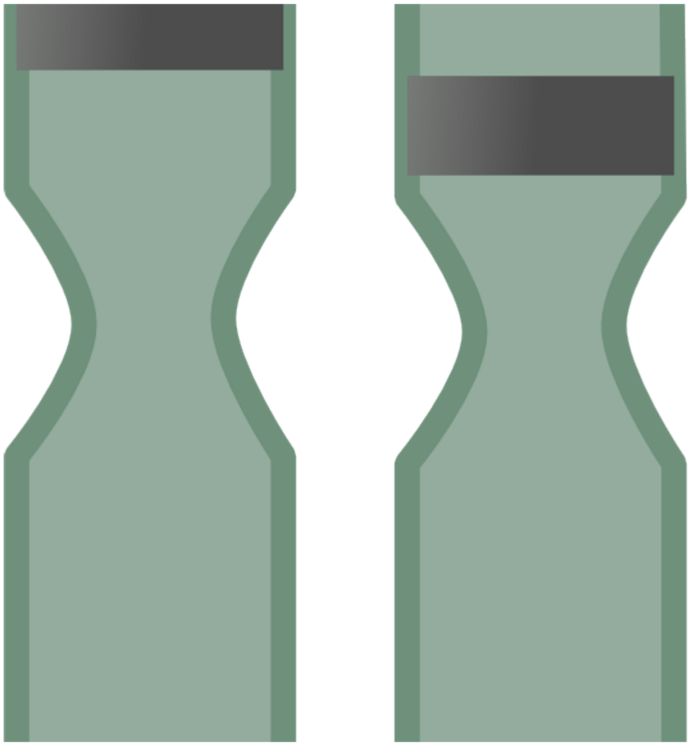
Protecting confidentiality in genomic studies

May 7 2018



Credit: CC0 Public Domain

Genome-wide association studies, which look for links between particular genetic variants and incidence of disease, are the basis of much modern biomedical research.

But databases of genomic information pose privacy risks. From people's raw genomic data, it may be possible to infer their surnames and perhaps even the shapes of their faces. Many people are reluctant to contribute their genomic data to biomedical research projects, and an organization hosting a large repository of genomic data might conduct a months-long review before deciding whether to grant a researcher's request for access.

In a paper appearing today in *Nature Biotechnology*, researchers from MIT and Stanford University present a new system for protecting the privacy of people who contribute their genomic data to large-scale biomedical studies. Where earlier cryptographic methods were so computationally intensive that they became prohibitively time consuming for more than a few thousand genomes, the new system promises efficient privacy protection for studies conducted over as many as a million genomes.

"As biomedical researchers, we're frustrated by the lack of data and by the access-controlled repositories," says Bonnie Berger, the Simons Professor of Mathematics at MIT and corresponding author on the paper. "We anticipate a future with a landscape of massively distributed genomic data, where private individuals take ownership of their own personal genomes, and institutes as well as hospitals build their own private genomic databases. Our work provides a roadmap for pooling together this vast amount of genomic data to enable scientific progress."

The first author on the paper is Hyunghoon Cho, a graduate student in electrical engineering and computer science at MIT; he and Berger are joined by David Wu, a graduate student in computer science at Stanford.

At the core of the system is a technique called secret sharing, which divides sensitive data among multiple servers. To store the number x, for instance, a secret-sharing system might send the random number r to one server and x-r to the other.

Neither server is independently able to infer x. Collectively, however, they can still perform useful operations. If one server stored a bunch of r's and added them together, and the other added up all the corresponding (x-r)'s, then sharing the results and adding them together would yield the sum of all the x's. Neither server, however, would ever observe the value of any one x.

If both servers are hacked, of course, the attacker could reconstruct all the x's. But so long as one server is trustworthy, the system is secure. Furthermore, that principle generalizes to multiple servers. If data are divided among, say, four servers, an attacker would have to infiltrate all four; hacking any three is insufficient to extract any data.

In this context, however, multiplication is more complicated than addition. Multiplying two x's requires the generation of three more random numbers—known as a Beaver triple, after the cryptographer Donald Beaver—in addition to the r's. Those three numbers, in turn, must be divided among servers using secret sharing. Adding the secret-shared components of those numbers to the x's and r's before multiplication gives rise to an algebraic expression in which all the added randomness can be filtered out, leaving only the product of the two x's.

Genome-wide association studies involve a massive table—or matrix—that maps the genomes in the database against the locations of genetic variations known as SNPs, for single-nucleotide polymorphisms. The SNPs will typically number about a million, so if the database contains a million genomes, the result will be a million-by-million matrix.

Finding useful disease correlations requires filtering out misleading correlations, a process known as population stratification correction. East Asians, for instance, are frequently lactose intolerant, but they also tend to be shorter than Northern Europeans. A naïve investigation of the genetic correlates of lactose intolerance might instead end up identifying those for height.

Population stratification correction typically relies on an algorithm called principal component analysis, which requires repeated multiplications involving the whole SNP-versus-genome matrix. If every entry in the matrix needed its own set of Beaver triples for each of those multiplications, analyzing a million genomes would be prohibitively time consuming.

But Cho, Berger, and Wu found a way to structure that sequence of multiplications so that many of the Beaver triples can be calculated only once and reused, drastically reducing the complexity of the computation.

They also use a couple other techniques to speed up their system. Because the Beaver triples must be shared secretly, each number in the Beaver triple has an associated random number: In the two-server scenario, one server would get the random number and the other would get the Beaver number minus the random number.

In Cho, Berger, and Wu's system, there's a server dedicated to generating Beaver triples and sharing them secretly. But while it needs to transmit the Beaver numbers minus the associated random numbers to the appropriate servers, it doesn't need to transmit the random numbers themselves. Instead, it simply shares the number it uses to "seed" an algorithm known as a pseudorandom number generator. The recipient servers can then generate the random numbers on their own, saving a huge amount of communication bandwidth.

Finally, when performing all its multiplications, the system doesn't actually use the whole million-by-million matrix. Instead, it uses an approximation technique called random projection to winnow the matrix down while preserving the accuracy of the final computation results.

Based on these techniques, Cho, Berger, and Wu's system accurately reproduced three published genome-wide association studies involving 23,000 individual genomes. The results of those analyses suggest that the system should scale efficiently to a million genomes.

Provided by Massachusetts Institute of Technology

Citation: Protecting confidentiality in genomic studies (2018, May 7) retrieved 25 April 2024 from https://phys.org/news/2018-05-confidentiality-genomic.html