genomeweb

Homomorphic Encryption Might Hold Key to Fast, Large-Scale GWAS

Jun 15, 2020 | Neil Versel

CHICAGO – A startup maker of data-encryption technology believes it has the answer to the security problem that has hampered researchers looking to perform massive genome-wide association studies with datasets from disparate sources, even when the records are located in different countries.

Results published last month in *Proceedings of the National Academy of Science* demonstrated the efficacy of homomorphic encryption technology from Duality Technologies in running complex biomedical computations on encrypted datasets of genomic variants. Homomorphic encryption is a method of encoding data as ciphertext to allow computation without decryption.

"You can keep the data encrypted while analyzing it without decrypting. That's the magic of homomorphic encryption," said Duality Chair and Cofounder Rina Shainski.

In the study, Newark, New Jersey-based Duality and the Dana-Farber Cancer Institute in Boston ran a GWAS on a dataset containing genotypic and phenotypic information on more than 25,000 patients. They produced results 30 times faster than a "state-of-the-art" method of multiparty computation described by computer scientists at Massachusetts Institute of Technology and Stanford University in 2018, without sacrificing accuracy, the company said.

"The 30X is really the difference between being [the technology being] practical and useful and potentially not being able to use it in practice," Shainski said. "We believe we overcame the barrier of practicality for this kind of analysis."

In the paper, the researchers cited earlier work that mentioned that it would be wonderful to apply homomorphic encryption to genomic data, but the estimated compute times were untenable. "I think this work really went somewhere that was not seen as possible in previous work," said one of the authors, Alexander Gusev, a quantitative geneticist at Dana-Farber and Harvard Medical School. "It was not just incremental."

The Duality and Dana-Farber researchers tested the technology with two techniques, allelic chi-square analysis and a logistic regression approximation (LRA), in a GWAS involving 12,461 cases age-related macular degeneration and 14,276 control subjects among a European population.

In the paper, the authors said that a hybrid analysis, applying the chi-squared method to all SNPs and retesting the 5 percent "most significant" SNPs by LRA, could be completed in 17 hours at the same level of accuracy, though they did not actually run such a trial.

They extrapolated that their technology could go through a GWAS of 100,000 people and 500,000 SNPs in 5.6 hours on a single server node, or in just 11 minutes on 31 parallel server nodes. Multiparty computation on the same scale would take 37 hours for association tests only or 193 hours including quality control and population stratification analysis, they reported.

The Duality LRA method would need 234 hours to run the extrapolated study.

Duality and academic partners including Dana-Farber won first place in the 2018 Integrating Data for Analysis, Anonymization and Sharing (iDASH) challenge for secure genome analysis for their homogenic encryption of LRA. The chi-square test turned out to be 40 times faster and uses one-sixth the memory of the LRA method, though it excludes covariates.

Homomorphic encryption allows people to contribute all or part of their genomic and phenotypic data to crowdsourced studies without researchers ever getting any information that could compromise individual privacy, even if there is a data breach years later, according to Gusev.

"For a lot of what we study, it's about understanding what's happening at the group level rather than at the individual level, so we don't even need the individual-level information at the end, but you have to go through it to do the statistics," he said. "This bypasses that entirely."

"Patients can now participate in the study without having to sign onto very open data-sharing protocols," Gusev said. "If there's a data breach or if somebody hacks the main database or [information] somehow accidentally gets released to the web, their data is still secure because it's encrypted the whole time. That box never gets opened."

While the concept dates to the late 1970s, homomorphic encryption really became feasible after a 2009 presentation to the Association for Computing Machinery.

"It was considered beautiful mathematics, but not very practical," Shainski said. Since then, some of Duality's founders "contributed greatly to the progress of making it much more practical and feasible," she said.

Cofounder and CEO Alon Kaufman is a former global director of data science and innovation at RSA Security, a data security firm that is now part of Dell Technologies. Cofounder and CTO Kurt Rohloff was a principal investigator of Programming Computation on Encrypted Data (PROCEED), a seminal research effort from the US Defense Advanced Research Project Agency (DARPA) that showed that it was practical to compute encrypted data.

Chief Scientist Shafi Goldwasser has a Turing Award and two Gödel Prizes for theoretical computer science and leads the Simons Institute for Theory of Computation at the University of California, Berkeley. Chief Cryptographer Vinod Vaikuntanathan is a former student of Goldwasser at Massachusetts Institute of Technology.

An IBM researcher first demonstrated in 2009 that computational analyses could be carried out on homomorphically encrypted data. The phrase has been in the lexicon of bioinformaticians since at least 2014, but maintaining encryption throughout the entire process has required far longer computational time than working with decoded data.

In a 2016 paper, Swiss researchers reported success with homomorphic encryption, but their method took about 12 minutes of computing time per patient.

In the last decade-plus, the computing power necessary for true homomorphic encryption has become more affordable, particularly on cloud platforms.

Marcelo Blatt, head of data science at Duality, went so far as to call homomorphic encryption the "Holy Grail of cryptography" because the data remains secure the entire time since only the data owner can decrypt the information.

"You could take your genomic data, send it straight to a laboratory [to] make an analysis," Blatt said. "This laboratory would not get any information about you because everything is encrypted, even the results. You will get back the results and you would be the only person to understand the result."

Shainski said this work will open many doors for collaboration in biomedical research.

"You have competitors that don't want you to see the fine data ingredients, but they're willing to share the results," explained. "That's also possible with this technology."

The high-level security allows for cross-border data use even when the data itself is not allowed to leave a particular country or region.

"Today it is extremely, extremely difficult," Shainski said, citing rules such as the European Union's General Data Protection Regulation (GDPR), HIPAA, and the California Consumer Privacy Act.

"If you want to apply some of the analysis that was developed in the US on healthcare data coming from Europe, it's probably impossible today … unless the data is in the public domain," Shainski said. "Being able to encrypt it in such a way that you can analyze it while encrypted opens the door for cross-border collaboration on this data."

Duality was founded in late 2016 by a group of cryptographers and data scientists. The company exists largely because of these developments, and Duality officials said that they have created more than a dozen crypto-engineering optimizations of previous techniques for their platform, called SecurePlus.

"We were able to take existing algorithms and tailor them or craft them again in order to be friendly with the cryptographic primitive that we have to deal with," Blatt said. Cryptographic primitives are simple, generic cryptographic instruments that serve as the building blocks for more sophisticated technologies.

"The magic of Duality is actually to make it possible not just to run simple computations, but to run complex computations efficiently, data science computations on encrypted data," Shainski said.

She said that this type of encryption is applicable anywhere there is sensitive data. "The more data we get, the more sensitive it gets, definitely in healthcare," according to Shainski.

Gusev said that Dana-Farber computer scientists have long been discussing how to maintain patient privacy and security when sharing data even internally.

He primarily works on algorithms involving large-scale datasets of tens or hundreds of thousands of patients, exactly the scale required for GWAS.

In its GWAS, Dana-Farber studies how germline mutations in individuals or comparative mutations at the tumor level can influence their treatment response or raise the risk of adverse events. The institution was looking to speed up processing and increase security for these studies.

"Without some kind of data-sharing solution, a lot of the research that I wanted to do was impossible," Gusev said. Duality was developing its technology as Dana-Farber was exploring this question, so he reached out to the vendor.

"It was clear that there was actually a research opportunity to demonstrate that this method worked, that this general approach worked in the genomics space," Gusev said.

"It doesn't require changing any existing practices and it doesn't require complicated data-sharing bureaucracy," he added. "You basically just encrypt your dataset. Nobody gets to look at it, and the fundamental computations that we're interested in can still be performed."

This technology eliminates the need to share decryption keys with users or verify the trustworthiness of a user. It also solves the problem of having to get all participating parties to reshare datasets when investigators change a study parameter. Prior to the study described in the *PNAS* article, Dana-Farber had not applied homomorphic encryption to GWAS. This research showed that this technology was feasible for this application, even at a large scale with real patient data, according to Gusev.

Gusev believes that homomorphic encryption might enable more crowdsourced oncology studies.

Currently, such studies go through what Gusev called "rapid transparency," by which all the data gets released to the public. "But that also comes with the limitations that you have to buy into having no privacy to participate," he said.

Duality and Dana-Farber said that the technology and the study have implications for COVID-19 research, as well.

GWAS might be able to identify variations or even environmental risk factors that could be associated with severe or mild responses to the SARS-CoV-2 coronavirus, according to Gusev.

Homomorphic encryption is particularly well suited for research into a pandemic because location data is as sensitive as health or genomic records, an issue growing in importance as contact tracing increases during the COVID-19 outbreak. "There is a lot of discussion how to balance between individual privacy and public health, and homomorphic encryption can be useful there as well," Shainski said.

Shainski said that SecurePlus is not yet broadly available, so nobody is using it specifically for COVID-19 research at this point. She did say that the company is in talks with unspecified pharmaceutical companies about this type of application and that the software should be generally available later this year.

Filed Under Informatics Cancer North America papers of note software developers macular degeneration Dana-Farber GWAS cloud computing genomic datasets