MedicalXpress

Researchers expose new vulnerabilities in the security of personal genetic information

January 17 2013, by Matt Fearer

Using only a computer, an Internet connection, and publicly accessible online resources, a team of Whitehead Institute researchers has been able to identify nearly 50 individuals who had submitted personal genetic material as participants in genomic studies.

Intent on conducting an exercise in "vulnerability research"—a common practice in the field of information security—the team took a multi-step approach to prove that under certain circumstances, the full names and identities of genomic research participants can be determined, even when their genetic information is held in databases in de-identified form.

"This is an important result that points out the potential for breaches of privacy in genomics studies," says Whitehead Fellow Yaniv Erlich, who led the research team. A description of the group's work is published in this week's *Science* magazine.

Erlich and colleagues began by analyzing unique genetic markers known as short tandem repeats on the Y chromosomes (Y-STRs) of men whose genetic material was collected by the Center for the Study of Human Polymorphisms (CEPH) and whose genomes were sequenced and made publicly available as part of the 1000 Genomes Project. Because the Y chromosome is transmitted from father to son, as are family surnames, there is a strong correlation between surnames and the DNA on the Y chromosome.

Recognizing this correlation, genealogists and genetic genealogy companies have established publicly accessible databases that house Y-STR data by surname. In a process known as "surname inference," the Erlich team was able to discover the family names of the men by submitting their Y-STRs to these databases. With surnames in hand, the team queried other information sources, including Internet record search engines, obituaries, genealogical websites, and public demographic data from the National Institute of General Medical Sciences (NIGMS) Human Genetic Cell Repository at New Jersey's Coriell Institute, to identify nearly 50 men and women in the United States who were CEPH participants.

Previous studies have contemplated the possibility of genetic identification by matching the DNA of a single person, assuming the person's DNA were cataloged in two separate databases. This work, however, exploits data between distant paternally-related individuals. As a result, the team notes that the posting of genetic data from a single individual can reveal deep genealogical ties and lead to the identification of a distantly-related person who may have no acquaintance with the person who released that genetic data.

"We show that if, for example, your Uncle Dave submitted his DNA to a genetic genealogy database, you could be identified," says Melissa Gymrek, a member of the Erlich lab and first author of the *Science* paper. "In fact, even your fourth cousin Patrick, whom you've never met, could identify you if his DNA is in the database, as long as he is paternally related to you."

Aware of the sensitivity of his work, Erlich emphasizes that he has no intention of revealing the names of those identified, nor does he wish to see public sharing of genetic information curtailed.

"Our aim is to better illuminate the current status of identifiability of

genetic data," he says. "More knowledge empowers participants to weigh the risks and benefits and make more informed decisions when considering whether to share their own data. We also hope that this study will eventually result in better security algorithms, better policy guidelines, and better legislation to help mitigate some of the risks described."

To that end, Erlich shared his findings with officials at the National Human Genome Research Institute (NHGRI) and NIGMS prior to publication. In response, NIGMS and NHGRI moved certain demographic information from the publicly-accessible portion the NIGMS cell repository to help reduce the risk of future breaches. In the same issue of *Science* in which the Erlich study appears, Judith H. Greenberg and Eric D. Green, the Directors of NIGMS and NHGRI, and colleagues author a perspective on this latest research in which they advocate for an examination of approaches to balance research participants' privacy rights with the societal benefits to be realized from the sharing of biomedical research data.

"Yaniv's work is a timely reminder that in this era in which massive amounts of genomic data are being generated rapidly and shared in the interest of scientific advancement, there is an increasing likelihood of privacy breaches," says Whitehead Institute Director David Page. "I'm delighted that, thanks to Yaniv's overture to NIH, we at Whitehead Institute have the opportunity to join policymakers at NHGRI and elsewhere in what will be a critical, ongoing dialog about the importance of safeguarding data, of sharing data, and the implications of failure in either endeavor."

More information: "Identifying Personal Genomes by Surname Inference" Melissa Gymrek et al. (*Science*, January 18, 2012.