# Genome analysis with near-complete privacy possible

*Stanford researchers used cryptography to cloak irrelevant genetic information in individuals' genomes while revealing disease-associated mutations. They say the technique could vastly improve patient privacy.*

It is now possible to scour complete human genomes for the presence of disease-associated genes without revealing any genetic information not directly associated with the inquiry, say Stanford University researchers.

This “genome cloaking” technique, devised by biologists, computer scientists and cryptographers at the university, ameliorates many concerns about genomic privacy and potential discrimination based on an individual's genome sequence.

Using the technique, the researchers were able to identify the responsible gene mutations in groups of patients with four rare diseases; pinpoint the likely culprit of a genetic disease in a baby by comparing his DNA with that of his parents; and determine which out of hundreds of patients at two individual medical centers with similar symptoms also shared gene mutations. They did this all while keeping 97 percent or more of the participants' unique genetic information completely hidden from anyone other than the individuals themselves.

“We now have the tools in hand to make certain that genomic discrimination doesn't happen,” said Gill Bejerano, PhD, associate professor of developmental biology, of pediatrics and of computer science. “There are ways to simultaneously share and protect this information. Now we can perform powerful genetic analyses while also completely protecting our participants' privacy.”

Bejerano shares senior authorship of the research, which was published Aug. 18 in *Science*, with Dan Boneh, PhD, professor of computer science and of electrical engineering. Graduate students Karthik Jagadeesh and David Wu share lead authorship of the study.

**Applying cryptography techniques**

The researchers hope that routine implementation of their technique will help individuals overcome any qualms about privacy that may keep them from sharing their genome sequences. In particular, people may be concerned that DNA sequences or genetic variants currently unassociated with diseases may in the future be linked with as-yet-unidentified increases in risk.

“These are techniques that the cryptography community has been developing for some time,” said Boneh, who is the Rajeev Motwani Professor in the School of Engineering. “Now we are applying them to biology. Basically, if you have 1 million people with genomic data they would like to keep private, this approach lets researchers analyze the data in aggregate and only report on findings that are pertinent. An individual might have dozens of anomalous genes, but the researchers and clinicians will only learn about the genes relevant to the study, and nothing else.”

When the human genome was fully sequenced in 2001, it was hailed as a remarkable achievement. For the first time, the 3 billion nucleotides that encode the approximately 20,000 genes that keep our bodies running smoothly were tidily listed as a string of letters. But every human has many variations from the published, consensus sequence. These individual differences are what make us unique, but they can also confer increased risk of genetic diseases.

More than 7,000 diseases are caused by variations in the sequence of a single gene. But in order to determine which variations cause the condition, it has been necessary until now to compare the genetic sequences of hundreds or thousands of individuals with and without the disease, letter by letter. Geneticists (or their computer software) then make a list of all the differences and identify which are found primarily in people with the disease under study but rarely in any unaffected people. Those variations are then considered to be prime disease-causing suspects.

“There is a general conception that we can only find meaningful differences by surveying the entire genome,” said Bejerano. “But these meaningful differences make up only a very tiny proportion of our DNA. There are now amazing tools in computer science and cryptography that allow researchers to pinpoint only these differences while keeping the remainder of the genome completely private.”

In 2008, President George W. Bush signed the Genetic Information Nondiscrimination Act, which prohibits discrimination in matters of health insurance and employment based on an individual’s genetic information. But there are many other arenas in which such discrimination could potentially occur, including the purchase of life or disability insurance or in the application for a loan.

**Giving power to the individual**

Jagadeesh and Wu worked together to adapt a cryptographic approach known as Yao’s protocol and cloud computing for use with human genomes. A key component of the technique is the involvement of the individual whose genome is to be studied. In particular, each individual encrypts their genome (with the help of a simple algorithm on their own computer or smart phone) into a linear series of values describing the presence or absence of the gene variants under study, without revealing any other information about their genetic sequence. The encrypted information is uploaded into the cloud and the researchers then use a secure, multi-party computation (a cryptographic technique that ensures the input data remain private) to conduct the analysis and reveal only those gene variants likely to be pertinent to the investigation.

“In this way, no person or computer, other than the individuals themselves, has access to the complete set of genetic information,” said Bejerano. In each case, the analysis was performed within seconds or minutes with moderate computing power. They hope to extend the technique to include diseases caused by combinations of multiple genetic variants or to handle tens of thousands of sequences such as those found in genomewide association studies.

Ultimately the goal is to find the best way to both share the genetic information with researchers while also protecting each patient’s privacy in order to advance medical knowledge.

“Often people who have diseases, or those who know that a particular genetic disease runs in their family, are the most reluctant to share their genomic information because they know it could potentially be used against them in some way,” said Bejerano. “They are missing out on helping themselves and others by allowing researchers and clinicians to learn from their DNA sequences.”

Bejerano is a member of Bio-X, the Stanford Child Health Research Institute, the Stanford Cancer Institute and the Stanford Neurosciences Institute.

Another Stanford study co-author is graduate student Johannes Birgmeier.

The study was funded by Stanford University fellowship grants, the National Science Foundation, the Defense Advanced Research Projects Agency, the David and Lucile Packard Foundation, Microsoft and the Simons foundation.

Stanford’s departments of Developmental Biology, of Pediatrics and of Computer Science also supported the work.

***Krista Conger***

Original on news.stanford.edu