MedicalXpress

Re-identifying faces from genomic data is more difficult than previously thought

November 18 2021

Credit: CC0 Public Domain

Direct-to-consumer genetic testing has enabled millions of individuals to determine their ancestry and gain insights about their genetic predisposition to inherited diseases. While individual genotyping

information is stored securely, some people consent to share their genomic data for further study.

This data sharing has raised some valid concerns about genomic privacy. For example, could hackers reidentify a person—perhaps construct a picture of their face—based on genotype data downloaded legally from open-source web platforms?

In 2017, genomics-based health intelligence company Human Longevity and other research groups reported that it was feasible to predict a person's facial appearance from their DNA.

Intrigued by the privacy risk implications of this work, Washington University in St. Louis faculty member Yevgeniy "Eugene" Vorobeychik, an expert in applying game theory to determine privacy risks in data sharing settings, undertook his own study.

"We wanted to see to what extent these results can generalize to the real world," said Vorobeychik, associate professor of computer science & engineering in the McKelvey School of Engineering. "We explored whether it was possible to demonstrate in a more practical situation that these concerns were real."

Vorobeychik and his co-authors—WashU graduate student Rajagopal Venkatesaramani and Vanderbilt University Biomedical Informatics Professor Bradley Malin—found the task of linking faces and genomes is much harder on average than previously reported. They published their findings in *Science Advances* Nov. 17, 2021.

In the study, they developed a method to calculate the risk of reidentifying individuals from a carefully curated dataset of 126 genomes obtained from the OpenSNP genome-sharing platform by linking these to publicly posted face images. Specifically, they used

neural network models to predict visible physical traits, such as hair, eye and skin color, as well as sex, and then used this information along with known genotype-trait correlations to score possible genome-face matches.

Earlier phenotype association studies used high-quality photos taken in a laboratory setting with professional quality lighting. Vorobeychik's team, on the other hand, conducted their research using real-world photographs found on social media sites.

"What we did was construct probabilistic models for these different kinds of visual characteristics and essentially connected the dots by scoring the matching quality between particular genomes and particular faces," Vorobeychik explained. "We then used that scoring system to predict which matches are most likely."

Overall, their results suggest that it's sometimes possible to link public face images with public genomic data, but the success rates are well below what prior research papers suggest in idealized settings.

"However, our observations are about average privacy risk for a collection of individuals; it is possible that for some people the privacy risk is indeed high," Vorobeychik said.

To protect those individuals' privacy, Vorobeychik's team created a method that alters a social media photo just enough to prevent the neural network from reliably identifying visible traits, and thereby reducing the risk of those who have publicly released their genomic data and whose image appears elsewhere online.

"Our method adds enough imperceptible noise to the image so it's difficult for a deep neural network to link the phenotype of the face to a particular genome," he said. "This carefully crafted noise doesn't change

one's perception of [the face] to the naked eye."

This tool could be further developed into image filters that individuals could use to protect their social media photos from hackers who might try to link their images to genetic data they've publicly shared on OpenSNP or other online sites.

More information: Rajagopal Venkatesaramani et al, Re-identification of individuals in genomic datasets using public face images, *Science Advances* (2021). DOI: 10.1126/sciadv.abg3296

Provided by Washington University in St. Louis