Efficient Privacy-Preserving String Search and an Application in Genomics

Kana Shimizu,[1,4,*] Koji Nuida[2,3] and Gunnar Rätsch[4,*]

[1] Biotechnology Research Institute for Drug Discovery, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi Koto-ku, Tokyo 135-0064, Japan
[2]Information Technology Research Institute, National Institute of Advanced Industrial Science and Technology, 2-4-7 Aomi Koto-ku, Tokyo 135-0064, Japan,
[3]Japan Science and Technology Agency (JST) PRESTO Researcher, Tokyo, Japan
[4]Computational Biology, Memorial Sloan Kettering Cancer Center, 1275 York, New York, NY, USA
*To whom correspondence should be addressed: shimizu-kana@aist.go.jp and ratschg@mskcc.org

Abstract

Motivation: Personal genomes carry inherent privacy risks and protecting privacy poses major social and technological challenges. We consider the case where a user searches for genetic information (e.g., an allele) on a server that stores a large genomic database and aims to receive allele-associated information. The user would like to keep the query and result private and the server the database.

Approach: We propose a novel approach that combines efficient string data structures such as the *Burrows-Wheeler transform* with cryptographic techniques based on additive homomorphic encryption. We assume that the sequence data is searchable in efficient iterative query operations over a large indexed dictionary, for instance, from large genome collections and employing the (positional) Burrows-Wheeler transform. We use a technique called *oblivious transfer* that is based on *additive homomorphic encryption* to conceal the sequence query and the genomic region of interest in positional queries.

Results: We designed and implemented an efficient algorithm for searching sequences of SNPs in large genome databases. During search, the user can only identify the longest match while the server does not learn which sequence of SNPs the user queries. In an experiment based on 2,184 aligned haploid genomes from the 1,000 Genomes Project, our algorithm was able to perform typical queries within ≈2 seconds and ≈20 seconds seconds for client and server side, respectively, on a laptop computer. The presented algorithm is at least one order of magnitude faster than an exhaustive baseline algorithm.

1 Introduction

String search is a fundamental task in the field of genome informatics, for which a large variety of techniques have been developed (see, for instance,[2, 15, 18]). Traditionally, those techniques have been optimized for accuracy and computational efficiency, however a recent boom of personal genome sequencing and analyses has spotlighted a new criteria, namely, privacy protection. As reported in many studies, a genome is considered to be one of the most critical pieces of information for an individual's privacy. In fact, it is largely different from any other personal information because it works as an identifier of an individual while it possesses the information that has strong correlation with the phenotype of the individual [25, 9]. Therefore, in principle, privacy protection is an inevitable problem when handling personal genomes. As a practice, the most popular approach is protecting genomes physically; genomic sequences have been kept at few collaborator sites, and only a limited number of researchers are allowed to access them. This conservative approach severely limits the great potential of existing genomic resources. In order to mitigate the stagnation caused by privacy issues, it appears crucial to develop practical methods that enable searching and mining genomic databases in a privacy-preserving manner.

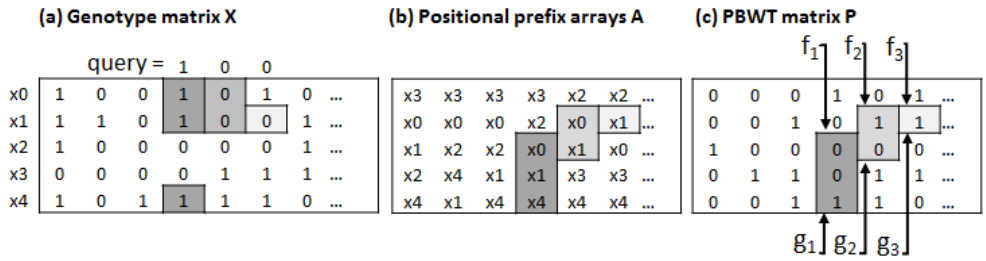

Figure 1: Outline of the search strategy with PBWT. A set of genotype sequences $X = \{x_0, \ldots, x_4\}$ illustrated in (a) is sorted by the algorithm described in [7] to obtain the positional prefix arrays A illustrated in (b). Each element $P_{i,j}$ of PBWT matrix illustrated in (c) is $(j+1)$-th letter of sequence $A_{i,j}$. By computing rank operations with regard to $(k+1)$-th query letter on $P_{\cdot,k}$, one can update an interval corresponding to $(k+1)$-mer match between the query and X. In this figure, the search starts from fourth allele. The first interval $[f_1, g_1]$ is initialized by rank operations on $P_{\cdot,3}$ with regard to first query letter '1'. The second interval $[f_2, g_2]$ is obtained by rank operations on $P_{\cdot,4}$ with regard to the second query letter '0' and $[f_1, g_1]$. Similarly, the third interval $[f_3, g_3]$ is obtained by rank operations on $P_{\cdot,5}$ with regard to the third query letter '0' and $[f_2, g_2]$. See Sections 2.2 and 3.3 for more details.

So far, several groups have tackled related problems. [14] developed secure multi-party computation protocols for computing edit distance. [5] proposed a protocol to search DNA string against a DNA profile represented by finite automata. [6] proposed a protocol to detect a match between two short DNA sequences for the purpose of genetic test. [4] also aimed for genetic test to develop a method for computing set intersection cardinality. [11] proposed a protocol for searching predefined keywords from databases. [22] proposed a substring search protocol for public databases while keeping user's query private. [3] developed a system by using several cryptographic techniques to find a subset of short reads which includes a fixed-length query string at specific position.

We propose a general approach which utilizes an efficient iteratively queriable data structure together with cryptographic techniques. Among many variations of such data structures, the Burrows-Wheeler Transform (BWT[16, 17, 19]) and related techniques such as the positional BWT (PBWT;[7]) have dramatically improved the speed of genomic database analyses. Those data structures commonly have an indexed dictionary called a rank dictionary. By referring to the rank dictionary in iterative operations, one can efficiently search the database. For the case of BWT, a match between query and database is reported as an interval $[f, g]$, and the interval is computed by the look-up of the rank dictionary. In our approach, we access the rank dictionary in privacy-preserving manner by using *additive homomorphic encryption* and *oblivious transfer* (OT).

Cryptographic approaches often require significant computational resources. The goal of this work is to illustrate that privacy-preserving queries are within reach when using current cryptographic techniques and standard computing hardware. We demonstrate that a typical query would only take about two seconds on the user side and ≈20 seconds on the server, while preserving privacy of the query string and the database.

The rest of the paper is organized as follows. In Approach, we describe the main ideas of our approach without going into technical details. In Methods, the detailed algorithm of recursive oblivious transfer is given followed by the description of a practical algorithm, named *Crypto-PBWT*, for privacy-preserving search in large-scale genotype databases. We also describe complexity and security properties of the proposed algorithm. We provide the more intricate details of a more efficient version of the algorithm in Supplementary Sections A-B. In Experiments, we evaluate the performance of *Crypto-PBWT* on datasets created from data of the 1,000 Genomes Project [27] and compare it to an alternative method for fixed-length k-mer search. Finally, we conclude our study in Section 5.

2 Approach

2.1 Problem Setup

We consider the setting in which a user would like to search a genomic sequence in a database with the aim to either determined whether this sequence exists in the queried database and/or to obtain additional information associated with the genomic sequence. An example is the use in a so-called *genomic beacon* (for instance, those created within the *Beacon Project* of the Global Alliance for Genome & Health (GA4GH).) Another application is the search of a specific combination of variants, for instance, in the BRCA1 or BRCA2 genes, with the aim to determine whether that combination of variants is known or predicted to be deleterious (see, for instance, GA4GH's *BRCA Challenge*). For privacy reasons, the user would like to conceal the queried sequence, which would be particularly relevant for the second example. For both examples it would be important that the server's database is protected.

In our algorithm using additive homomorphic encryption, the user generates two keys: one is a public key which is used for encryption and the other is a secret key which is used for decryption. The public key is sent to the server before the search starts. After computing a search result in encrypted form, the sever sends it to the user. By using the secret key, only the user decrypts the result. Figure 2 illustrates a flow of such algorithm for the case of computing linear communication size OT. Sections 3.1-3.2 describe more details about additive homomorphic encryption and OT.

2.2 Recursive Search Data Structures

PBWT is an efficient data structure of M aligned (genomic) sequences of length N [7]. PBWT stores information very efficiently and still allows computations (this is a property of Succinct Data Structures, see [13]). To search for a query string q over the alphabet Σ, one iteratively operates on intervals that can later be used to identify the matching genomic regions based on the PBWT. A substring match starting at a pre-specified position t is represented by an interval $[f, g]$. The number of matches is given by the length of the interval $g - f$. It is known that the $(k+1)$-th interval $[f_{k+1}, g_{k+1}]$ corresponding to a $(k+1)$-mer match can be updated from the k-th interval $[f_k, g_k]$ and the $(k+1)$-th letter of the query q. See Figure 1 for an illustration.

We will provide more details on how to update f and g in Section 3.3. To understand the key ideas, it is sufficient to understand that the updates can be written in the form of

$$f_{k+1} = v_c[f_k] \quad \text{and} \quad g_{k+1} = v_c[g_k],$$

where $c = q[k+1]$ and $v_c \in \mathbb{N}^{M \times N}$ is a large, static lookup table. Hence, the iterative algorithm of updating $[f_k, g_k]$ by using the query q, can be written as a recursive algorithm:

$$f_{k+1} = v_{q[k+1]}[v_{q[k]}[v_{q[k-1]}[\ldots v_{q[1]}[f_0]]]].$$

This can be done analogously for g_{k+1}. Other algorithms, including the non-positional Burrows-Wheeler transform (for instance,[10, 17]), can be expressed in similar ways. In this work we will refer to data structures that can be queried in the recursive way described above as *recursive search data structures*.

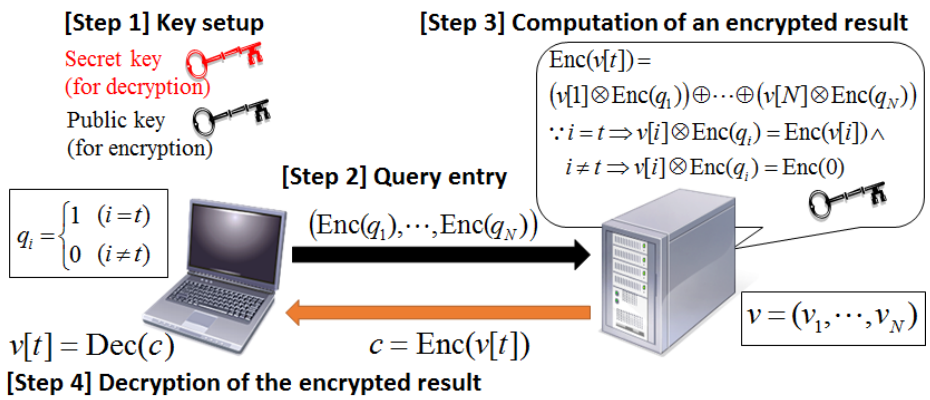

Figure 2: Outline of Oblivious Transfer (OT) based on additive homomorphic encryption. See Sections 2.3 and 3.2 for more details.

2.3 Oblivious Transfer for Privacy-Preserving Search

Concealing the Query: A key idea of our approach is to look-up elements of $\boldsymbol{v}_c$ by using a cryptographic technique called *oblivious transfer* (OT). By employing OT, the sender ("user") learns only the t-th element in the receiver's ("server") vector without leaking any information about t to the receiver [24]. Among several efficient algorithms for this scheme [20, 21, 29], we used those which are based on additive homomorphic encryption. By this property, our approach enables the user to obtain the interval $[f_{k+1}, g_{k+1}]$ without revealing $[f_k, g_k]$ to the server. See Figure 2 for an illustration. We explain homomorphic encryption and OT in some detail in Sections 3.1-3.2.

Concealing the Database: While this approach protects a user's privacy, the server leaks information of $\boldsymbol{v}_c$ which may be sufficient to reconstruct parts of the genotypes in the database. In order to rigorously protect the server's privacy, we propose a technique that allows for recursive oblivious transfer where the user does not learn intermediate results but only if a unique match was found. It is based on a bit-rotation technique which enables the server to return $\hat{f}_k := R(f_k)$ and $\hat{g}_k := R'(g_k)$ which are random values to the user. Only the server can recover f_k and g_k *in encrypted form* (i.e. the server does not see f_k and g_k when recovering them), and thus the user can recursively access $v_c[f_k]$ and $v_c[g_k]$ correctly. The details of this approach are given in Section 3.2.

In this work, we designed an algorithm based on these techniques that can be used for for privacy-preserving search in large genotype databases.

3 Methods

3.1 Additively homomorphic encryption

Our main cryptographic tool in this paper is an additive-homomorphic public-key encryption scheme $(\mathsf{KeyGen}; \mathsf{Enc}; \mathsf{Dec})$, which enables us to perform additive operations on *encrypted* values. Here, the algorithm KeyGen generates a public key pk and a secret key sk; $\mathsf{Enc}(m)$ denotes a cipher text obtained by encrypting message m under the given pk; and $\mathsf{Dec}(c)$ denotes the decryption result of cipher text c under the given sk. The scheme also has the following additive-homomorphic properties:

- Given two cipher texts $\mathsf{Enc}(m_1)$ and $\mathsf{Enc}(m_2)$ of integer messages m_1 and m_2, $\mathsf{Enc}(m_1 + m_2)$ can be computed without knowing m_1, m_2 and the secret key (denoted by $\mathsf{Enc}(m_1) \oplus \mathsf{Enc}(m_2)$).
- Given a cipher text $\mathsf{Enc}(m)$ of a message m and an integer e, $\mathsf{Enc}(e \cdot m)$ can be computed without knowing m and the secret key (denoted by $e \otimes \mathsf{Enc}(m)$). In particular, $\mathsf{Enc}(-m)$ can be computed in this manner.

This scheme should have semantic security; that is, a cipher text leaks no information about the original message [12]. For example, we can use either the Paillier cryptosystem [23] or the "lifted" version of the ElGamal cryptosystem [8]; now the second operation $\otimes$ can be realized by repeating the first operation $\oplus$. It goes beyond the scope of this paper to review the details of these cryptographic techniques and the reader is referred to a book [28] on homomorphic encryption. A typical addition operation in the ElGamal cryptosystem takes about $2 \cdot 10^{-7}$ seconds on a single CPU based on AIST's ElGamal library [1].

3.2 Recursive Oblivious Transfer by Random Rotations

To protect the privacy of the database, we propose a technique for recursively querying a data structure without obtaining information about intermediate results. Let us define the recursive oblivious transfer problem as follows:

Model 1 *A user has a private value $1 \leq x_1 \leq N$ and a server has a private vector $\boldsymbol{v}$ of length N. Let us denote $x_{k+1} = v[x_k]$ and the user is allowed to access the server $\ell - 1$ times. After the calculation, the user learns only x_ℓ and the server learns nothing about $x_1, \ldots, x_\ell$.*

Let us explain our idea by extending a simple linear communication size OT where the user aims to know the t-th element of the server's vector $\boldsymbol{v}$. In the first step, the user creates a bit vector:

$$\boldsymbol{q} = (q_1 = 0, \ldots, q_t = 1, \ldots, q_N = 0) \ ,$$

and sends the following encrypted vector to the server.

$$\vec{\mathsf{Enc}}(\boldsymbol{q}) = (\mathsf{Enc}(q_1) \ldots, \mathsf{Enc}(q_N))$$

The server computes

$$c = \bigoplus_{i=1}^{N} \big(v[i] \otimes \mathsf{Enc}(q_i)\big),$$

and sends c to the user. The user computes $\mathsf{Dec}(c)$ and obtains $v[t]$.

Now we consider the case that the server does not leak $v[t]$, but allows the user to access $v[v[t]]$. Our idea is that the server generates a random value $r \in \{0, 1, \ldots, N-1\}$ and returns the cipher text

$$\hat{c} = \bigoplus_{i=1}^{N} \big(((v[i] + r)_{\mathsf{mod}\ N}) \otimes \mathsf{Enc}(q_i)\big) = \mathsf{Enc}((v[t] + r)_{\mathsf{mod}\ N}),$$

where $(a + b)_{\mathsf{mod}\ N}$ denotes addition in a number field modulo N. The user decrypts $\hat{c}$ to know a randomized result $(v[t] + r)_{\mathsf{mod}\ N}$, and performs the next query:

$$\hat{\boldsymbol{q}} = (\hat{q}_1 = 0, \ldots, \hat{q}_{((v[t]+r)_{\mathsf{mod}\ N})} = 1, \ldots, \hat{q}_N = 0).$$

Note that $\hat{\boldsymbol{q}}$ is the r-rotated permutation of the 'true' query:

$$\boldsymbol{q}' = (q'_1 = 0, \ldots, q'_{v[t]} = 1, \ldots, q'_N = 0).$$

Therefore, denote $\mathsf{Perm}(\boldsymbol{q}, r)$ as the permutation of a vector $\boldsymbol{q}$ such that i-th element moves to $((i-r)_{\mathsf{mod}\ N}))$-th position, the server can correctly recover 'true' query $\boldsymbol{q}'$ in its encrypted form by the following permutation: $\vec{\mathsf{Enc}}(\boldsymbol{q}') = \mathsf{Perm}(\vec{\mathsf{Enc}}(\hat{\boldsymbol{q}}), r)$. In this way, the server correctly computes an encrypted $v[t]$-th element by

$$\mathsf{Enc}(v[v[t]]) = \bigoplus_{i=1}^{N} \big(v[i] \otimes \mathsf{Enc}(q'_i)\big),$$

without learning any information about the user's query.

By recursively applying these calculations, the user can obtain x_{k+1} according to Model 1. The complete algorithm implementing this idea is given in Algorithm 1. It uses a function ROT for rotating the servers results to conceal intermediate query results in order to protect the database.

3.3 *Crypto-PBWT*: Privacy-preserving search on genotype databases

In this section, we introduce a practical genotype database search based on recursive oblivious transfer and PBWT. We only introduce the algorithm to search the longest match starting from t-th column, however, variations are possible and would allow for a variety of different search types (see also[7]).

To formulate the problem, let us consider a set X of M haplotype sequences x_i, $i = 1, \ldots, M$ over N genomic positions indexed by $k = 1, \ldots, N$, and a query q which is a user's haplotype sequence over the same N genomic positions. We denote k-th allele of a sequence x_i by $x_i[k]$. Given two indices k_1 and k_2, we say that there is a match between q and x_i from k_1 to k_2, if $q[k_1] \ldots q[k_2 - 1] = x_i[k_1] \ldots x_i[k_2 - 1]$. We say that the match is set-longest at k_1 if there is no match between q and any sequence x_j (possibly with $j = i$) from k_1 to $k_2 + 1$.

The goal is to find a set-longest match at a given position t between q and X in a privacy-preserving manner. Here, we consider the case that the user's private information is the query string and the position t is not the user's private information. We later introduce the case that the both the query string and t are user's private information. The formal description of the model is described as follows:

Model 2 *The user is a private haplotype sequence holder, and the server is a holder of a set of private haplotype sequences. The user learns nothing but a set-longest match at a given position t between the query and the database while the server learns nothing about the user's query. t is not a user's private information and the server knows it.*

Let us remember how to search the set-longest match in *non*-privacy-preserving manner. PBWT involves a matrix $P \in \mathbb{N}^{M \times N}$ that stores well-compressible information in an efficiently searchable form. It is created from the genotype matrix X by algorithms described in [7]. By using rank dictionary operations on P (see below), one can search a match between a query and X. When operating on P one computes updates of intervals using the following two quantities (see[7]for more details): i) The rank dictionary for sequence S for letter $c \in \Sigma$ at position t:

$$\mathsf{Rank}_c(S, t) = |\{\ j \mid S[j] = c\ ,\ 1 \leq j \leq t\ \}|\ ,$$

where Σ is the alphabet of S. ii) The table CF counting occurrences of letters that are lexicographically smaller than c in S by

$$\mathsf{CF}_c(S) = \sum_{r<c} \mathsf{Rank}_r(S, N) \ .$$

Based on these two quantities, we can compute the updates $[f_{k+1}, g_{k+1}]$ using two simple operations

$$f_{k+1} = \mathsf{CF}_c(P_{\cdot,k}) + \mathsf{Rank}_c(P_{\cdot,k}, f_k) \ ,$$

$$g_{k+1} = \mathsf{CF}_c(P_{\cdot,k}) + \mathsf{Rank}_c(P_{\cdot,k}, g_k) \ ,$$

where we denoted the k-th column vector by $P_{\cdot,k}$. Let us define a look-up vector $\boldsymbol{v}_c$ for the column k where

$$\boldsymbol{v}_c[i] = \begin{cases} \mathsf{CF}_c(P_{\cdot,k}) & (i = 0) \\ \mathsf{CF}_c(P_{\cdot,k}) + \mathsf{Rank}_c(P_{\cdot,k}, i) & (1 \leq i \leq M) \end{cases} \tag{1}$$

for $c \in \Sigma$. Then, updating an interval is equivalent to two look-ups in the vector $\boldsymbol{v}_c$:

$$f_{k+1} = \boldsymbol{v}_c[f_k] \quad \text{and} \quad g_{k+1} = \boldsymbol{v}_c[g_k]. \tag{2}$$

Given a position t and a PBWT P of the database sequences, the first match is obtained as an interval $[f_1 = v_c[0], g_1 = v_c[M]]$ where $c = q[1]$ and $\boldsymbol{v}_c$ is a look-up vector for $(t-1)$-th column of P (see the definition of $\boldsymbol{v}_c$ in equation 1). For the case of $t = 1$, a column ${}^t(x_1[0], \ldots, x_M[0])$ is added to P as 0-th column, in order to compute the first match. The match is extended by one letter by an update of the interval. The update from the k-th interval to $(k+1)$-th interval is conducted by specifying $c = q[k+1]$, re-computing v_c for $(k+1)$-th column of P and referring $v_c[f_k]$ and $v_c[g_k]$ as f_{k+1} and g_{k+1} (see (2)). The set-longest-match is found when $f = g$.

In order to achieve the security described in the model 2, for each update, the user has to specify c without leaking c to the server, and obtain only $v_c[f]$ and $v_c[g]$ without leaking f and g. To satisfy the second requirement, the user accesses the server's $\boldsymbol{v}_c$ through the function ROT, which allows the user to obtain a specific element in the specified vector. To achieve the first requirement, the server computes all possible intervals (i.e., computing $[f,\ g]$ for the all case of $c = 1, \ldots, |\Sigma|$). This allows the user to obtain the correct interval, however, the sever leaks extra information (i.e., intervals for $c \neq q[k]$). To avoid this, the user sends $\mathsf{Enc}(q[k])$, and the server adds a conditional randomization factor $r \times (q[k] - c)$ to f and g with different random value r for all $c \in \Sigma$. Note that this factor becomes equivalent to 0 iff. $q[k] = c$, and user only obtains the interval for $c = q[k]$.

In order to identify the set-longest match, the user has to know if $f = g$. The user cannot compute the identity of f and g directly from the server's return, because ROT returns a value which is a random value to the user (but the 'true' value is recovered in encrypted form only at the server side). Therefore, the server also sends an encrypted flag d which shows whether or not $f = g$. Since f and g are represented as indices of $\boldsymbol{q}'_f = \mathsf{Perm}(\boldsymbol{q}_f, r'^{(f)})$ and $\boldsymbol{q}'_g = \mathsf{Perm}(\boldsymbol{q}_g, r'^{(g)})$ (see the functions $\mathsf{PrepQuery}$ and ROT), the server computes d by following:

$$d = \bigoplus_{i=1}^{M} \mathsf{Enc}\big(r_i \times (q'_f[i] - q'_g[i])\big)$$

where r_i is a random value. $\mathsf{Dec}(d)$ is equal to 0 iff. $\boldsymbol{q}_f = \boldsymbol{q}_g$. See Supplementary Algorithm 7 which defines a function $\mathsf{isLongest}$. In addition to finding a set-longest match at t, it is convenient

to find a longest substring which matches to at least ϵ sequences. This operation enables to avoid detecting unique haplotype and provides ϵ-anonymity result and is implemented by replacing the function: $\mathsf{isLongest}$ by another function: $\mathsf{isLongestGT}$ ϵ which computes flags each of which shows if the interval matches to $0, \ldots, \epsilon - 1$ respectively and returns shuffled flags, and the user knows the result by checking if there is a flag which is equal to zero. See Supplementary Algorithm 7 for more details.

The detailed algorithm of *Crypto-PBWT* is shown in Algorithm 2.

3.4 Concealing the Search Position

By the algorithm introduced above, the match position t needs to be provided to the server. Let us consider the case that t needs to be concealed (e.g., the used would not like to reveal which gene is analyzed). In practical genotype database search, it is often sufficient for the user to hide t in a set of multiple columns. Therefore, here we assume following security model.

Model 3 *The user is a private haplotype sequence holder, and the server is a holder of a set of private haplotype sequences. The user has a vector of D positions $T = (t_1, \ldots, t_D)$. The user learns nothing but a set-longest match at a given position $t \in \{t_1, \ldots, t_D\}$ between the query and the database while the server learns nothing about the user's query string. The server knows T but cannot identify which element the user queries.*

Conceptually, the user could query multiple positions at the same time to conceal the search position. In the extreme case the user would query all search positions to avoid leaking any information about t. However, every answered query would leak more information from the database and querying would become computationally prohibitive. We therefore propose joint processing using OT that simultaneously uses multiple search positions. Let us define V_c as another look-up vector for a letter c as follows:

$$V_c[o_j + i] = \begin{cases} \mathsf{CF}_c(P_{\cdot,(t_j+k)}) + o_j & (i = 0) \\ \mathsf{CF}_c(P_{\cdot,(t_j+k)}) + \mathsf{Rank}_c(P_{\cdot,(t_j+k)}, i) + o_j & (i \neq 0) \end{cases}$$
$$(1 \leq j \leq D, 0 \leq i \leq M)$$

where $o_j = (j-1)(M+1)$ is an offset and k is an index which is initialized by -1 and incremented by 1 in each iteration of the recursive search. Note that $(V_c[o_j], \ldots, V_c[o_j + M])$ corresponds to $\boldsymbol{v}_c$ for t_j-th column. The algorithm for the Model 3 is designed by replacing the lookup tables $\boldsymbol{v}_c$ by V_c (see Step 2a, item 1 in Algorithm 2) and initialize f and g by o_x and $o_x + M$, respectively, where $t = t_x$ (see Step 1 in Algorithm 2). As a result the tables get D times larger which has an impact on computing requirements and data transfer size (see Section 3.7). We therefore suggest using this algorithm for small D.

3.5 Reducing Communication Size

As we will describe in the Complexity analysis in the following section, the *Crypto-PBWT* algorithm using standard OT requires $\mathcal{O}(M|\Sigma|)$ in communication size per iteration in the best case, which makes core algorithm less practical. We propose to use an algorithm for sublinear-communication OT (SC-OT) proposed in [29]. Using this approach we can reduce the communication size of *Crypto-PBWT* to $\mathcal{O}(\sqrt{M|\Sigma|})$ (best case). Here, we only outline the key ideas of SC-OT and its

adaptation of *Crypto-PBWT*. In the SC-OT, the one encodes the position t by in a two dimensional representation: $t_0 = t\ /\lceil\sqrt{N}\ \rceil + 1$, $t_1 = (t)_{\mathsf{mod}\ \lceil\sqrt{N}\ \rceil} + 1$, where $\lceil\cdot\rceil$ denotes the ceil of the argument. The user sends $\mathsf{Enc}(t_0)$ and $\vec{\mathsf{Enc}}(\boldsymbol{q})$ to the server, where

$$\vec{\mathsf{Enc}}(\boldsymbol{q}) = (\mathsf{Enc}(q_1 = 0) \ldots, \mathsf{Enc}(q_{t_1} = 1), \ldots, \mathsf{Enc}(q_{\lceil\sqrt{N}\ \rceil} = 0)).$$

The server obtains random values $r_k, k = 1, \ldots, \lceil\sqrt{N}\ \rceil$, and computes

$$c_k = \bigoplus_{i=1}^{\lceil\sqrt{N}\ \rceil} \big(v[(k-1) \times \lceil\sqrt{N}\ \rceil + i] \otimes \mathsf{Enc}(q_i)\big) \oplus \big(r_k \otimes \mathsf{Enc}(t_0 - k)\big),$$

and sends $\boldsymbol{c} = (c_1, \ldots, c_{\lceil\sqrt{N}\ \rceil})$ to the user. The user knows the result by the decryption: $\mathsf{Dec}(c_{t_0})$. Note that $\mathsf{Enc}(t_0 - k) = \mathsf{Enc}(0)$ iff. $t_0 = k$, therefore the decryption of c_i becomes a random value when $i \neq t_0$.

In order to apply bit-rotation technique naturally to SC-OT, the server needs to return $v[t]$ in the same two dimensional representation. The key idea here is that the server creates $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ where $v_0[i] = v[i]/\lceil\sqrt{N}\ \rceil + 1$ and $v_1[i] = (v[i])_{\mathsf{mod}\ \lceil\sqrt{N}\ \rceil} + 1$, $i = 1, \ldots, N$, and searches on both $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$. Similar to the linear communication size function ROT, the removable random factors are added to server's returns. More details on SC-OT is given in Section A. The complete algorithm for privacy-preserving search based on SC-OT is given in Supplementary and Algorithm 4.

3.6 An Exhaustive Baseline Algorithm

Since there is no obvious candidate to compare our algorithm to, we briefly describe a baseline algorithm based on exhaustive enumeration of k-mers. In order to identity the match, the user queries the server about the presence of a k-mer. Here, the server stores all k-mers, there are $\mathcal{O}(|\Sigma|^k)$ of them, and we use SC-OT. Such a strategy is efficient for short queries as $|\Sigma|^k$ is not too large. However, the resource requirements will be dominated by queries for large k and the algorithm quickly gets intractable.

3.7 Complexity

Most of the computing and transfer on server and user side is related to the encryption/decryption and the computational cost of the search is negligible. While PBWT requires essentially $\mathcal{O}(1)$ to update the intervals per iteration, *Crypto-PBWT* needs to conceal the query and requires $M|\Sigma|$ operations on the server, where M is the number of sequences in the database and $|\Sigma|$ is the size of the alphabet. When multiple queries are performed at the same time, i.e., $D > 1$, the effort increases linearly in D, i.e., the server sides compute effort is $\mathcal{O}(MD|\Sigma|)$ per iteration. When using SC-OT, the communication size and effort for the user is $\mathcal{O}(\sqrt{MD|\Sigma|})$ (see Section 3.5 and Supplementary Section A for details).

Table 1 summarizes the time, data transfer overhead and space complexities of the *Crypto-PBWT*, when the server's PBWT is $M \times N$ matrix consisting of a set of alphabet letters Σ and the user's query length is ℓ and the number of queries positions is D (including $D - 1$ decoy positions; see Section 3.4 for details). For the purpose of comparison, we consider the method outlined in Section 3.6 that achieves the same security and utility as *Crypto-PBWT*. Since the complexity of the exhaustive approach is exponential to the query length, its performance deteriorates quickly for

long matches. On the other hand, the time and data transfer overhead complexity of the *Crypto-PBWT* are linear and sub-linear to the query length, which enables the user to find a long match efficiently.

Table 1: The summary of the time, communication and space complexities of *Crypto-PBWT* (CP) and an exhaustive method (EX). Both algorithms use SC-OT. M is the number of haplotype sequences (server side), D is the number of queried positions (including $D-1$ decoy position to conceal the query position), ℓ is the length of query and $|\Sigma|$ is the alphabet size.

	Time	Communication	Space
CP (user)	$O(\ell\sqrt{MD\|\Sigma\|})$	$O(\ell\sqrt{MD\|\Sigma\|})$	$O(\sqrt{MD\|\Sigma\|})$
CP (server)	$O(\ell MD\|\Sigma\|)$	$O(\ell\sqrt{MD\|\Sigma\|})$	$O(MD\|\Sigma\|)$
EX (user)	$O(\sqrt{D\|\Sigma\|^{\ell}})$	$O(\sqrt{D\|\Sigma\|^{\ell}})$	$O(\sqrt{D\|\Sigma\|^{\ell}})$
EX (server)	$O(D\|\Sigma\|^{\ell})$	$O(\sqrt{D\|\Sigma\|^{\ell}})$	$O(D\|\Sigma\|^{\ell})$

3.8 Security Notion

In this paper, we assume the security model called *Semi-honest model* where both parties follow the protocol, but an adversarial one attempts to infer additional information about the other party's secret input from the legally obtained information. The semantic security of the encryption scheme used in the protocol (see Section 3.1) implies immediately that the server cannot infer any information about the user's query q during the protocol. Also, the user can not infer any information about server's return except for the result.

We do not consider *Malicious model* where an adversarial party cheats even in the protocol (e.g., by inputting maliciously chosen invalid values) in order to illegally obtaining additional information about the secret, however, we would like to mention that it is possible to design an algorithm for the Malicious model with a small modification by using a known cryptographic technique [26].

4 Experiments

In this section, we evaluate the performance of the proposed method on the datasets created from the chromosome 1 data from the 1,000 Genomes Project phase 1 data release which consists of 2,184 haploid genomes [27]. In our experiments and as in [7], we used alleles having SNPs, but we did consider indel variants.

We implemented the proposed algorithm in C++ based on an open source C++ library of *elliptic curve ElGamal encryption* [1]. We used the standard parameters called *secp192k1*, according to the recommendation by The Standards for Efficient Cryptography Group. For comparison, we also implemented an exhaustive baseline method (see Section 3.6) that achieves the same security and utility as *Crypto-PBWT*. In order to perform a fair comparison, both *Crypto-PBWT* and the exhaustive method used the same SC-OT module where computation of c_k (see Algorithm 1) is simply parallelized by OpenMP. Since our implementation is a prototype, it does not support communication over the network, instead, the data is transferred by file I/O which is also included in run time.

In the first experiment, we used all the 2,184 genomes of original data and the user selected a true start position together with 49 decoys. In this experiment, both *Crypto-PBWT* and the exhaustive method were run on the same laptop computer equipped with Intel Core(TM) i7 3.00GHz CPU (four cores with hyper-threading) and 16GB memory. The user side programs used a single

thread while the server side programs used eight threads. Figures 3 and 4 show run time and data transfer overhead of *Crypto-PBWT* and of the exhaustive method. The observed run time and data transfer size of *Crypto-PBWT* is linear in the query length, while that of the exhaustive approach is exponential. For query lengths larger than 30 bit, the computation of the exhaustive method did not finish within 24 h. These results fit the theoretical complexity described in Section 3.7.

The user's run time of *Crypto-PBWT* is relatively small making it suitable for a practical case where computation power in a server side is generally stronger than that of user side. Since the memory usage of *Crypto-PBWT* does not depend on query length, it used less than 60 MB while that of the exhaustive method exponentially increases according to the query length and required 12 GB when the query length is 25 bit.

Although the exhaustive method is efficient for short queries, we consider that *Crypto-PBWT* is more practical when taking into account that the bit length of a unique substring for a human genome is greater than 31 bits. Moreover, since there are large linkage blocks, even queries with more than 100 bits would not always lead to unique matches in the 1,000 genomes data. Hence, the exhaustive search strategy would either not always be able to return a unique match or would be very inefficient. The proposed iterative privacy-preserving technique is efficient also for long queries.

In the second experiment, we evaluated the performance of the run time of *Crypto-PBWT* on a laptop with one CPU socket (4 cores with hyper-threading) and a compute node equipped with four CPU sockets (Intel Xeon 2.40GHz CPU; total of 32 cores with hyper-threading). In this experiment, all the 2,184 genomes were also used and the user selected search position of interest together with 0, 4, 9, 14 and 49 additional decoy positions (see Section 3.4 for details). Although the current implementation is a prototype and there is room for improvement in terms of parallelization, the server's run time was at an acceptable level in practical configurations (Table 2). We note, that with improvements in parallelization, the server run time may be reduced to 3-4 seconds.

5 Conclusion

In this paper, we have proposed a novel approach for searching genomic sequences in a privacy-preserving manner. Our approach combines an efficient data structure that can be used for recursive

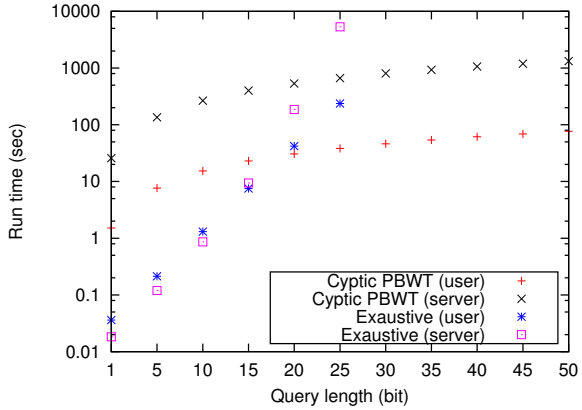

Figure 3: Run time of *Crypto-PBWT* and the exhaustive method on 2,184 aligned haploid genomes on a laptop with 4 cores. The user selected 49 decoy positions for concealing the true query position. The server used 8 threads while the user used a single thread.

Table 2: The run time of a typical query with *Crypto-PBWT* on M=2,184 aligned haploid genomes on a laptop with 4 cores and a compute node with up to 32 cores (both with hyper-threading) and a query length of 25 SNP positions. Wall time includes server (89%) and user time (11%). D is the number of positions queried simultaneously to conceal the query position (if required). The current prototype implementation supports only a limited degree of parallelization and we observe a close to linear speedup only for large D and a relatively small number of cores.

Compute System	Laptop	Compute node			
Parallel Compute Cores	4	4	8	16	32
Run time (sec) with $D = 1$	22	27	20	17	16
Run time (sec) with $D = 5$	78	88	57	47	36
Run time (sec) with $D = 10$	144	161	98	69	54
Run time (sec) with $D = 20$	283	306	180	118	89
Run time (sec) with $D = 50$	698	757	425	277	190

search and a novel approach for recursive oblivious transfer. It achieves high utility and has strong security features and requires acceptable compute and communication resources.

The developed novel algorithm can find the longest match between a query and a large set of aligned genomic sequences indexed by PBWT. We implemented our algorithm and tested on the dataset created from the 1,000 Genomes Project data [27]. Compared to an exhaustive baseline approach, our algorithm, named *Crypto-PBWT*, was orders of magnitude more efficient both in run time and data transfer overhead for practical query sizes. When the prototype program was run on a laptop machine, the combined user's and server's run time was 22 sec for searching on 2,184 genomes without concealing the query position. Searches with with concealed query position using a compute node took between 36 and 190 seconds depending on the level of privacy.

As the original data structure supports many useful search options such as wild card search and set maximal search, *Crypto-PBWT* could also support those options by using the same techniques used in the original structures in combination with cryptographic techniques, including OT. Moreover, the approach could be easily applied for BWT and has a potential to be applied for other recursively searchable data structures.

To the best of our knowledge, the proposed algorithm is the first that is allows set-maximal search of genomic sequences in a privacy-preserving manner for user and database. We note that

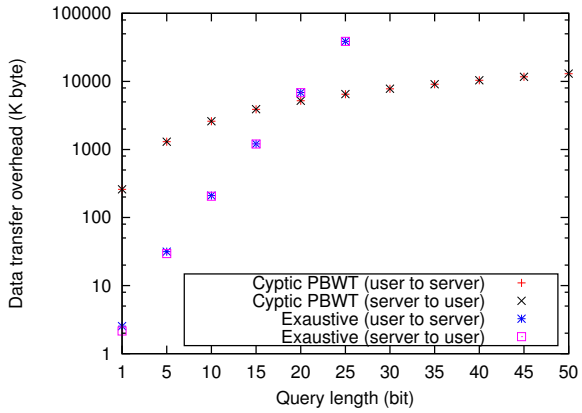

Figure 4: Data transfer overhead of *Crypto-PBWT* and the exhaustive method on 2,184 aligned haploid genomes on a laptop with 4 cores. The user selected 49 decoy positions for concealing the true query position.

the implementation can still be improved and the overall run time can likely be reduced to not more than a few seconds per query. This would make it practical to use our approach in a genomic Beacon (see GA4GH's Beacon Project) that would allow the privacy-preserving search for combinations of variants. It also appears practical to use our approach to enable the search by a user that has access to his/her genomic sequence and would like to query the database, for instance, for information related to disease risk without sharing this information with anybody. Finally, the algorithm can also be used to facilitate sharing of genetic information across institutions and countries in order to identify large enough cohorts with a similar genetic backgrounds. This is in spirit of the mission of the Global Alliance for Genome and Health.

Acknowledgement

We are thankful to Stephanie Hyland for proof-reading the manuscript. We would also like to acknowledge an encouraging discussion by Richard Durbin. We gratefully acknowledge funding from AIST (to K.S.), Memorial Sloan Kettering Cancer Center (to G.R.) and NIH (grant 1R01CA176785-01A1). This study was also supported by the Japan-Finland Cooperative Scientific Research Program of Japan Science and Technology Agency (JST; to K.S.).

A The sublinear communication size recursive oblivious transfer

In this section, we describe the detailed algorithm of the sublinear communication size recursive oblivious transfer. In Section 3.2, we introduced the bit-rotation technique for the case of the linear communication size oblivious transfer. As mentioned in Section 3.5, the same technique is also applied for the $O(\sqrt{N})$-communication size oblivious transfer (SC-OT).

A.1 The sublinear communication size oblivious transfer

Let us review the SC-OT algorithm. In the SC-OT, the one encodes the position t by in a two dimensional representation: $t_0 = t\ /\lceil\sqrt{N}\ \rceil + 1$, $t_1 = (t)_{\mathsf{mod}\ \lceil\sqrt{N}\ \rceil} + 1$, where $\lceil\cdot\rceil$ denotes the ceil of the argument. The user sends $\mathsf{Enc}(t_0)$ and $\vec{\mathsf{Enc}}(\boldsymbol{q})$ to the server, where

$$\vec{\mathsf{Enc}}(\boldsymbol{q}) = (\mathsf{Enc}(q_1 = 0)\ldots, \mathsf{Enc}(q_{t_1} = 1), \ldots, \mathsf{Enc}(q_{\lceil\sqrt{N}\ \rceil} = 0)).$$

The server obtains random values r_k for $k = 1, \ldots, \lceil\sqrt{N}\ \rceil$, and computes

$$c_k = \bigoplus_{i=1}^{\lceil\sqrt{N}\ \rceil} \big(v[(k-1) \times \lceil\sqrt{N}\ \rceil + i] \otimes \mathsf{Enc}(q_i)\big) \oplus \big(r_k \otimes \mathsf{Enc}(t_0 - k)\big),$$

and sends $\boldsymbol{c} = (c_1, \ldots, c_{\lceil\sqrt{N}\ \rceil})$ to the user. The user knows the result by the decryption: $\mathsf{Dec}(c_{t_0})$. Note that $\mathsf{Enc}(t_0 - k) = \mathsf{Enc}(0)$ iff. $t_0 = k$, therefore the decryption of c_i becomes a random value when $i \neq t_0$. See the function SCOT in Algorithm 3 for detailed description.

A.2 Bit-rotation technique for the sublinear communication size oblivious transfer

In order to apply bit-rotation technique naturally to SC-OT, the server needs to return $v[t]$ in the same two dimensional representation. The key idea here is that the server creates $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ where

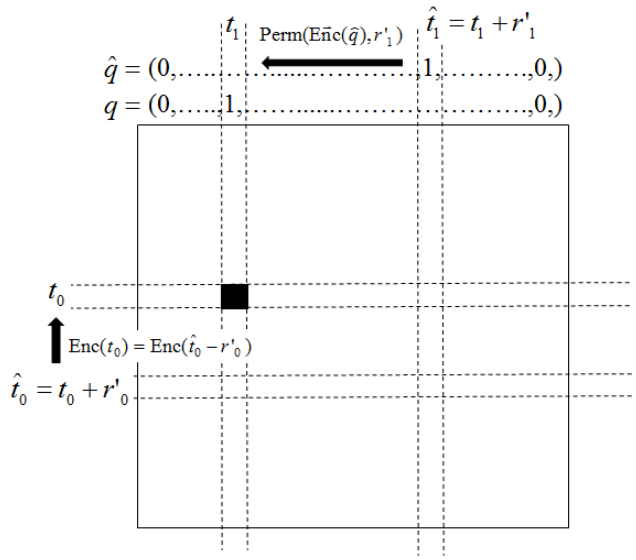

Figure 5: The illustration of the remove of random factors in the server side. $\boldsymbol{q}$ and t_0 show the plain text of the user's 'true' query while $\hat{\boldsymbol{q}}$ and $\hat{t_0}$ show the plain text of the user's query. The server recovers correct t_1 by computing -r'_1 rotated permutation of the server's query $\hat{\boldsymbol{q}}$. It also recovers correct t_0 by the homomorphic encryption: $\mathsf{Enc}(\hat{t}_0 - r'_0)$.

$v_0[i] = v[i]/\lceil\sqrt{N}\,\rceil + 1$ and $v_1[i] = (v[i])_{\mathsf{mod}\,\lceil\sqrt{N}\,\rceil} + 1$, $i = 1, \ldots, N$, and searches on both $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$. We designed the server's function SCROT which adds the removable random factors to the server's returns based on the approach similar to the function ROT for the linear communication size algorithm. SCROT which is described in Algorithm 3 takes nine arguments: user's query $\mathsf{Enc}(\hat{t_0})$, $\vec{\mathsf{Enc}}(\hat{\boldsymbol{q}}) = (\hat{q}_1 = 0, \ldots, \hat{q}_{(\hat{t_1})_{\mathsf{mod}\,L_1}} = 1, \ldots, \hat{q}_N = 0)$, a target vector $\boldsymbol{v}_x$ ($x \in \{0, 1\}$), a random value r for randomizing the result, random values r'_0 and r'_1 which were used for randomizing 'true' values t_0 and t_1 (i.e., $\hat{t_0} = t_0 + r'_0$ and $\hat{t_1} = (t_1 + r'_1)_{\mathsf{mod}\,L_1}$) and row length L_0 and column length L_1 of the two dimensional representation (i.e., $L_0 = L_1 = \lceil\sqrt{N}\,\rceil$ for this case). Figure 5 illustrates the server process for removing random factors previously added to the server's return. Since $\mathsf{Enc}(\hat{t_0} - r'_0)$ causes the position shift from $\hat{t_0}$ to $(\hat{t_0} - r'_0)_{\mathsf{mod}\,L_0}$ in server's return $\boldsymbol{c}$, the server also needs another permutation $\mathsf{Perm}(\boldsymbol{c},\ -r'_0)$ before returning the result. See Algorithm 3 for detailed description. By this function SCROT, the server can add *removable* random factor to the result, and therefore it enables user to search $\boldsymbol{v}$ recursively.

A.3 Solving the problem caused by the ambiguity of the server's results

In the function SCROT, the server generates random value r and conducts randomization by:

$$(v[i] + r)_{\mathsf{mod}\,\lceil\sqrt{N}\,\rceil},$$

and returns $\mathsf{Enc}((v[i] + r)_{\mathsf{mod}\,\lceil\sqrt{N}\,\rceil})$ to the user.

Since the modulo operation yields different results for the same r according to the two conditions:

$$v[i] + r \leq \lceil\sqrt{N}\,\rceil$$

and

$$v[i] + r > \lceil \sqrt{N} \rceil \ ,$$

and neither the user nor the server knows which condition is applied (note that the user's choice $v[i]$ and server's random value are their private information), the server needs to return two results assuming both conditions in the next round. For the case of computing $\mathsf{Enc}(t_0)$, the sever needs to compute both

$$\boldsymbol{c}_0 \leftarrow \mathsf{SCROT}(\mathsf{Enc}(t_0), \vec{\mathsf{Enc}}(\boldsymbol{q}), \boldsymbol{v}_0, r_0, \lceil \sqrt{N} \rceil, r'_0, r'_1, \lceil \sqrt{N} \rceil, \lceil \sqrt{N} \rceil)$$

and

$$\boldsymbol{c}'_0 \leftarrow \mathsf{SCROT}(\mathsf{Enc}(t_0), \vec{\mathsf{Enc}}(\boldsymbol{q}), \boldsymbol{v}_0, r_0, \lceil \sqrt{N} \rceil, (r'_0 - \lceil \sqrt{N} \rceil), r'_1, \lceil \sqrt{N} \rceil, \lceil \sqrt{N} \rceil) \ .$$

Since only one of c_{0,t_0} and c'_{0,t_0} becomes an encryption of a correct result and the other becomes an encryption of a random value, user is able to obtain the next t_0 by checking if $1 \leq \mathsf{Dec}(c_{0,t_0}) \leq \lceil \sqrt{N} \rceil$ or $1 \leq \mathsf{Dec}(c'_{0,t_0}) \leq \lceil \sqrt{N} \rceil$ (see the function: $\mathsf{ChooseDec}$ in Algorithm 3). In similar way, the user also obtains t_1. Algorithm 4 shows the full description of sublinear communication size recursive oblivious transfer algorithm taking into account of the above problem.

B The sublinear communication algorithm for $Crypto - PBWT$

In Section 3.3, the linear size communication algorithm for $Crypto - PBWT$ is introduced. Here we introduce the sublinear communication size algorithm by adapting $SC - ROT$ to the search by $PBWT$. The goal is to find a set-longest match at a given position t between a query S and a set of genotype sequences X in a privacy-preserving manner. in this section, we consider that both t and S are private information and use the following model which is the same model as Model 3 in Section 3.4.

Model 4 *The user is a private haplotype sequence holder, and the server is a holder of a set of private haplotype sequences. The user has a vector of D positions $T = (t_1, \ldots, t_D)$. The user learns nothing but a set-longest match at a given position $t \in \{t_1, \ldots, t_D\}$ between the query and the database while the server learns nothing about the user's query string. The server knows T but cannot identify which element the user queries.*

Similar to the linear size communication algorithm for $Crypto - PBWT$, the server creates $\boldsymbol{v}_c$ which is a look-up vector for a letter c as follows:

$$v_c[o_j + i] = \begin{cases} \mathsf{CF}_c(P_{\cdot,(t_j+k)}) + o_j & (i = 0) \\ \mathsf{CF}_c(P_{\cdot,(t_j+k)}) + \mathsf{Rank}_c(P_{\cdot,(t_j+k)}, i) + o_j & (i \neq 0) \end{cases}$$
$$(1 \leq j \leq D, 0 \leq i \leq M)$$

where $o_j = (j-1)(M+1)$ is an offset and k is an index which is initialized by -1 and incremented by 1 in each iteration of recursive search. In order to search $\boldsymbol{v}^{(c)}$ by SC-ROT, the server converts each element in $\boldsymbol{v}^{(c)}$ into the two dimensional representation and stores them in $\boldsymbol{v}_0^{(c)}$ and $\boldsymbol{v}_1^{(c)}$. In addition, we designed an algorithm such that a single SC-ROT is conducted for the search of all letter tables in stead of conducting SC-ROT for each letter table in order to minimize communication size. To this end, all the letter tables $\boldsymbol{v}^c$ for $c \in \Sigma$ are concatenated into one single vector. When updating the interval to extend matches by a letter $S[i]$, the user needs to specify the region of

the single vector, which corresponds to a letter table $\boldsymbol{v}^{(S[i])}$. In our algorithm, we designed row length L_0 and column length L_1 for the two dimensional representation (L_0 and L_1 are not the matrix size of PBWT) such that elements of the same position in the different letter tables should be placed in the same column after concatenating all the tables (i.e., $(i)_{\mathsf{mod}\ L_1} = (i + |\boldsymbol{v}_0^{(1)}|)_{\mathsf{mod}\ L_1} = (i + |\boldsymbol{v}_0^{(1)}| + |\boldsymbol{v}_0^{(2)}|)_{\mathsf{mod}\ L_1}, \ldots, = (i + \sum_{c\in\{1,\ldots,|\Sigma|-1\}} |\boldsymbol{v}_0^{(c)}|)_{\mathsf{mod}\ L_1}$) in order that the user can specify the letter table by controlling an offset of row value of the query. For this purpose, the server configures $L_1 = \sqrt{D(M+1)|\Sigma|}$, an offset unit size $L_0' = D(M+1)/L_1 + 1$, $L_0 = L_0' \times |\Sigma|$, and $|\boldsymbol{v}_0^{(1)}| =, \ldots, = |\boldsymbol{v}_0^{(|\Sigma|)}| = |\boldsymbol{v}_1^{(1)}| =, \ldots, = |\boldsymbol{v}_1^{(|\Sigma|)}| = L_0' L_1$. The server creates vectors $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ by the concatenations $\boldsymbol{v}_0 = \boldsymbol{v}_0^{(1)}, \ldots, \boldsymbol{v}_0^{(|\Sigma|)}$ and $\boldsymbol{v}_1 = \boldsymbol{v}_1^{(1)}, \ldots, \boldsymbol{v}_1^{(|\Sigma|)}$. Figure 6 is a graphical view of the rearrangement of $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$.

Now the user is able to search $\boldsymbol{v}^{(c)}$ recursively in an oblivious manner by using SC-ROT. In PBWT, the match is reported as an interval $[f, g]$ and the number of matches is equivalent to $g - f$. Since the user wants to start the search from t_x-th column on PBWT, user initialized f and g by $f = o_x,\ g = o_x + M$ where $o_j = (j-1)(M+1)$ and computes two dimensional representation of them: $f_0 = f/L_1 + 1,\ f_1 = (f)_{\mathsf{mod}\ L_1} + 1,\quad g_0 = g/L_1 + 1,\ g_1 = (g)_{\mathsf{mod}\ L_1} + 1$. Then the user recursively searches $\boldsymbol{v}^{(c)}$ for updating f and g until he/she finds the match. For the i-th round of the recursive search, meaning that the user updates the interval for finding matches ending with $S[i]$, he/she adds an offset to f_0 and g_0 in order to specify $S[i]$ by: $f_0 \leftarrow f_0 + (S[i]-1)L_0'$, $g_0 \leftarrow g_0 + (S[i]-1)L_0'$. For each round, the server also computes an encrypted flag whose plain text is equal to 0 iff. $f = g$. The detailed description of this part is described in the function $\mathsf{isSCLongest}$ in Algorithm 5. Finally, the user learns the set-longest match at t by $\mathsf{Dec}(d)$. In order to hide the length of the set-longest match to the server, the user keep sending decoy queries until it reaches to ℓ-th round. Algorithm 6 and Algorithm 5 show a detailed algorithm of *Crypto-PBWT*.

B.1 Modification of SCOT and SCROT

Due to the randomization method described in Section A, the server computes SC-ROT with two configurations for each update of f_0, f_1, g_0 and g_1. For the case of updating f_0, the sever needs to computes both:

$$\mathsf{SCROT}(\mathsf{Enc}(f_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_f), \boldsymbol{v}_0, r_0^{(f)}, L_0', r_0'^{(f)}, r_1'^{(f)}, L_0, L_1),$$

and

$$\mathsf{SCROT}(\mathsf{Enc}(f_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_f), \boldsymbol{v}_0, r_0^{(f)}, L_0', (r_0'^{(f)} - L_0'), r_1'^{(f)}, L_0, L_1)\ .$$

Since the look-up table $\boldsymbol{v}_0$ is a concatenation of all the letter tables $\boldsymbol{v}_0^{(c)}$ for $c = \{1, \ldots, |\Sigma|\}$, the configuration which does not match the user's true query leaks an unnecessary element of $\boldsymbol{v}_0$. If the user's true request is t, the server leaks either $v_0[t + L_0' L_1]$ or $v_0[t - L_0' L_1]$.

To avoid leaking extra information, we slightly generalize the SCOT algorithm to support additional filtering of the result, taking a hash function H and a hash value for auxiliary filtering γ (see the function $\mathsf{SCOT}^\dagger$ in Algorithm 5 for details). According to the modification of the function SCOT, the function SCROT also takes H and γ as arguments (see the function $\mathsf{SCROT}^\dagger$ in Algorithm 5 for details). Algorithm 6 shows full description of $Crypto-PBWT$ with auxiliary filtering.

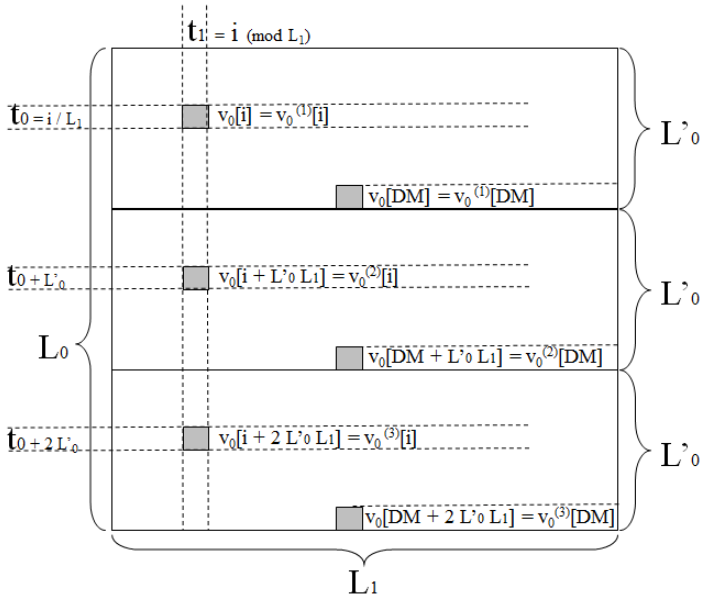

Figure 6: The arrangement of elements of $\boldsymbol{v}_0$ when $\Sigma = \{1, 2, 3\}$. The length of $\boldsymbol{v}_0^{(c)}$ for $c \in \Sigma$ is designed such that $v_0^{(1)}[i]$, $v_0^{(2)}[i]$ and $v_0^{(3)}[i]$ are aligned in the same column after the concatenation. The elements of $\boldsymbol{v}_1$ is also arranged in the same manner.

References

[1] C++ library implementing elliptic curve elgamal crypto system [8]. `https://github.com/aistcrypt/Lifted-ElGamal`, 2015. URL accessed April 13, 2015.

[2] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myer, and David J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[3] Erman Ayday, Jean Louis Raisaro, Urs Hengartner, Adam Molyneaux, and Jean-Pierre Hubaux. Privacy-preserving processing of raw genomic data. In *Data Privacy Management and Autonomous Spontaneous Security - 8th International Workshop, DPM 2013, and 6th International Workshop, SETOP 2013, Egham, UK, September 12-13, 2013, Revised Selected Papers*, pp. 133–147, 2013.

[4] Pierre Baldi, Roberta Baronio, Emiliano De Cristofaro, Paolo Gasti, and Gene Tsudik. Countering GATTACA: efficient and secure testing of fully-sequenced human genomes. In *Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011, Chicago, Illinois, USA, October 17-21, 2011*, pp. 691–702, 2011.

[5] Marina Blanton and Mehrdad Aliasgari. Secure outsourcing of DNA searching via finite automata. In *Data and Applications Security and Privacy XXIV, 24th Annual IFIP WG 11.3 Working Conference, Rome, Italy, June 21-23, 2010. Proceedings*, pp. 49–64, 2010.

[6] Fons Bruekers, Stefan Katzenbeisser, Klaus Kursawe, and Pim Tuyls. Privacy-preserving matching of dna profiles. *IACR Cryptology ePrint Archive*, 2008:203, 2008.

[7] Richard Durbin. Efficient haplotype matching and storage using the positional burrows-wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272, 2014.

[8] T ElGamal. A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms. *IEEE Transactions on Information Theory*, 31(4):469–472, 1985.

[9] Yaniv Erlich and Arvind Narayanan. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, 15:409–421, 2014.

[10] Paolo Ferragina and Giovanni Manzini. Indexing compressed text. *Journal of the ACM (JACM)*, 52(4):553, 2005.

[11] Michael J. Freedman, Yuval Ishai, Benny Pinkas, and Omer Reingold. Keyword search and oblivious pseudorandom functions. In *Theory of Cryptography, Second Theory of Cryptography Conference, TCC 2005, Cambridge, MA, USA, February 10-12, 2005, Proceedings*, pp. 303–324, 2005.

[12] Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *J. Comput. Syst. Sci.*, 28(2):270–299, 1984.

[13] G. J. Jacobson. *Succinct static data structures*. PhD thesis, Carnegie Mellon University Pittsburgh, PA, USA, 1988. ACM Order number AAI8918056.

[14] Somesh Jha, Louis Kruger, and Vitaly Shmatikov. Towards practical privacy for genomic computation. In *IEEE Symposium on Security and Privacy*, pp. 216–230, 2008.

[15] W. James Kent. BLAT - the BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, 2002.

[16] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[17] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[18] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

[19] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–7, 2009.

[20] Helger Lipmaa. An oblivious transfer protocol with log-squared communication. In *Information Security, 8th International Conference, ISC 2005, Singapore, September 20-23, 2005, Proceedings*, pp. 314–328, 2005.

[21] Helger Lipmaa. New communication-efficient oblivious transfer protocols based on pairings. In *Information Security, 11th International Conference, ISC 2008, Taipei, Taiwan, September 15-18, 2008. Proceedings*, pp. 441–454, 2008.

[22] Ken Naganuma, Yoshino Masayuki, and Hisayoshi Sato. Private string search using the block-sorting algorithm. In *The proceedings of SCIS 2012 (In Japanese)*, 2012.

[23] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *Proceedings of the 17th international conference on Theory and application of cryptographic techniques*, EUROCRYPT'99, pp. 223–238, Prague, Czech Republic, 1999.

[24] Michael O. Rabin. How to exchange secrets with oblivious transfer. Technical Report 81, Harvard University, 1981.

[25] Patricia A. Roche and George J. Annas. Protecting genetic privacy. *Nature Reviews Genetics*, 2:392–396, 2001.

[26] Yusuke Sakai, Keita Emura, Goichiro Hanaoka, Yutaka Kawai, and Kazumasa Omote. Methods for restricting message space in public-key encryption. *IEICE Transactions*, 96-A(6):1156–1168, 2013.

[27] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, November 2012.

[28] Xun Yi, Russell Paulet, and Elisa Bertino. *Homomorphic Encryption and Applications*. Springer Briefs in Computer Science. Springer, 2014.

[29] Bingsheng Zhang, Helger Lipmaa, Cong Wang, and Kui Ren. Practical fully simulatable oblivious transfer with sublinear communication. In *Financial Cryptography and Data Security - 17th International Conference, FC 2013, Okinawa, Japan, April 1-5, 2013, Revised Selected Papers*, pp. 78–95, 2013.

Algorithm 1 Recursive oblivious transfer

function $\mathsf{PrepQuery}(t,\, N)$
  $\boldsymbol{q} = (q_1 = 0, \ldots, q_t = 1, \ldots, q_N = 0)$
  $\vec{\mathsf{Enc}}(\boldsymbol{q}) = (\mathsf{Enc}(q_1), \ldots, \mathsf{Enc}(q_N))$
  return $\vec{\mathsf{Enc}}(\boldsymbol{q})$
end function

function $\mathsf{ROT}(\vec{\mathsf{Enc}}(\hat{\boldsymbol{q}}),\, \boldsymbol{v},\, r,\, r',\, N)$
  $\vec{\mathsf{Enc}}(\boldsymbol{q}') = \mathsf{Perm}(\vec{\mathsf{Enc}}(\hat{\boldsymbol{q}}), r')$
  $\hat{c} = \bigoplus_{i=1}^{N} \big(((v[i] + r)_{\mathsf{mod}\ N} \otimes \mathsf{Enc}(q'_i)\big)$
  return $\hat{c}$
end function

$\boldsymbol{v}$ is a server's private vector of length N.
x_1 is a user's private value.
x_ℓ is the value of user's interest.
ℓ is known to both user and server.
User's initialization: $t \leftarrow x_1$
Server's initialization: $r' \leftarrow 0$
Common initialization: $i \leftarrow 1$
while $i < \ell$ **do**
  The user computes: $\vec{\mathsf{Enc}}(\boldsymbol{q}) \leftarrow \mathsf{PrepQuery}(t, N)$
  if $i == (\ell - 1)$ **then**
    Server sets: $r = 0$
  else
    Server generates random value r
  end if
  Server computes: $\hat{c} \leftarrow \mathsf{ROT}(\vec{\mathsf{Enc}}(\boldsymbol{q}),\, \boldsymbol{v},\, r,\, r',\, N)$
  Server sets: $r' \leftarrow r$
  Server sends $\hat{c}$ to user
  User computes: $t \leftarrow \mathsf{Dec}(\hat{c})$
end while
User obtains $x_\ell = t$.

Algorithm 2 The detailed description of *Crypto-PBWT* finding a set-longest match at position t.

- Public input: The length of column M, the length of row N and a set of alphabet letters $\Sigma = \{1, 2, .., |\Sigma|\}$, the position $t \in \{1, \ldots, N\}$ at which the search starts
- Private input of user: A query sequence S of length ℓ
- Private input of server: PBWT matrix $P \in \mathbb{N}^{M \times N}$

0. *(Key setup of cryptosystem)* User generates key pair $(\mathsf{pk}, \mathsf{sk})$ by key generation algorithm KeyGen for additive-homomorphic cryptosystem and sends public key pk to server (only user knows secret key sk).
1. *(User initialization)* Set initial interval $[f, g]$ by $f = 0,\ g = M$.
2. *(Recursive search)*
   Initializes query and position index: $i \leftarrow 1$; $k \leftarrow t - 1$
   while $(i \leq \ell)$ **do**
   (a) *(Query entry)* The user performs the following steps:
   - Prepares next query:
     $\vec{\mathsf{Enc}}(\boldsymbol{q}_f) \leftarrow \mathsf{PrepQuery}(f, M)$
     $\vec{\mathsf{Enc}}(\boldsymbol{q}_g) \leftarrow \mathsf{PrepQuery}(g, M)$
   - Sends $\mathsf{Enc}(S[i]),\ \vec{\mathsf{Enc}}(\boldsymbol{q}_f),\ \vec{\mathsf{Enc}}(\boldsymbol{q}_g)$ to the server.

   (b) *(Search)* The server performs the following steps:
   - Compute look-up tables for all $c \in \Sigma$:

$$\boldsymbol{v}_c[j] = \begin{cases} \mathsf{CF}_c(P_{\cdot,k}) & (j = 0) \\ \mathsf{CF}_c(P_{\cdot,k}) + \mathsf{Rank}_c(P_{\cdot,k}, j) & (1 \leq j \leq M) \end{cases}$$

   - Obtain random values $r^{(f)},\ r^{(g)}$
   - Set $r'^{(f)} = r'^{(g)} = 0$ iff. $i == 0$
   - Compute next possible intervals for all $c \in \Sigma$:
     $e_c^{(f)} \leftarrow \mathsf{ROT}\big(\vec{\mathsf{Enc}}(\boldsymbol{q}_f), \boldsymbol{v}_c, r^{(f)}, r'^{(f)}, M\big)$
     $e_c^{(g)} \leftarrow \mathsf{ROT}\big(\vec{\mathsf{Enc}}(\boldsymbol{q}_g), \boldsymbol{v}_c, r^{(g)}, r'^{(g)}, M\big)$
   - Randomize return values except for user's target interval by computing followings for all $c \in \Sigma$
     Generate temporary random values r_0, r_1
     $e_c^{(f)} \leftarrow e_c^{(f)} \oplus \mathsf{Enc}\big(r_0 \times (S[i] - c)\big)$
     $e_c^{(g)} \leftarrow e_c^{(g)} \oplus \mathsf{Enc}\big(r_1 \times (S[i] - c)\big)$
   - Compute an encrypted flag showing if match is longest
     $d \leftarrow \mathsf{isLongest}\ \big(\vec{\mathsf{Enc}}(\boldsymbol{q}_f),\ \vec{\mathsf{Enc}}(\boldsymbol{q}_g),\ r'^{(f)},\ r'^{(g)}\big)$
   - Store random values $r'^{(f)} \leftarrow r^{(f)},\ r'^{(g)} \leftarrow r^{(g)}$
   - Send d, $e^{(f)}$, $e^{(g)}$ to the user

   (c) *(Decryption of encrypted flag and randomized interval)* The user performs the following steps:
   if $(\mathsf{Dec}(d) == 0)$
   Sends decoy queries to server until $i == \ell$
   Reports result $S[1, \ldots, i-2]$
   else
   Computes $f \leftarrow \mathsf{Dec}(e_{S[i]}^{(f)}),\ g \leftarrow \mathsf{Dec}(e_{S[i]}^{(g)})$,
   end if

   $i \leftarrow i + 1, \quad k \leftarrow k + 1$
   end while

Algorithm 3 Building blocks for sublinear communication size recursive oblivious transfer and $Crypto-PBWT$

function $\mathsf{SCPrepQuery}(t_0,\, t_1,\, L_1)$
  $\boldsymbol{q} = (q_1 = 0, \ldots, q_{t_1} = 1, \ldots, q_{L_1} = 0)$
  $\vec{\mathsf{Enc}}(\boldsymbol{q}) = (\mathsf{Enc}(q_1) \ldots, \mathsf{Enc}(q_{L_1}))$
  return $\mathsf{Enc}(t_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q})$
end function

function $\mathsf{SCOT}(\mathsf{Enc}(t_0),\, \vec{\mathsf{Enc}}(\boldsymbol{q}),\, \boldsymbol{v},\, L_0,\, L_1)$
  for $k = 1$ to L_0 **do**
    Generate random value r_k
    $x = (k-1) \times L_1$
    $c_k = \bigoplus_{i=1}^{L_1} \big(v[x+i] \otimes \mathsf{Enc}(q_i)\big) \oplus r_k \otimes \mathsf{Enc}(t_0 - k)$
  end for
  return $\boldsymbol{c} = (c_1, \ldots, c_{L_0})$
end function

function $\mathsf{SCROT}(\mathsf{Enc}(\hat{t_0}),\, \vec{\mathsf{Enc}}(\hat{\boldsymbol{q}}),\, \boldsymbol{v},\, r,\, r'_0,\, r'_1,\, L,\, L_0,\, L_1)$
  $\hat{\boldsymbol{v}} \leftarrow \boldsymbol{v} + (r)_{\mathsf{mod}\ L}$
  $\boldsymbol{c} \leftarrow \mathsf{SCOT}(\mathsf{Enc}(\hat{t_0} - r'_0),\, \mathsf{Perm}(\vec{\mathsf{Enc}}(\hat{\boldsymbol{q}}),\, r'_1),\, \hat{\boldsymbol{v}},\, L_0,\, L_1)$
  $\boldsymbol{c} \leftarrow \mathsf{Perm}(\boldsymbol{c}, -r'_0)$ $\triangleright$ recovering the original position
  return $\boldsymbol{c} = (c_1, \ldots, c_{L_0})$
end function

function $\mathsf{ChooseDec}(c_0,\, c_1,\, L)$
  for $x = 0$ to 1 **do**
    $m \leftarrow \mathsf{Dec}(c_x)$
    if $(1 \le m \le L)$
      return m
    end if
  end for
end function

Algorithm 4 The detailed protocol of the sublinear communication size recursive oblivious transfer.

- Public input: the database size N, query length ℓ
- Private input of a user: a start position t
- Private input of a server: a vector $\boldsymbol{v}$ of length N

0. *(Key setup of cryptosystem)* The user generates a key pair $(\mathsf{pk}, \mathsf{sk})$ by the key generation algorithm KeyGen for the additive-homomorphic cryptosystem and sends public key pk to the server (the user and the server share public key pk and only the user knows secret key sk).
1. *(Server initialization)* The server computes $v_0[i] = v[i]/\lceil\sqrt{N}\rceil + 1,\ \ v_1[i] = (v[i])_{\mathsf{mod}\ \lceil\sqrt{N}\rceil} + 1$ for $i = 1, \ldots, N$.
2. *(User initialization)* The user computes $t_0 = t/\lceil\sqrt{N}\rceil + 1,\ \ t_1 = (t)_{\mathsf{mod}\ \lceil\sqrt{N}\rceil} + 1$.
3. *(Recursive search)*
   Initializes the index by $i \leftarrow 1$

while $(i \leq \ell)$ **do**

(a) *(Query entry)* The user performs the following steps:

- Prepare query
  if $(i \neq 1)$
  $t_0 \leftarrow \mathsf{Dec}(c_{0,t_0}),\ \ t_1 \leftarrow \mathsf{Dec}(c_{1,t_0})$
  end if
  $\mathsf{Enc}(t_0),\ \ \vec{\mathsf{Enc}}(\boldsymbol{q}) \leftarrow \mathsf{SCPrepQuery}(t_0,\ t_1,\ \lceil\sqrt{N}\rceil)$
- Sending $\mathsf{Enc}(t_0),\ \ \vec{\mathsf{Enc}}(\boldsymbol{q})$ to the server.

(b) *(Searching)* The server performs the following steps:

if $(i \neq \ell)$
Generating random values $r_0,\ \ r_1$
else
$r_0 = 0,\ r_1 = 0$
end if
$\triangleright$ ROT removes r'_0, r'_1 from a query and add r_0 or r_1 to each result.
$\boldsymbol{c}_0 \leftarrow \mathsf{SCROT}(\mathsf{Enc}(t_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}),\ \boldsymbol{v}_0,\ r_0,\ \lceil\sqrt{N}\rceil,\ r'_0,\ r'_1, \lceil\sqrt{N}\rceil,\ \lceil\sqrt{N}\rceil)$
$\boldsymbol{c}'_0 \leftarrow \mathsf{SCROT}(\mathsf{Enc}(t_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}),\ \boldsymbol{v}_0,\ r_0,\ \lceil\sqrt{N}\rceil,\ (r'_0 - \lceil\sqrt{N}\rceil),\ r'_1, \lceil\sqrt{N}\rceil,\ \lceil\sqrt{N}\rceil)$
$\boldsymbol{c}_1 \leftarrow \mathsf{SCROT}(\mathsf{Enc}(t_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}),\ \boldsymbol{v}_1,\ r_1,\ \lceil\sqrt{N}\rceil,\ r'_0,\ r'_1, \lceil\sqrt{N}\rceil,\ \lceil\sqrt{N}\rceil)$
$\boldsymbol{c}'_1 \leftarrow \mathsf{SCROT}(\mathsf{Enc}(t_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}),\ \boldsymbol{v}_1,\ r_1,\ \lceil\sqrt{N}\rceil,\ (r'_0 - \lceil\sqrt{N}\rceil),\ r'_1, \lceil\sqrt{N}\rceil,\ \lceil\sqrt{N}\rceil)$
$r'_0 \leftarrow r_0,\ \ r'_1 \leftarrow r_1$

- Sending $\boldsymbol{c}_0,\ \boldsymbol{c}'_0,\ \boldsymbol{c}_1,\ \boldsymbol{c}'_1$ to the user.

$i \leftarrow i + 1$

end while

4. *(Decryption of the result)* The user performs the following steps to obtain result x.
   $t_0 \leftarrow \mathsf{ChooseDec}(c_{0,t_0}, c'_{0,t_0}, \lceil\sqrt{N}\rceil),\ \ t_1 \leftarrow \mathsf{ChooseDec}(c_{1,t_0}, c'_{1,t_0}, \lceil\sqrt{N}\rceil)$
   $x = t_0 \times (\lceil\sqrt{N}\rceil - 1) + t_1$

Algorithm 5 Building blocks for sublinear communication size $Crypto - PBWT$

function isSCLongest($\vec{\mathsf{Enc}}(\boldsymbol{q}_f)$, $\vec{\mathsf{Enc}}(\boldsymbol{q}_g)$, $r'^{(f)}$, $r'^{(g)}$)
  $\vec{\mathsf{Enc}}(\boldsymbol{q}'_f) = \mathsf{Perm}(\vec{\mathsf{Enc}}(\boldsymbol{q}_f), r'^{(f)})$
  $\vec{\mathsf{Enc}}(\boldsymbol{q}'_g) = \mathsf{Perm}(\vec{\mathsf{Enc}}(\boldsymbol{q}_g), r'^{(g)})$
  for $i = 1$ to M **do**
    Generating random value r
    $d = d \oplus \mathsf{Enc}\big(r \times (q'_f[i] - q'_g[i])\big)$
  end for
  return d
end function

function isSCLongestGTϵ($\vec{\mathsf{Enc}}(\boldsymbol{q}_f)$, $\vec{\mathsf{Enc}}(\boldsymbol{q}_g)$, $r'^{(f)}$, $r'^{(g)}$, ϵ)
  for $k = 1$ to ϵ **do**
    $\vec{\mathsf{Enc}}(\boldsymbol{q}'_f) = \mathsf{Perm}(\vec{\mathsf{Enc}}(\boldsymbol{q}_f), r'^{(f)})$
    $\vec{\mathsf{Enc}}(\boldsymbol{q}'_g) = \mathsf{Perm}(\vec{\mathsf{Enc}}(\boldsymbol{q}_g), r'^{(g)})$
    $\vec{\mathsf{Enc}}(\boldsymbol{q}'_g) = \mathsf{Perm}(\vec{\mathsf{Enc}}(\boldsymbol{q}_g), k)$ ▷ $\boldsymbol{q}'_f = \mathsf{Perm}(\boldsymbol{q}'_g, k)$ iff. $(g - f) = k$
    for $i = 1$ to M **do**
      Generating random value r
      $d = d \oplus \mathsf{Enc}\big(r \times (q'_f[i] - q'_g[i])\big)$
    end for
    $d_k = d$
  end for
  $\boldsymbol{d} = (d_1, \ldots, d_\epsilon)$
  Shuffling order of elements in $\boldsymbol{d}$
  return $\boldsymbol{d}$
end function

function $\mathsf{SCOT}^\dagger$($\mathsf{Enc}(t_0)$, $\vec{\mathsf{Enc}}(\boldsymbol{q})$, $\boldsymbol{v}$, L_0, L_1, H, $\mathsf{Enc}(\gamma)$)
▷ H is a hash function, γ is a hash value for auxiliary filtering
▷ (it becomes ordinary OT when $H \equiv \gamma$ (or H is ignored))
  for $k = 1$ to L_0 **do**
    Generate random values r_k, r'_k
    $x = (k - 1) \times L_1$
    $c_k = \bigoplus_{i=1}^{L_1} \big(v[x + i] \otimes \mathsf{Enc}(q_i)\big) \oplus r_k \otimes \mathsf{Enc}(t_0 - k)$
    $c_k \leftarrow c_k \oplus r'_k \otimes \mathsf{Enc}(\gamma - H(k))$ ▷ auxiliary filtering
  end for
  return $\boldsymbol{c} = (c_1, \ldots, c_{L_0})$
end function

function $\mathsf{SCROT}^\dagger$($\mathsf{Enc}(\hat{t_0})$, $\vec{\mathsf{Enc}}(\hat{\boldsymbol{q}})$, $\boldsymbol{v}$, r, L, r'_0, r'_1, L_0, L_1, H, $\mathsf{Enc}(\gamma)$)
  $\boldsymbol{c} \leftarrow \mathsf{SCOT}^\dagger(\mathsf{Enc}(\hat{t_0} - r'_0), \mathsf{Perm}(\vec{\mathsf{Enc}}(\hat{\boldsymbol{q}}), r'_1), (\boldsymbol{v} + r)_{\mathsf{mod}\ L}, L_0, L_1, H, \mathsf{Enc}(\gamma))$
  $\boldsymbol{c} \leftarrow \mathsf{Perm}(\boldsymbol{c}, -r'_0)$ ▷ recovering the original position
  return $\boldsymbol{c} = (c_1, \ldots, c_{L_0})$
end function

Algorithm 6 The detailed description of sublinear communication size $Crypto-PBWT$ for finding a set-longest match at position t.

- Public input: The length of column M, a set of alphabet letters $\Sigma = \{1, 2, .., |\Sigma|\}$ and a set of $(D-1)$ decoy positions and true position $T = (t_1, \ldots, t_D)$.
- Private input of a user: A starting column $t_x \in T$, a query sequence S of length ℓ
- Private input of a server: PBWT matrix $P \in \mathbb{N}^{M \times N}$

0. *(Key setup of cryptosystem)* The user generates a key pair $(\mathsf{pk}, \mathsf{sk})$ by the key generation algorithm KeyGen for the additive-homomorphic cryptosystem and sends public key pk to the server (while only the user knows secret key sk).
1. *(Server initialization)*
   - The server computes $L_1 = \sqrt{D(M+1)|\Sigma|}$, $L'_0 = D(M+1)/L_1 + 1$, $L_0 = L'_0 \times |\Sigma|$ and announces L_0, L_1 and L'_0 to the user.
   - The server defines a hash function H by $H((\gamma - 1)L'_0 + a) = \gamma$ for any $\gamma \in \Sigma$ and $1 \le a \le L'_0$.
2. *(User initialization)*
   - The user set initial interval $[f, g]$ by $f = o_x,\ g = o_x + M$ where $o_j = (j-1)(M+1)$.
   - The user computes two dimensional representation of $[f, g]$ by $f_0 \leftarrow f/L_1 + 1,\ f_1 \leftarrow (f)_{\mathsf{mod}\ L_1} + 1,\quad g_0 \leftarrow g/L_1 + 1,\ g_1 \leftarrow (g)_{\mathsf{mod}\ L_1} + 1$
3. *(Recursive search)* Initializes the indices by $i \leftarrow 1 \quad k \leftarrow -1$
   while $(i \le \ell)$ **do**
   (a) *(Query entry)* The user performs the following steps:
   - Prepare next query:
     $f_0 \leftarrow f_0 + (S[i] - 1)L'_0$, $g_0 \leftarrow g_0 + (S[i] - 1)L'_0$ $\quad \triangleright$ Setting offset to search matches ending with $S[i]$
     $\big(\mathsf{Enc}(f_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}_f)\big) \leftarrow \mathsf{SCPrepQuery}(f_0,\ f_1,\ L_1)$, $\big(\mathsf{Enc}(g_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}_g)\big) \leftarrow \mathsf{SCPrepQuery}(g_0,\ g_1,\ L_1)$
   - Sending $\mathsf{Enc}(f_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}_f),\ \mathsf{Enc}(g_0),\ \vec{\mathsf{Enc}}(\boldsymbol{q}_g),\ \mathsf{Enc}(S[i])$ to the server.

   (b) *(Searching)* The server performs the following steps:
   - Computes vectors $\boldsymbol{v}^{(c)}$ of length $D \times (M+1)$ for all $c \in \Sigma$:
   $$v_c[o_j + u] = \begin{cases} \mathsf{CF}_c(P_{\cdot,(t_j+k)}) + o_j & (u = 0) \\ \mathsf{CF}_c(P_{\cdot,(t_j+k)}) + \mathsf{Rank}_c(P_{\cdot,(t_j+k)}, u) + o_j & (1 \le u \le M) \end{cases}$$
     where $o_j = (j-1)(M+1)$ for $j = 1, \ldots, D$.
   - Creates vectors $\boldsymbol{v}_0^{(c)}, \boldsymbol{v}_1^{(c)}$ of length $L'_0 \times L_1$ for $c = 1, \ldots, |\Sigma|$.
   - Computes $v_0^{(c)}[i] = v^{(c)}[i]/L_1 + 1,\ v_1^{(c)}[i] = (v^{(c)}[i])_{\mathsf{mod}\ L_1} + 1$ for $i = 1, \ldots, D(M+1)$ and $c = 1, \ldots, |\Sigma|$.
   - Creates vectors $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ by concatenating $\boldsymbol{v}_0 = \boldsymbol{v}_0^{(1)}, \ldots, \boldsymbol{v}_0^{(|\Sigma|)}$ and $\boldsymbol{v}_1 = \boldsymbol{v}_1^{(1)}, \ldots, \boldsymbol{v}_1^{(|\Sigma|)}$.
   - Generates random values $r_0^{(f)}, r_1^{(f)},\ r_0^{(g)}, r_1^{(g)}$
   - Computes next intervals and an encrypted flag showing if the match is the longest
     $\boldsymbol{c}_0^{(f)} \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(f_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_f), \boldsymbol{v}_0, r_0^{(f)}, L'_0, r'^{(f)}_0, r'^{(f)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$,
     $\boldsymbol{c}'^{(f)}_0 \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(f_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_f), \boldsymbol{v}_0, r_0^{(f)}, L'_0, (r'^{(f)}_0 - L'_0), r'^{(f)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$
     $\boldsymbol{c}_1^{(f)} \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(f_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_f), \boldsymbol{v}_1, r_1^{(f)}, L_1, r'^{(f)}_0, r'^{(f)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$,
     $\boldsymbol{c}'^{(f)}_1 \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(f_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_f), \boldsymbol{v}_1, r_1^{(f)}, L_1, (r'^{(f)}_0 - L'_0), r'^{(f)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$
     $\boldsymbol{c}_0^{(g)} \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(g_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_g), \boldsymbol{v}_0, r_0^{(g)}, L'_0, r'^{(g)}_0, r'^{(g)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$,
     $\boldsymbol{c}'^{(g)}_0 \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(g_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_g), \boldsymbol{v}_0, r_0^{(g)}, L'_0, (r'^{(g)}_0 - L'_0), r'^{(g)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$
     $\boldsymbol{c}_1^{(g)} \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(g_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_g), \boldsymbol{v}_1, r_1^{(g)}, L_1, r'^{(g)}_0, r'^{(g)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$,
     $\boldsymbol{c}'^{(g)}_1 \leftarrow \mathsf{SCROT}^\dagger(\mathsf{Enc}(g_0), \vec{\mathsf{Enc}}(\boldsymbol{q}_g), \boldsymbol{v}_1, r_1^{(g)}, L_1, (r'^{(g)}_0 - L'_0), r'^{(g)}_1, L_0, L_1, H, \mathsf{Enc}(S[i]))$
     $d \leftarrow \mathsf{isSCLongest}(\mathsf{Enc}(f_0 - r'^{(f)}_0), \mathsf{Enc}(g_0 - r'^{(g)}_0), \mathsf{Perm}(\vec{\mathsf{Enc}}(\boldsymbol{q}_f), -r'^{(f)}_1), \mathsf{Perm}(\vec{\mathsf{Enc}}(\boldsymbol{q}_g), -r'^{(g)}_1))$
   - Storing random values $r'^{(f)}_0 \leftarrow r_0^{(f)},\ r'^{(f)}_1 \leftarrow r_1^{(f)},\ r'^{(g)}_0 \leftarrow r_0^{(g)},\ r'^{(g)}_1 \leftarrow r_1^{(g)}$
   - Sending $\boldsymbol{c}_0^{(f)}, \boldsymbol{c}'^{(f)}_0, \boldsymbol{c}_1^{(f)}, \boldsymbol{c}'^{(f)}_1, \boldsymbol{c}_0^{(g)}, \boldsymbol{c}'^{(g)}_0, \boldsymbol{c}_1^{(g)}, \boldsymbol{c}'^{(g)}_1, d$ to the user

   (c) *(Decryption of the encrypted flag and the randomized interval)* The user performs the following steps:
   if $(\mathsf{Dec}(d) == 0)$
   Reports the result $S[1, \ldots, i-2]$ and sending the server decoy queries until $i == \ell$
   else
   Computes $f_0 \leftarrow \mathsf{ChooseDec}(c^{(f)}_{0,f_0}, c'^{(f)}_{0,f_0}, L'_0)$, $g_0 \leftarrow \mathsf{ChooseDec}(c^{(g)}_{0,g_0}, c'^{(g)}_{0,g_0}, L'_0)$,
   $f_1 \leftarrow \mathsf{ChooseDec}(c^{(f)}_{1,f_0}, c'^{(f)}_{1,f_0}, L_1)$, $g_1 \leftarrow \mathsf{ChooseDec}(c^{(g)}_{1,g_0}, c'^{(g)}_{1,g_0}, L_1)$ $\quad \triangleright$ for choosing correct results
   end if

   $i \leftarrow i + 1 \quad k \leftarrow k + 1$
   end while

Algorithm 7 Building blocks for linear communication size $Crypto - PBWT$

```
function isLongest(Enc⃗(q_f), Enc⃗(q_g), r'^(f), r'^(g))
    Enc⃗(q'_f) = Perm(Enc⃗(q_f), r'^(f))
    Enc⃗(q'_g) = Perm(Enc⃗(q_g), r'^(g))
    for i = 1 to M do
        Generating random value r
        d = d ⊕ Enc(r × (q'_f[i] − q'_g[i]))
    end for
    return d
end function

function isLongestGTϵ(Enc⃗(q_f), Enc⃗(q_g), r'^(f), r'^(g), ϵ)
    for k = 1 to ϵ do
        Enc⃗(q'_f) = Perm(Enc⃗(q_f), r'^(f))
        Enc⃗(q'_g) = Perm(Enc⃗(q_g), r'^(g))
        Enc⃗(q'_g) = Perm(Enc⃗(q_g), k)                ▷ q'_f = Perm(q'_g, k) iff. (g − f) = k
        for i = 1 to M do
            Generating random value r
            d = d ⊕ Enc(r × (q'_f[i] − q'_g[i]))
        end for
        d_k = d
    end for
    d = (d_1, ..., d_ϵ)
    Shuffling order of elements in d
    return d
end function
```