# Private DNA Sequencing: Hiding Information in Discrete Noise

Kayvon Mazooji, Roy Dong, Ilan Shomorony
Department of Electrical and Computer Engineering
University of Illinois, Urbana-Champaign
mazooji2@illinois.edu, roydong@illinois.edu, ilans@illinois.edu

***Abstract*—When an individual's DNA is sequenced, sensitive medical information becomes available to the sequencing laboratory. A recently proposed way to hide an individual's genetic information during this process is to mix in DNA samples of other individuals, which act as "noise" for the sequencing lab. Motivated by this idea, we study the problem of hiding a binary random variable $X$ (a genetic marker) with the additive noise provided by mixing DNA samples, using mutual information as a privacy metric. This is equivalent to the problem of finding a worst-case noise distribution for recovering $X$ from the noisy observation among a set of feasible discrete distributions. We characterize upper and lower bounds to the solution of this problem, which are empirically shown to be very close. The lower bound is obtained through a convex relaxation of the original discrete optimization problem, and yields a closed-form expression. The upper bound is computed via a greedy algorithm for selecting the mixing proportions.**

***Index Terms*—DNA sequencing, genetic privacy, additive discrete noise, worst-case noise distribution.**

## I. Introduction

Advances in DNA sequencing technologies have led to the generation of human genetic data at an unprecedented rate [1]. This offers exciting prospects for biomedical research, and recent studies have leveraged the genetic data of hundreds of thousands of individuals to identify genetic markers associated with many traits and diseases [2–5].

Genetic testing for disease predisposition [6] and popular direct-to-consumer genomics services [7, 8] can provide us with important and actionable information about our health. However, these services require the submission of a blood or saliva sample, making an individual's entire DNA available to the testing center. This raises significant privacy concerns surrounding genetic data [9], particularly regarding the potential use of this information by insurance companies [10].

Given the potential privacy risks of DNA sequencing, an important question is whether it is possible to alter a physical DNA sample prior to submitting it to a laboratory, in order to "hide" some of its genetic information. One possible way to alter a sample could be to *mix* it with the DNA of other individuals. Upon sequencing, the lab would then observe a mixture of the data from the different samples, which would hinder its ability to retrieve individual genetic variants.

The general idea of mixing samples to attain genetic privacy was proposed in [11] (and later extended in [12]). Suppose Alice wants to have her DNA sequenced and has at her disposal the DNA samples of $K$ other people who *already know* their DNA sequence. Alice can then mix all $K+1$ DNA samples and send them to the sequencing lab. From the lab's perspective, the DNA of the $K$ additional individuals plays the role of noise, impairing the lab's ability to recover Alice's DNA. However, upon receiving the sequencing data back, the contribution of the "noise individuals" can be removed, and Alice can recover her DNA sequence information. This approach is illustrated in Figure 1.

Motivated by this idea, we study how to *optimally mix* DNA samples in order to maximize the privacy achieved. We focus on a single *biallelic site* $s$ on the genome; i.e., a location on the human genome that admits two possible alleles, and can thus be modeled as a single variable $X \in \{0,1\}$. This could model, for example, the presence of the mutation on the BRCA2 gene that increases the likelihood of breast cancer [13] and many other disease genetic markers.

In order to hide her genotype $X$ from the sequencing lab, Alice mixes into her sample the samples of $K$ individuals using proportions $\alpha_0, \alpha_1, ..., \alpha_K$, where $\sum_{i=0}^{K} \alpha_i = 1$. We model the lab's observation of site $s$ as $Y = \alpha_0 X + \sum_{i=1}^{K} \alpha_i Z_i$, where $Z_i \in \{0,1\}$ is the allele value of the $i$th noise individual. This is motivated by the fact that, if the lab uses

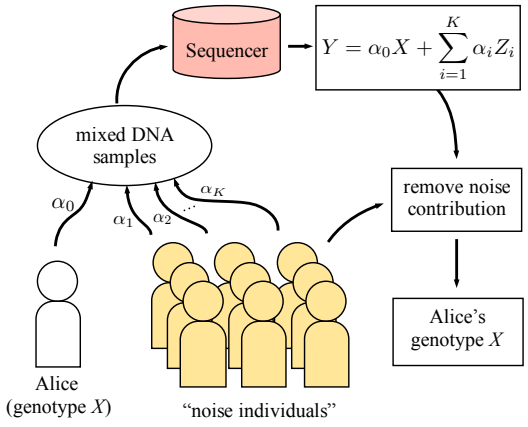


Fig. 1. In order to hide her genotype $X$ at a given locus $s$, Alice mixes her DNA sample with that of $K$ individuals in amounts $\alpha_1, ..., \alpha_K$. Upon receiving the sequencing data from the lab, Alice can remove the contribution from the "noise individuals" (whose genotype at $s$ is known) to recover $X$.

shotgun sequencing technologies [14], each reading of site $s$ is effectively a $\mathrm{Ber}\left(\alpha_0 X + \sum_{i=1}^{K} \alpha_i Z_i\right)$ random variable. Via repeated readings of $s$, $Y$ can thus be obtained with arbitrary accuracy, so we assume for simplicity that $Y$ is observed exactly. We refer to the long version of this manuscript [15] for a more detailed discussion on this model.

Following [11], we model $X, Z_1, ..., Z_K$ as i.i.d. random variables with $\Pr(X = 1) = p \in [0, 0.5]$. We refer to $p$ as the minor allele frequency, a parameter that is known in practice for genetic loci of interest. As in [11], we utilize the mutual information as our privacy metric. If we let $\alpha = (\alpha_0, ..., \alpha_K)$ and $Z_\alpha = \sum_{k=1}^{K} \alpha_k Z_k$, our goal is thus to solve

$$\min_{\alpha} I\left(X; \alpha_0 X + Z_\alpha\right), \tag{1}$$

i.e., choose the mixing coefficients $\alpha_0, ..., \alpha_K$ to minimize the mutual information between $X$ and the lab's observation.

The problem in (1) is equivalent to maximizing the conditional entropy $H(X|Y)$; i.e., the residual uncertainty in $X$ after observing $Y = \alpha_0 X + Z_\alpha$, and can thus be understood as maximizing the privacy of $X$. It can also be thought of as the problem of finding a worst-case noise $Z_\alpha$ among those of the form $\sum_k \alpha_k Z_k$, with $Z_k$ being i.i.d. $\mathrm{Ber}(p)$. Our main result is that the solution to (1) is lower-bounded as

$$\min_{\alpha} I\left(X; \alpha_0 X + Z_\alpha\right) \geqslant I\left(X; X + G\right), \tag{2}$$

where $G \sim \mathrm{Geom}\left((1-p)^K\right)$ and $G$ is independent of $X$. The right-hand side of (2) can be computed explicitly as a function of $p$ and $K$. Moreover, we verify empirically that this lower bound is very close to an upper bound provided by a greedy algorithm that selects $\alpha_1, \alpha_2, ..., \alpha_K$ sequentially to minimize the resulting mutual information, establishing $I\left(X; X + G\right)$ as a good approximation to the solution of (1). We derive (1) via a convex relaxation of (1), and use KKT conditions to show that its solution is lower bounded by $I\left(X; X + G\right)$.

## II. Problem Setting and Preliminaries

Our goal is to characterize the mixing proportions $\alpha_0, ..., \alpha_K$ that minimize the mutual information in (1). Notice that we do not need to constrain the mixing proportions to add up to 1, since scaling the observation $Y$ does not change the mutual information. We restrict $\alpha_0, ..., \alpha_K$ to be rational numbers for simplicity (but our results hold in the case where they are arbitrary real numbers as discussed in the longer version [15]). Finally, we notice that a rational-valued solution $(\alpha_0, ..., \alpha_K)$ can be converted into an integer-valued one with the same value of (1) by scaling by the least common denominator. As a result, we can restrict ourselves to solving the discrete optimization problem

$$\min_{\alpha \in \mathbb{N}^{K+1}} I(X; \alpha_0 X + Z_\alpha), \tag{3}$$

where $\alpha = (\alpha_0, ..., \alpha_K)$, $Z_\alpha = \sum_{i=1}^{K} \alpha_i Z_i$, and $X, Z_1, ..., Z_K$ are independent $\mathrm{Ber}(p)$ random variables.

The optimization problem in (3) is surprisingly complex. The symmetry between the variables $\alpha_0, ..., \alpha_K$ may suggest that $\alpha_i = 1$ for $i = 0, ..., K$ would be an optimal solution. However, a brute-force solution to (3) for small values of $K$ shows that optimal solutions $(\alpha_0, ..., \alpha_K)$ vary widely for different values of $p$, as illustrated in Figure 2. Observe that the curve appears to be only piecewise smooth. At $p = 0.5$, the optimal solution is given by $[1, 1, 2, 4, 8, 16]$, at $p = 0.25$, the optimal solution is given by $[1, 1, 1, 2, 3, 4]$, and at $p = 0.01$, the optimal solution is given by $[1, 1, 1, 1, 1, 1]$.

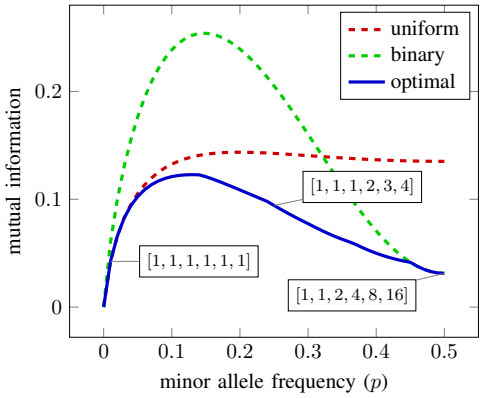


Fig. 2. Optimal value of (3) compared to the uniform scheme, for $K = 5$ and $p \in [0, 0.5]$. At $p = 0.5$, the optimal scheme is $\alpha = [1, 1, 2, 4, 8, 16]$. At $p = 0.25$, the optimal scheme is $\alpha = [1, 1, 1, 2, 3, 4]$. At $p = 0.01$, the optimal scheme is $\alpha = [1, 1, 1, 1, 1, 1]$.

As it turns out, the optimal solution to (3) can be exactly characterized in the two extremes cases of $p$. More precisely, if we define the *uniform* scheme to be $\alpha_i = 1$ for $i = 0, ..., K$, and we define the *binary* scheme to be $\alpha_0 = 1$ and $\alpha_i = 2^{i-1}$ for $i = 1, ..., K$, we have the following result.

**Theorem 1.** *Fix some $K \in \mathbb{N}$. Then there exists some $p^* > 0$ such that the uniform scheme is optimal for $p < p^*$. Moreover, the binary scheme is optimal for $p = 0.5$.*

We prove Theorem 1 in the longer version of this paper [15]. Aside from the cases $p = 0.5$ and $p \approx 0$, there does not appear to be a simple expression for the optimal solution $\alpha$.

**Notation:** Throughout the paper we use $\mathbb{N}$ to denote the set of natural numbers excluding 0, $\mathbb{N}_0$ to denote the set of natural numbers including 0, and $[N]$ to denote the set of natural numbers in $\{1, 2, ..., N\}$ for an integer $N$.

## III. Main Results

In order to tackle the discrete optimization problem in (3) in the entire interval $p \in [0, 0.5]$, we seek to bound its optimal solution. Our main result is the following lower bound.

**Theorem 2.** *For any $K \in \mathbb{N}$ and $p \in [0, 0.5]$, we have*

$$\min_{\alpha \in \mathbb{N}^{K+1}} I(X\ ;\ \alpha_0 X + Z_\alpha) \geqslant I(X\ ;\ X + G), \tag{4}$$

*where $G$ is a geometric random variable with $\Pr(G = i) = (1-p)^K(1-(1-p)^K)^i$ for $i = 0, 1, 2, ....$*

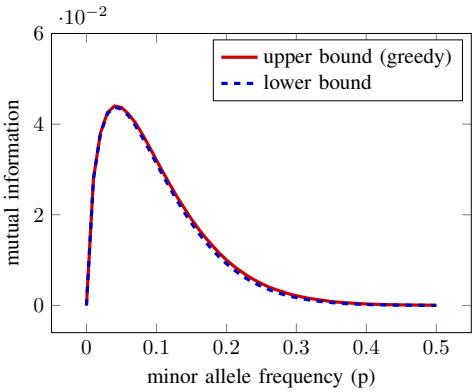


Fig. 3. Comparison between the lower bound from (4) and the upper bound provided by the greedy algorithm for $K = 15$.

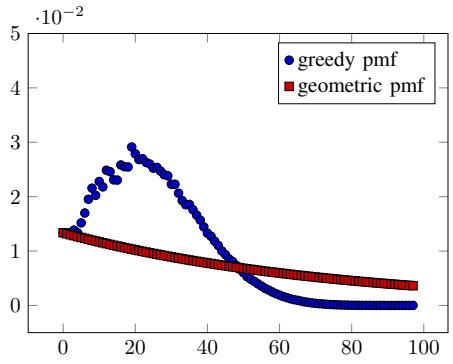


Fig. 4. Comparison of the pmf of $Z_\alpha$ produced by the greedy algorithm ($\alpha = [1, 1, 1, 1, 2, 2, 3, 4, 5, 6, 7, 8, 10, 12, 16, 19]$) and the (truncated) Geometric pmf in the lower bound (4), for $K = 15$ and $p = 0.25$.

Intuitively, Theorem 2 says that the noise distribution of $G$ is worse than the worst-case noise $Z_\alpha$. This lower bound can in fact be explicitly computed (described in detail in [15]) as

$$\begin{aligned} &I(X\,;\,X+G) = H(p) \\ &-(p-1)\left((1-p)^K-1\right)\log\left(\frac{1-(1-p)^{K+1}}{(p-1)\left((1-p)^K-1\right)}\right) \\ &-p\log\left(\frac{1-(1-p)^{K+1}}{p}\right). \end{aligned} \tag{5}$$

Observe that this formula is quickly computable for any value of $K$ and $p$, making it attractive from a computational standpoint. To assess how tight the lower bound is, we empirically compare it to an upper bound that is computed with a greedy algorithm in Figure 3. For a given $p$ and $K$, the greedy algorithm chooses $\alpha_0 = 1,\ \alpha_1 = 1$, and sequentially chooses

$$\alpha_j = \underset{\alpha\in\mathbb{N}\,:\,1\leqslant\alpha\leqslant 1+\sum_{i=1}^{j-1}\alpha_i}{\arg\min} I(X\,;\,X+\alpha Z_j+\sum_{i=1}^{j-1}\alpha_i Z_i) \tag{6}$$

for $2 \leqslant j \leqslant K$. At the $j$th step we consider all $\alpha$'s between 1 and $1+\sum_{i=1}^{j-1}\alpha_i$ because $1+\sum_{i=1}^{j-1}\alpha_i$ is the highest choice for $\alpha$ that does not increase the number of support values $t$ where $H(X \mid X+\alpha Z_j+\sum_{i=1}^{j-1}\alpha_i Z_i = t) = 0$.

As seen in Figure 3, $I(X\,;\,X+G)$ serves as a tight lower bound when compared with the upper bound. A similar picture can be obtained for other values of $K$. This is surprising, because for a given $K$, it is not possible in general to choose $\alpha_i$s to make the pmf of $\sum_{i=1}^{j-1}\alpha_i Z_i$ look like the pmf of $G$ (or a shifted version of it), as illustrated in Figure 4.

While finding the greedy solution $(\alpha_0, ..., \alpha_K)$ requires $\Omega(2^K)$ time in the worst case, similar plots to Figure 3 can be obtained for larger values of $K$ using a more computationally efficient variation of the greedy algorithm that is discussed in the longer version of this paper [15].

At a high level, we prove the lower bound in Theorem 2 by forming a convex relaxation of the minimization problem, perturbing the relaxation to form a problem that is analytically solvable using KKT conditions, and using perturbation analysis to find a lower bound on the relaxation.

## IV. Proof of Theorem 2

We obtain the lower bound on $I(X\,;\,\alpha_0 X+Z_\alpha) = H(p) - H(X \mid \alpha_0 X+Z_\alpha)$ by finding a lower bound on $-H(X \mid \alpha_0 X+Z_\alpha)$. Therefore, we consider

$$\min_{\alpha_0,\ldots,\alpha_K\in\mathbb{N}} -H\left(X \;\middle|\; \alpha_0 X+\sum_{k=1}^{K}\alpha_k Z_k\right). \tag{7}$$

Observe that the pmf of the random variable $Z_\alpha = \sum_{i=1}^{K}\alpha_i Z_i$ has probability $(1-p)^K$ at its lowest support value. More precisely, 0 is the minimum value that $Z_\alpha$ can take, which occurs with probability $(1-p)^K$. A relaxation to (7) can then be obtained by ignoring all constraints on the pmf of $Z_\alpha$ except the constraint on the minimum pmf value. Thus, for a fixed value of $\alpha_0$, a relaxation to (7) is given by

$$\begin{aligned} \min_{Q}\ & -H\left(X \mid \alpha_0 X+Q\right) \\ \text{subject to: } & Q \text{ is an integer-valued random variable} \\ & Q \text{ is independent of } X \\ & \Pr(Q=0) = (1-p)^K \\ & \Pr(Q=i) = 0 \ \text{ for } i<0. \end{aligned} \tag{8}$$

Furthermore, as we prove in Lemma 3 in Section V, fixing $\alpha_0 = 1$ in (8) above does not change the optimal value, and we assume throughout that $\alpha_0 = 1$.

Let $q_{(i)}$ be the pmf of $Q$; i.e., $q_{(i)} = \Pr(Q=i)$ for $i \geqslant 0$. In order to write (8) explicitly in terms of $q$, we define

$$\begin{aligned} g_j(q) &\triangleq -H(X|X+Q=j)\Pr(X+Q=j) \\ &= \Pr(Q+X=j) \\ &\quad\times\left[\frac{(1-p)q_{(j)}}{\Pr(Q+X=j)}\log\left(\frac{(1-p)q_{(j)}}{\Pr(Q+X=j)}\right)\right. \\ &\quad\left.+\frac{pq_{(j-1)}}{\Pr(Q+X=j)}\log\left(\frac{pq_{(j-1)}}{\Pr(Q+X=j)}\right)\right] \\ &= (1-p)q_{(j)}\log\left(\frac{(1-p)q_{(j)}}{(1-p)q_{(j)}+pq_{(j-1)}}\right) \\ &\quad+pq_{(j-1)}\log\left(\frac{pq_{(j-1)}}{(1-p)q_{(j)}+pq_{(j-1)}}\right). \end{aligned} \tag{9}$$

Written in terms of $g_j(q)$ and with $\alpha_0 = 1$, (8) is given by

$$\begin{aligned} \min_{q_{(i)} \geqslant 0:\ i \in \mathbb{N}_0} \quad & \sum_{j \in \mathbb{N}} g_j(q) \\ \text{subject to:} \quad & q_{(0)} = (1-p)^K \\ & \sum_{j=0}^{\infty} q_{(j)} = 1. \end{aligned} \tag{10}$$

Observe that (10) is a convex minimization problem with infinitely many variables. For such problems, to the best of our knowledge, a solution to the KKT conditions is not in general guaranteed to yield an optimal solution. For that reason, we do not seek to directly solve the KKT conditions and, instead, we consider a *support-constrained* version of (10), where the support of the pmf of $Q$ is restricted to $\{0, ..., n\}$, and let $n \to \infty$. The support-constrained version of (10) is given by

$$\begin{aligned} \min_{q_{(i)} \geqslant 0:\ i \in \{0,\ldots,n+1\}} \quad & \sum_{j=1}^{n+1} g_j(q) \\ \text{subject to:} \quad & q_{(0)} = (1-p)^K, \quad q_{(n+1)} = 0 \\ & \sum_{j=0}^{n+1} q_{(j)} = 1. \end{aligned} \tag{11}$$

Note that this is no longer a relaxation to the original problem.

In order to ensure that the derivative of the objective function exists at all feasible $q_{(i)}$, and ultimately find a lower bound on the optimal value of (11) through perturbation analysis, we change the constraints in (11) from $q_{(i)} \geqslant 0$ to $q_{(i)} > 0$ and from $q_{(n+1)} = 0$ to $q_{(n+1)} = \epsilon$ where $\epsilon > 0$, obtaining

$$\begin{aligned} \min_{q_{(i)} > 0:\ i \in \{0,\ldots,n+1\}} \quad & \sum_{j=1}^{n+1} g_j(q) \\ \text{subject to:} \quad & q_{(0)} = (1-p)^K, \quad q_{(n+1)} = \epsilon \\ & \sum_{j=0}^{n+1} q_{(j)} = 1. \end{aligned} \tag{12}$$

Let $V_n^*$ be the optimal value of (11) and let $V_{n,\epsilon}^*$ be the optimal value of (12) for a given $\epsilon$. Due to the continuity of the objective function in (11), we have $V_n^* \geqslant \inf_{\epsilon > 0} V_{n,\epsilon}^*$. More precisely, for any solution to (11), it follows that (12) can get arbitrarily close to the corresponding value as $\epsilon \to 0$ due to the continuity of the objective function.

While we do not know the solution to (11) or (12), we will perturb (12) to form a problem we can solve analytically, then use perturbation analysis to lower bound the optimal value of (12), and ultimately lower bound the optimal value of (11). Let $\beta \triangleq (1-p)^K$. The perturbed version of (12) is given by

$$\begin{aligned} \min_{q_{(i)} > 0:\ i \in \{0,\ldots,n+1\}} \quad & \sum_{j=1}^{n+1} g_j(q) \\ \text{subject to:} \quad & q_{(0)} = \beta, \quad q_{(n+1)} = \beta(1-\beta)^{n+1} \\ & \sum_{j=0}^{n+1} q_{(j)} = 1 - (1-\beta)^{n+2}. \end{aligned} \tag{13}$$

Observe that as $n$ increases, we have that $\beta(1-\beta)^{n+1} \to 0$ and $1-(1-\beta)^{n+2} \to 1$. In other words, the constraints in (13) approach the constraints in (11). We can solve (13) by solving the KKT conditions since the problem is convex, Slater's condition holds, and the objective function and constraint functions are differentiable on the domain $q_{(i)} > 0$. Let $f_0(q)$ be the objective function of (13). The Lagrangian is given by

$$\begin{aligned} L(q, v, \lambda) = f_0(q) &+ v_1 \left( \sum_i q_{(i)} - 1 + (1-\beta)^{n+2} \right) \\ &+ v_2 \left( q_{(0)} - \beta \right) \\ &+ v_3 \left( q_{(n+1)} - \beta(1-\beta)^{n+1} \right). \end{aligned} \tag{14}$$

The perturbation values in (13) are carefully chosen so that the KKT conditions yield an optimal solution given by

$$q_{(i)} = \beta(1-\beta)^i \text{ for } i \in \{0, ..., n+1\}. \tag{15}$$

This corresponds to the first $n+2$ terms of the pmf of a Geometric distribution. The derivation of (15) and the optimal Lagrange multipliers $v_1^*, v_2^*, v_3^*$ are provided in Lemma 4.

Let $U_n^*$ be the solution to (13), obtained by plugging in (15). Using the perturbation analysis from Section 5.6.1 of [16], we see that the optimal value of (12), $V_{n,\epsilon}^*$, is lower bounded as

$$V_{n,\epsilon}^* \geqslant U_n^* - v_1^* \left( (1-\beta)^{n+2} \right) - v_3^* \left( -\beta(1-\beta)^{n+1} + \epsilon \right), \tag{16}$$

where $v_1^*$ and $v_3^*$ are the optimal Lagrange multipliers for (13) described in Lemma 4 in Section V. Taking the infimum of (16) over $\epsilon > 0$ then yields

$$V_n^* \geqslant U_n^* - v_1^* \left( (1-\beta)^{n+2} \right) - v_3^* \left( -\beta(1-\beta)^{n+1} \right). \tag{17}$$

The sequence $U_n^*$ of optimal values returned by (11) is non-increasing in $n$. Thus, letting $n \to \infty$ in (17) implies that $\lim_{n\to\infty} U_n^*$ is a lower bound to $-H(X \,|\, \alpha_0 X + Z_\alpha)$ for any choice of $\alpha_i$'s. Notice that, as $n \to \infty$, $q_{(j)}$ in (15) converges to the pmf of a Geometric random variable $G$ with $\Pr(G = i) = \beta(1-\beta)^i = (1-p)^K(1-(1-p)^K)^i$ for $i \in \mathbb{N}_0$. Since the objective function of (13) is $-H(X|X+Q)$ where $Q$ has pmf $q_{(j)}$, this concludes our proof.

## V. Auxiliary Lemmas

**Lemma 3.** *Fixing $\alpha_0 = 1$ in the optimization problem (8) does not change the optimal value.*

*Proof:* Let $\alpha_0 \in \mathbb{N}$. Let $Q$ be any random variable such that its pmf has integral support and minimum support value at $t = 0$. Let $\hat{Q} = Q + S$, where $S = (-Q) \mod \alpha_0$. Then

$$\begin{aligned} I(X;\ \alpha_0 X + Q) &= I(X;\ \alpha_0 X + \hat{Q} - S) \\ &\overset{(i)}{=} I(X;\ \alpha_0 X + \hat{Q},\ S) \\ &\geqslant I(X\ ;\ \alpha_0 X + \hat{Q}) \end{aligned} \tag{18}$$

where $(i)$ follows since from $\alpha_0 X + \hat{Q} - S$, we can compute

$$\begin{aligned} S &= (-(\alpha_0 X + \hat{Q} - S)) \mod \alpha_0, \\ \alpha_0 X + \hat{Q} &= (\alpha_0 X + \hat{Q} - S) + S. \end{aligned}$$

Thus, any solution $Q$ to (8) with $\alpha_0 = a$ can be transformed into a solution $\hat{Q}$ whose pmf has support values in $\{ta : t \in \mathbb{Z}\}$, without increasing the mutual information. Therefore, for a fixed $\alpha_0$, it suffices to only consider pmfs that have support values in $\{t\alpha_0 : t \in \mathbb{Z}\}$ in the optimization. If $\alpha_0 = a$, any solution random variable $Q$ with support values in $\{ta : t \in \mathbb{Z}\}$ can be transformed into a solution with the same mutual information, $\alpha_0 = 1$, and support values in $\{t : t \in \mathbb{Z}\}$. This is accomplished by dividing $Q$ by $a$. Therefore it suffices to fix $\alpha_0 = 1$ in the optimization. ∎

**Lemma 4.** *A solution to the KKT conditions for (13) is*

$$\begin{aligned} q_{(i)} &= \beta(1-\beta)^i \text{ for } i \in \{0, ..., n+1\} \\ v_1 &= -p \log\left(\frac{p}{(1-p)(1-\beta)+p}\right) \\ &\quad - (1-p)\log\left(\frac{(1-p)(1-\beta)}{(1-p)(1-\beta)+p}\right) \\ v_2 &= (1-p)\log\left(\frac{(1-p)(1-\beta)}{(1-p)(1-\beta)+p}\right) \\ v_3 &= p\log\left(\frac{p}{(1-p)(1-\beta)+p}\right). \end{aligned} \tag{19}$$

*Proof:* Let $f_0(q)$ be the objective function of (13). The Lagrangian of (13) is given in (14). As described in detail in [15], the derivative of $f_0(q)$ with respect to $q_{(j)}$ is given by

$$\begin{aligned} &p\log\left(\frac{pq_{(j)}}{(1-p)q_{(j+1)} + pq_{(j)}}\right) \\ &\quad + (1-p)\log\left(\frac{(1-p)q_{(j)}}{pq_{(j-1)} + (1-p)q_{(j)}}\right) \end{aligned} \tag{20}$$

for $j \in \{1, ..., n\}$. For $j = 0$ and $j = n+1$, the derivative can be similarly computed and the KKT conditions are

$$\begin{aligned} &q_{(0)} = (1-p)^K, \\ &q_{(n+1)} = (1-p)^K(1-(1-p)^K)^{n+1}, \\ &\sum_{j=0}^{n+1} q_{(j)} = 1 - (1-(1-p)^K)^{n+2}, \\ &\text{for } j = 1, ..., n, \quad p\log\left(\frac{pq_{(j)}}{(1-p)q_{(j+1)} + pq_{(j)}}\right) \\ &\quad + (1-p)\log\left(\frac{(1-p)q_{(j)}}{pq_{(j-1)} + (1-p)q_{(j)}}\right) + v_1 = 0, \\ &p\log\left(\frac{pq_{(0)}}{(1-p)q_{(1)} + pq_{(0)}}\right) + v_1 + v_2 = 0, \\ &(1-p)\log\left(\frac{(1-p)q_{(n+1)}}{pq_{(n)} + (1-p)q_{(n+1)}}\right) + v_1 + v_3 = 0. \end{aligned} \tag{21}$$

The last three conditions can be rewritten as

$$\begin{aligned} &\text{for } j = 1, ..., n, \quad p\log\left(\frac{p}{(1-p)\frac{q_{(j+1)}}{q_{(j)}} + p}\right) \\ &\quad + (1-p)\log\left(\frac{(1-p)}{p\frac{q_{(j-1)}}{q_{(j)}} + (1-p)}\right) + v_1 = 0, \end{aligned}$$

$$p\log\left(\frac{p}{(1-p)\frac{q_{(1)}}{q_{(0)}} + p}\right) + v_1 + v_2 = 0, \tag{22}$$

$$(1-p)\log\left(\frac{(1-p)}{p\frac{q_{(n)}}{q_{(n+1)}} + (1-p)}\right) + v_1 + v_3 = 0. \tag{23}$$

Notice that, if the ratio $\frac{q_{(j+1)}}{q_{(j)}}$ between consecutive values of $q_{(j)}$ is the same for all $j$, then $v_1$, $v_2$, $v_3$ can be chosen so that these derivatives equal 0 for all $j$. Setting

$$\frac{q_{(j+1)}}{q_{(j)}} = (1-(1-p)^K),$$

we have that (19) is a solution to all equations in (21). ∎

## VI. Concluding Remarks

We derived a lower bound on the best possible mutual information achievable by a private DNA sequencing scheme, and showed that it is tight by comparing it with the scheme generated by a greedy algorithm. Directions for follow-up work include considering multiple, possibly correlated, genetic loci simultaneously, and characterizing the optimal tradeoffs between privacy and the ability to correctly recover the genotype $X$ from the data received from the lab.

## References


[1] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: Astronomical or genomical?," *PLOS Biology*, vol. 13, pp. 1–11, 07 2015.

[2] P. van der Harst and N. Verweij, "Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease," *Circulation research*, vol. 122, no. 3, pp. 433–443, 2018.

[3] A. Mahajan *et al.*, "Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps," *Nature genetics*, vol. 50, no. 11, 2018.

[4] R. S. Desikan *et al.*, "Genetic assessment of age-associated alzheimer disease risk: Development and validation of a polygenic hazard score," *PLoS medicine*, vol. 14, no. 3, 2017.

[5] E. T. Cirulli and D. B. Goldstein, "Uncovering the roles of rare variants in common disease through whole-genome sequencing," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 415–425, 2010.

[6] W. D. Hall, K. I. Morley, and J. C. Lucke, "The prediction of disease risk in genomic medicine: Scientific prospects and implications for public policy and ethics," *EMBO reports*, vol. 5, no. S1, pp. S22–S26, 2004.

[7] S. Hogarth, G. Javitt, and D. Melzer, "The current landscape for direct-to-consumer genetic testing: legal, ethical, and policy issues," *Annu. Rev. Genomics Hum. Genet.*, vol. 9, pp. 161–182, 2008.

[8] M. A. Allyse, D. H. Robinson, M. J. Ferber, and R. R. Sharp, "Direct-to-consumer testing 2.0: emerging models of direct-to-consumer genetic testing," *Mayo Clinic Proceedings*, vol. 93, no. 1, pp. 113–120, 2018.

[9] Y. Erlich and A. Narayanan, "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics*, vol. 15, no. 6, 2014.

[10] N. Martin, "How DNA Companies Like Ancestry And 23andMe Are Using Your Genetic Data," *Forbes*, Dec. 2018.

[11] A. Gholami, M. A. Maddah-Ali, and S. A. Motahari, "Private shotgun DNA sequencing," in *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 171–175, IEEE, 2019.

[12] A. Gholami, M. A. Maddah-Ali, and S. A. Motahari, "Private shotgun DNA sequencing: A structured approach," in *2019 Iran Workshop on Communication and Information Theory (IWCIT)*, pp. 1–6, IEEE, 2019.

[13] R. Wooster *et al.*, "Identification of the breast cancer susceptibility gene BRCA2," *Nature*, vol. 378, 1995.

[14] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature methods*, vol. 5, no. 1, pp. 16–18, 2008.

[15] K. Mazooji, R. Dong, and I. Shomorony, "Private DNA sequencing: Hiding information in discrete noise," *available at https://ilanshom.github.io/papers/private.pdf*.

[16] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge: Cambridge University Press, 2004.