse read-overlap graphs," *Bioinformatics*, vol. 32, no. 17, pp. i494–i502, Sep. 2016.

[11] H. Si, H. Vikalo, and S. Vishwanath, "Information-theoretic analysis of haplotype assembly," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3468–3479, Jun. 2017.

[12] O. Milenkovic and B. Vasic, "Information theory and coding problems in genetics," in *Proc. Inf. Theory Workshop*, Oct. 2004, pp. 60–65.

[13] H. M. Kiah, G. J. Puleo, and O. Milenkovic, "Codes for DNA sequence profiles," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3125–3146, Jun. 2016.

[14] A. Gholami, M. A. Maddah-Ali, and S. Abolfazl Motahari, "Private shotgun DNA sequencing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 171–175.

[15] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.

[16] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM J. Appl. Algebra Geometry*, vol. 1, no. 1, pp. 647–664, 2017.

[17] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

[18] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[19] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Oct. 2018, pp. 173–179.

[20] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2032–2043, Apr. 2020.

[21] S. Simmons, C. Sahinalp, and B. Berger, "Enabling privacy-preserving GWASs in heterogeneous human populations," *Cell Syst.*, vol. 3, no. 1, pp. 54–61, Jul. 2016.

[22] S. E. Fienberg, A. Slavkovic, and C. Uhler, "Privacy preserving GWAS data sharing," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, Dec. 2011, pp. 628–635.

[23] H. Cho, S. Simmons, R. Kim, and B. Berger, "Privacy-preserving biomedical database queries with optimal privacy-utility trade-offs," *Cell Syst.*, vol. 10, no. 5, pp. 408–416, May 2020.

[24] S. R. Browning and B. L. Browning, "Haplotype phasing: Existing methods and new developments," *Nature Rev. Genet.*, vol. 12, no. 10, pp. 703–714, Oct. 2011.

[25] P.-R. Loh *et al.*, "Reference-based phasing using the haplotype reference consortium panel," *Nature Genet.*, vol. 48, no. 11, p. 1443, 2016.

[26] S. Das *et al.*, "Next-generation genotype imputation service and methods," *Nature Genet.*, vol. 48, no. 10, p. 1284, 2016.

[27] N. Li and M. Stephens, "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data," *Genetics*, vol. 165, no. 4, pp. 2213–2233, Dec. 2003.

[28] S. McCarthy *et al.*, "A reference panel of 64,976 haplotypes for genotype imputation," *Nature Genet.*, vol. 48, no. 10, pp. 1279–1283, Oct. 2016.

[29] E. C. Chi, H. Zhou, G. K. Chen, D. O. Del Vecchyo, and K. Lange, "Genotype imputation via matrix completion," *Genome Res.*, vol. 23, no. 3, pp. 509–518, Mar. 2013.

[30] J. Y. Dutheil, G. Ganapathy, A. Hobolth, T. Mailund, M. K. Uyenoyama, and M. H. Schierup, "Ancestral population genomics: The coalescent hidden Markov model approach," *Genetics*, vol. 183, no. 1, pp. 259–274, Sep. 2009.

[31] F. Ye, C. Naim, and S. E. Rouayheb, "ON-OFF privacy in the presence of correlation," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 7438–7457, Nov. 2021.

[32] F. Ye, C. Naim, and S. El Rouayheb, "ON-OFF privacy against correlation over time," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2104–2117, 2021.
[33] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2879–2887.
[34] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2009.
[35] Y. S. Song, "Na Li and Matthew stephens on modeling linkage disequilibrium," *Genetics*, vol. 203, no. 3, p. 1005, 2016.
[36] C. Sudlow *et al.*, "U.K. biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, no. 3, 2015, Art. no. e1001779.
[37] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
[38] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.* Berlin, Germany: Springer, 2008, pp. 1–19.
[39] A. Harmanci, X. Jiang, and D. Zhi, "Haplohide: A data hiding framework for privacy enhanced sharing of personal genetic data," pp. 1–49, Sep. 2019, *bioRxiv:786517*, doi: 10.1101/786517.
[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

Fangwei Ye (Member, IEEE) received the B.Eng. degree in information engineering from Southeast University in 2013 and the Ph.D. degree from the Department of Information Engineering, The Chinese University of Hong Kong, in 2018. From 2018 to 2020, he was a Post-Doctoral Associate with the Department of Electrical and Computer Engineering, Rutgers University. He is currently with the Broad Institute of MIT and Harvard. His research interests include information theory and its applications to privacy, bioinformatics, and coding opportunities in learning.

Hyunghoon Cho (Member, IEEE) received the B.S. and M.S. degrees in computer science from Stanford University in 2013 and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in 2019. He is currently a Schmidt Fellow with the Broad Institute of MIT and Harvard. His research interests include computational biology and biomedical data privacy. He was a recipient of the NIH Director's Early Independence Award.

Salim El Rouayheb (Senior Member, IEEE) received the Diploma degree in electrical engineering from the Faculty of Engineering, Lebanese University, Roumieh, Lebanon, in 2002, the M.S. degree from the American University of Beirut, Lebanon, in 2004, and the Ph.D. degree in electrical engineering from Texas A&M University, College Station, in 2009. He was a Post-Doctoral Research Fellow with UC Berkeley from 2010 to 2011 and a Research Scholar with Princeton University from 2012 to 2013. He was an Assistant Professor with the ECE Department, Illinois Institute of Technology, from 2013 to 2017. He is currently an Associate Professor with the ECE Department, Rutgers University, New Brunswick, NJ, USA. His research interests are in the broad area of information theory and coding theory with applications to reliability, security, and privacy in distributed systems. He was a recipient of the NSF Career Award.