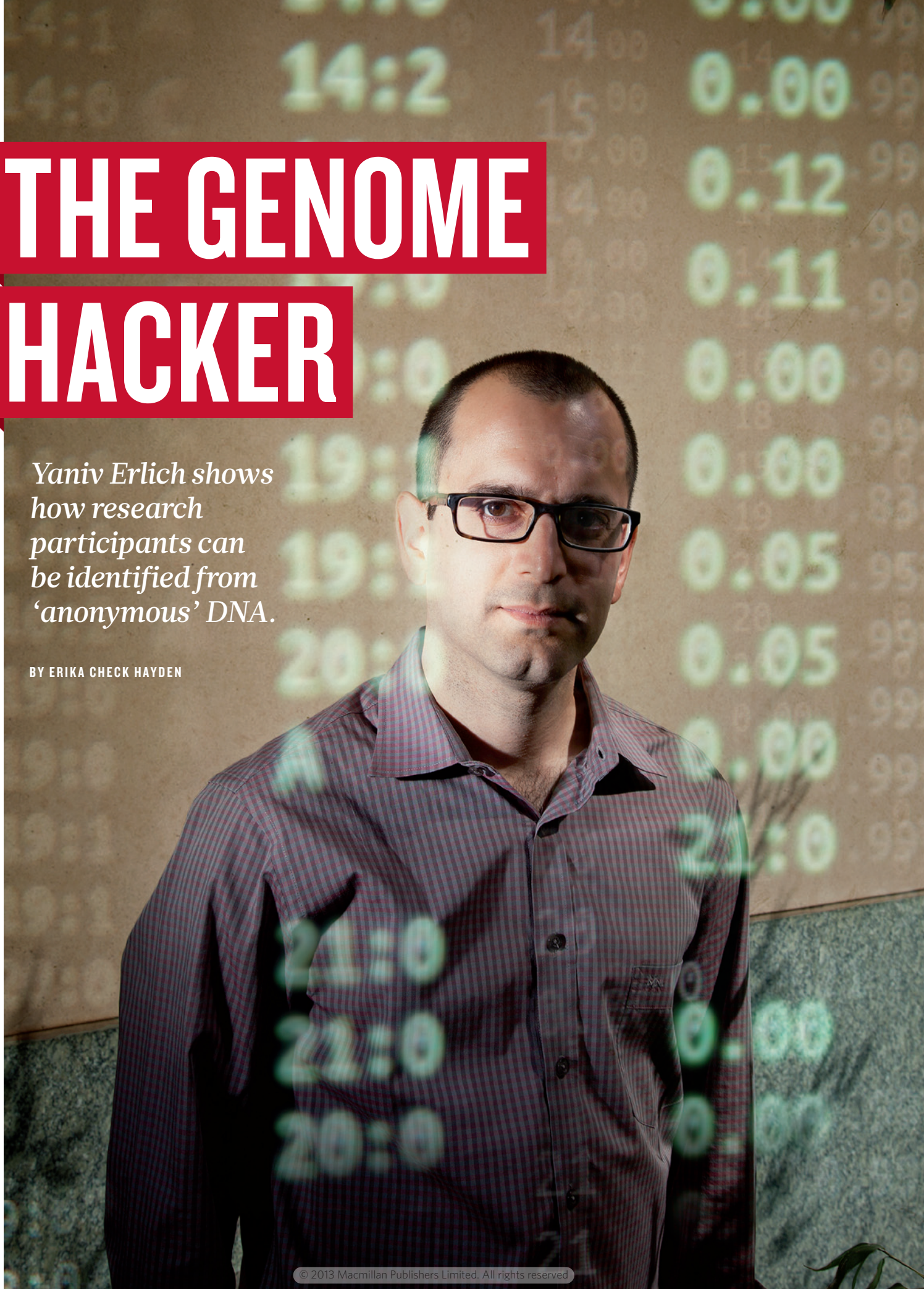
THE GENOME HACKER

Yaniv Erlich shows how research participants can be identified from 'anonymous' DNA.

BY ERIKA CHECK HAYDEN



Late at night, a video camera captures a man striding up to the locked door of the information-technology department of a major Israeli bank. At this hour, access can be granted only by a fingerprint reader — but instead of using the machine, the man pushes a button on the intercom to ring the receptionist's phone. As it rings, he holds his mobile phone up to the intercom and presses the number 8. The sound of the keypad tone is enough to unlock the door. As he opens it, the man looks back to the camera with a shrug: that was easy.

Yaniv Erlich — the star of this 2006 video — considers this one of his favourite hacks. Technically a "penetration exercise" conducted to expose the bank's vulnerabilities, it was one of several projects that Erlich worked on during a two-year stint with a security firm based near Tel Aviv. Since then, the 33-year-old computational biologist has been bringing his hacker ethos to biology. Now at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, he is using genome data in new ways, and in the process exposing vulnerabilities in databases that hold sensitive information on thousands of individuals around the world.

In a study published in January[1], Erlich's lab showed that it is possible to discover the identities of people who participate in genetic research studies by cross-referencing their data with publicly available information. Previous studies had shown that people listed in anonymous genetic data stores could be unmasked by matching their data to a sample of their DNA. But Erlich showed that all it requires is an Internet connection.

Erlich's work has exposed a pressing ethical quandary. As researchers increasingly combine patient data with other types of information — everything from social-media posts to entries on genealogy websites — protecting anonymity becomes next to impossible. Studying these linked data has its benefits, but it may also reveal genetic and medical information that researchers had promised to keep private — and that, if made public, might hurt people's employability, insurability or even personal relationships.

Such revelations may make the scientific community uncomfortable and undermine the public's trust in medical research. But Erlich and his colleagues see their work as a way to alert the world about flawed systems, keep researchers honest and ultimately strengthen science. In March, for instance, the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, claimed that the genome sequence that it had published for the HeLa cell line would not reveal anything about Henrietta Lacks — the source of the cells — or her descendants. Erlich issued a tart response: "Nice lie EMBL!" he tweeted. The sequence was later pulled from public databases, and the EMBL admitted that it would indeed be possible to glean information about the Lacks family from it, even though much of the HeLa genetic data had already been published as part of other studies.

"Most scientists would not go anywhere close to these questions, out of a sense of what it might mean for the field, or for them personally," says David Page, director of the Whitehead Institute, who has advised Erlich about his research. "But this is not about publicity-seeking — this is about fearlessness, and a kind of interest in how all the parts of the Universe fit together that mark all of Yaniv's work."

GAMING THE SYSTEM

Erlich was inspired to teach himself programming as a child in Israel after seeing the 1983 film *WarGames*, in which a teenager accidentally hacks into government computer systems and nearly launches "global thermonuclear war". Erlich thought that he would study maths and physics at university, but after a friend told him that there was a lot of maths in biology, he decided to major in computational neuroscience. In 2006, following his graduation, Erlich moved to the United States to earn his PhD in genetics at Cold Spring Harbor Laboratory in New York.

Under his adviser, molecular biologist Greg Hannon, Erlich devised what he called "DNA Sudoku": a sequencing method that could be used on tens of thousands of specimens analysed simultaneously. It allowed scientists to use computational techniques to find a gene carrying a rare mutation from this mixed batch of DNA and assign it to the right specimen[2]. Erlich is now using the technique to find disease-causing mutations in young Ashkenazi Jews to inform their decisions about potential marriage partners.

"PEOPLE WERE CONCERNED THAT THE NIH WOULD SHUT DOWN ITS DATABASES OR THAT THE PUBLIC WOULD STOP DONATING."

In 2011, as Erlich was setting up his first independent lab as a Whitehead Fellow, he met a Colorado-based woman, Wendy Kramer, whose son had managed to track down his father — an anonymous sperm donor — by searching a consumer-focused genetic-genealogy database for people with DNA similar to his own.

Erlich wondered whether a computer program that he had been working on with an undergraduate student, Melissa Gymrek, might enable a similar trick using de-identified genome data from human research studies. The software, called lobSTR, scours sequences and generates a profile of repetitive genetic markers called short tandem repeats (STRs), which are often used in genealogy to identify people. Could Erlich extract STRs from the anonymous data, and then hunt through public genealogy databases for a match and a name? "I had my background in security, and I had lobSTR in hand, and I thought, 'Is this going to affect personal genomes?'"

Erlich and his team tested the idea on a man's full genome that had been published in 2007 (ref. 3). They used lobSTR to determine the STR profile of the man's Y chromosome, and then searched a consumer genealogy database called Ysearch until they had matches with a few likely surnames. Public records on one of these surnames linked it to a man fitting the geographic location and age listed in the paper: the genomics pioneer J. Craig Venter. Venter had, in fact, already revealed himself as the donor — one reason Erlich chose that genome was that he thought he could do no harm in revealing Venter's identity. But there was no reason to believe that this process would not work for others.

PROOF OF PRINCIPLE

When Erlich submitted his paper to *Science*, the reviewers wanted proof that a completely anonymous donor could be identified. So his team extended its analysis to men whose genomes had been sequenced as part of the international 1000 Genomes Project. Extensive information about these men, including their ages and detailed family pedigrees, was available on the website of the Coriell Institute for Medical Research in Camden, New Jersey, which distributes cell lines made from their tissues to researchers.

Erlich's team used lobSTR to infer the men's STRs from their 1000 Genomes data, and then searched Y-chromosome databases to find linked last names. After that, it was relatively easy to search public records databases to find men with those last names who were the right age, came from the right place and had similar family trees. The team identified nearly 50 people, including DNA donors and their relatives. When he first saw the results, Erlich said later, he was so shocked at how easily the method worked that he had to go outside and take a walk.

Geneticists elsewhere had already revealed security flaws in anonymized genetic data. In 2008, for instance, David Craig, a computational biologist at the Translational Genomics Research Institute in Phoenix, Arizona, reported that he could use information from an

DANA SMITH

individual's DNA sample to confirm whether that person had contributed to a genome-wide association study (GWAS), even if the study reported only summary statistics on hundreds or thousands of participants[4].

This and other studies prompted policymakers at the US National Institutes of Health (NIH) to pull GWAS data from public databases, and to require investigators to obtain permission to access it. Many researchers resent this move, because it makes it difficult to pool data from different studies.

Erlich's study upped the stakes, because it showed that it was possible to identify people from their genetic data by linking not to other sources of research data, but to information freely available on the Internet. He realized that publishing these results might stoke public anger, so he consulted lots of other researchers and ethicists first. "People were concerned that the NIH would shut down its databases or that the public would stop donating their material," says Erlich. He contacted NIH officials about his findings, and met some of them in Bethesda, Maryland, last December. The NIH's National Institute of General Medical Sciences, which funds the Coriell repository, decided to remove the ages of participants from public view.

INFORMATION WITHHELD

When Erlich published the results of his work in January[1], he revealed no research participants' names. Neither did he spell out all the steps he had taken to find their identities: "There is an obvious tension, because as a scientist you want to tell everything about how you did the work. On the other hand, you can't do that, because it will expose people's identities to the world," says Erlich.

The question remains of how to handle privacy in future. Removing information after loopholes are revealed — what some call the whack-a-mole response[5] — does not seem to satisfy anyone.

Some geneticists argue that the public is becoming more accustomed to sharing personal information, and that no harm has ever been done to anyone identified from genetic studies. But many, including Brad Malin a privacy researcher at Vanderbilt University in Nashville, Tennessee, consider that a weak argument. "A lot of people say that because information flows much more freely today than it did 10 years ago, that privacy is dead, and this is certainly not the case," he says. People still expect some information — especially health and medical data — to be private, says Malin. And so far, none of the people identified from anonymous genetic data sets has been named publicly, so it is perhaps too early to say that they are not at risk.

Eric Green, director of the US National Human Genome Research Institute (NGHRI) in Bethesda, says that the NIH is trying to balance access and privacy. "One value is to make the data as widely available and unencumbered as possible, but then you're trading that off against concerns about how data is being used, and maintaining privacy and confidentiality," he says. "We're constantly exploring models that put us between those two extremes."

CAREFUL SCRUTINY

Currently, anyone with an Internet connection can access data from the 1000 Genomes Project. Researchers must apply for access to genetic data from most other studies, and must usually submit a new access request for information from each one. That makes it onerous to analyse data from different sources together.

Many large data-holders around the world take this approach; the EMBL's European Bioinformatics Institute in Hinxton, UK, for instance, relies on data-access committees to determine what uses of data are appropriate given the consent terms of any particular study. "It's difficult to imagine how else one would do it, since most of these studies are built around consent agreements," says Paul Flicek, head of DNA resources at the institute.

Some researchers say that genetic data should be deposited with central data-hosting agencies that then grant broad access to trusted users. This would mean that the data would be off-limits to the public, but researchers would not have to ask for permission to access every data set. Laura Rodriguez, director of the NGHRI's division of policy, communications and education, says that NIH committees on data use have concerns about this idea: "We've seen investigators request access to large swathes of data, and it's clear from their proposed-research statement that they haven't read the use limitations of the data they're requesting."

Erlich argues that genetic data should be broadly available, but that scientists should be more honest about the difficulty of guaranteeing anonymity. Amy McGuire, a lawyer and ethicist at Baylor College of Medicine in Houston, Texas, with whom Erlich consulted on his publication, agrees. But she is not sure that informing people of the risk of re-identification is enough. It may be difficult for someone signing up for a research study to understand all the ways in which their data might be used in the future, let alone to weigh the risks when researchers themselves do not necessarily know them. "There are challenges to putting so much weight on informed consent," she says.

Scientists should explore further ways to protect research participants, says Erlich, such as encrypting the data before they are deposited, allowing researchers who possess the decryption key to work with them freely without jeopardizing privacy. But Green is concerned that researchers might not be able to work as freely with encrypted data as they can with unencrypted data.

There are no simple answers, but researchers give Erlich credit for forcing these issues onto the public stage. Page warns that this could be a double-edged sword for a young scientist: "This piece of work represents only a slice of Yaniv's broader interests, and the danger could be the risk of being completely consumed by this debate," he says.

Erlich seems happy to be consumed. In a new project that he calls Genetic Epidemiology 2.0, for example, he is working with Daniel MacArthur, a geneticist at Massachusetts General Hospital and Alkes Price, a biostatistician at Harvard School of Public Health, both in Boston, to mine social networks for information that might yield insight into the genetic basis for complex human traits. The project focuses on genealogy-based social networks on which members post extensive family trees — a potentially rich source of information about inherited traits.

Erlich is aware of the ethical complexities of such a study. To start with, he is focusing on public information about deceased people, to minimize the risk that anyone will be harmed by the work. But if the project succeeds, he may go on to ask members of the networks whether they want to upload other types of information — such as medical records, which could yield insight into a wider range of disease traits.

It is a project that plays to Erlich's strengths, says Hannon. "When Yaniv says, 'What data is out there?' he doesn't think, 'What data is out there in the literature?' He thinks about what data is out there holistically." If the technique works, it would use information in the public domain to tackle one of the most difficult problems facing genetic researchers: how to assemble the enormous groups of related individuals needed to illuminate the complex genetic underpinnings of human biology. "Yaniv believes nothing is impossible," says Hannon.

Of course, it could expose all kinds of new vulnerabilities. That may not be such a bad thing, says Erlich, harking back to his penetration testing on banks. "As a client of a US bank, I'm sure you are happy that they undergo these tests. You wouldn't want to say, 'Let's not find something we won't like.'" ■

"YANIV BELIEVES NOTHING IS IMPOSSIBLE."

Erika Check Hayden *writes for* Nature *from San Francisco, California.*

1. Gymrek, M., McGuire, A. L., Golan, D., Halperin, E. & Erlich, Y. *Science* **339,** 321–324 (2013).
2. Erlich, Y. *et al. Genome Res.* **19,** 1243–1253 (2009).
3. Levy, S. *et al. PLoS Biol.* **5,** e254 (2007).
4. Homer, N. *et al. PLoS Genet.* **4,** e1000167 (2008).
5. Ohm, P. *UCLA Law Rev.* **57,** 1701 (2010).