Forensic DNA Profiles: 
Database Expansion, Familial Search, and a Radical Solution

Michael R. Seringhaus

Knight Law & Media Scholar and Information Society Project Fellow
Yale Law School
127 Wall St., New Haven CT 06511
michael.seringhaus@yale.edu

Abstract

This paper outlines the substantial privacy risks posed by the human genome, and explains how sensitive information is omitted from FBI DNA profiles. It then details the quiet expansion of law enforcement DNA databases, including the controversial technique of “familial DNA search,” and argues for a population-wide, universal DNA database.

Genomes, Profiles and Privacy Risks

The human genome is potentially a privacy nightmare. Citizens and scholars alike are increasingly concerned about the threat of highly personal information being exploited to invade individual privacy.

Genetic privacy is worthy of concern: the genome is an information-rich data warehouse that poses a considerable privacy threat for two reasons.

First, human genomic DNA encodes the biological blueprint necessary to create an individual person. As such it necessarily includes a vast amount of sensitive information, from outwardly observable characteristics, to sensitive health information such as genetic predispositions to disease. Indeed, it is difficult to imagine any information more deeply personal. Moreover, the genome carries substantial information about relatives, particularly close relatives like parents, children and siblings.

Second, we all routinely shed copies of our genomic DNA. A human being consists of some 50-100 trillion cells, almost all of which carry a complete copy of that individual’s genome. And we deposit copies of our genomes everywhere—in shed skin cells, in saliva on soda cans and cigarette butts, and so on—free for the taking (and analysis). This reality will surely pose challenging legal questions in coming years.

However, this paper concerns a particular and very narrow use of genomic information: the forensic DNA profile, as used in Combined DNA Index System (CODIS), the FBI DNA database system. These profiles ignore all the sensitive genomic information mentioned above, and distill just 26 two-digit numbers from any genomic sample.

As such, they pose little if any privacy risk, and accordingly different standards should be brought to bear in evaluating their collection and retention by government. This paper presents the central argument of a longer paper currently in preparation: namely, the argument for a population-wide, universal database of CODIS-style DNA profiles.

The Genome: Scientific Background

Your genome is composed of some 3 billion nucleotides (chemicals represented by the letters A, C, G and T) divided among 23 chromosomes. Like all humans, you are a diploid organism—meaning that each somatic cell in your body carries two copies of each chromosome—and thus you have 46 chromosomes per cell. In sequence terms, your genome is roughly 99.5% identical to that of any other human. That is, when our genome sequences are aligned, you and I should differ at only about 0.5% of our 3 billion sequence positions, or 15 million nucleotides in all.

It appears that only a small fraction of the genome, about 2%, encodes genes. (In general terms, genes are nucleotide sequences that are “read” to create cellular components called proteins.) The remaining 98% consists of what was once called “junk DNA.” (No more: this term has fallen out of favor in recent years as researchers uncovered functionality in this vast repository of non-coding DNA.) Nonetheless, it remains true that the bulk of the genome does not directly encode genes, and whatever other functions this sequence may perform, it almost certainly also includes vast stretches of true “junk”: meaningless repeated DNA left over from thousands of years of viral invasion and duplication.

Despite being freighted with seemingly meaningless extra sequence, the genome is your biological blueprint, and an immensely potent information repository. When decoded by cellular machinery, the genome provides
information necessary and sufficient to create a unique organism of astounding complexity: you.

Of course, to state that your genome encodes sensitive health-related information is a considerable understatement. Provided scientists continue to decode patterns of genomic information, the genome will someday be able to divulge the genetic component of virtually every biological fact about you.

Humans have always carried their genomic information in DNA. But this has only recently become a concern for privacy advocates because technology has now enabled us to read the genome. DNA sequencing has evolved rapidly over the past decade, and is today a $10,000-per-genome proposition. In the next decade, that figure will continue to plummet—the next target is $1000 per individual genome sequenced—especially as new chemical-free sequencing technologies such as IBM’s sequencing chip come online. However, it is not necessary to sequence the entire genome to learn anything of value from it. Indeed, a very small target area of interest can be sequenced (at greatly reduced cost), or cheaper still, specific regions can be examined via other methods such as PCR. DNA profiles, for instance, do not involve any sequencing at all.

What DNA Profiles Are (and Aren’t)

DNA profiles (as used in CODIS) are to be sharply distinguished from the genomic information from which they arise.

Whereas the genome is a massive dataset rich with biologically meaningful information, DNA profiles are very small data strings that discard 99.999996% of all genomic information and loosely paraphrase the rest.

Profiles are computed by quantifying a select group of biologically meaningless sequence repeats, members of the class of true “junk” viral leftovers detailed above.

Scientists have classified DNA repeats into various groups according to their characteristics. The group of interest to forensic scientists is the Short Tandem Repeat (STR). These are repeated DNA sequences wherein the repeated block is 2-10 nucleotides long (e.g., “ACGTC” is 5 nucleotides long), and wherein that block is repeated end-to-end anywhere from 5 to 100 times (e.g., ACGTCACGTACGTCACGTACGTCACGTC is simply this block repeated 6 times in series). STRs are useful as identifiers for a simple reason: different people have different numbers of repeats at a given STR site. As a special bonus, the precise number of repeats one carries at a given site has no known biological relevance (and is unlikely to in future).

Thus, by tabulating just the number of repeats present at various known sites, researchers can generate a DNA “fingerprint” for an individual. Examining a number of sites in this way creates a surprisingly robust identifier.

More than 10,000 STR sites exist in the human genome. To generate a DNA profile, the FBI characterizes just 13. Because each individual carries 2 copies of each chromosome, examining 13 chromosomal STR sites will generate 26 numbers per individual—one distinct sites for each copy of each chromosome—where the numbers correspond to the number of times the repeated sequence block repeats at each site.

In sum, a DNA profile is a set of 26 numbers. These numbers act as an identifier, allowing a unique numerical tag to be assigned to each human being (with the notable exception of identical twins, who share identical genomes). Even using just 13 STR loci, these identifiers are surprisingly powerful: the odds of two random individuals sharing the same number of repeats at all 26 positions is often quoted as one in several hundred billion.

These are the strong statistics upon which courtroom forensic DNA identification is based.

Importantly, to recap, a DNA profile contains no sequence information. It is incapable of disclosing health-related information, or any other biologically meaningful data (aside from providing a clue to two individuals’ relatedness).

A DNA profile is a means of abstracting a unique individual identifier string from a complex data set. The subtleties and biological information content of the underlying genomic information are lost in the process.

The Legal Justification for DNA Collection

CODIS has expanded rapidly since its founding in 1998. Today it houses over 7 million DNA profiles of the type described above.

Compulsory DNA collection constitutes a search under the Fourth Amendment.

As such, to be constitutionally permissible, this search must be reasonable. That said, U.S. Circuit Courts have widely upheld compulsory DNA collection from convicts against Fourth Amendment challenges, using convicts’ diminished expectation of privacy to justify applying a lenient standard of analysis: the totality of the circumstances balancing test. In this test, the invasion of privacy is weighed against any legitimate government interest served by the search.

The convicts’ diminished expectation of privacy not only permits the application of this lower standard test, but also skews its result in favor of government interests.

Courts have also upheld, on similar reasoning, collection from parolees and prisoners on other forms of supervised release.

Recent Developments in DNA Profiling

Arrestee DNA

CODIS contains profiles of two types: forensic profiles, generated from crime scene samples, and offender profiles, taken largely from incarcerated offenders. Originally, only certain classes of convicted violent felons qualified for inclusion in CODIS, but the list of qualifying offenses has swelled over the past decade. Most recently, in early 2009,
the federal government announced that it would join several states in including DNA profiles from arrestees not yet convicted of any crime. The legal justification for including arrestee DNA was not immediately clear, at least when juxtaposed with the “diminished expectation” reasoning applied to collection from convicts. After all, a person convicted of a crime expects to surrender certain liberties, but one merely arrested and awaiting trial may well be acquitted or have charges dropped.

State courts have been mixed in their treatment of arrestee DNA profiling. Federal courts first considered the issue in May 2009, in U.S. v. Pool (E.D. Cal., May 27, 2009), and held that inclusion of arrestee DNA profiles was permissible on the theory that the prior finding of judicial probable cause necessary to effect the arrest in the first place was also sufficient to justify the compulsory collection of DNA.

DNA collection from an ever-broader swath of offenders is currently underway, with the obvious goal of expanding database coverage.

**Familial DNA Search**

Familial DNA search represents a de facto expansion of the DNA database to include close relatives of profiled offenders.

Suppose that individual A commits a crime, and a forensic DNA sample is recovered from the crime scene. Individual A has no prior arrest record and is not profiled in CODIS. As such, a search of the forensic sample against the database returns no matches.

Now, suppose that individual A has a biological brother B, who has been arrested and convicted of a violent felony and is profiled in CODIS. Again using the crime scene sample, the law enforcement agency can perform a “familial search” of the database—a search seeking not a perfect match, but profiles that match only at some fraction of the 26 alleles. This search is much more likely to return a partial match to individual B than to a random person. This is because first-degree relatives (such as siblings, parents and children) share on average at least half their DNA. Searching with A’s crime scene sample will return a partial match to B.

When partial matches are returned by accident, they are called “partial matches”—when deliberately sought, the same technique is called “familial DNA search.”

Returning a partial profile match to individual B may be useful. Police know B did not commit the crime—the profiles are not an exact match—but they also know it is likely one of his relatives did. As such, the police can use B to reach A.

The legal status of this search technique remains unclear. As of this writing (December 2009) the FBI does not perform familial search, although it now permits states to do so. Several states have implemented official familial search policies, including Colorado and California. (So far, only Maryland has banned the practice.) Familial search is used routinely in the United Kingdom, where it has been used to solve several high profile cases. U.S. courts have yet to hear the issue.

On its face familial DNA search seems difficult to justify, at least under the “diminished expectation” reasoning used to support compulsory DNA collection from convicts and arrestees. After all, relatives of criminals surely do not have a diminished expectation of privacy simply by virtue of their relatives’ misdeeds.

On the other hand, it is not clear that mere inclusion in the database constitutes a Fourth Amendment search at all. If the “search” consists of actually taking a biological sample (such as a cheek swab or a blood vial) to extract DNA, then family members have not been searched at all. If instead the search is the querying of a (perhaps virtual) DNA profile against crime scene DNA samples, then familial search does subject family members to a Fourth Amendment search, and that search would need to be justified differently from current DNA collection.

From a technical standpoint, familial DNA search also does not work particularly well. In practice, partial matches return not only relatives of the individual providing the crime scene sample, but also a large number of false positives. The odds of matching 50% of all positions (13/26 alleles) in the general population are roughly 3%, equivalent to returning 200,000 records in a single “partial match” CODIS search. Moreover, CODIS software is not optimized to spot the characteristic DNA inheritance patterns that relatives share. Thus, whereas 26 alleles certainly suffice for robust sample-to-sample exact matches, they are ill-suited to effectively matching relatives.

**Racial Bias in CODIS**

Members of certain races are overrepresented in CODIS. Although this is simply a reflection of national crime statistics, it is nonetheless a source of complaint about the database. Familial search, however, threatens to amplify this bias still further. By effectively expanding database coverage to include close relatives of an already skewed profiled population, the database may approach universal coverage for certain races while almost entirely omitting others. Even if this may ultimately prove legally justifiable under equal protection jurisprudence, it may nonetheless prove socially and politically problematic.

**Mounting Problems**

The general trend in DNA profiling is towards increasing population coverage. The government seeks to expand the number of individuals profiled in CODIS, and for good reason: a larger profiled population means better success at linking criminals to crimes, particularly when exotic new techniques such as familial search are used.

However, this expansion comes at a cost: CODIS is fast outstripping its original mandate—to serve as a database of certain classes of convicted felons—and moreover threatens to overstep its legal authorization.
Familial search stands on questionable legal footing and is an overextension of DNA profile technology, fraught with false positives and of limited practical use. Moreover, it threatens to skew even further the racial bias in CODIS.

A (Radical) Solution:
A Universal DNA Database

A possible solution to these increasing problems is a population-wide universal DNA database.

This proposal is tenable specifically because of the limited information content of DNA profiles themselves: whereas the genome poses a serious privacy risk, DNA profiles do not. This section presents the characteristics of the proposal, as well as arguments in favor of the universal database.

The Proposal in Detail

The universal DNA database should contain DNA profiles (13 STR, CODIS-style) for all U.S. citizens.

DNA samples could be acquired at birth, or later in life. U.S. newborns already submit to a heel-stick blood test as a matter of course for test to various genetic diseases. Adding a DNA profiling step to this process would be straightforward. Mandatory or quasi-mandatory profiling of adults could be effected if collection were implemented as a necessary corollary to obtaining a driver’s license, social security card, or other such government identification. Such collection could be justified under existing Fourth Amendment special needs / suspicionless search jurisprudence: beyond law enforcement use, there may exist compelling government justifications for obtaining DNA samples for identity verification, identification of remains in traffic accidents, and so forth.

A crucial point: use of the biological sample must be limited to generating only a DNA profile. No phenotypic prediction, biological trait analysis, or other profiling is suggested or endorsed here.

To further insure against misuse, biological samples should not be retained. Once the DNA profile is generated from the biological source material, the sample should be destroyed. (This is true only for individual identifying samples, not forensic crime scene evidentiary samples). This could be accomplished, for instance, by the use of a dedicated machine that automatically processes biological samples into CODIS profiles and then automatically destroys them.

The proposed sample-free database stands in sharp contrast to current CODIS practices. The typical argument for sample retention—that labs must retain samples to test for errors—is outweighed both by the ease of collecting new samples and also by the vast privacy risk posed by government access to full biological samples.

The genome is a wealth of deeply personal information, and DNA profiles are an acceptable use precisely because they do not include this information. Because access to DNA samples is virtually impossible to control, the best approach is to limit their use.

With luck, privacy legislation will soon finish what GINA started, and will close the door entirely on intrusive sifting of genetic information. Until that time, however, we are simply safer if sample retention is avoided.

Access to and use of a universal database is also a concern. From a law enforcement perspective, use of the database itself should be limited to investigating serious and specific crimes. This will prevent unwanted DNA dragnets, blind “fishing” searches, unwarranted intrusions such as tracking personal movements via DNA samples, probing parentage relationships, and the use of DNA evidence for trivial infractions such as littering.

Arguments in Support

The chief argument in support of the universal database is the same argument that currently undergirds the continued expansion of CODIS: namely, better database coverage translates to better success in solving crimes. The value of this should not be understated: if society benefits when criminals are caught, then a universal database should be a boon indeed.

This database would permit police to more quickly apprehend criminals who typically offend many times before first getting caught (burglars, for instance).

Universal coverage would also dispense with the questionable notion that when a crime is committed, the only meaningful suspects are prior offenders.

The universal database would be an effective deterrent against crime, in particular for first-time offenders. As others have noted, it would also help prevent wrongful convictions.

The second main argument in support of universal coverage is that it immediately would solve the issues of familial search and racial bias. If every individual is recorded in the database, there is no longer any need to conduct partial searches of their relatives. This also permits law enforcement to stick with powerful exact matches as opposed to error-prone partial matches when searching database records. Including everybody’s profile should make available an exact match for virtually every crime scene sample—regardless of race—as opposed to subjecting close family members of inmates of certain races to often-needless scrutiny.

Third, as this short paper has outlined, DNA profiles represent a non-invasive use of DNA: they distill complex sequence information down to just 26 numerical values with no known medical or biological meaning. In this, DNA profiles resemble a biologically derived Social Security Number: a unique numerical identifier that can be reconstructed from crime-scene biological samples. As such, otherwise valid concerns about genomic privacy ought not to apply to the particular case of CODIS-style DNA profiles.

Fourth, genomic DNA is routinely shed. It is difficult or impossible to control access to this information—cellular DNA can be isolated from discarded soda cans and a
plethora of similar sources—so it makes sense going forward to assume universal access to this information and control instead its use.

The universal database provides a structured means to do this.

Fifth, DNA collection need not amount to a Fourth Amendment search. Just as taking a DNA sample from an abandoned soda can or cigarette butt would not constitute such a search, the taking of DNA via cheek swab for the sole purpose of generating a non-invasive DNA profile similarly should not constitute a Fourth Amendment search. Even if such collection is deemed a Fourth Amendment search, however, courts could potentially still find these searches reasonable under existing suspicionless search or special needs jurisprudence.

Finally and most important, if implemented properly, a universal DNA database should pose no significant privacy risk. None of the sensitive biological or health information typical of the genome is preserved in DNA profiles.

Moreover, profiling every voting citizen will very quickly bring the issue of appropriate use of genetic material into the public spotlight—where it most certainly belongs. Under the current system, wherein only criminals are profiled, it is far easier to justify substantially more invasive use of DNA data.

The current trend of quiet expansion with little cause for public concern will in the long term pose a more substantial threat to privacy than creating a carefully considered and narrowly tailored universal DNA database today.