

PLEASE, LET'S BURY THE JUNK: THE CODIS LOCI AND THE REVELATION OF PRIVATE INFORMATION

*D.H. Kaye**

In a recent essay, Professor Simon Cole asks “Is the ‘Junk DNA’ designation bunk?”¹ He concludes that in one sense, it is not. There is no scientific evidence that the specific DNA variations used to identify the sources of crime-scene DNA perform any biological functions. Nonetheless, he contends that this fact, in and of itself, does not obviate the concern that the specific STR profiles stored in law enforcement databases of offenders (and sometimes arrestees) might be used to extract medically or socially sensitive information. I agree and have said as much in the past.²

Professor Cole also writes that “[t]he privacy threat posed by forensic STRs may not be great,”³ but he does not explain the basis for this view, and many of his remarks could be construed as being more consistent with the opposite conclusion—that the privacy threat may well be great. He criticizes the assurances of forensic scientists and human geneticists that, at present, “forensic DNA has no predictive value or medical significance”⁴ as “misleading” and “not fully informative.”⁵ He proposes that the records of the STR types of offenders contained in existing law enforcement databases “may, in fact, be precisely the kind of ‘predictive medical information’ that

* Freeman Foundation Visiting Professor of Law, Hopkins-Nanjing Center for Chinese-American Studies; Regents’ Professor, ASU Sandra Day O’Connor College of Law; Professor, ASU School of Life Sciences; Fellow, ASU Center for the Study of Law, Science, and Technology. I am indebted to Bruce Budowle, John Butler, James Crow, Elliot Goldstein, and the law review editors for noting errors, omissions, and ambiguities in a draft of this essay.

¹ Simon A. Cole, *Is the “Junk” DNA Designation Bunk?*, 102 NW. U. L. REV. COLLOQUY 54 (2007), <http://www.law.northwestern.edu/lawreview/colloquy/2007/23/> (link).

² See *infra* note 16 and text accompanying note 41.

³ Cole, *supra* note 1, at 63.

⁴ Originally, scientists provided such assurances with respect to the VNTRs used in forensic DNA typing from approximately 1985–1995. *E.g.*, Randall S. Murch & Bruce Budowle, *Are Developments in Forensic Applications of DNA Technology Consistent with Privacy Protections?*, in GENETIC SECRETS: PROTECTING PRIVACY AND CONFIDENTIALITY IN THE GENETIC ERA 212, 224–25 (Mark Rothstein ed., 1997); NATIONAL COMMISSION ON THE FUTURE OF DNA EVIDENCE, NATIONAL INSTITUTE OF JUSTICE, THE FUTURE OF FORENSIC DNA TESTING: PREDICTIONS OF THE RESEARCH AND DEVELOPMENT WORKING GROUP 37 (2000) [hereinafter NCFDNA], available at <http://www.ncjrs.gov/pdffiles1/nij/183697.pdf> (link).

⁵ Cole, *supra* note 1, at 59, 61.

concerns privacy advocates,”⁶ and he refers to STRs as potential “markers” having “predictive utility.”⁷ In particular, he asserts that “the forensic STRs . . . correlate with . . . disease-causing genes”⁸ and “phenotypically perceived race.”⁹ He concludes that “[i]f some forensic STRs are correlated with genes that cause physical traits, . . . the public can [and should] be informed of that fact”¹⁰ so that it “can decide for itself whether and to what extent the privacy risk offsets the benefits of genetic databases.”¹¹ The genetically influenced physical traits that he proposes are discernible from the DNA sequences used in criminal identification databases in the United States include diseases that would be of interest to insurance companies or employers and physical features associated with conventional racial categories.

These remarks require clarification. Just as the argument that nonfunctional DNA cannot be a threat to privacy is superficial, it would be incomplete and misleading simply to inform the public that an STR profile contains information that is correlated to physical traits such as disease and possibly behavioral predispositions and hence could be used to predict whether an individual will develop a disease. By innuendo, this formulation suggests that these nonfunctional loci, which are very weakly associated (if at all) with disease or behavior, are comparable to the loci used in much more powerful modern genetic testing for the DNA sequences of mutations that do cause disease.

This Colloquy Essay therefore analyzes in greater depth the medical and biological implications of the DNA records in the National DNA Index System (NDIS) and its local and state components. It explains why the STR profiles are useless as a “genetic test to screen for any particular disease.”¹² No one can say for certain what the future of genetics holds, but based on current knowledge and practice, the information coded in the databases is and will remain, with the limited exceptions noted below,¹³ useful only for identification.

To develop these points, Part I briefly describes the four possible ways in which genetic loci could possess predictive or diagnostic value with regard to diseases and explains why these mechanisms have not led, and probably cannot lead, to useful screening tests with the Convicted Offender DNA Index System (CODIS) profiles in national, state, and local databases.

⁶ *Id.* The response here focuses on the STR profiles in the context of offender databases. I discuss more fully the privacy implications of expanding these databases to other groups in D.H. KAYE, DNA IDENTIFICATION AND THE THREAT TO CIVIL LIBERTIES (forthcoming 2008).

⁷ *Id.* at 59.

⁸ *Id.* at 59, 62.

⁹ *Id.* at 62.

¹⁰ Cole, *supra* note 1, at 63.

¹¹ *Id.*

¹² *Id.* at 59. *But see infra* Part I.D.

¹³ *See infra* Parts I.D, II.

Part II considers the “physical traits” and familial relationships that the CODIS STRs can be used to identify. That the profiles carry limited information about an individual’s race and familial relationships has long been part of the public dialogue, and Part II places the resulting privacy issues in perspective. Part III comments on analogies between STR types and fingerprints, social-security numbers, and the like, employed when discussing these issues in the public forum.

I. THE CODIS STR PROFILES AS A SOURCE OF HEALTH INFORMATION

There are only four mechanisms through which a genetic locus could have predictive or diagnostic value: (1) mutations at the locus itself; (2) physical linkage of this locus to a locus at which a disease-causing mutation is present; (3) population structure; and (4) trisomies. In the first situation, certain DNA sequences at the locus cause a disease. In the other three situations, no alleles at the locus cause a disease, but they could be correlated with a disease. This Part explains why, for the CODIS STR loci, none of these mechanisms can be exploited to produce a valid and useful disease-screening test in the offender databases and why this situation is unlikely to change.

A. *Gene Mutations*

As the first mechanism, certain DNA sequences are transcribed into RNA that either regulates gene expression or is translated into proteins. Some of these DNA-sequence variations (“alleles” is the technical term) at various locations on the chromosomes could be harmful. For example, individuals afflicted with von Hippel-Lindau disease (VHL) develop tumors or cysts in the eyes, brain, spinal cord, kidneys, or a few other sites.¹⁴ The VHL gene normally is transcribed into RNA molecules that are translated into a tumor-suppressor protein. The protein stops cells from forming tumors. VHL disease occurs when a cell has two defective copies of the gene. A man born with a normal copy of the gene on one chromosome and a defective copy on the other produces the tumor suppressor because he has a functioning gene. No VHL tumors occur. If this gene mutates in just one cell out of the trillions in his body, however, the cell no longer produces the tumor-suppressing protein and can turn cancerous and proliferate. If the DNA sequence of the original, defective mutation is known, it may be possible to develop a genetic test that recognizes that specific sequence. Di-

¹⁴ The information on VHL presented here can be found in an educational module on genetics for high school students prepared at the National Institutes of Health, titled *VHL: A Genetic Disease* by Ruth Levy Guyer, and disseminated at <http://science-education.nih.gov/home2.nsf/Educational+ResourcesTopicsGenetics/4C2BAD0D0ED8F6C985256CCD00701E43> (link).

rectly testing for the mutation thus identifies asymptomatic individuals at risk to develop VHL.¹⁵

As Cole recognizes, there is currently no indication that the CODIS STRs are transcribed into RNA that would affect gene expression.¹⁶ Since the CODIS STR alleles are not disease-causing mutations, they cannot be the basis of a genetic test that directly detects such mutations. Furthermore, even if these STRs someday prove to be functional through an as-yet-unknown mechanism, this would not necessarily confer predictive medical value on them. For instance, hypothetically, STRs could be essential to embryonic development: an embryo with *no* STRs will not survive in the womb. In this case, the STRs would be functional, but this functionality would have no privacy implications for the offender databases. Variations in these biologically significant DNA sequences would convey no information about the health status or any other trait of any living human being.

Cole does not dispute the fact that the mutations that have given rise to the many alleles in CODIS STR loci do not affect phenotypes. He suggests, however, that a nontranscribed DNA sequence might be statistically associated with a disease-causing allele, and this correlation might be exploited to provide a useful genetic screening test.¹⁷ To assess this speculation, we need to examine the remaining three ways in which an association can arise.

B. *Physical Linkage*

The second possible mechanism for an association between an STR locus and disease status or propensity is physical linkage between a functional gene and an STR. For example, consider the same man who was born with a normal VHL allele (which we can designate as VHL⁻) and a mutant allele (VHL⁺). If some STR locus is very near the VHL locus then the two DNA sequences will tend to be inherited as a package.¹⁸ In particular, suppose that the man has an STR allele consisting of twelve repeats on the VHL⁺ chromosome and a different STR allele on the VHL⁻ chromosome. This

¹⁵ If other mutations in the population also inactivate the gene, then the test will not be very specific. It will miss those individuals who are at risk due to these other mutations.

¹⁶ See D.H. Kaye, *Science Fiction and Shed DNA*, 101 NW. U. L. REV. COLLOQUY 62 (2006), <http://www.law.northwestern.edu/lawreview/colloquy/2006/7/> (link). Other STRs have been implicated in certain inheritable diseases. See generally GENETIC INSTABILITIES AND HEREDITARY NEUROLOGICAL DISEASES (Robert D. Wells & Stephen T. Warren eds., 1997); Roger N. Rosenberg, *DNA-Triplet Repeats and Neurologic Disease*, 335 NEW ENG. J. MED. 1222 (1996). These STRs differ from the forensic STRs in that the core sequences are triplets (usually with both a G and a C in these repeated units), and the disease-related alleles have more repeats than the CODIS STRs.

¹⁷ This thought is not new. Cole accuses me of blurring or eliding “the distinction between causal and predictive significance,” but I address this distinction in David H. Kaye, *Two Fallacies About DNA Data Banks for Law Enforcement*, 67 BROOK. L. REV. 179 (2001) (link). It is not repeated in Kaye, *supra* note 15, because that essay is a short rebuttal of a claim regarding causation, not correlation, in Elizabeth Joh, *Reclaiming “Abandoned” DNA: The Fourth Amendment and Genetic Privacy*, 100 NW. U. L. REV. 857 (2006) (link).

¹⁸ ELAINE J. MANGE & ARTHUR P. MANGE, BASIC HUMAN GENETICS 194–98 (1st ed. 1994).

man will transmit the 12-allele along with VHL+ mutation to about half his children, while the other half will inherit the other STR allele along with the VHL- allele. Assuming that the mother does not have a 12-allele, their children with the 12-allele are much more likely to have inherited the father's VHL+ allele than the normal VHL-. These children are at high risk for tumors. This STR is not functional, but in this one family it is a marker for VHL. If the physical linkage between the STR and the VHL gene occurs throughout the population, and if the 12-allele is associated with the VHL mutation throughout the population, then the STR could be used to predict the occurrence of VHL tumors in the population.

Cole apparently believes that physical linkages with disease loci are known to exist and that they could be the basis for a useful screening test. He quotes two sentences from an article in the *Journal of Forensic Sciences*¹⁹ that supposedly "state[] that some forensic STRs are already *predictive*, though not *causal*, of disease."²⁰ However, neither this article nor other scientific literature asserts that "the forensic STRs . . . may be useful for tracking which individuals have the disease-causing genes."²¹ The two sentences allude only to the possibility that "many or possibly most STRs will eventually be shown to be useful in following a genetic disease or other genetic trait within a family" and that "a number of the core STR loci . . . have been reported to be useful in tracking various genetic diseases through loss of heterozygosity or allelic imbalance."²² These remarks sound ominous, but this sort of "following" and "tracking" is not a privacy problem for law enforcement databases because neither family studies nor loss-of-heterozygosity (LOH) studies would be useful in discerning the disease status of individuals in these databases via examination of their STR alleles. To see why, we need to understand how STRs are used in these two types of research studies.

1. *LOH studies.* An LOH study is useless for discerning disease status from the NDIS records. This type of study tracks the progression of cancer in a patient by gross changes in the DNA of a patient's cells. The technique requires tissue *from cancerous tumors*.²³ Needless to say, the law

¹⁹ John M. Butler, *Genetics and Genomics of Core Short Tandem Repeat Loci Used in Human Identity Testing*, 51 J. FORENSIC SCI. 253 (2006).

²⁰ Cole, *supra* note 1, at 59 (emphasis in original).

²¹ *Id.*

²² For the full quotation, see *id.*

²³ If a tumor-suppressor gene has been inactivated by a gross deletion of a portion of the chromosome containing the gene, it is likely that any nearby STRs also will be missing. If the patient's normal cells are heterozygous—the two chromosomes have two different alleles at the same STR marker locus—then there has been a loss of heterozygosity (LOH) in the diseased cells. Only the one remaining allele shows up when the tumor cells are typed. Thus, by comparing cells from a heterozygous patient's normal tissue to those from the same patient's tumors, clinicians may be able to learn something about the progression of the cancer. For an interesting example, see JAMES R. DAVIE ET AL., UNIV. OF PITTSBURG SCH. OF MED., MOLECULAR DIAGNOSTICS: CASE 287—MEDICOLEGAL CASE OF 56 YEAR

enforcement database records do not include the genetic profiles of an offender's normal tissues and of his tumors.

2. *Family studies.* Gene hunters use STRs in studies of specific families with rare, single-locus disease in order to zero in on the location of the disease-causing mutation within the family. If the gene is located, then the mutation can be sequenced, and a genetic test for the mutated sequence itself can be devised.²⁴ For this purpose, STRs have played an important role in biomedical research. But the STRs are not used in the genetic tests, and the fact that an STR locus was helpful in localizing a gene does not make that locus predictive of a disease in the general population. In the earlier example of the hypothetical STR marker for the VHL gene, the 12-allele is informative for the family studied. In the general population, however, the same correlation will not exist. This is because STRs have a high mutation rate, and each different STR allele is fairly common.²⁵ The number of people who have the 12-allele but do not have the VHL mutation will swamp the tiny number who have 12-allele along with the mutation. Learning that an unrelated individual happens to have the 12-allele does not appreciably increase the probability that this person has the VHL+ allele or bring it to a level that would have any screening utility. Some crude calculations suggest that a positive result on the 12-allele "test" for a VHL mutation in the NDIS database might be correct in less than one case in a thousand.²⁶ Parties with access to the database cannot use the recorded

OLD FEMALE WITH ONE LYMPH NODE METASTASIS AND TWO ORAL SQUAMOUS CELL TUMORS (2001), <http://path.upmc.edu/cases/case287.html> (link). LOH also can be used as a research tool to infer the location of previously unknown tumor-suppressor genes. P. Bennett, *Demystified . . . Microsatellites*, 53 MOLECULAR PATHOLOGY 177, 181–82 (2000).

²⁴ Even the direct test generally will not be useful for general population screening because each Mendelian disease is rare, diminishing the predictive value of a positive test result, and there may be many disease-causing mutations in the population, diminishing the predictive value of a negative result.

²⁵ See, e.g., JOHN M. BUTLER, FORENSIC DNA TYPING: BIOLOGY, TECHNOLOGY, AND GENETICS OF STR MARKERS (2d ed. 2005).

²⁶ Today, NDIS contains STR profiles of about 4.7 million convicted offenders. NDIS Statistics, <http://www.fbi.gov/hq/lab/codis/clickmap.htm> (link) (last visited Sept. 9, 2007). If the 12-allele frequency in the population were, say, 5%, we would expect to find slightly less than 10% of the people in this database with either one or two copies of this allele. (This relationship assumes that the population is in Hardy-Weinberg equilibrium. This is not correct for the mixture of the major population groups reflected in NDIS, but it is reasonable as a first approximation for the rough calculation here. See COMMITTEE ON DNA TECHNOLOGY IN FORENSIC SCIENCE: AN UPDATE, NATIONAL RESEARCH COUNCIL, THE EVALUATION OF FORENSIC DNA EVIDENCE 92 (1996). It also assumes that the general population and the convicted-offender population in the database have the same STR allele frequencies.) VHL is rare—fortunately, there are only about 7,000 people in the United States who are afflicted with the condition. For the sake of argument, let us assume that fully 10% of them are convicted offenders whose STR profiles are in the database. Finally, assume that notwithstanding the high mutation rates of STRs, every one of these 7,000 people are descendants of a common ancestor who had the 12-allele near the mutant VHL allele, so that they too have this association. If we were to use the 12-allele to predict the VHL allele for everyone in the database, we would identify some 10% of 4.5 million, or 450,000 people as positive for VHL. Of these 450,000 positive predictions, we would be correct in only 700

STR profiles to pick out people who carry the VHL allele or have developed the heritable disease. This exemplifies a well known statistical phenomenon: even a fairly accurate test for a disease with a small prevalence in the population has low predictive value.²⁷

One might think that there could be greater predictive value if an STR locus is physically linked to a gene involved in a *common* disease such as diabetes or coronary disease. But common diseases and health problems are affected by multiple genes and strongly influenced by environmental factors. The same is true of the behavioral traits that Cole “more speculatively” proposes some day might be the target of a screening test in NDIS.²⁸ Even if a CODIS STR turned out to be tightly linked to one gene in such a system—something that is unlikely given that there are only thirteen such STRs and 20–25,000 genes sprinkled across billions of base pairs—the correlation with the disease will be highly diluted. Again, a genetic test using such an STR will have very little predictive value.²⁹ As such, the medical community cannot simply use forensic STRs or some other genetic test to screen for a particular disease. Nature is not this generous.

C. Population Stratification

Third, a statistical association theoretically could arise without physical linkage but through population structure. The U.S. population is composed of many subpopulations that have somewhat different allele frequencies at various loci. These frequency differences disappear over time as people mate outside their ancestral groups, but there are some disease-related genetic differences across different subpopulations. As a group, African-Americans, for example, are at greater risk than Caucasians for sickle-cell

cases. For any particular positive prediction, the probability of the VHL mutation is only $700/450,000 = 0.001$. This probability is known as the “positive predictive value” (PPV) of the screening test. The real PPV will be even closer to zero because the correlation between the marker and the gene within the family will not be perfect. So despite the high correlation in the family study, the STR locus has essentially no predictive value in the general population.

²⁷ As a result, it is not easy for “the medical community [to] ultimately choose[] to use forensic STRs or some other genetic test to screen for any particular disease.” Cole, *supra* note 1, at 59. Geneticists cannot simply take a CODIS STR locus that may have shown some correlation with some disease in a family study and use it as a screening test that will reveal which convicted offenders have the disease. And, the idea that they can use “some other genetic test to screen” is quite impossible—the only genetic data in the database records are the alleles at the 13 STR loci. Of course, one could go back to the stored DNA *samples* and retype them with the “other genetic test.” That is why sample retention poses a scientifically real privacy risk, as Cole clearly recognizes.

²⁸ *Id.*

²⁹ The STRs are sufficiently common that many individuals in the population with a particular STR allele would not carry the gene mutation. In other words, the population correlation between the STR type and a disease-related gene is likely to be weak. Furthermore, even if the correlation were strong, the individual gene to which the STR is linked would contribute a small fraction to the occurrence of the polygenic, environmentally influenced disease.

disease; conversely, cystic fibrosis occurs more often in Caucasians.³⁰ Inasmuch as certain combinations of CODIS STRs occur more frequently in African-Americans than in Caucasian-Americans and vice versa,³¹ one can conceive of STRs being exploited to produce a probability that an individual is an African-American or a Caucasian.

It is not clear whether Cole believes that putting together these two sets of correlations will result in a valid screening test for disease, but he expresses concern about using STRs to infer race, and he writes that this possibility takes us “back to disease prediction.”³² If the argument is that even in the absence of physical linkage, STRs will be useful in predicting disease because of their correlation to self-reported or socially-perceived race, there is little cause for alarm. Not only is the correlation between the CODIS STRs and racial classifications weak, but even if it were perfect, no one can make useful predictions about an individual’s disease prospects merely on the basis of that person’s race. Knowing that a person in the database is white or black warrants no prediction or diagnosis of sickle-cell disease or cystic fibrosis. That an individual has an STR profile that is more often seen in one census category than another is even less useful as a test for disease.

D. Abnormal Chromosome Numbers

Finally, genetic tests with three of the CODIS STRs could indicate the presence of certain chromosome-number abnormalities.³³ Cole does not note this possibility, and it is not a significant threat to the privacy of individuals with profiles in the NDIS database. Two of the conditions (trisomy 18 and 13) are irrelevant to the offender database-privacy issue because they are so debilitating.³⁴ The third abnormality (trisomy 21, or Down syndrome) is more common and less incapacitating, but given the health problems and physical appearance of individuals with this syndrome, in virtually

³⁰ See, e.g., KEITH WAILOO & STEPHEN PEMBERTON, *THE TROUBLED DREAM OF GENETIC MEDICINE: ETHNICITY AND INNOVATION IN TAY-SACHS, CYSTIC FIBROSIS, AND SICKLE CELL DISEASE 1* (2006).

³¹ NCFDNA, *supra* note 4, at 60.

³² Cole, *supra* note 1, at 62.

³³ See, e.g., S.K. Dey & Sujoy Ghosh, *PCR-Based Detection of Parental Origin of Extra Chromosome 21 in Down Syndrome*, 5 INT’L J. HUM. GENETICS 183 (2005).

³⁴ “Trisomy” refers to the presence of an extra chromosome (or part of one) in an individual’s cells. Only five to ten percent of the babies with Trisomy 18 or 13 survive the first year of life. There are a few reports of survival into the teens, Lucile Packard Children’s Hospital, Medical Genetics: Trisomy 18 & 13, <http://www.lpch.org/DiseaseHealthInfo/HealthLibrary/genetics/trisomy.html> (link) (last visited Sept. 10, 2007), but these children are exceedingly unlikely to be swept into an offender database, and even if that were to occur, their physical condition would be obvious without genetic testing. See also NAT’L LIBRARY OF MED., MEDLINE PLUS, MEDICAL ENCYCLOPEDIA, <http://www.nlm.nih.gov/medlineplus/ency/article/001660.htm> (link) (discussing Trisomy 13); *id.*, <http://www.nlm.nih.gov/medlineplus/ency/article/001661.htm> (link) (discussing Trisomy 18).

all of the very rare cases among convicted offenders, the existence of the condition would already be known to the government and the public.³⁵

II. THE CODIS STR PROFILES AS A SOURCE OF INFORMATION ON OTHER PHENOTYPES AND GENETIC RELATIONSHIPS

As Professor Cole remarks, at issue in the lingering debate over the biology of STRs “is what is meant by the term ‘medical significance.’”³⁶ With regard to the privacy threat of law enforcement databases, it is essential to ask how the data—the STR profiles—sitting inside the FBI’s computers could be misused. Can the government, employers, or insurers employ the identification profiles to predict or infer something useful about an individual’s health status? As we have seen, the “medical significance” of STRs (and other classes of markers) in biomedical research does not make the STR profiles contained in the law enforcement databases medically significant in the sense of revealing health status or disease risks. At present, the CODIS STR profiles cannot be used in this way.

But what about other phenotypes? Cole implies that forensic STRs are “socially or medically significant [because] [t]hey . . . predict . . . phenotypically perceived race.”³⁷ We already have discussed the lack of medical significance of an STR profile as an indicator of race-related diseases. As for social significance, it would be very peculiar for the police to want to use STRs to draw any conclusion about the race of an offender who already has been convicted.³⁸ Having seen him, they already have perceived his race. They have his photograph and probably his own statement as to race or ethnicity. As one forensic biologist noted, a “photograph reveals a lot more about the person’s physical, social and maybe even mental state than the anonymous patterns in genetic fingerprints.”³⁹ Why use the STRs to make uncertain inferences about race when there are simpler methods?

Professor Cole gives an answer when he observes that the “memories of Japanese internment in the United States are not so old. A government agency ordered to round up individuals of a certain ethnic descent could, conceivably, perform ancestry testing on a genetic database to automate the

³⁵ See NAT’L LIBRARY OF MED, MEDLINE PLUS, MEDICAL ENCYCLOPEDIA, <http://www.nlm.nih.gov/medlineplus/ency/article/000997.htm> (link) (discussing Down syndrome).

³⁶ Cole, *supra* note 1, at 54.

³⁷ Cole, *supra* note 1, at 62 (internal quotations removed).

³⁸ STRs also are not likely to be used in inferring the probable racial or ethnic group of the source of a crime-scene sample because there are other loci that are much more informative of ancestry. See, e.g., Mark Shriver et al., Letter to the Editor, *Getting the Science and the Ethics Right in Forensic Genetics*, 37 NATURE GENETICS 449, 449 (2005). These “ancestry informative markers” have been used in rare cases in which police have turned to geneticists to evaluate the probable race of a serial rapist or other unidentified criminal.

³⁹ Mark Benecke, *Coding or Non-Coding, That Is the Question*, 3 EMBO REPORTS 498, 500 (2002).

process.”⁴⁰ Even if the prospect of a new internment program (of dubious constitutional provenance) were deemed realistic, it is doubtful that the 13 CODIS STRs could be used to pick out Japanese-Americans, Iranian-Americans, or some other conceivable xenophobic target group. Inasmuch as the CODIS STRs were chosen in part because most population groups contain the same alleles, it would be surprising if they were to prove useful for distinguishing between, say, Japanese-Americans and Chinese-Americans. A substantial fraction of people would be misclassified, and many of them easily could prove that they are from other groups. If so, the automated round-up would be both costly and inaccurate compared to looking at other records on file.

Another fact that could be exposed from STR database profiles is kinship. Although Professor Cole does not pursue this concern⁴¹ and it is not part of his theory of statistical correlations to genes that cause physical traits, STRs, like other inherited characteristics, could be used to try to ascertain whether particular individuals are related. For example, a population-wide database would make it possible to determine if a given individual is an illegitimate child. In a small, local database of convicted offenders it might be possible to identify some parent-child and sibling pairs. For the five million or so individuals represented in NDIS, however, a mere 13 loci will yield a very large list of people with an above-average chance of being related in these ways.⁴² Most of them will not be relatives after all, and the profiles of many of these nonrelatives will give a stronger appearance of relatedness than the profiles of the true relatives.⁴³ In assessing the privacy threat from relatedness testing, it also should be noted that most parent-child and sibling relationships are not private facts, but matters of public knowledge and official records. Still, in some instances unsuspected relationships could be revealed. To this extent, the STR profiles recorded in offender databases could be used to uncover and expose private information.

⁴⁰ Cole, *supra* note 1, at 55.

⁴¹ He merely notes that “[p]rivacy advocates contend that ‘DNA samples can provide insights into personal family relationships, disease predisposition, physical attributes, and ancestry.’” *Id.* at 55.

⁴² Virtually all close relatives have an unusually high number of matching alleles. In particular, ignoring mutations, a child is identical by descent with one parent at one allele at every locus, and siblings have a 25% chance of matching by descent at both alleles at each locus. But a nonrelative occasionally can be the source of an equally good or better partial match. This has a small probability in each case, but in a large database, there are a vast number of opportunities for coincidental partial matches to occur.

⁴³ In repeated searches for matches in a simulated database of 50,000 unrelated individuals and a profile of one child, researchers found that the profile of the true parent emerged as the most likely candidate only about half the time. Frederick R. Bieber et al., *Finding Criminals Through DNA of Their Relatives*, 312 *SCIENCE* 1315, 1315 (2006). The larger the database, the more opportunities there are for nonrelatives to emerge, by chance, as good partial matches.

III. INFORMING THE PUBLIC

The preceding discussion has probed the ways in which a law enforcement database of STR profiles might reveal private information—both today and in the future. As Professor Cole emphasizes, the mere fact that the CODIS STR loci are not implicated in gene expression does not logically establish that the STR profiles are useful only for identification. The full argument is complex, and the genetics and statistics of the situation are subtle. In a world of sound bites and editorials, this poses a problem for experts asked to opine on the dangers of the STR profiles. They can repeat the full explanation provided here—and even this exposition is oversimplified—but not many reporters or readers will listen to or read all of it, and even fewer will understand it.

To cut to the chase, it can be helpful to use certain metaphors that place the true privacy risk of DNA databases in perspective. Commentators have suggested that the 13 STR loci used in state and federal convicted-offender DNA databases are no more revealing of personally sensitive information than a fingerprint and that they are much more like a passport, social-security, or license-plate number than a medical record.⁴⁴ Professor Cole worries that these analogies can be misunderstood, and that is certainly possible. Nevertheless, such similes are roughly accurate in the context of disease screening. To scientists, the CODIS STRs are of negligible value for ascertaining information about an individual's health. The simplest way to say this is to refer to the loci as having “no predictive value,” for that is what it means to the scientist seeking to exploit such data. Technically, a genetic marker may have a non-zero correlation with some disease, in which case it is literally true that it is “predictive.” However, the same can be said of a fingerprint and a social security number. The former reflects both genetics and influences in utero. The latter is correlated with age, which is correlated with the incidence of many diseases. Predictive utility, however, depends on the magnitude of the association, and the predictive power of any plausible associations of the CODIS STRs with disease in the general population is trivial.

Outside the context of disease-screening, the metaphors are less apt. As I cautioned in an article contemplating a population-wide database, “the profile can be used in investigations of kinship, such as parentage determinations. This biological fact makes DNA profiles potentially more revealing than fingerprints or social security numbers”⁴⁵ Perhaps it would

⁴⁴ See Cole, *supra* note 1, at 60 (citing commentary).

⁴⁵ D.H. Kaye & Michael E. Smith, *DNA Identification Databases: Legality, Legitimacy, and the Case for Population-wide Coverage*, 2003 WIS. L. REV. 413, 432 n.59 (link). As previously indicated, this is much less of an issue in a convicted-offender database. To test for marital infidelity, for example, the investigators would need to have the profiles of the child and the putative mother and father. All three profiles are not likely to be in the database. In contrast, this privacy concern is more significant when DNA typing is done in mass disaster cases to identify the remains of missing people. In that situation, investigators compare suitable alleles in the remains to those of known family members.

better to draw an analogy to other biological variations that have been used in forensic science for decades. Arguably, scientists, politicians, lawyers, criminologists, and advocacy groups could communicate more accurately by saying that the information content in a person's STR-identification profile is no more threatening than that of a blood group or tissue type. Blood groups and tissue types are correlated to the incidence of certain diseases, they vary by ancestry, and they are useful in kinship testing. Nonetheless, like the CODIS STRs, they do not provide useful predictive or diagnostic tests for diseases, and this situation is not likely to change.

* * *

Toward the end of his essay, Professor Cole indicates that his fundamental concern is not so much whether the CODIS STRs really possess current or future predictive or diagnostic validity and utility, but rather whether scientists will be deluded into thinking that they have found predictive value even when it does not exist.⁴⁶ The sordid history of genetic determinism and genetic theories of racial inferiority should give us pause, as he advocates. Fortunately, genetics has progressed since the enactment of the eugenics laws of the 1920s. As I see it, the scenarios 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. And, as Cole notes, the debate about the information content in an STR *profile* is a distraction from the scientifically tenable claim that the DNA *molecules* in a sample are a threat to privacy. It is time to move on from the debate over "junk DNA" and to address realistically the true privacy problems posed by the growing repositories of DNA samples.

⁴⁶ Cole, *supra* note 1, at 62.