

GENETIC PRIVACY

Whole-Genome Data Not Anonymous, Challenging Assumptions

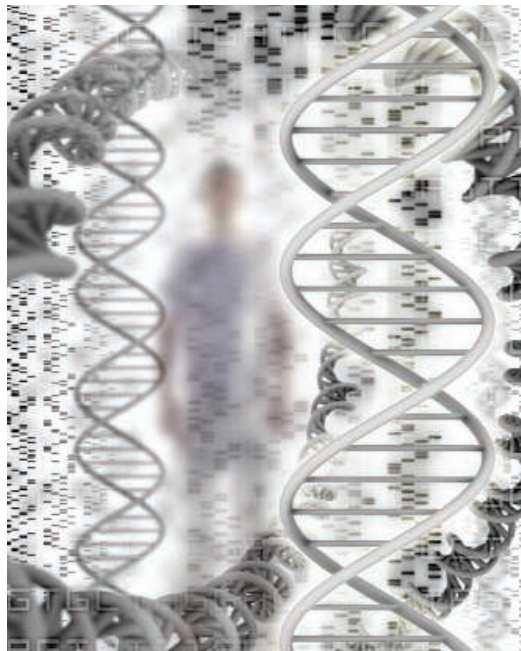
Last week, scientists learned that a type of genetic data that is widely shared and often posted online can be traced back to individuals who proffered up their DNA for research. The revelation, in a paper published in *PLoS Genetics*, prompted the National Institutes of Health (NIH) in Bethesda, Maryland, and the Wellcome Trust in the United Kingdom to strip some genetic data from their publicly accessible Web sites, and NIH to recommend that other institutions do the same.

The concern is with studies in which researchers pool genetic data from hundreds of people to look for broad patterns of genetic inheritance. Because the pool consists of DNA from so many people, the assumption has been that it would be impossible to identify any one individual's DNA. The new study suggests that's not the case. NIH officials and others agree that the likelihood of a breach of privacy is low, largely because the pooled data must be matched against a particular person's isolated DNA—something that, currently, only researchers generally have access to.

But the discovery that these DNA pools don't protect anonymity is still troubling, especially because no one had considered that a possibility. The first response to the results "is, 'You're crazy,'" says David Craig, a geneticist at the Translational Genomics Research Institute in Phoenix, Arizona, who conducted the work. Less than 9 months ago, NIH was so confident in the anonymity of pooled genetic data that it recommended it be made public for all researchers to use.

Craig found this confidence misplaced, for a simple reason: Geneticists now routinely examine hundreds of thousands of DNA variants, called single-nucleotide polymorphisms (SNPs), at a time, instead of hundreds as they did just a few years ago. As a result, they're gathering enough information about the pattern of SNPs in a pooled sample that it's feasi-

ble to deduce whether a particular individual, with her own unique SNP blueprint, is represented in a much bigger pool of DNA—even if that person's DNA was less than 1% of the mix. Craig and his colleagues managed to do this by ascertaining the distribution pattern of every single SNP—essentially, asking the same question 500,000 times. They were successful because, it turns out, every individual shifts a genetic pool subtly in certain directions, and studying enough SNPs unveils the pattern of those shifts. The biggest chance of error comes from false posi-



Faceless no more. A new study shows that individuals can be pinpointed in pooled DNA.

tives from relatives whose DNA may also appear in the pool, says Craig.

NIH officials were startled when Craig notified them of his findings about 2 months ago; they had their own statisticians repeat the experiments. "They said, 'Yup, this works,'" says Elizabeth Nabel, head of NIH's Heart, Lung, and Blood Institute. "We still consider the risk to the individual relatively low," she continues,

but "there's a window of vulnerability."

The greatest concern is that identifying an individual this way could reveal sensitive health information. Genome-wide association studies compare data from people with and without a particular disease, so knowing which pool a person falls into can convey whether they have, say, cancer, or diabetes, or multiple sclerosis. "We have a false sense of security with pooled data," says Pablo Gejman, a psychiatric geneticist at Northwestern University in Evanston, Illinois. "There is sensitive information" here.

The Wellcome Trust has pulled data on about a dozen common diseases, and NIH has pulled data from nine genetic studies off two sites, dbGaP, which includes genome-wide association studies, and CGEMS, a site for cancer genetics work. The seven affected studies on dbGaP had been downloaded by about 1000 people all told, says James Ostell, who oversees that and other NIH databases.

NIH officials are informing geneticists about the policy change through e-mails and their Web site; the Broad Institute in Cambridge, Massachusetts, has followed suit and removed pooled data from its site. This is "a logical choice, a necessary choice," says Michael Boehnke, a statistical geneticist at the University of Michigan, Ann Arbor, whose data from a diabetes study was taken down from NIH.

Nabel says that NIH is considering a new policy in which the pooled data will be released to researchers who apply, as is now the case with data traditionally considered much more sensitive.

Still, Ostell and others say the current privacy risk is minimal. It could be of more concern 5 or 10 years from now, as genetic information proliferates. One possible scenario is that law enforcement agencies might turn to pooled data to determine whether their suspect is present—and even demand that the researcher help them identify him.

Craig's work could help future forensic investigators in another way: Currently, they're unable to identify a suspect's DNA in a mixed sample—say, a sample of blood from several people—if the suspect's blood is less than 10% of the total. "A lot of forensic crime samples do have small contributions from people of interest, [and] right now we can do essentially nothing," says Bruce Weir, a biostatistician who studies genetics and forensics at the University of Washington, Seattle. **—JENNIFER COUZIN**