Mopping Up After

Coming Clean About "Junk DNA"

D.H. Kaye
Sandra Day O'Connor College of Law
School of Life Sciences
Center for the Study of Law, Science, and Technology
ASU

22 November 2007

In Coming Clean About "Junk DNA," Professor Simon Cole (2007b) puts a final spin on the exchange of views about the privacy implications of law enforcement DNA databases in Joh (2006), Kaye (2006), Cole (2007a), and Kaye (2007). He interprets such phrases as "contain predictive medical information" and "have predictive value" as applied to the CODIS STR types recorded in convicted-offender databases to mean "only a very small amount of predictive value, at least currently." He also criticizes my emphasis on the fact that there are strong reasons to question the prediction that the CODIS profiles ever will constitute valuable medical information as unduly "presentist." These remarks merit a brief surrebuttal. I shall quote (omitting footnotes and citations) portions of his essay that seem problematic, then respond.

I. What Can be Said Without Oversimplifying?

Statement:

[T]his exchange has not been a useless exercise. It began when I discovered Professors Joh and Kaye’s contributions during my effort to better understand—and, therefore, more clearly convey in my own writing—the state of scientific knowledge concerning the claim that the information held in law enforcement genetic databases is innocuous from a privacy standpoint. Professor Joh asserted that the claim of innocuousness was not true, and Professor Kaye countered that Professor Joh’s claim was flatly “false.” Under such circumstances, I was at a loss as to what to tell my own readers. Therefore, I traced back Professor Kaye’s key source, and offered my own contribution to the debate, suggesting that both authors had engaged in a certain degree of oversimplification.

Corrections:

Joh (2006) wrote that "some markers now thought to be meaningless may be (and have been) found to contain predictive medical information . . . ." In an attempt to clarify the state of current biological knowledge, I wrote that "any claim that the DNA profiles currently used for identification constitute ‘predictive medical information’ is false." Whether the broader claim that "the information held in law enforcement genetic databases is innocuous from a privacy standpoint" is another matter.

Regarding the single proposition that I designated as false, I asserted that "no forensic STR locus has been found to be predictive." (Kaye, 2006). Unlike such risk factors as blood pressure or cholesterol level, the forensic STR loci do not indicate that an individual is at a meaningfully greater risk for any disease. Cole (2007) denied this in
part because he misunderstood several sentences (or their implications) in what he thought was a "key source." This contribution to the debate necessitated a more detailed explanation of why the database profiles are not "predictive medical information" as those words ordinarily are used.

That the identification profiles have no value in medical practice does not resolve the question of whether the databases are completely "innocuous," and I did not write that all assertions of danger are "flatly false." Neither did I maintain that it is wrong to consider the possibility that future discoveries will change the fact that DNA database profiles do not constitute "predictive medical information." I merely disputed the suggestion that meaningful predictions can be made today, and I explained that "emerging theories do not imply that STRs 'contain predictive medical information.' " Cole could have told his readers that (1) claims that database profiles indicate whether an individual will succumb to a particular disease are false, and (2) who knows, maybe someday this situation will change.

II. What Does "Contain Predictive Medical Information" Mean?

Statement:

Professor Kaye’s most recent contribution to this exchange brings further clarity to the issue. As his meticulous exposition of the precise mechanisms behind contemporary genetic screening demonstrates, when he and other proponents of forensic DNA databases say that forensic DNA profiles have “no predictive value,” they actually mean that the profiles have predictive value, but that it is so small as to be practically useless. Likewise, when Professor Joh and other opponents of such databases say that forensic DNA profiles “contain predictive medical information,” they also mean that forensic STRs have only a very small amount of predictive value, at least currently.

Question:

"Predictive value" is both a technical term (as in "positive predictive value") and a lay term. No individual prediction of future disease status can be made from an STR profile in a law enforcement DNA database. As noted above, they are not known risk factors for any disease. How can "opponents of [DNA] databases" say that the STR profiles in databases "contain predictive medical information" if what they mean is that the database profiles do not provide information that can be used to predict any diseases, but they might (or might not) become informative at some future date?

Statement:

Professor Kaye’s response clarifies his declarations that law enforcement DNA profiles “have no meaning,” “reveal nothing about propensities for disease, behavioral traits, or the like,” “can tell nothing about a person,” and are “as meaningless as fingerprints,” and explains how his claims that “no forensic STR locus has been found to be predictive” and that “any claim that the DNA profiles currently used for identification constitute ‘predictive medical information’ is false,” over the course of his substantial
body of work on the subject were shorthand for the more complex explanation contained in his response. His response makes clear that forensic STRs contain predictive information, but that he cannot envision feasible exploitations of this information given the current state of genetic knowledge.

**Observation:**

Apparently, Cole concedes that in today’s world no medically useful predictions can be made from profiles in a database. I hope that my explanation of this point did not suggest, let alone "make clear" that the database records "contain predictive information . . . given the current state of genetic knowledge." My reading of the scientific literature and my exchanges with geneticists and forensic scientists indicate that given our current state of knowledge, the records can be used for individual identification — not for disease prediction.

### III. What Are Science Fiction, Extrapolation, and "Presentism"?

**Statement** (footnote 14):

I still find Professor Kaye’s account excessively presentist. I understand Professor Kaye’s aversion to what he calls the “speculation” about how it may be possible to exploit genetic databases as scientific knowledge progresses. However, it seems incorrect to label any extrapolation of future knowledge as “science fiction,” as he has done twice in this exchange.

**Comments and corrections:**

Plainly, I did not "label any extrapolation of future knowledge as 'science fiction' . . ." In fact, I never applied this term to any scientific hypothesis standing alone. Instead, I identified the following as science fiction:

*The year is 2025. The population is 325 million, and the FBI has the DNA profiles of all of them. Unlike fingerprints, these profiles reveal vital medical information. This universal database arrived surreptitiously. First, the Department of Defense’s repository of DNA samples from all military personnel, established to identify remains of soldiers missing in action, was given to the FBI. Then local police across the country shadowed individuals, collecting their shed DNA for the databank. On the way, thousands of innocent people were imprisoned because they had the misfortune to have race-based crime genes in their DNA samples. Sadly, it did not have to be this way. If only we had passed laws against collecting and using shed DNA . . . .*

I explicitly acknowledged that this science-fiction story — and I have to insist that it is fiction — was only "loosely inspired by an essay . . . by Professor Elizabeth Joh." (Kaye 2006). Dismissing this story of a covert, universal database as unduly speculative emphasizes — fairly, I think — the attenuated relationship between the problem of shed DNA and the establishment of a population-wide DNA database.
In my response to Cole, I suggested that "the scenarios [described in his essay] for the misuse by the government, insurers, or employers of the STR-identification profiles in NDIS and other law enforcement databases border on science fiction." (Kaye 2007). For instance, Cole's mention of using a repressive federal government subverting the DNA database of arrestees and convicted offenders to identify Japanese-Americans or other groups for internment would be good background for a science fiction story. My point, of course, is not that the extrapolation of existing trends into the future is impermissible. It is simply that we need to be clear about what is speculative and what is probable.

Statement (note 14, cont'd):

The term “science fiction” implies untruth, whereas in fact, given enough science fiction scenarios, one future scenario must turn out to be correct. In other words, while any particular prediction of the future state of genetic knowledge may be unlikely to be correct, we do know that genetic knowledge is likely to advance in some way as yet unforeseeable. Therefore, a minimal assumption that genetic knowledge will advance seems appropriate. Insisting, as Professor Kaye does, that any potential uses of genetic profiles must plausibly proceed from our current understanding of genetics knowledge (or effectively doing the same by refusing to “speculate” about such advances) is no less “science fiction” than assuming any particular scenario. Professor Kaye’s prediction that “the information coded in the databases is and will remain, with . . . limited exceptions . . . useful only for identification,” is itself only one of many possible extrapolations of the future, a science fiction scenario. Professor Kaye’s insistence on labeling all extrapolations of the future state of genetic knowledge that cannot be supported by reference to current theory as “science fiction” puts opponents of DNA databases in an unfair bind because it essentially demands solid evidence of the state of future knowledge, something no one can produce.

Presumably, Professor Kaye would respond that his extrapolation of the future is more defensible than others because it is “based on current knowledge and practice.” It may be more defensible, but that does not mean it is any more likely to be correct. Would the current capability of genetics have been predictable from the state of knowledge and practice in 1960? If not, there is no reason to assume that the capabilities of genetics in 2050—when the law enforcement DNA databases we are building today will likely still be in place and encompass a large portion of the population—must be wholly predictable from the current state of theory and knowledge.

Comments and Corrections:

Plainly, I did not assert that future developments "must be wholly predictable from the current state of theory and knowledge" to merit attention. Moreover, I am more than willing to entertain the "minimal assumption that genetic knowledge will advance." What I am unwilling to do is treat all conceivable "advances" as equally probable and imminent. Police rushing to a crime scene to analyze DNA on the spot with a handheld DNA analyzer is science fiction, but the probability that the police will have this ability is very high. Consequently, this possibility deserves serious attention in formulating policy.
Cole minimizes the fact that one prediction can be "more defensible" than others on the curious ground that this "does not mean it is any more likely to be correct." One would hope that, on average, more defensible theories are more likely to be true than less defensible ones. If not, how they can be defended?

If we are willing to try to rank risks and dangers, then we must ask how "defensible" is Cole’s theory of "correlations" with disease being used to stigmatize or disadvantage people whose STR profiles are in an offender database. The theory is not easily reconciled with facts of statistics (the nature of positive predictive value) and genetics (that the more common diseases are influenced by quantitative trait loci that involve many genes). These ideas have not changed much since 1960. They date back to Bayes and Galton. They are not likely to change by 2050. If Cole thinks that I have misapplied these ideas, then it behooves him to explain how.

Rather than examine the scientific basis for the various future worlds under discussion, however, Cole criticizes the effort as "extremely presentist." He cites the general principle that the future need not resemble the present. But arguing at this level of generality is vacuous. Would it be unfair to dismiss a story of a future with perpetual motion machines as science fiction rather than to regard it as sound extrapolation?

That said, the reminder that scientific dogma can be overturned is welcome. It is possible that the STR profiles in today’s databases eventually will turn out to "contain predictive medical information" and to threaten personal privacy for that reason. In devising public policy, however, we should try to distinguish, as best we can, between extrapolation and guesswork.

REFERENCES


