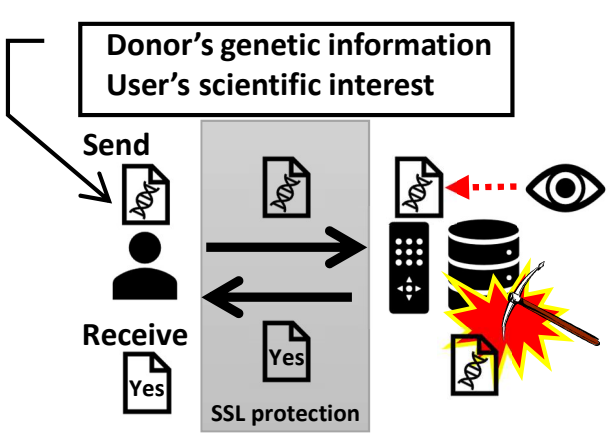
Privacy-preserving Search for Sharing Genetic Variants

Masanobu Jimbo[1, 2], Nobutaka Mitsuhashi[3], Shigeo Mitsunari[4] , Shin Kawano[5], Toshiaki Katayama[5], Kiyoshi Asai[6], Kana Shimizu[1]

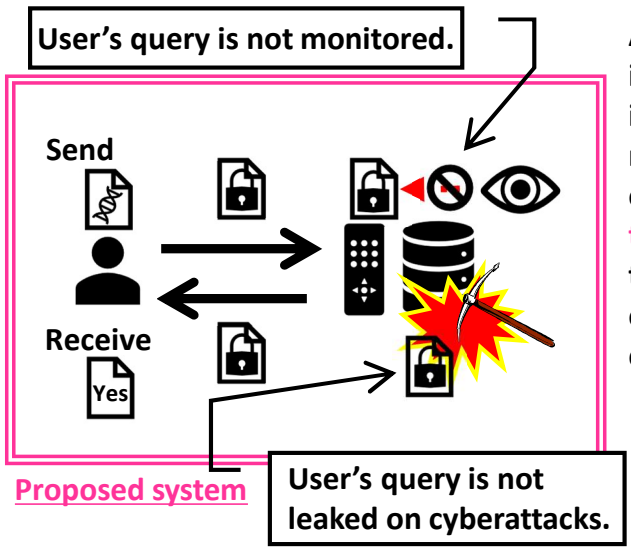
1. Waseda University, 2. AIST-Waseda University CBBD-OIL, 3. National Bioscience Database Center, 4. Cybozu Labs, Inc., 5. Database Center for Life Science, 6. The University of Tokyo

Beacon of the Global Alliance for Genomics and Health (GA4GH) is an important international platform for sharing genetic variants. The Beacon search is described as follows: a user sends a pair consisting of a genomic position and an allele (e.g., Chr21:122211 and 'A') to a server, and the server returns only a binary response "yes" or "no" to let the user know whether or not the variant exists in the database. This simple search enables global discovery of the genetic variants and facilitates the sharing of the genomic data resources from various cohort studies. Since the Beacon query often includes private information (such as patient's genomic variant), there is a need to protect query privacy. We introduce a novel search system called ***Crypto Beacon*** which enables the user to hide the query from the server while successfully obtaining the results. In our system, the query is encrypted on the user's browser; the server conducts the search without decrypting the query and returns the encrypted result. Only the user can decrypt the result, i.e., our system ensures that the server never monitors the user's query. Hence, the user can send a sensitive query in full privacy. The encryption is done by using a cryptographic technique, called homomorphic encryption, which enables the addition/multiplication in the encrypted space.

Motivation & Aim



User-privacy risk of Beacon search

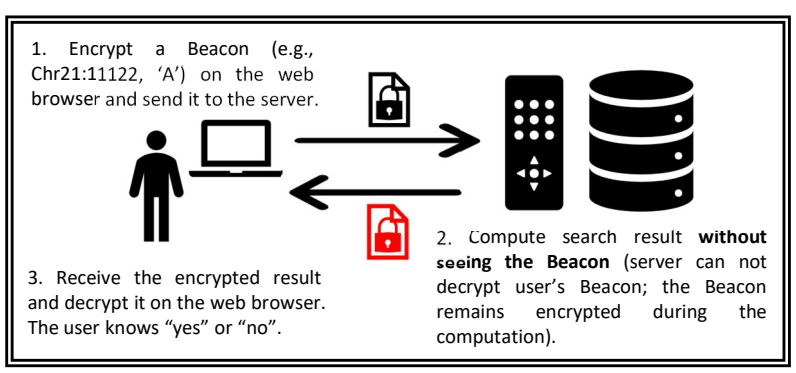


Proposed system

A Beacon query is a piece of genetic information of the donor and is regarded as a scientific or medical interest of a user. Most institutions that provide Beacon search do not have a positive reason to monitor a query. This work highlights user-side privacy of Beacon and proposes the novel genome-scale Beacon search that protects a query. Our system, named Crypto Beacon enables the user to hide the query even from the server while successfully obtaining the results. Crypto Beacon, in which all the queries stay encrypted even on the server, is robust against cyberattacks.

Sharing genetic information by Beacon search : A user sends a genomic position and an allele. The server returns only a binary response "yes" or "no" to let the user know whether the variant exists in the database.

Crypto Beacon



Installation-free system & real-time search

Crypto Beacon is the first system that can ***process privacy-preserving search through a web browser without installing a specific software package.*** Generally, such a privacy-preserving data-mining system is based on some complex and heavy cryptographic operations; therefore, it always requires an installation of a specific stand-alone software that uses a lot of CPU power, which deteriorates its usability. In our system, state-of-the-art web technology called WebAssembly is used for performance optimization. We implemented all the necessary cryptographic algorithms as WebAssembly modules and the client application call them through JavaScript, so ***the system can work even on a mobile tablet such as iPhone***.

Provide strong privacy protection

Crypto Beacon is the first public service that protects genome privacy for large-scale databases. The system is based on an oblivious transfer implemented by a homomorphic encryption.

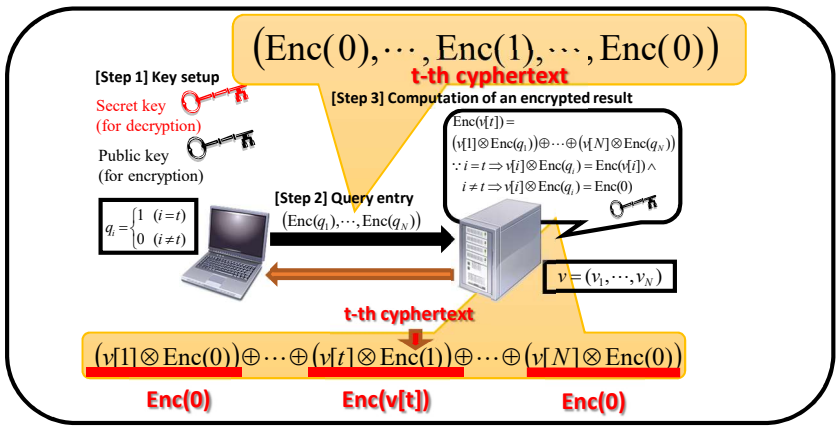
Homomorphic encryption

The cryptosystem that enables to run additive/multiplicative operations of plaintexts without decryption.

$$Enc(m1 + m2) = Enc(m1) \oplus Enc(m2)$$

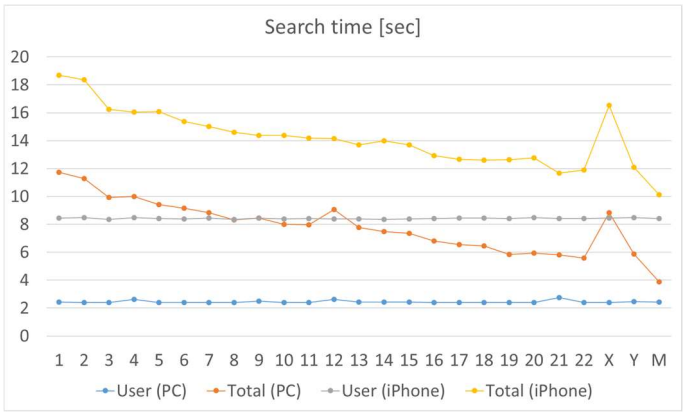
Oblivious transfer

Assume that User has an index ***t*** and Server has a vector ***v***. OT is the cryptographic protocol that enables User to know only ***v[t]*** without disclosing ***t*** to Server.



Performance test 1:

User's query: Beacon and an ID of the target chromosome.
Computation:
User: Laptop PC with 1 CPU or iPhone
Server: PC with 4 CPU cores
Communication: Tested within the same network. PC for wired, iPhone for wireless network.
Dataset: Research ID at NBDC: hum0013, genotype counts of 2.5 million SNPs for Japanese genomes.



Performance test 2:

User's query (test 2.1): Beacon.
User's query (test 2.2): Beacon and an ID of the target chromosome.
Computation (test 2.1):
User: Laptop PC with 1 CPU
Server: NDBC's cluster machine (16 CPU cores)
Computation (test 2.2):
User: Laptop PC with 1 CPU
Server: NDBC's cluster machine (4 CPU cores)

	Test 2.1	Test 2.2 (Average)
Initialization on the browser (sec)	1.16	1.16
User time (sec)	4.96	2.54
Server time (sec)	188.03	4.16
Transfer time (sec)	2.85	5.71

Communication: Tested over the internet between Waseda University at Tokyo and NDBC at Chiba.
Dataset: Research IDs at NBDC: hum0013, hum0014 (455,781 SNPs), hum0015 (4,301,546 SNPs) and hum0029(449,205 SNPs)

Screen copy of the web interface →