CHANCE: NEW DIRECTIONS FOR STATISTICS AND COMPUTING 3(3):26-32, 1990

A genetic pattern matching technique is being used to identify individuals. An assumption of statistical independence yields extremely small matching probabilities. Can this assumption be believed?

DNA Fingerprinting: What (Really) Are the Odds?

Joel E. Cohen

The Issue of Statistical Independence

Question: Why is "DNA fingerprinting" like the Declaration of Independence?

Answer: Both were promulgated on July 4: DNA fingerprinting in 1985 in the British scientific journal Nature, *the Declaration of Independence in 1776 in Philadelphia.*

Response: True, but that's not the answer I wanted. Try again.

Answer: Okay. For both DNA fingerprinting and the Declaration of Independence, it takes more than a simple declaration of independence to make independence a reality. That's true whether independence is statistical or political.

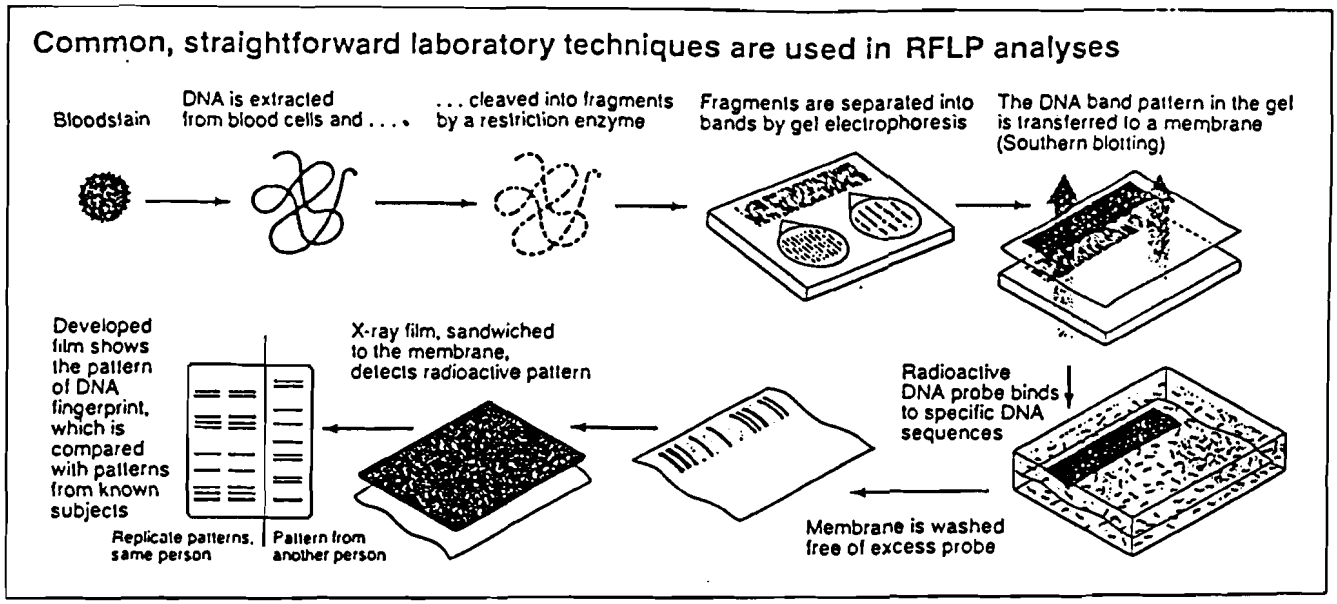
Response: Go to the head of the class; do not go to jail.

The serious point of this article is that some people may be going to jail because statistical independence has been declared in forensic applications of DNA fingerprinting without anyone ever collecting the data required to justify it.

DNA fingerprinting is a name given by Alec Jeffreys, a British geneticist, to a biochemical technique. The technique transforms DNA, the genetic material of most living cells, into a visible pattern something like the bar chart on grocery items. The details of the technique (see Figure 1) matter less than two properties of the pattern it produces (see Figure 2): stability and diversity. The pattern seems to be stable within an individual person, that is, the same whether derived from hair, blood, semen, skin or other tissues, and the same at different times. The pattern also seems to be highly variable from one unrelated person to another.

These two properties led Jeffreys and colleagues to propose (on the 4th of July, 1985) that the technique be used for identifying individuals. For example, if the pattern from blood or semen found at the scene of a crime matched that of a suspect, and if the probability of a match by chance alone were sufficiently low, then the involvement of the suspect would be proved with high probability.

Jeffreys called the biochemical technique a form of "fingerprinting" to associate it with the aura of specificity and infallibility that most people attribute to ordinary fingerprinting. In order not to prejudge the effectiveness of the tech-



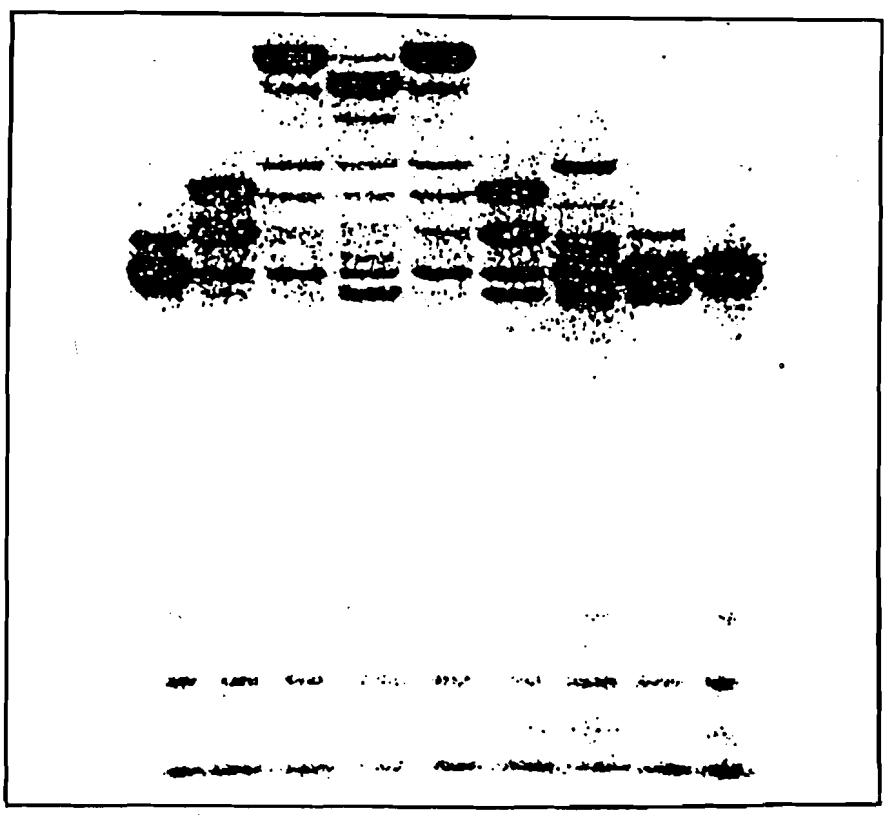


Figure 1. (above) How genetic pattern matching is done, using the technique based on restriction fragment length polymorphisms (RFLP).

Figure 2. (left) Bar-charts produced by the RFLP technique. Each vertical column represents the genetic pattern of one specimen or individual; each horizontal bar, or band, within a column represents DNA fragments of a particular size. When bands in different columns occur at (roughly) the same height, they are said to match. When all the bands in one column match all the bands in another, the patterns are said to match. Both illustrations reprinted with permission from Thornton, J.I., *Chem. Eng. News*, November 20, 1989, 67(47), pp 18–30. Copyright 1989 American Chemical Society.

nique as a means of identifying individuals in forensic applications, I prefer a more neutral term. For example, the term "genetic pattern matching" does not prejudge how reliably the matching identifies an individual. I will use the abbreviation GPM for the rapidly growing family of biochemical techniques that use genetic material for forensic identification.

Given a match between the pattern obtained from a specimen and the pattern of an individual, how could one calculate the probability that this match could have arisen at random? An obvious possibility is to record the complete genetic pattern of each individual in a population survey and count how many complete patterns match that of the specimen. If S individuals are surveyed, and one individual's pattern matches the specimen's pattern, this approach cannot yield a match prob-

ability lower than 1/*S*.

Jeffreys and his colleagues obtained the genetic patterns of 20 white volunteers at the University of Leicester. These individuals were not randomly sampled from any well-specified population. From an analysis of these 20 patterns, Jeffreys argued, using the summary data in Table 1, that the chance that any two random individuals would have matching genetic patterns is less than five in ten million million million (10^{19}). An extrapolation from 20 nonrandomly selected individuals to odds of matching that are less than one in a million million million has to be one of the most heroic leaps in science. Such a leap could be supported only by a powerful theory or by a powerful assumption.

Table 1. "Similarities of DNA fingerprints between random pairs of individuals" (from Jeffreys et al. 1985b, p. 76) (though the 20 individuals involved were volunteers, and were not sampled randomly in the statistical sense)

Probe	DNA fragment size (*kb*)	No. of fragments per individual ± s.d.	Probability *x* that fragment in *A* is present in *B*	Maximum mean allelic frequency/ homozygosity
33.6	10–20	2.8 ± 1.0	0.11	0.06
	6–10	5.1 ± 1.3	0.18	0.09
	4–6	5.9 ± 1.6	0.28	0.14
33.15	10–20	2.9 ± 1.0	0.08	0.04
	6–10	5.1 ± 1.1	0.20	0.10
	4–6	6.7 ± 1.2	0.27	0.14

In this case, the powerful assumption was statistical independence. The assumption of statistical independence gained its power by being applied, not to complete patterns like those in Figure 2, but to each separate horizontal stripe, or "band" within the pattern of a single individual or a single specimen. A band represents the presence of a DNA fragment of a particular size, and it is the pattern of fragment sizes that is supposed to characterize an individual.

Jeffreys and his colleagues, and apparently everyone who has used GPM in court, assumed that DNA fragments match up independently. They assumed independence between different radioactive probes (except in the rare case where different probes identified the same gene); between DNA fragments of different size classes; and between DNA fragments within a given size class.

When two events are statistically independent, the probability that both of them will occur is simply the product of the probability of each event separately. For example, if 0.11 is the probability of a match for a band made by DNA fragments between 10 and 20 kilobases long detected by using Jeffreys' probe 33.6 (as in the upper half of Table 1), then, assuming statistical independence, the probability that two bands will match in that size range is $0.11 \times 0.11 = 0.0121$. Given the assumption of statistical independence, if the probability of matching is less than one, then producing astronomically high odds against matching genetic patterns by chance alone requires only matching enough bands.

Empirical evidence to support the assumption of statistical independence between bands does not exist, to my knowledge. Such data could and should be gathered. If future data show that different bands are roughly statistically independent, I will be relieved that those people convicted on the basis of assumed statistical independence were not wrongly convicted (at least in that respect). But if, as some evidence suggests, it turns out that the assumption of statistical independence is not really justified, then some corrective action may be required.

Henceforth, let me use independence to mean statistical independence. The balance of this article is devoted to explaining why independence between bands in GPM should not be assumed until direct statistical evidence proves it to be true. In concentrating on the question of independence, I do not mean to suggest that the assumption of independence is the only problem (if indeed it is a problem) with GPM. On the contrary, there are many crude practical problems. Among them are the problems of collecting uncontaminated specimens at the scene of a crime, degradation of materials prior to analysis, use of internal controls and mixture experiments in electrophoretic gels, and the necessity for "blind" judgments and probabilistic assessment of a match between bands in different lanes of a gel or on different gels (see the accompanying article by Berry). There are problems of laboratory protocol such as the chain of custody of samples and quality assurance. There are thorny legal issues, such as what probability constitutes "beyond a reasonable doubt," which I leave aside altogether. There are even other statistical problems with current techniques of analysis; these will be briefly mentioned later.

Population Heterogeneity and Statistical Dependence of Alleles

To a geneticist or molecular biologist unaccustomed to the subtleties of populations, the assumption of independence between bands seems appealing. One of the basic laws of Mendelian genetics is that genes on different chromosomes assort, or are inherited, independently. This means that a given sperm (or egg) is supposed to get its genes on chromosome 1 independently of its genes on chromosome 2. So independence is built into the genetic mechanism, according to current theory. There is sometimes linkage (or statistical association) between two genes that are on the same chromosome, meaning that genes located close together on a chromosome are more often inherited together. But linked genes may approach linkage equilibrium, or statistical independence, in a population if, for example, mating is random (with respect to the genes under consideration) and there are no differences in survival between individuals (or gametes) in which certain alleles of the genes are positively coupled compared with individuals in which they are negatively coupled. Many geneticists consider these conditions plausible.

Unfortunately, life need not be so simple. The independence of different bands in GPM can be disturbed by a common phenomenon known as population heterogeneity in gene frequencies. Let me explain. Consider a large human population (e.g., Britain). Suppose a GPM technique is used that measures bands produced by just two genes, and suppose each gene has just two alleles. Call the alleles of the first gene A and a, and the alleles of the second gene B and b. No dominance between alleles is implied: A, a, B, and b each are assumed to correspond to well-defined distinct bands, so that the genotype of an individual (for these two genes) can be unequivocally determined from inspection of a gel (such as Figure 2). Suppose also that the two genes are located on different autosomes (chromosomes other than the sex chromosomes), or far enough apart on the same autosome so that the recombination fraction between genes is 1/2; in other words, assume that, within an *individual*, the two genes are inherited independently. Suppose also that inheritance at each gene is strictly Mendelian. So far this is a textbook model of two genes with two alleles.

Now suppose that Britain contains two subpopulations, such as individuals of European origin and individuals of non-European origin. Call the two subpopulations F and G. Suppose that within each subpopulation the two genes are in linkage equilibrium and there is random mating and no selection (i.e., no differences in reproduction or survival) with respect to both genes. Let $p(A, F)$ denote the gene frequency of allele A in subpopulation F. More generally, let $p(i, H)$ denote the gene frequency of allele i in subpopulation H, where $i = A, a, B, b$ and $H = F, G$. Thus $p(A, F) + p(a, F) = 1$, $p(B, F) + p(b, F) = 1$, and similarly with F replaced by G.

Under the preceding assumptions, within each subpopulation, each band (allele) is statistically independent: The genotype frequencies within a subpopulation are simply the products of the appropriate gene frequencies within that subpopulation. For example, if $f(AABb, F)$ denotes the relative frequency of the AA homozygote at the first gene and the Bb heterozygote at the second gene in subpopulation F, then $f(AABb, F) = 2p(A,F)^2p(B, F)p(b, F)$.

If the gene frequencies are different in the two subpopulations, it is not in general true that each band (allele) is statistically independent in the population as a whole. On the contrary, in the population, different bands may be positively or negatively associated, depending on the proportions of people in the different subpopulations and the differences of the allele frequencies at each gene.

A numerical example, using completely arbitrary figures, illustrates the phenomenon. Suppose the fractions of the whole population that belong to subpopulations F and G are given by $\pi(F) = 0.9$ and $\pi(G) = 0.1$. Suppose that $p(A, F) = 0.3$, $p(A, G) = 0.6$, $p(B, F) = 0.4$, and $p(B, G) = 0.8$. The overall relative frequency of the A allele in the population is $p(A) = p(A, F)\,\pi(F) + p(A, G)\pi(G) = 0.33$, whence $p(a) = 1 - p(A) = 0.67$ is the overall relative frequency of the a allele. Similarly, $p(B) = 0.44$ and $p(b) = 1 - p(B) = 0.56$. The actual relative frequency of the $AABB$ genotype in the population is

$$\begin{aligned} &f(AABB) \\ &= p(A, F)^2 p(B,F)^2 \pi(F) + \\ &\quad p(A, G)^2 p(B, G)^2 \pi(G) \\ &= 0.036 \end{aligned}$$

However, if the relative frequency of the $AABB$ genotype in the population is calculated assuming independence between alleles, the estimate is $f^*(AABB) = p(A)^2p(B)^2 = 0.021$.

To make the example slightly more realistic, assume that, *within each subpopulation*, the alleles $A_1, A_2, \ldots, A_{10}$ have the same allele frequency as A (so that $a_1, a_2, \ldots, a_{10}$ have the same allele frequency as a) and that the alleles $B_1, B_2, \ldots, B_{10}$ have the same allele frequency as B (so that $b_1, b_2, \ldots, b_{10}$ have the same allele frequency as b) and that the different alleles are statistically independent. The actual relative frequency of the homozygous genotype $A_1A_1A_2A_2 \ldots A_{10}A_{10}B_1B_1B_2B_2 \ldots B_{10}B_{10}$ in the

whole population is

$$[p(A, F)^2 p(B, F)^2]^{10} \pi(F) + [p(A, G)^2 p(B, G)^2]^{10} \pi(G) = 4.2 \times 10^{-8}$$

However, if the relative frequency of the genotype in the population is calculated assuming independence between alleles, the estimated relative frequency of the homozygous genotype $A_1A_1A_2A_2 \ldots A_{10}A_{10}B_1B_1B_2B_2 \ldots B_{10}B_{10}$ is $[p(A)^2 p(B)^2]^{10} = 1.7 \times 10^{-17}$. Now suppose a forensic specimen is determined to have the homozygous genotype $A_1A_1A_2A_2 \ldots A_{10}A_{10}B_1B_1B_2B_2 \ldots B_{10}B_{10}$ by GPM, and a suspected individual is identified whose genotype exactly matches that of the specimen. In this case, the estimated probability of a match, assuming independence of alleles, is lower than the true probability of a match, allowing for the heterogeneity of subpopulations, by a factor of more than 10^9. The estimated probability, being lower than the true probability, exaggerates the significance of a match and unnecessarily incriminates the suspect. The numerical values in this example were chosen in advance for simplicity, rather than to illustrate a worst case.

The point is that attributes (alleles or bands, in this case) may be positively or negatively associated in a population as a result of pooling the frequency of the attributes in subpopulations in which the attributes are independent. An association that arises in this way is called spurious correlation. Spurious correlation has been familiar to most statisticians at least since Yule pointed it out in 1903. Between 1972 and 1975, no fewer than four groups of population geneticists (Sinnock and Singh; Prout; Nei and Li; and Feldman and Christiansen) demonstrated using simple mathematical models that linkage disequilibrium (spurious correlation between pairs of loci) could arise from the mixing of subpopulations, each of which is in linkage equilibrium. Such linkage disequilibrium would not vanish in one generation even under random mating, and could be sustained by continued migration. The example given above contains no new genetics or statistics, but the issues it raises appear to have been overlooked in forensic applications of GPM.

This hypothetical example resembles reality in that there is likely to be significant genetic heterogeneity in real populations. The allele frequencies of genes of medical interest differ from one human subpopulation to another. Other genes that serve as markers of disease-related genes, including many of those used in GPM, are likely to share that heterogeneity. The DNA probes used for forensic identification detect alleles that have heterogeneous allele frequencies: Lander, using Wahlund's formula, found excess homozygosity (relative to Hardy-Weinberg equilibrium) at genes identified by DNA probes in the Hispanic population used as the reference population in a murder trial, demonstrating "the presence of genetically distinct subgroups within the Hispanic sample."

The hypothetical example given above differs from reality in that neither the assumption of just two subpopulations nor the particular allele frequencies and subpopulation frequencies assumed are likely to be realistic. The actual effect of subpopulation heterogeneity could be larger or smaller than that in this example.

Implications of the Example

The example demonstrates three points. First, the population used to obtain estimates of allele frequencies is crucial for subsequent applications of match probabilities to individual cases. Future studies should carefully define a reference population (what statisticians call a sampling universe) and should make explicit the procedure (such as systematic sampling or random sampling) that is used to sample from this population. Procedures for proper sampling are well known to statisticians.

There are good practical and scientific reasons for giving serious attention to sampling. A practical reason is that a study of matching probabilities for GPM that is based on an ill-specified sample can be challenged in court, on the grounds that the study sample is not the sample most appropriate to the case. A scientific reason for paying serious attention to sampling is that GPM provides a means of assessing the genetic heterogeneity of populations. Studies of well-defined samples offer an opportunity to compare the genetic variability of different populations and could be of potential interest to students of human evolution.

A second conclusion from the above example is that alleles (bands) may be significantly statistically associated in a population if there is heterogeneity between subpopulations in the allele frequencies, even though the genes involved may be strictly Mendelian, unlinked, and at linkage equilibrium within each subpopulation. Wherever subpopulations are heterogeneous, true random samples of populations and direct tests of association between alleles or bands are required to measure directly whether bands really are independent.

How can association between alleles be tested? First, the presence or absence of bands in specified narrow bins or intervals of molecular weight could be determined for each individual in a large sample. (Specific molecular weights may be determined by markers of known molecular weight.) Then, in principle, log-

linear models for multidimensional contingency tables could be used to determine appropriate models for the possible independence or dependence of DNA fragments (see Cohen 1990 for details). In practice, it may be necessary to resort to alternative approaches, such as methods for the analysis of large sparse contingency tables or contingency tables with incompletely classified data, or to tests for pairwise independence of bands. Whatever the particulars of the statistical technique, it is clear that without appropriate data, no amount of statistical theory can say whether different fragments are statistically independent or are statistically associated in a population.

Third, the statistical association of alleles, though undetectable in terms of chromosomal mechanisms or within homogeneous subpopulations, may induce significant errors in estimates of match probabilities if the estimates ignore the statistical association.

The gratuitous assumption of independence is well known among statisticians as a source of superficially persuasive arguments for the existence of miracles, which correspond in the present situation to extravagantly small alleged probabilities of obtaining a match at random.

Other Statistical Problems

In addition to the assumption of independence and the lack of a well-defined reference population, some previous statistical analyses of GPM data have several other problems. (For details see Cohen 1990.) Some analyses assumed, for example, within a size class of DNA fragments identified by a given probe, that the probability of a match is constant for all fragments in the size class, and that there is no variability in the number of fragments per specimen or per person in the size class. It is obvious, on the contrary, that there is variability in the number of fragments per person in each size class because the standard deviations in Table 1 are positive. It can be shown that the assumption that the match probabilities are constant within a size class is not consistent with the data in Table 1 under the assumption of independence between fragments. Some analyses mistakenly used the geometric mean rather than the arithmetic mean to estimate matching probabilities. Even taking the assumption of independence as valid for the sake of argument, examples show that, with data like those in Table 1, the use of the geometric mean could underestimate the true probability of matching, and therefore exaggerate the effectiveness of GPM in identifying individuals, by a factor of ten thousand.

Conclusion

Scientific data and statistical analyses play increasing roles in the courtroom. It is the responsibility of scientists and statisticians to provide measurements, analyses, and conclusions that justify lawyers' faith in scientific techniques. When such faith is not justified, it is scientists' responsibility to provide clear warning labels to the contrary. Since human lives and liberty are at stake in uses of GPM for forensic identification, there should be little room for doubt about the assumptions underlying the analysis and interpretation of GPM data, including their statistical analysis and statistical interpretation.

The difficulty in establishing the statistical basis of GPM for forensic identification lies in assuring that the assumptions implicit in the calculations are justified by evidence or theory and that any simplifying approximations made give conservative estimates (that is, overstatements) of match probabilities. Apparently, no one yet has collected and analyzed data to support the common assumption that the bands produced by GPM techniques occur statistically independently. Just declaring independence does not make it so; think of 1776!

Genes that are statistically independent in subpopulations may be statistically associated in the population as a whole if there is heterogeneity in gene frequencies between subpopulations. In the populations where GPM is used for forensic applications, the assumption that DNA fragments occur statistically independently for different probes, different genes or different fragment size classes lacks supporting data so far; there is some contrary evidence. Statistical association of alleles may cause estimates based on the assumption of statistical independence to understate the true matching probabilities by many orders of magnitude. The conclusion is that some astronomically small probabilities of matching by chance, which have been claimed in forensic applications of GPM, presently lack substantial empirical and theoretical support.

Many of the same problems arise in paternity testing through GPM. A child's pattern is supposed to contain bands drawn from the patterns of either its mother or its father. Because of the possible relatedness of individuals in paternity testing, the genetic formulas are more complicated than the formulas used in identifying unrelated individuals. However, the underlying hypothesis of statistical independence is usually invoked, and most of the same caveats apply.

Future experiments and analyses could provide a firmer foundation for GPM by careful attention to sampling and possible statistical dependence among fragments. GPM can be the basis of a useful method of identifying individuals, provided that claims for it are not exaggerated.

Additional Reading

Christiansen, F.B. (1987), The deviation from linkage equilibrium with multiple loci varying in a stepping-stone cline, *J. Genet.* (India) 66, 45–67. (Genetic models for linkage disequilibrium, or spurious correlation, between loci in populations with heterogeneous subpopulations.)

Cohen, J.E. (1990), DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability, *Am. J. Hum. Genet.* 46, 358–368. (Apparently the first detailed critique of statistical assumptions made in forensic applications of GPM; portions are reproduced here.)

Ford, S., Thompson, W.C. (1990), A question of identity: some reasonable doubts about DNA "fingerprinting," *The Sciences* (NY Acad Sci) 30(1), 37–43. (A skeptical account of the pitfalls of the biochemical procedures in GPM.)

Jeffreys, A.J., Wilson, V., Thein, S.L. (1985), Individual-specific "fingerprints" of human DNA, *Nature* 316, 76–79. (Started a wave of forensic applications of GPM.)

Lander, E.S. (1989), DNA fingerprinting on trial, *Nature* 339, 501–505. (The limitations of GPM in practice.)

Neufeld, P.J., Colman, N. (1990), When science takes the witness stand, *Sci. Amer.* 262(5), 46–53. (Practical, legal and institutional problems in using DNA technology in court.)

Thornton, J.I. (1989), DNA profiling: New tool links evidence to suspects with high certainty, *Chem. and Engrg. News*, Nov. 20, 18–30. (An exposition of the chemical procedures, by a true believer in their effectiveness.)

Joel E. Cohen has been professor of populations and head of the Laboratory of Populations at Rockefeller University since 1975. He earned a Ph.D. in applied mathematics in 1970 and a Dr.P.H. in population sciences and tropical public health in 1973, both from Harvard University. He is a former Fellow of King's College Cambridge, the Center for Advanced Study in the Behavioral Sciences, the John Simon Guggenheim Foundation, and the John D. and Catherine T. MacArthur Foundation, and a Fellow of ASA, the American Association for the Advancement of Science, and the American Academy of Arts and Sciences. He is a trustee or director of the Russell Sage Foundation, the Societal Institute of the Mathematical Sciences, and the Black Rock Forest Preserve.

This article was supported in part by National Science Foundation Grant BSR–87–05047.