POLICY FORUM

The alleged Golden State Killer, Joseph DeAngelo, appears at his arraignment in Sacramento, California, in late April.

PHOTO: CHAIWAT SUBPRASOM/REUTERS PICTURES

SCIENCE AND LAW

Is it time for a universal genetic forensic database?

Bias and privacy concerns cloud police use of genetics

By **J. W. Hazel[1,2], E. W. Clayton[1,2,3], B. A. Malin[2,4,5,6], C. Slobogin[2,3]**

DNA is an increasingly useful crime-solving tool. But still quite unclear is the extent to which law enforcement should be able to obtain genetic data housed in public and private databases. How one answers that question might vary substantially, depending on the source of the data. Several countries—the United Kingdom, Kuwait, and Saudi Arabia among them—have even toyed with creating a "universal" DNA database, populated with data from every individual in society, obviating the need for any other DNA source (*1*). Although this move would be controversial, it may not be as dramatic as one might think. In the United States, for example, the combination of state and federal databases (containing genetic profiles of more than 16.5 million arrestees and convicts) and public and private databases (containing genetic data of tens of millions of patients, consumers, and research participants) already provides the government with potential access to genetic information that can be linked to a large segment of the country, either directly or through a relative (*2*, *3*). We discuss here how, if correctly implemented, a universal database would likely be more productive and less discriminatory than our current system, without compromising as much privacy.

Current law enforcement methods of genetic investigation are both haphazard and underregulated. In early 2018, U.S. law enforcement officers investigating the Golden State Killer case were able to home in on a suspect after querying GEDmatch, a publicly accessible database that encourages consumers to upload genetic data coupled with personal identifiers in order to gain insights into their genealogy. Without authorization from a court, law enforcement simply pretended to be the donor of what was, in fact, crime scene DNA. Through that ruse, officers found a match to a person in the database who was distantly related to Joseph DeAngelo, the man ultimately arrested for the crimes. Since these revelations came to light last spring, multiple law enforcement agencies have used similar long-range familial searches of publicly accessible databases to close 13 cold cases, including several murders (*2*, *4*).

In the Golden State case, the government could justify its action by pointing to the fact that GEDmatch is advertised as a publicly accessible database—one that does not specifically ban the type of deception police used in that case. But publicly accessible databases are not the only source of genetic information that law enforcement might query. For instance, if accessing such a database fails to yield a useful result, which will often be the case, law enforcement could resort to private databases, such as those maintained by direct-to-consumer (DTC) companies, e.g., 23andMe and Ancestry.com. Although these databases are not as easily exploited as databases meant to be accessed by the public, in most jurisdictions in the United States and throughout the world a subpoena is all that law enforcement needs to force those companies to determine whether they have a match with crime scene data. A subpoena only

[1]*Center for Biomedical Ethics and Society, Vanderbilt University, Nashville, TN 37203, USA.* [2]*Center for Genetic Privacy and Identity in Community Settings, Vanderbilt University Medical Center, Nashville, TN 37203, USA.* [3]*Vanderbilt University Law School, Nashville, TN 37203, USA.* [4]*Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA.* [5]*Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN 37203, USA.* [6]*Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN 37203, USA. Email: james.w.hazel.1@vumc.org*

requires showing that the data sought are relevant to an investigation and is therefore much simpler to get than a warrant based on probable cause.

Until now law enforcement has largely focused its efforts on targeting publicly accessible resources such as GEDmatch. But requests for privately maintained data are likely to become much more frequent in the future, given the increasing value of genetic data to law enforcement, the low level of justification required for a subpoena, and the tremendous amount of effort that can be associated with long-range familial searching by using a resource such as GEDmatch, which might generate dozens or hundreds of possible leads in a given case (*2*).

If publicly accessible databases and DTC companies are of no help, law enforcement might try to access genetic data in the possession of healthcare providers and researchers. Again, only a subpoena is needed to obtain genetic information contained in patients' electronic medical records under the U.S. Health Insurance Portability and Accountability Act of 1996 (*5*). And although biomedical research efforts are often protected by government-issued Certificates of Confidentiality (*6*), which purport to assure participants that research data are immune from court orders, the enhanced protections recently conveyed by the U.S. 21st Century Cures Act of 2016 remain largely untested in the courts. Further, because Certificates of Confidentiality typically apply only to research funded by the National Institutes of Health and other agencies within the U.S. Department of Health and Human Services, genetic research funded by other sources remains largely unprotected unless a request for a Certificate of Confidentiality is made and granted.

Last, in addition to these public and private resources, a government interested in using DNA to help solve crimes can maintain its own database. In the United States, many states and the federal government maintain DNA profiles not only of convicted felons but also of those simply arrested for a felony or, in some cases, even a misdemeanor (*1*). The U.S. Supreme Court has given its imprimatur to such databases (*7*). As we explain below, this development is one of the most potent reasons for considering establishment of a more comprehensive genetic database.

UNIVERSAL DATABASE

The first obvious benefit of a universal database is its potential for solving or deterring serious crimes such as murder, rape, robbery, and burglary. As both research and anecdotal reports indicate, DNA matches have often been crucial in catching the perpetrators of such crimes and useful in identifying bodies and remains as well (*8*, *9*). Unfortunately, from law enforcement's perspective, forensic databases that contain only genetic data of arrestees and those convicted of crimes have serious limitations, a fact demonstrated by law enforcement's increasing reliance on publicly accessible and private databases, composed almost entirely of "innocent" individuals. And when law enforcement chooses the latter route, a match is by no means guaranteed; additionally, considerable inefficiency is likely if the effort to find a match requires consulting numerous companies, all of which may need to re-analyze their sample to generate the relevant profile.

Just as important, a universal database would eliminate or reduce problems associated with the current haphazard genetic investigative regime. First, such a database would virtually erase the government's incentive to conduct long-range familial DNA searches of the type used in the Golden State Killer case. It would thus markedly alleviate the impact on innocent people who happen to be related to criminals and whom police are likely to treat as suspects unless and until countervailing evidence surfaces.

Second, a universal database would eliminate the temptation on the part of law enforcement to use public, DTC, or research databases for investigative purposes. Indeed, for reasons we give below, universal database legislation should prohibit law enforcement officials from trawling nongovernmental DNA sources such as GEDMatch, Ancestry.com and research-oriented databases. That in turn might enhance research into diseases, treatments, and other socially beneficial avenues because studies indicate that many people, especially those of color, are reluctant to provide genetic information to researchers out of fear it will be misused by the government (*10*, *11*).

Last, a universal database would be less discriminatory than the government's current method of compelling genetic samples. If the government collects DNA only from convicted individuals or only from individuals arrested for serious crimes—as is true as a matter of law in the United Kingdom and as a practical matter with the U.S. Combined DNA Index System (CODIS)—there is real concern that the resulting databases will be skewed against the disadvantaged because they are the ones most likely to be the focus of such convictions and arrests.

The situation in the United States has been exacerbated by federal, state, and local governments now creating "shadow" databases (*9*)—not only of people arrested for any crime but also of people who are merely stopped on suspicion of having committed a crime without being arrested (the so-called "stop-and-spit" and "swab-and-go" practices). As a result, arrest-based DNA databases contain a huge proportion of the young nonwhite male population and a much smaller representation of other groups (*9*, *12*). Indeed, that is why police had to rely on a publicly accessible database to catch the Golden State Killer, a white former police officer; in sharp contrast to government DNA caches such as CODIS, public and DTC databases tend to contain the genetic data of predominately white individuals, generally from higher income brackets.

Despite these advantages of a universal database, many concerns have been raised about its privacy implications and the associated potential for misuse of genetic information. As a result, in some countries a universal database is clearly prohibited. In *S. and Marper* v. *United Kingdom* (*13*), the European Court of Human Rights concluded that the indefinite retention of biological samples and profiles (including not only genetic data but also fingerprints and other biological information) is a violation of the right to privacy protected under the European Convention of Human Rights. That not only spells doom for universal genetic databases, it also prohibits long-term databases composed of profiles of people who are arrested but not convicted. In response to Marper, the United Kingdom, which had been retaining the DNA samples of virtually all arrestees, now destroys such samples immediately if collected from individuals charged with minor crimes and after 3 years for those arrested for serious crimes. Although Marper applies only in the Council of Europe's 47 member countries, many other countries follow its dictates (*1*).

ALLAYING CONCERNS

To some extent, the decision in Marper is based on fear that those in the database will be associated with criminality. But that drawback is specific to databases that focus on arrestees; the criminal stigma of being in a database is eliminated if everyone's DNA is acquired. More relevant is Marper's objection that broad collection of genetic material might increase "the risk of abuse and arbitrariness" (*13*). These concerns would clearly be raised by the establishment of a universal database, but they can be allayed in a number of ways.

Most important to recognize is that a forensic database would only require a subset of genetic markers with little to no medical relevance. Profiles would consist of a few dozen short-tandem repeats, with perhaps a modest expansion of the 20 CODIS loci currently used to improve the identification of degraded samples or the addition of a limited subset of "forensic" single-nucleotide polymorphisms to enhance the identification of more distant relatives in the rare instances in which familial searches were still needed (*3*). As a result, these law enforce-

ment profiles would reveal substantially less sensitive information than the thousands (or hundreds of thousands) of genetic variants, often coupled with individual and family medical histories, that are found in the healthcare, research, or DTC ecosystems that law enforcement might otherwise be tempted to commandeer.

Many other protections against misuse of DNA databases can and should be created by the relevant legislative body (which, in the United States, would be Congress, given the nationwide impact of the law). For instance, legislation could require that genetic data not only be uncoupled from any personal identifiers within the system, as it is in CODIS, but also establish a more robust "unmasking" process that limits law enforcement access to any personal information until an association has been made and confirmed (a procedure better monitored through one central system than state-by-state or company-by-company). In further contrast to the current system, legislation might limit access to the database to specific circumstances, such as investigations into felonies and identification of missing persons' remains.

> ***"...legislators, their kin, and their key constituents will be affected just like everyone else."***

Universal database legislation should also require that the DNA database be housed in an independent agency and that access to it be authorized by a warrant (not just a subpoena) based on probable cause to believe a match will produce a perpetrator (a showing that is usually impossible with a database that is not universal). Most important, the law should require that the physical samples analyzed to create the database be destroyed after obtaining the relevant genetic information, to mitigate the risk that the sample will be subjected to further analysis or used for purposes other than populating the database.

Additional privacy protection could be realized through emerging cryptographic protocols that control access to genomic data through multiple keys. Where more than one organization is required to "turn the key" to decrypt any record, the risk of a rogue individual or agency misusing the resource is substantially mitigated (*14*). Simultaneously, because law enforcement needs would be fully met, Congress should (and probably would) severely restrict the ability of law enforcement to search other health-care, research, or DTC databases, increasing trust in these activities and avoiding government access to the more complete genetic information housed there.

Whatever its precise structure, the most important point is that the population-wide nature of the database would all but guarantee the adoption of strong security measures such as those just described, as well as the enactment of harsh penalties for abuse of the type currently associated with misuse of data in CODIS (a fine of up to $250,000 or imprisonment for up to 1 year). That is because members of Congress would know that government DNA harvesting would no longer focus solely on out-groups but would also sweep in their own DNA. As the ubiquity of federal and state legislation strictly regulating the privacy of communications records and tax information suggests, slippery-slope concerns about government collection and exploitation of every citizen's full genetic makeup dissipate in a regime in which legislators, their kin, and their key constituents will be affected just like everyone else.

These concerns are further minimized in jurisdictions such as the United States and Europe that, unlike many countries that have considered a universal database, have codified basic privacy protections that would mitigate the potential for abuse or misuse of the data (for example, the Privacy Act of 1974 and the Genetic Information Nondiscrimination Act in the U.S., and the General Data Protection Regulation in Europe).

IMPLEMENTATION ISSUES

There remain implementation issues that would need to be debated by the public and ultimately resolved by Congress, including whether a universal database should be populated by obtaining samples from all newborns or instead through a census-style effort (or a combination of both), how to collect the DNA of visitors from other countries, and the appropriate incentives to promote compliance with the program.

The ethical objections to mandating forensic profiling of newborns and/or compelling every citizen or visitor to submit to a buccal swab or to spit in a cup when they have done nothing wrong are not trivial. But newborns are already subject to compulsory medical screening, and people coming from foreign countries to the United States already submit to fingerprinting. It is also worth noting that concerns about coercion or invasions of privacy did not give pause to legislatures (or, for that matter, even the European Court) when authorizing compelled DNA sampling from arrestees, who should not forfeit genetic privacy interests simply by virtue of being arrested.

A universal database would not be cheap; extrapolating from a $20- to $40-per-profile estimate for the existing CODIS system calculated in 2010 (*15*), compiling a database of ~350 million people could cost between $7.5 billion and $15 billion dollars. Although this figure does not include implementation costs (which are difficult to estimate), the economies of scale associated with a universal system, coupled with the declining cost of forensic profiling, would likely drive this figure lower.

In addition, the societal and economic benefits that could be derived from the system could easily offset these costs. Criminal activity is extremely expensive for private citizens (both monetarily and in terms of intangible harms to victims), third parties (such as businesses and insurers), and the government (for police investigations and incarceration). There is evidence that existing forensic databases have more than made up for their initial costs by increasing the efficiency, accuracy, and success rate of ongoing criminal investigations and by deterring would-be criminals (*15*).

At the very least, putting the idea of a universal forensic database on the table would spur a long overdue debate about the deficiencies of the current system and, more broadly, our societal commitment to privacy, fairness, and equal protection under the law. ■

REFERENCES AND NOTES

1. H. M. Wallace, A. R. Jackson, J. Gruber, A. D. Thibedeau, *Egyptian J. Forens. Sci.* **4.3**, 57 (2014).
2. Y. Erlich, T. Shor, I. Pe'er, S. Carmi, *Science* **362**, 690 (2018).
3. J. Kim, M. D. Edge, B. F. B. Algee-Hewitt, J. Z. Li, N. A. Rosenberg, *Cell* **175**, 848 (2018).
4. A. Regalado, *MIT Technol. Rev.* (13 September 2018); www.technologyreview.com/s/612001/hundreds-of-crimes-will-soon-be-solved-using-dna-databases-genealogist-predicts.
5. 45 Code of Federal Regulations 164.512(f).
6. 42 U.S. Code 241(d)(1)(E).
7. *Maryland* v. *King*, 569 U.S. 435 (2013).
8. H. Safir, *Forens. Mag.* (1 September 2007); www.forensicmag.com/article/2007/10/dna-technology-effective-tool-reducing-crime..
9. T. Jones, *Guardian* (9 April 2018); www.theguardian.com/politics/2010/apr/10/dna-analysis-crime-solving.
10. E. W. Clayton, C. M. Halverson, N. A. Sathe, B. A. Malin, *PLOS ONE* 10.1371/journal.pone.0204417 (2018).
11. L. G. Landry, N. Ali, D. R. Williams, H. L. Rehm, V. L. Bonham, *Health Aff.* **37**, 780 (2018).
12. R. Brame, S. D. Bushway, R. Paternoster, M. G. Turner, *Crime Delinq.* **60**, 471 (2014).
13. *S. and Marper* v. *the United Kingdom*, Grand Chamber of the European Court of Human Rights, Strasbourg (4 December 2008); https://rm.coe.int/168067d216.
14. H. Cho, D. Wu, B. Berger, *Nat. Biotech.* **36.6**, 547 (2018).
15. J. L. Doleac, *Am. Econ. J. Appl. Econ.* **9.1**, 165 (2017).

ACKNOWLEDGMENTS

The authors thank members of the Center for Genetic Privacy and Identity in Community Settings (GetPreCiSe) for their thoughtful comments and feedback. **Funding:** This work was funded by the Center for Genetic Privacy and Identity in Community Settings (NIH/NHGRI RM1HG009034). **Author contributions:** All authors contributed equally to this work. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All data is available in the main text.

10.1126/science.aav5475

Is it time for a universal genetic forensic database?

J. W. Hazel, E. W. Clayton, B. A. Malin and C. Slobogin

Science **362** (6417), 898-900.
DOI: 10.1126/science.aav5475