THE DNA TYPING CONTROVERSY AND NRC II

LAURENCE D. MUELLER*

The application of modern molecular biological techniques of DNA (deoxyribonucleic acid) typing to forensic science has been a major advance. These techniques aid in the identification of people who may be the sources of biological tissue samples left at the scene of crimes. The application of these techniques in forensic science has not been without controversy. One consequence of this controversy has been the creation of two different committees by the NRC (National Research Council), a branch of the National Academy of Sciences, to review these problems. The first committee issued their report in April of 1992 (NRC, 1992) and suggested some major changes in the methods used to report statistics by forensic DNA typing laboratories. Four years later another NRC committee issued a second report (NRC, 1996) which suggested that many of the recommendations of the first NRC committee were unnecessary. In this paper I review some of the population genetic and statistical issues which have been at the heart of the debate in forensic DNA typing. I also analyze, for the first time, forensic PCR (polymerase chain reaction) databases from five different laboratories. Finally, I discuss some of the specific recommendations made by the second NRC committee (also called NRC II).

1. Why there is a controversy with forensic DNA typing. There are two scientific disciplines whose techniques and theories are used in the development of DNA typing technology today: molecular genetics and population genetics. For restriction fragment length polymorphism (RFLP) based techniques the theories and principles of molecular genetics are used for developing the techniques to isolate DNA from evidence samples, to break up the DNA into small fragments and then to finally visualize these fragments on x-ray films or autorads. Once these patterns are visible and it has been determined that DNA from an evidence sample and DNA from a suspect "match" the theories and principles of population genetics are used to estimate how rare people with matching profiles might be in the potential suspect population. This statistic is often called the match probability (or genotype frequency) and is most often computed by a method known as the product rule.

With DNA typing techniques one can easily get predicted frequencies of one in millions to one in trillions as estimates of match probabilities. If these procedures yielded match probabilities of 1 in 100, say, then a moderate number of people with matching profiles would be expected in a database of several thousand people. An excess or deficiency of people matching the profile would serve as empirical refutation of the product rule.

*Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525. Email: ldmuelle@uci.edu.

This type of independent check is not possible if most genetic profiles are not present in the existing data bases. Consequently, there is a far greater need to verify that the techniques which generate these very rare predicted frequencies are accurate. Certainly, the estimated frequencies should not exaggerate the rarity of the genetic pattern.

2. Investigating the assumptions and reliability of the product rule. A DNA profile is composed of genetic information from several different sites on the hereditary material, called DNA. These locations are referred to as genes, loci or (incorrectly) probes. There are two copies of each gene, a maternally derived copy and a paternally derived copy. When there exists, within a *population*, different forms of a gene the different forms are called alleles. As an example for a particular gene, say A, if there exist ten different forms, they can be represented by the symbols $A_1, A_2, ..., A_{10}$. An individual may be either a heterozygote at the A locus, meaning the individual has two different copies of the gene (e.g. A_2A_6) or the individual may be a homozygote, meaning the individual has two copies of the same allele (e.g. A_7A_7). A profile usually consists of several loci, say A, B, C and D, thus a single individuals profile may be represented as, $A_iA_jB_iB_jC_iC_jD_iD_j$.

The first step in the product rule is to estimate the frequency, within an appropriate reference population, of the combination of alleles at a single locus (e.g. the frequency of people having the C_iC_j combination of alleles). This is accomplished by a relationship known in population genetic theory as the Hardy-Weinberg law. This law states that for heterozygotes the frequency of people with that particular combination of alleles is twice the product of the constituent allele frequencies (thus if the C_i allele occurs with frequency 0.1 in the reference population and the C_j allele with frequency 0.2 then people with the C_iC_j combination should occur with frequency $2 \times 0.1 \times 0.2 = 0.04$). If the individual is a homozygote the frequency of this pattern is given by the square of the individuals allele frequency (thus if the C_i allele occurs with frequency 0.1 in the reference population then people with the C_iC_i combination should occur with frequency $0.1 \times 0.1 = 0.01$).

The next step in the product rule is the multiplication of each of the Hardy-Weinberg results from each locus by each other. This multiplication is only permissible if the population of interest obeys the condition of linkage equilibrium (thus if the Hardy-Weinberg frequencies at the A, B, C and D loci were $0.1, 0.15, 0.2$ and 0.25 respectively, the final profile frequency would be, $0.1 \times 0.15 \times 0.2 \times 0.25 = 0.00075$).

Neither Hardy-Weinberg nor linkage equilibrium are guaranteed to hold from first principles. On the contrary many biological phenomenon can invalidate these assumptions (Hartl and Clark, 1989, pgs. 31–32; Lewontin and Hartl, 1991). The biological phenomenon that poses the most serious threat to the use of the product rule is the possibility that the reference populations (usually African Americans, Caucasians and Hispanics) are composed of population subgroups that are genetically different from each other (e.g. if Hispanic databases were samples from Cuban Hispanics and Mexican Hispanics and these were genetically different from each other). Population subgroups typically appear because mating between members of these groups has been limited for some period in the past. Consequently, there have been several attempts to investigate the accuracy of the product rule by testing its key assumptions of Hardy-Weinberg and linkage equilibrium.

2.1. The hypothesis testing paradigm. The appropriateness of assuming Hardy-Weinberg and linkage equilibrium may be investigated by statistical hypothesis tests. Some tests of the Hardy-Weinberg equilibrium simply compared expected and obeserved frequencies of total homozygotes and heterozygotes (Odelberg et al., 1989; Budowle et al., 1991; Weir, 1992a). When used with RFLP databases these tests generally suggested an excess of homozygotes in the samples. Devlin et al. (1990) developed a goodness of fit test of Hardy-Weinberg which was designed to take into account one artifact of the RFLP technique which could generate the observed excess of homozygotes. Geisser and Johnson (1993) use a method based on quantiles and have found evidence of departures from Hardy-Weinberg equilibrium in the FBI databases. Weir (1992a) developed likelihood ratio tests for independence within and between loci. Weir (1992a,b) typically found departures from independence when the entire database was analyzed but few departures when the analysis was restricted to heterozygotes. This latter technique was justified because some heterozygotes may appear as homozygotes due to technical limitations of the RFLP process. Many of the tests described above can not be easily extended to multiple loci because of the limited size of the existing databases and the large number of multilocus genotypes that are possible with these genes.

However, there are some tests which permit the detection of multilocus dependence even with small databases. Detecting independence at three or more loci is important since it is theoretically possible that all pairs of loci may be in linkage equilibrium but the combination of three or more may not be (Feldman *et al.*, 1974). The number of heterozygous loci for each individual forms the basic observation for the test developed by Brown and Feldman (1980). This test, while easy to apply, is not very powerful especially since there are many configurations of linkage disequilibrium that will not be detected by this technique. More recently, there has been work on applying Fisher's exact tests to genetic systems with many alleles and many loci (Guo and Thompson, 1991; Maiste and Weir, 1995; Zaykin, Zhivotovsky and Weir, 1995). These tests work by computing the probability of the observed sample, under the hypothesis of independence, to randomly shuffled databases. The frequency of shuffled databases that are less likely to occur than the observed database constitute the p-value of the test.

2.1.1. Limitations of the hypothesis testing paradigm. The value of the hypothesis testing results are limited for several different reasons. If the product rule is used with four loci, say, then one must show that there is simultaneous linkage equilibrium for all four loci. Presently only the exact tests and the Brown and Feldman tests can be used with more than two loci for most existing databases. At the time the second NRC report was issued there had been no application of the exact tests to two or more loci and only a few instances of the less powerful Brown and Feldman test (Budowle et al., 1995). Interestingly, in the Budowle et al. (1995) paper the exact test is used for within locus tests of independence but not for tests across multiple loci. The only test used by Budowle et al. for more than two loci was the Brown and Feldman test. For some databases (especially very small ones) and for some statistical tests the product rule could be false but the test would fail to detect this due to low power. This issue will be discussed in more detail in the next section. With RFLP testing techniques it is possible for certain heterozygotes to appear as homozygotes (Devlin et al., 1990; Mueller, 1991). This makes testing for independence more difficult (Zaykin et al., 1995).

Another issue which needs to be considered is the way to handle datasets which have been subjected to a large number of multiple tests of independence. Several studies have sought to protect the level of type I errors, even with as few as nine different tests, with simultaneous tests statistics such as the Bonferroni inequality (Devlin et al., 1990; Budowle et al., 1995). There has been little discussion in these publications about the power of the resulting tests (see next section), although it is clear that one ultimately sacrifices statistical power by utilizing simultaneous test statistics (Miller, 1966).

2.1.2. Applications of hypothesis tests to forensic RFLP databases. There have been many studies now attempting to test Hardy-Weinberg, linkage-equilibrium or both in a variety of forensic and non-forensic databases (Odelberg *et al.*, 1989; Devlin *et al.*, 1990; Budowle *et al.*, 1991; Chakraborty and Daiger, 1991; Devlin and Risch, 1992; Risch and Devlin, 1992; Edwards *et al.*, 1992; Weir, 1992a, 1992b; Geisser and Johnson, 1993, 1995; Slimowitz and Cohen, 1993; Maiste and Weir, 1995). The results of these studies have been mixed. Some studies have yielded results inconsistent with either Hardy-Weinberg or linkage equilibrium (Odelberg et al., 1989; Budowle et al., 1991; Edwards et al., 1992; Weir, 1992a; Geisser and Johnson, 1993, 1995; Slimowitz and Cohen, 1993; Maiste and Weir, 1995), while others have generally supported these assumptions (Devlin *et al.*, 1990; Chakraborty and Daiger, 1991; Devlin and Risch, 1992; Risch and Devlin, 1992; Weir, 1992a, 1992b).

Some of these tests suffer from a clear lack of power. For instance Devlin *et al.* (1990) were the first to develop tests specifically for these RFLP data sets and claim their results appeared to conform to the Hardy-

Weinberg expectations. Devlin et al. examined three databases (Lifecodes) and three probes for a total of nine different tests of the Hardy-Weinberg expectations. These tests produced a number, z, which measured the departure from Hardy-Weinberg. If this number was greater than -2.8 and less than 2.8 the results were judged consistent with Hardy-Weinberg. Devlin et al.'s test statistic incorporated the simultaneous test protection of the Bonferroni inequality. The problem with using the Bonferroni inequality is the potential to reduce the statistical power of the tests.

This problem was raised by several commentators on Devlin *et al.'s.* work (Cohen, *et al.*, 1991; Green and Lander, 1991). In response to these criticisms Devlin *et al.* (1991) conducted numerical studies which seemed to suggest that the original test was quite powerful. However the numerical studies by Devlin *et al*'s. (1991) used a different statistical test than in their original paper. The new test both removed the Bonferroni protection and changed the original two-tailed test to a one-tailed test. With this altered test z was now required to be greater than -1.65 and less than 1.65 to be considered consistent with Hardy-Weinberg. If this version of their test is applied to their original results, the Lifecodes databases show two significant departures from the Hardy-Weinberg expectations.

More recently Zaykin et al. (1995) have analyzed the power of the exact test to detect departures from independence caused by population substructure. The power of these tests was reasonable when the coancestry coefficient, θ, (which measures the extent of population substructure and which the second NRC committee suggested would typically be between 0.01 and 0.03 in most U. S. populations) was in the range of 0.05 to 0.1. In fact the power gets better with increasing numbers of loci. However, when the exact test is applied to the FBI's VNTR database there are numerous examples of departures from Hardy-Weinberg and linkage equilibrium (two-loci, Maiste and Weir, 1995). Maiste and Weir attribute these departures to the existence of small missing bands which cause some heterozygotes to be misclassified as homozygotes. They thus reanalyzed these data, examining only the heterozygotes and omiting the homozygotes. Their conclusion was that this type of analysis showed that, ".. there is overall evidence for independence." However, this conclusion stands in marked contrast to the conclusions Zaykin *et al.* (1995). In that paper the power of the exact test with homozygotes excluded was barely at the nominal level of 5% for tests with two-loci and θ at 0.05 (see table 4, Zaykin *et al.*, 1995). The power of these tests actually gets worse with increasing number of loci, eventually having power less the nominal level of 5%.

2.1.3. Application of hypothesis tests to forensic DQ-A and polymarker databases. In addition to the RFLP DNA tests are a variety of DNA tests which utilize the polymerase chain reaction (PCR). This technique permits the amplification of very small amounts of DNA in an evidence sample until there is enough to be tested by a variety of techniques.

The net result is that samples too small to provide results by RFLP techniques may be analyzed by PCR techniques. The DQ-A gene and five genes known as the polymarker genes are amplified by PCR techniques and used in forensic DNA typing.

Budowle et al. (1995) analyzed the FBI's African American, Caucasian and Hispanic databases for independence and concluded there was general agreement with the assumptions of the product rule. I present results below which question this conclusion. My own analysis of just the FBI database has recently turned up one error in the results published by the FBI (Mueller, 1998). It appears that application of the exact test to the FBI's Caucasian database and the HBGG locus yields a highly significant departure from equilibrium ($p = 0.008$) which was reported by Budowle et al. as non-significant ($p = 0.887$). Budowle et al.'s result can only be obtained if one changes the genotype of a *CC* homozygote (the only individual with a *C*-allele in the FBI database) to an *AA* homozygote.

A major limitation of the FBI database for assessing independence is its relatively small size, with samples sizes between 94 and 148 people. This shortcoming can be remedied by pooling data from several other forensic laboratories which have analyzed the same loci and presumably the same populations. I have done this for databases obtained from the FBI, Cellmark, the Minnesota Bureau of Criminal Apprehension, Perkin-Elmer and the Virginia State Crime Laboratory. All five databases include samples from Caucasians and African American's. Only three laboratories had Hispanic databases, the FBI, Perkin-Elmer and the Virginia State Crime Lab. For this analysis the Florida and Texas FBI Hispanic databases were combined. All test results for independence within loci are shown in Tables 1-3 along with the significant departures from multilocus equilibrium.

I have not set as a pre-condition for these tests a comparison of the allele frequencies in the pooled databases. Thus, its possible that genetic differences that exist among the sampled populations contribute to the lack of independence. However, one must remember that the pooling of these samples is in fact no different than the sampling process that originally gave rise to these databases. Since there has been no attempt to insure these sample are random it is important to also determine if ad hoc samples pooled together obey the independence laws. If they do not, then all the current databases would need to be recreated with closer attention paid to the sampling process.

There are only two significant departures from Hardy-Weinberg equilibrium, but they are both in the Caucasian population (HBGG and GC, Table 2). There are a total of 57 different multilocus tests that were performed on each database. Since there were two different tests done (Fisher's and the chi-square exact, see Zaykin et al., 1995) there are actually a total of 114 tests per population. Assuming all tests are independent (which of course they are not) we might expect about 5 or 6 significant results. For the African American population (Table 1) there were 15 significant departures (including the combined polymarker and DQ-A loci) while for the Caucasian population (Table 2) there were 11. However, for the Hispanic population (Table 3) there were only two significant results. It would appear that both the Caucasian and African American population are yielding evidence of significant departures from independence while the Hispanic population is not. At this point it is also worth noting that the Hispanic database is about half the size of the other two databases and hence we expect the tests of independence in the Hispanic populations to be less powerful.

TABLE 1

Summary of results from exact tests of independence for the combined African America dataset. All single locus results are shown. Only those multilocus tests with at least one significant ($p < 0.050$) departure from equilibrium are shown. There were a total of 57 different multilocus tests performed. Estimated p-values are based on 3200 permutations of the database, thus an approximate confidence interval on a p-value of 0.05 is about ±0.0076. For the tests within loci the permutations broke up allele combinations at a single locus. For the tests between loci allele combinations at a single locus were preserved and genotypes between loci were randomly shuffled. This latter test procedure appears to the most powerful for detecting departures from linkage equilibrium and will not be influenced by departures from Hardy-Weinberg at the constituent loci (Zaykin et al., 1995).

Loci	N	Fisher's Test, p	Chi-Square Test, p
DQ-α	638	0.23	0.23
LDLR	638	0.31	0.27
GYPA	638	0.88	0.93
HBGG	638	0.78	0.81
D7S8	638	0.31	0.28
GC	638	0.28	0.28
Loci Combination			
DQ-α/HBGG/GC		0.0041	0.013
LDLR/HBGG/GC		0.0088	0.053
DQ-α/LDLR/HBGG/GC		0.033	0.0053
DQ-α/GYPA/HBGG/GC		0.089	0.014
DQ-α/HBGG/D7S8/GC		0.063	0.043
LDLR/GYPA/HBGG/GC		0.0052	0.011
LDLR/HBGG/D7S8/GC		0.028	0.056
DQ-α/LDLR/GYPA/HBGG/GC		0.30	0.0053
DQ-α/LDLR/HBGG/D7S8/GC		0.079	0.0094
DQ-α/GYPA/HBGG/D7S8/GC		0.22	0.048
LDLR/GYPA/HBGG/D7S8/GC		0.18	0.011
DQ-α/LDLR/GYPA/HBGG/D7S8/GC		0.24	0.011

As mentioned previously it has been suggested that techniques, like the Bonferroni inequality, be used in the situations of multiple testing to insure a type-I error of no greater than 5%. My own feeling is that this

TABLE 2

Summary of results from exact tests of independence for the combined Caucasian dataset. All single locus results are shown. Only those multilocus tests with at least one significant ($p < 0.050$) departure from equilibrium are shown. There were a total of 57 different multilocus tests performed. The test procedure follows the methods in table 1.

Loci	N	Fisher's Test, p	Chi-Square Test, p
DQ-α	624	0.93	0.93
LDLR	624	0.28	0.30
GYPA	624	0.82	0.88
HBGG	624	0.0028	0.0013
D7S8	624	0.73	0.68
GC	624	0.010	0.010
Loci Combination			
LDLR/HBGG		0.049	0.51
HBGG/GC		0.20	0.025
DQ-α/LDLR/HBGG		0.0091	0.61
DQ-α/HBGG/D7S8		0.021	0.53
LDLR/D7S8/GC		0.45	0.044
GYPA/HBGG/GC		0.056	0.013
HBGG/D7S8/GC		0.39	0.034
DQ-α/LDLR/HBGG/D7S8		0.029	0.56
GYPA/HBGG/D7S8/GC		0.18	0.025
LDLR/GYPA/HBGG/GC		0.088	0.035
LDLR/GYPA/HBGG/D7S8/GC		0.14	0.049

TABLE 3

Summary of results from exact tests of independence for the combined Hispanic dataset. All single locus results are shown. Only those multilocus tests with at least one significant ($p < 0.050$) departure from equilibrium are shown. There were a total of 57 different multilocus tests performed. The test procedure follows the methods in table 1.

Loci	N	Fisher's Test, p	Chi-Square Test, p
DQ-α	381	0.090	0.12
LDLR	381	0.92	0.92
GYPA	381	0.13	0.14
HBGG	381	0.71	0.76
D7S8	381	0.74	0.74
GC	381	0.90	0.87
Loci Combination			
LDLR/GYPA		0.023	0.033

sacrifice in power can't be justified. For instance if the Bonferroni inequality were used on the 57 multilocus tests in Tables 1-3 the p-value needed for significance would be 0.0009. To assess the implications of such a procedure I have performed Fisher's exact test on three biallelic loci from the FBI's Caucasian database, LDLR, D7S8 and GYPA. However, in these tests I actually saved the lowest 5% of the shuffled database based on their probability of being observed and I saved the lowest 0.09% (out of 10,000 permutations). These extreme databases all showed a deficiency of heterozygotes. The magnitude of this deficiency can be quantified with the parameter θ from the following relationship,

$$[\text{extreme heterozygote frequency}] = (1 - \theta)[\text{Hardy-Weinberg frequency}]$$

The results (Table 4) show that in general adding the Bonferroni protection requires that the magnitude of the departures from Hardy-Weinberg be almost twice as large as tests without the Bonferroni protection.

TABLE 4

The magnitude of population substructure, as measured by θ that is necessary to cause rejection of the Hardy-Weinberg null hypothesis via Fisher's exact test. The databases used are from the FBI's Caucasian polymarker database.

	θ value necessary for	
Locus	Rejection at the 5% level	Rejection at the 0.09% level
D7S8	0.17	0.26
LDLR	0.15	0.26
GYPA	0.15	0.26

Another class of genetic markers which can be amplified by PCR techniques are called short tandem repeats (STR's). Several forensic laboratories have started to use these genetic markers in case work. In Tables 5-6, I show tests of independence for the Cellmark African American (Table 5) and Caucasian (Table 6) databases. These tests include DQ-α, polymarker and the three STR loci HUMCSF1PO, HUMTPOX, and HUMTHO1 for nine loci altogether. Each of these databases is rather small (100-103 individuals) and there were a total of 502 multilocus tests done per population. The African American database yielded 21 significant departures which is in the ball park of the nominal number expected if all tests are considered independent. The Caucasian database yieided more than twice as many significant results (49), including the combinations of $HUMCSF1PO$ and $HUMTPO1$ and the non-STR loci. Taken altogether, the results in Tables 1-6 suggest that for many PCR based systems there appears to be evidence of departures from statistical independence.

2.2. Investigation of population subgroups. Another methodology to test the suitability of the product rule is to investigate the underlying assumptions of Hardy-Weinberg and linkage equilibrium. An important underlying assumption is that the populations utilized are homogeneous, randomly mating populations. They should not be composed of genetically differentiated subgroups. In an attempt to study this problem the FBI has collected data from laboratories throughout the United States and from many other countries (Budowle *et al.*, 1994a, b). Many of these databases

TABLE 5

Exact test of independence for the African American Cellmark database (100 people). These tests were evaluated by using the chi-square statistic (Zaykin, Zhivotovsky and Weir, 1995). The test procedures were in Table 1. The loci used are numerically identified as follows. **1** *- LDLR,* **2**- *GYPA,* **3** *- HBGG,* **4**- *D7S8,* **5** *- GC,* **6** *- DQ-a,***7** *- HUMCSF1P0,* **8** *- HUMTPOX,* **9** *- HUMTHO1.*

Probability (p-value)	Loci (locus)
0.013	1
0.045	1/5/7
0.036	1/6/7
0.021	2/3/8
0.013	2/3/9
0.048	3/6/9
0.0084	5/6/7
0.048	1/2/5/7
0.039	1/2/6/7
0.0041	1/5/6/7
0.035	2/3/4/9
0.046	2/3/6/9
0.0066	2/5/6/7
0.015	3/5/6/7
0.043	4/5/6/7
0.0034	1/2/5/6/7
0.013	1/3/5/6/7
0.025	1/4/5/6/7
0.020	2/3/5/6/7
0.014	1/2/3/5/6/7
0.027	1/2/4/5/6/7

are not samples of population subgroups, rather they are simply samples from heterogeneous groups (like Caucasians) from different geographic regions of the United States.

TABLE 6

Exact tests of independence for the Caucasian Cellmark database (103 people). These tests were evaluated by using the chi-square statistic (Zaykin, Zhivotovsky and Weir, 1995). The loci used are numerically identified as follows. **1**-*LDLR,* **2** *- GYPA,* **3** *- HBGG,* **4** *- D7S8,* **5** *- GC,* **6** *- DQ - a,* **7** *- HUMCSF1PO,* **8** *- HUMTPOX,* **9** *- HUMTHO1.*

Probability (p-value)	Loci (locus)
0.029	5
0.041	9
0.0044	1/3
0.033	2/6
0.038	1/3/8
0.036	1/5/7
0.033	2/3/8
0.046	2/5/7

0.049	3/4/8
0.033	3/5/8
0.0094	3/6/8
0.041	1/2/3/8
0.013	1/2/5/7
0.035	1/3/5/8
0.0097	1/3/6/8
0.036	2/3/5/8
0.0084	2/3/6/8
0.044	3/4/5/8
0.0097	3/4/6/8
0.0088	3/5/6/8
0.038	3/6/8/9
0.040	1/2/3/5/8
0.011	1/2/3/6/8
0.041	1/2/4/5/7
0.043	1/3/4/5/8
0.012	1/3/4/6/8
0.0084	1/3/5/6/8
0.048	1/3/6/7/8
0.042	1/3/6/8/9
0.044	2/3/4/5/8
0.011	2/3/4/6/8
0.012	2/3/5/6/8
0.050	2/3/6/7/8
0.040	2/3/6/8/9
0.014	3/4/5/6/8
0.045	3/5/6/7/8
0.046	3/5/6/8/9
0.046	1/2/3/4/5/8
0.013	1/2/3/4/6/8
0.012	1/2/3/5/6/8
0.047	1/2/3/6/7/8
0.039	1/2/3/6/8/9
0.012	1/3/4/5/6/8
0.048	1/3/5/6/7/8
0.015	2/3/4/5/6/8
0.041	2/3/5/6/8/9
0.015	1/2/3/4/5/6/8
0.043	1/2/3/4/6/8/9
0.045	1/2/3/5/6/7/8
0.041	1/2/3/5/6/8/9
0.050	1/2/3/4/5/6/8/9

However, some of the samples do come from different population subgroups, e.g. Swiss Caucasians, and East Indians and can be used to directly assess the assumption of no genetic differentiation. This is done most directly by simply comparing the frequency of the different forms of these genes in two populations. Figure 1 shows one such example from the FBI's world wide study.

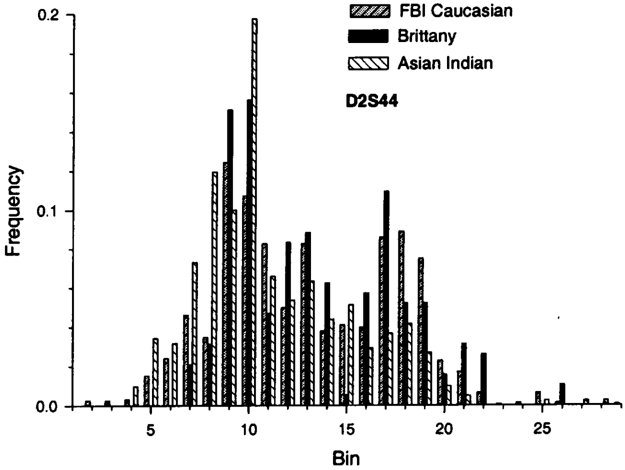

FIG. 1. *The frequency of the FBI fixed bins in three different populations. The differences that exist in these three populations is statistically significant meaning it is too great to be due to chance events associated with the sampling process itself.*

This example is only one of many. Recently, Sawyer *et al.* (1996) reported that about 70% of all such comparisons in the Caucasian populations in the world wide study show significant differences. Data from other studies (Balzas *et al.*, 1992; Deka *et al.*, 1991; Krane *et al.*, 1992; Sanjantila *et al.*, 1992) have also confirmed the general conclusion that many population subgroups are significantly differentiated for these VNTR genes.

I have recently completed a similar analysis as Krane *et al.*, with the Hinfl databases and the U.S. Caucasian database of Cellmark. Many populations subgroups are not represented in this study. For instance there were no subgroups of Hispanics studied e.g. Cubans, Puerto Ricans, Haitians, Mexican Americans, etc. Likewise the study of Asian Americans was very superficial consisting of two Chinese samples neither with more than 20 people, a number wholly inadequate for any reasonable inferences. However, the world wide study does contain some reasonable data on Caucasians which I have analyzed.

There are several different ways one might analyze the significance of population subgroups. One could, in principle compare the frequencies of a number of five locus profiles to see if they differ from one population to the next. However, this is practically impossible since the Cellmark databases contain very few people which have been typed at all five loci used by Cellmark. For instance in Cellmark's African American population there are only two people with complete five locus profiles. The FBI has studied subgroups by estimating five locus profile frequencies by using the product rule. This is not a reliable procedure since there is no way to independently verify the accuracy of product rule frequencies.

The most direct and reliable way to assess whether there are genetic differences between population subgroups is to compare the observed frequency of DNA fragments of different sizes. I have used this procedure with the Cellmark Caucasian database and a variety of other population groups (Table 7). The FBI fixed bins were used as a standard so that different populations could be compared to each other. There is no easy way to do this comparison using the floating bin procedure. For each pair of populations a likelihood ratio test was used to assess whether the fixed bin frequencies were significantly different. If the difference is very unlikely to be due to chance alone it is then reasonable to conclude that the genetic differences between the two populations are real.

In Table 7, 32 of 49 comparisons show significant differences or 65% of all contrasts. Thus, these data support the existence of widespread population substructure. Consequently, one needs to be concerned about the lack of random sampling on the part of Cellmark since these data suggest that the frequencies of genetic variants detected by the RFLP techniques of Cellmark do vary significantly from one population to the next. In some cases these differences are large. In figure 2, I graphically show some of these differences.

While these studies confirm the widespread existence of allele frequency differences among population subgroups the more difficult question is whether there are differences among the multilocus profiles. While it seems improbable that there can be allele frequency differences and not multilocus profile differences determining the magnitude of this effect is difficult.

One approach taken by the FBI is to use the product rule to compute the frequency of a single profile in multiple databases. Observing that this procedure typically gave different, but rather rare frequencies the FBI has concluded (Budowle et al., 1994 a,b) that population subgroups lead to no "forensically" significant differences among the profile frequencies. While the phrase "forensically significant" sounds like "statistically significant" the basis of the concepts is quite different. Statistical significance has an objective meaning which can be easily understood while forensic significance might mean something different to every person. In principle two frequency estimates would be considered forensically significant if a juror would render a different verdict depending on which number they heard. Of course determining when differences are forensically significant can only

TABLE 7

Results from a likelihood ratio test comparing the bin frequencies in the Cellmark Caucasian data base and one of the data bases below. The first number in each cell is the G statistic, the subscript is the degrees of freedom and the probability of observing this statistic is given in parentheses. The differences between the two populations are judged significant when the probability of observing the test results is 5% or less. The significant results are listed in boldface type.

Population	D1S7	D2S44	D7S21	D7S22	D12S11
Alsatian		29_{17} **(0.031)**			24_{13} **(0.033)**
Danish	42_{24} **(0.011)**	31_{18} **(0.028)**	21_{14} (0.11)		31_{13} **(0.004)**
English	52_{25} **(0.001)**	32_{18} **(0.022)**	24_{18} (0.17)	131_{25} **(<0.001)**	17_{14} (0.26)
Finnish		29_{14} **(0.01)**			
German	60_{27} **(<0.001)**	50_{17} **(<0.001)**	19_{15} (0.22)	34_{24} (0.09)	51_{14} **(<0.001)**
Italian		33_{18} **(0.018)**			42_{14} **(<0.001)**
New Zealander		22_{15} (0.12)			17_{12} (0.14)
Norwegian		24_{17} (0.11)	19_{12} (0.092)	73_{24} **(0.028)**	37_{12} **(<0.001)**
South Europe	36_{22} **(0.031)**		20_{13} (0.087)		36_{13} **(0.001)**
Spanish		25_{18} (0.12)	18_{13} (0.15)		46_{13} **(<0.001)**
Swiss		46_{18} **(<0.001)**	19_{16} (0.26)	46_{24} **(0.004)**	76_{15} **(<0.001)**
Swedish	37_{25} (0.055)	24_{18} (0.15)	17_{12} (0.14)	31_{24} (0.16)	52_{12} **(<0.001)**
Asian Indian	67_{27} **(<0.001)**	47_{16} **(<0.001)**	36_{19} **(0.01)**	151_{26} **(<0.001)**	132_{14} **(<0.001)**
Maori		102_{14} **(<0.001)**			207_{12} **(<0.001)**
Pacific Islander		184_{15} **(<0.001)**			372_{13} **(<0.001)**

be done if the scientist can determine and quantify how "jurors" evaluate and weigh evidence. There is no agreed upon process for making these conclusions (Koehler *et al.*, 1995).

As discussed earlier the preferred method for evaluating the accuracy of the product rule would be to compare the predicted frequencies to the observed. In some cases this can be done. The profiles in table 8 are for two individuals, one from the Maya population in Mexico and the second from the Surui population in Brazil (Kidd *et al.*, 1991).

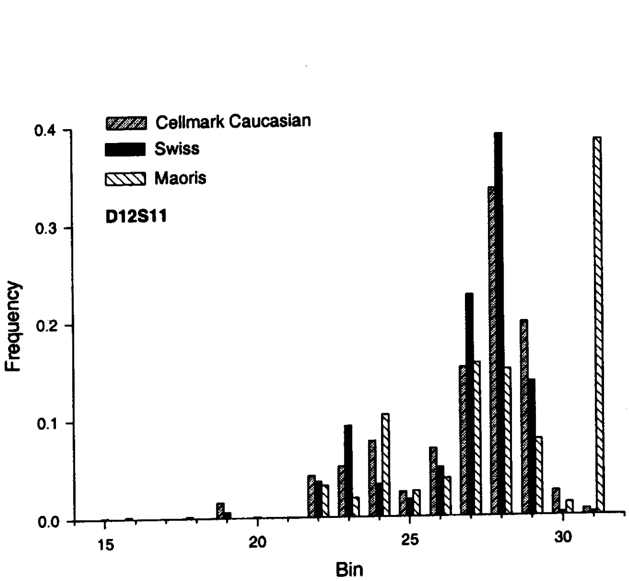

FIG. 2. *The frequency of fixed bins in three populations for the D12S11 locus. The frequencies of these fixed bins in the Maoris and Swiss populations are significantly different than the Cellmark Caucasian database.*

TABLE 8

The size (in base pairs) of each allele at six loci for two unrelated individuals from the Surui and Maya populations.

Individual	D2S44	D17S79	D14S1	D14S13	D18S27	LILA5
Surui	10,600	3,570	4,570	12,790	4,750	7,330
	10,290	3,320	3,790	5,210	4,520	7,330
Maya	10,820	3,560	4,550	12,780	4,750	7,320
	10,270	3,290	3,790	5,180	4,550	7,320

The product rule predicts that the profile the two individuals in table 8 share in common should have a frequency of 1 in 96 million, although the observed frequency is 1 in 37. This large discrepancy stands in contrast to the FBI and NRC II's assertions that subgroups typically alter profile frequencies by only a factor of 10-100.

2.4. How rare are matches between unrelated people? Another research protocol is to determine how common matching profiles are in forensic databases (Risch and Devlin, 1992). This is done by comparing the profile of each person to every other person in the database using the laboratories quantitative match criteria. For example, if there are 1000 people in the database, person number 1 is compared to person number 2 and then to person number 3 etc., then person number 2 is compared to person number 3 etc. In this way the total number of pairwise comparisons can be much greater than the total number of people in the database. In the previous example of a database with 1000 people there are a total of 499,500 pairwise comparisons.

The results of pairwise comparisons will only be useful if the databases examined are random samples with respect to the frequency of multilocus matches. These protocols also depend critically on using a correct match criteria for RFLP based techniques which will declare matches between samples in the database with the same frequency as if they had been compared to each other in a forensic test. This means that if a forensic laboratory uses a quantitative match criteria which will declare matches between samples from the same source 99% of the time (no realistic criteria will catch 100% of all matches) then the criteria used to find matching profiles in the database should also quantitatively match exactly 99% of all matches, no more or no less.

Both of these criteria have been violated in the current studies. For instance the Risch and Devlin study utilized FBI databases. However, before these databases were supplied to Risch and Devlin the FBI had already surveyed them for matching profiles. The FBI search had in fact found numerous matches. In some cases (data from A. Eisenberg, at the Texas College of Osteopathic Medicine) examination of the original records was able to verify that duplicates of the same persons blood had been sent to the FBI. The FBI appropriately deleted these matching profiles. In other cases there was no way to definitively determine if the matches were duplicates or not. Nevertheless, the FBI deleted these matches, since they were working under the assumption that such matches should not occur. A database from a black population in South Carolina contained 60 matching profiles that could not be accounted for (memorandum of 8/31/90 from J.J. Kearney to Mr. Hicks). The entire South Carolina black database was deleted and replaced with data from different populations. This problem is not limited to the FBI but has also happened in other forensic laboratories. The Metropolitan police Department Laboratory in St. Louis found four seven probe matches and one six probe match in their database and deleted these (memo of 6 May 1996, by Donna Becherer).

The upshot of this is these databases are biased samples. In many instances cited above matching samples have been deleted simply because of a belief that such matches shouldn't occur. **These databases can not then be used as research tools to investigate how rare multilocus matches are.**

A second more difficult issue concerns which match criteria should be used when searching databases for matching profiles. For instance in Risch and Devlin's study of the FBI's database they used a match criteria which was only half the size of the FBI's match criteria. To illustrate this problem consider the FBI's match criteria. When evidence and suspect DNA are compared on the same gel and autorad, the two samples are judged to match when the difference in their measured size is less than or equal to 5% of their average size. This match criteria should include greater than 99% of all matching profiles. Database samples are included on many different autorads and collected over a very long period of time. Consequently, two samples from the same source will tend show a greater separation in the database than they would had they been tested at the same time. This variability has been quantified (Geisser, 1996) and the database samples show about 1.8 times as much variation as samples on the same autorads. That is the standard deviation for replicated measurements made on different gels and autorads is 1.8 times larger than the standard deviation for measurements made on the same gel and autorad. This means that to find matches in a database with the same frequency as they would be found in casework requires using a match window which is 1.8 times larger than the standard match criteria. Using this criteria Geisser (unpublished) has found six four locus matches in the FBI's Caucasian database and one six locus match even after the filtering process mentioned earlier.

3. Why the random match probability isn't a sufficient statistic for evaluating the weight of DNA evidence. In the typical forensic case involving DNA evidence, DNA from an evidence sample is compared to DNA from a suspect. When these two samples match the inference that is usually made is that the DNA in the evidence came from the suspect. There are two different phenomena which can make this inference invalid. One phenomenon that can invalidate this inference is if the DNA came from a person other than the suspect who happens to have the same genetic profile as the suspect. Secondly, the DNA profiles in the evidence and the suspect may be different but due to a laboratory error a match has been declared.

Both of these phenomenon must be evaluated and quantified. The overall probability of an erroneous match would then equal the sum of the probability of a coincidental match between unrelated people and the probability of a laboratory error. If one of these probabilities is very rare and one very common the overall probability of an erroneous match will simply be equal to the more common probability. For example if the random match probability is 1 in 1 million but the laboratory error is 1 in 500 the chance of an erroneous match is about 1 in 500. In fact the only situation under which it would be appropriate to use 1 in 1 million to weight a DNA match is in a laboratory in which the rate of error was demonstrably less than 1 in 1 million. Not only have false matches been made in forensic laboratories but all the available evidence suggests that rates of false matches are in the range of 1 in hundreds to 1 in thousands (Mueller, 1993; Koehler, 1993; Koehler *et al.*, 1995). To only provide a jury with a very rare random match probability vastly overstates the weight which should be accorded a single DNA match (Hagerman, 1990; Lempert, 1991; Thompson, 1995).

Recent empirical research (Koehler *et al.*, 1995) has shown that these ideas are not well understood. Lay people who participated in mock trials tended to base their evaluations on the rarer of two statistics not the more common of the two as is appropriate. Thus, people who heard a very rare match probability and a very common laboratory error rate tended to convict at the same rate as people who only heard only the rare match probability. However, those who heard only the common laboratory error rate were much less likely to convict. Consequently, if the laboratory error rate and the match probability are presented separately the connection between the two numbers needs to be explained in some detail.

Some laboratories have suggested that their error rates are zero since they have made no errors in all the proficiency tests they have taken. The problem with this assertion is that in the context of match probabilities, usually, of the order 1 in millions or 1 in billions, a zero error rate means lab errors must be much less than the match probability. For a laboratory that has only performed 50 proficiency tests it is impossible to claim that the error rate is less than one in billions. In fact using widely accepted statistical methods the best that can be done for the laboratory performing all 50 tests correctly is to place an upper bound on how common errors might be (in this case that upper bound is 1 in 17 with 95% certainty). The only way to lower this bound is for the laboratory to do additional proficiency tests.

A logical question is, what if the laboratories error rate is really 1 in millions? If this were true then it would take an enormous effort for any laboratory to demonstrate their proficiency. At this point we can step back and take a larger view of all laboratories which do DNA testing. From this vantage point it is clear that false matches do not occur at the exceeding rare rate of 1 in millions or billions but are much more likely to be on the order of 1 in thousands. While it is certainly the case that some laboratories may be better than another there are no reason to believe that some labs may have error rates of 1 in 100's while others have error rates of 1 in billions for basically the same testing procedure. Consequently, laboratories should use the upper 95% confidence interval for their own individual estimate of error when they have a perfect record, or if the number of test they have done is very small the laboratory may use an estimate take from the industry in general.

4. The second report on DNA typing from the national research council. The first report of the National Research Council (NRC, 1992) suggested a number of changes to the standard procedures used by forensic laboratories for computing match probabilities. Many of these suggestions were aimed at resolving some of the uncertainties generated by population subgroups and some suggestions dealt directly with the proper reporting and quantification of laboratory error rates.

There were numerous complaints about several recommendations of the first NRC committee and most of the important recommendations were never implemented by forensic laboratories. In response to this situation and requests directly from agencies like the FBI a second committee was created by the National Research Council to reconsider mainly issues of statistics and population genetics. This second committee issued their report in May, 1996 (NRC, 1996). Below I review some of the important conclusions in that report.

A major reason for the creation of this committee was the suggestion by the forensic community that since 1992 there had been new scientific studies and data collected which obviated the need for many of the first NRC's recommendations. An important part of this new research were a number of studies examining statistical independence itself. Many of these studies have been previously cited in this paper. Since these studies tend to come to very different conclusions an important role for the new NRC committee would have been to review this research and provide its own scientific evaluations of the pros and cons of the competing claims.

In fact the new committee does consider these studies (NRC, 1996, chapter 4). However, the analysis is superficial and incomplete. While they clearly are disposed to accept the studies which have supported claims of independence they don't offer any scientific evaluation for why they support one set of hypotheses tests and not another. The most striking example of published papers which reach opposite conclusions regarding independence are Weir (1992) and Slimowitz and Cohen (1993) since they both analyzed exactly the same databases but come to different conclusions. The NRC report not only fails to analyze the different methodologies and conclusions in these papers but the report even fails to cite Slimowitz and Cohen as a paper making an important contribution to this issue.

When assessing the role of population substructure the committee uncritically accepts the inferences from the FBI's world wide study The problems with this type of analysis have already been discussed. While the committee does acknowledge that subgroups exist and may cause results from the product rule to be unreliable the committee offers only a partial solution. They suggest the incorporation of Wright's F-statistic when computing frequencies of homozygotes from PCR tests. This recommendation can correct for departures from Hardy-Weinberg equilibrium but not linkage equilibrium. The committee assumes departures from linkage equilibrium will be negligible.

The committee also correctly discusses a glaring and important design flaw in the FBI's fixed bin system (NRC, 1996, chapter 5; Fung, 1996). They point out that the size of many of the FBI's fixed bins are substantially less than their match criteria and thus biased to produce statistics that suggest matching profiles are rarer than they really are. To remedy this problem the first NRC committee had made the very reasonable suggestion that in those cases were calculations depend on fixed bins which are too small that they be made larger by merging them to a neighboring bin. The second NRC committee concludes this solution is not needed since, hopefully, profiles will consist of bins which are also larger than they need to be thus canceling out any errors created by bins which are too small. Since there is no guarantee that a DNA profile will always contain the appropriate mixture of bins which are too small and too large its hard to justify abandoning the first NRC's recommendation.

The committee also does not feel it is appropriate to use proficiency tests to attempt to estimate laboratory error rates (pgs. 3-10- 3-11). It appears that this conclusion is based on two different sorts of logic: (1) no proficiency test will ever fully replicate all the unusual features of real case work and hence are not relevant to real case work and (2) although several false positives were made in 1988 and 1989 in proficiency tests, since then no documented false positives have occurred, this signaling a time dependent trend in which error rates have now gone to zero or very close to it and thus can be safely ignored.

There are many compelling reason to reject these arguments and accept the original NRC's call to use proficiency tests to estimate laboratory error rates, here I review only some of the most obvious. If the NRC's logic about testing were correct than it should be impossible to assess the general level of competence of a lawyer by his performance on a bar exam since these involve the discussions of cases and facts which will certainly always differ from the details of "real' cases he will later deal with. In fact since most forensic proficiency tests are much less challenging than case work, error rates from proficiency tests could be considered a lower bound on case work error rates.

The arguments about the decline in error rates over time is also difficult to take seriously. As discussed earlier only if laboratory error rates were less than one in millions do they become unimportant. It seems unlikely that error rates, which may have been as high as 1 in 50 at Cellmark in late 1980's, have now dropped to 1 in millions. There is certainly no objective evidence to support this claim. Very recently a false match has been identified in actual case work. In late 1995 Cellmark mistakenly identified a know sample from a victim as the known sample from a defendant. This error than lead to the mistaken conclusion that the suspect matched DNA in an evidence sample which in fact only matched the victim (testimony of C. Word at California v. John Ivan Kocak, 17 November 1995, No. SCD110465).

A second example is found in testimony of Ms. Donna Dowden, a forensic scientist with the California Department of Justice DNA laboratory (California v. Noah Isaiah Wright, SC-078796A). In this testimony Ms. Dowden describes a false positive she made in an internal proficiency test utilizing DQ-α. In addition to Ms. Dowden's error in her report was an error by the personnel who reviewed the report since they failed to detect the error the first time the report was reviewed. In this proficiency test the false match was actually due to an error in transcribing data from the original tests results to the final report (and of course of the supervisors failure to detect the error). Nevertheless, when Ms. Dowden was asked in court about her performance on proficiency tests she stated that she had performed satisfactorily. The basis for this opinion is Ms. Dowden's belief that since the molecular biological part of the tests gave the correct result her inability to correctly report this result was of no consequence. In fact Cellmark employee's have also suggested that the false match in the Koncak case does not constitute a false positive since the source of the error was clerical rather than a failure in the molecular biology (testimony of C. Word, California v. Bishop, 21 August 1996). Ultimately, a laboratory report which incorrectly declares a match between evidence and defendant DNA has the same misleading effect, whether the error is clerical in nature or represents a failure of the basic scientific procedures.

Even if one thinks proficiency tests can only provide rough estimates of laboratory error rates, a range of error rates could be utilized to show how they effect the ultimate weight of a DNA match. The failure of the second NRC report to make reasonable recommendations about estimating and reporting laboratory error rates has now been criticized repeatedly (Balding, 1997; Koehler, 1997; Lempert, 1997; Thompson, 1997).

There continues to exist a diversity of opinions on the proper manner to summarize the weight of DNA evidence. In attempting to respond to some of the criticisms of the first NRC report the second committee has contradicted several very useful recommendations in the first report. The application of DNA typing techniques in forensic science requires a balance between reasonable inference and tenuous speculation. That balance has not been achieved in the second report of the National Research Council.

Acknowledgements. I thank an anonymous referee for helpful comments on the manuscript and Dr. Paul Lewis for use of his software to carry out the exact tests.

REFERENCES

[1] BALDING, D.J., *Errors and misunderstandings in the second NRC report*, Jurimetrics 37:469–476, 1997.

[2] BALZAS, I., J. NEUWEILER, P. GUNN, J. KIDD, K.K. KIDD, J. KUHL, AND L. MINGJUN, *Human population genetic studies using hypervariable loci 1. Analysis of Assamese, Australian, Cambodian, Caucasian, Chinese and Melanesian populations*, Genetics 131:191–198, 1992.

[3] BROWN, A.H.D. AND M.W. FELDMAN, *Population structure of multilocus associations*, Proc. Natl. Acad. Sci. USA 78: 5913–5916, 1981.

[4] BUDOWLE, B., A.M. GUISTI, J.S. WAYE, F.S. BAECHTEL, R.M. FOURNEY, D.E. ADAMS, L.A. PRESLEY, H.A. DEADMAN, AND K.L. MONSON, *Fixed bin analysis for statistical evaluation of continuous distribution of allelic data from VNTR loci, for use in forensic comparisons*, Amer. J. Hum. Genet. 48:841–855, 1991.

[5] BUDOWLE, B., J.A. LINDSEY, J.A. DECOU, B.W. KOONS, S.M. GUISTI, C.T. COMEY, *Validation and population studies of the loci LDLR, GYPA, HBGG, D7S8 and Gc (PM loci), and HLA-DQα using a multiplex amplification and typing procedure*, J. Forensic Sci. 40:45–54, 1995.

[6] BUDOWLE, B., K.L. MONSON, A.M. GIUSTI, AND B.L. BROWN, *The assessment of frequency estimates of Hae III-generated VNTR profiles in various reference databases*, J. Forensic Sci. 39:319–352, 1994a.

[7] BUDOWLE, B., K.L. MONSON, A.M. GIUSTI, AND B.L. BROWN, *Evaluation of Hinf I generated VNTR profile frequencies determined using various ethni databases*, J. Forensic Sci. 39:988–1008, 1994b.

[8] CHAKRABORTY, R. AND S.P. DAIGER, *Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah*, Human Biology 63:571–587, 1991.

[9] COHEN. J., M. LYNCH AND C.E. TAYLOR, *Forensic DNA tests and Hardy-Weinberg equilibrium*, Science 253:1037–1038, 1991.

[10] DEKA, R., R. CHAKRABORTY, AND R.E. FERRELL, *A population genetic study of six VNTR loci in three ethnically defined populations*, Genomics 11:83–92, 1991.

[11] DEVLIN, B. AND N. RISCH, *A note on Hardy-Weinberg equilibrium of VNTR data by using the Federal Bureau of Investigation's fixed-bin method*, Am. J. Hum. Genet. 51:549–553, 1992.

[12] DEVLIN, B., N. RISCH, AND K. ROEDER, 1990. *No excess of homozygosity at loci used for DNA fingerprinting*, Science 249:1416–, 1990.

[13] DEVLIN, B., N. RISCH, AND K. ROEDER, *Forensic DNA tests and Hardy-Weinberg equilibrium*, Science 253:1039–1041, 1991.

[14] EDWARDS, A.H.A. HAMMOND, L. JIN, T. CASKEY, AND R. CHAKRABORTY, *Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups*, Genomics 12:241–253, 1992.

[15] FELDMAN, M.W., I. FRANKLIN, AND G.J. THOMSON, *Selection in complex genetic systems I. The symmetric equilibria of the three-locus symmetric viability model*, Genetics 76:135–162, 1974.

[16] FUNG, W.K., *A numeric 10% or 5% match window in DNA profiling?* Forensic Sci. Int. 78:111–118, 1996.

[17] GEISSER, S., *Some statistical issues in forensic DNA profiling*, In: *Modelling and Prediction: honoring Seymour Geisser*, J.C. Lee, W.O. Johnson, A. Zellned (eds.), Springer, New York, 1996.

[18] GEISSER, S. AND W. JOHNSON, *Testing independence of fragment lengths within VNTR loci*, Am. J. Hum. Genet. 53:1103–1106, 1993.

[19] GEISSER, S. AND W. JOHNSON, *Testing independence when the form of the bivariate distribution is unspecified*, Statistics in Medicine 14:1621–1639, 1995.

[20] GREEN, P. AND E.S. LANDER, *Forensic DNA tests and Hardy-Weinberg equilibrium*, Science 253:1038–1039, 1991.

[21] GUO, S.W. AND E.A. THOMPSON, *Performing the exact test of Hardy-Weinberg proportion for multiple alleles*, Biometrics 48:361–372, 1992.

[22] HAGERMAN, P.J., *Letter to the editor*, Am. J. Hum. Genet. 47:876–877, 1990.

[23] HARTL, D.L. AND A.G. CLARK, *Principles of Population Genetics*, 2nd Ed., Sunderland, MA., Sinauer Associates, 1989.

[24] KIDD, J.R., F.L. BLACK, K.M. WEISS, I. BALAZS, AND K.K. KIDD, *Studies of three Amerindian populations using nuclear DNA polymorphisms*, Human Biology 63:775–794, 1991.

[25] KOEHLER, J.J., *Error and exaggeration in the presentation of DNA evidence at trial*, Jurimetrics 34:21–39, 1993.

[26] KOEHLER, J.J., *Why DNA likelihood ratios should account for error (even when a National Research Council Report say they should not)*, Jurimetrics 37:425–437, 1997.

[27] KOEHLER, J.J., A. CHIA AND S. LINDSEY, *The random match probability (RMP) in DNA evidence: irrelevant and prejudicial?*, Jurimetrics, 35:201–219, 1995.

[28] KRANE, D.E., R.W. ALLEN, S.A. SAWYER, D.A. PETROV AND D.L. HARTL, *Genetic differences at four DNA typing loci in Finnish, Italian and mixed Caucasian populations*, Proc. Natl. Acad. Sci. USA 89:10583–10587, 1992.

[29] LANDER, E. AND B. BUDOWLE, *DNA fingerprinting dispute laid to rest*, Nature 371:735–738, 1994.

[30] LEMPERT, R., *Some caveats concerning DNA as criminal identification evidence: with thanks to the Reverend Bayes*, Cardozo Law Review 13:303–341, 1991.

[31] LEMPERT, R., *After the DNA wars : skirmishing with NRC II*, Jurimetrics 37:439–468, 1997.

[32] LEWONTIN, R.C. AND D.L. HARTL, *Population genetics in forensic DNA typing*, Science 254:1745–1750, 1991.

[33] MAISTE, P.J. AND B.S. WEIR, *A comparison of tests for independence in the FBI RFLP data bases*, Genetica 96:125–138, 1995.

[34] MILLER, R.G., *Simultaneous Statistical Inference*, McGraw-Hill: New York, 1966.

[35] MORTON, N.E., *Genetic structure of forensic populations*, Proc. Natl. Acad. Sci. USA 89:2556–2560, 1992.

[36] MUELLER, L.D., *Population genetics of hypervariable human DNA*, In: *Forensic DNA Technology*, M.A. Farley, J.J. Harrington (eds.), Lewis Publishers, Chelsea, MI, 1991.

[37] MUELLER, L.D., *The use of DNA typing in forensic science*, Accountability in Research 3:55–67, 1993.

[38] MUELLER, L.D., *Letter to the Editor*, J. Forensic Sci. 43:446–447, 1998.

[39] National Research Council, *DNA Technology in Forensic Science* , National Academy Press, Washington, D. C., 1992.

[40] National Research Council *The Evaluation of Forensic DNA Evidence*, National Academy Press, Washington, D. C., 1996.

[41] ODELBERG, S.J., R. PLATKE, J.R. ELDRIDGE, L. BALLARD, P. O'CONNELL, Y. NAKAMURA, M. LEPPERT, J.M. LALOUEL, AND R. WHITE, *Characterization of eight VNTR loci by agarose gel electrophoresis*, Genomics 5:915–924, 1989.

[42] RISCH, N.J. AND B. DEVLIN, *On the probability of matching DNA fingerprints*, Science 255:717–720, 1992.

[43] SAJANTILA, A., B. BUDOWLE, M. STROM, V. JOHNSSON, M. LUKKA, L. PELTONEN, AND C. EHNHOLM, *PCR amplification of alleles at the D1S80 locus: comparison of a Finnish and a North American Caucasian population sample, and forensic casework evaluation*, Am. J. Hum. Genet. 50:816–825, 1992.

[44] SAWYER, S., D. KRANE, A. PODLESKI AND D. HARTL, *DNA fingerprinting loci do show population differences - comments on Budowle et al.*, Am. J Hum. 59:272–274, 1996.

[45] SLIMOWITZ, J.R. AND J.E. COHEN, *Violations of the ceiling principle: exact conditions and statistical evidence*, Am. J. Hum. Genet. 53:314–323, 1993.

[46] THOMPSON, W.C., *Subjective interpretation, laboratory error, and the value of forensic DNA evidence: three case studies*, Genetica 92:153–168, 1995.

[47] THOMPSON, W.C., *Accepting lower standards: the National Research Council's second report on forensic DNA evidence*, Jurimetrics 37:405–424, 1997.

[48] WEIR, B.S., *Independence of VNTR alleles defined as fixed bins*, Genetics 130:873–887, 1992a.

[49] WEIR, B.S., *Independence of VNTR alleles defined as floating bins*, Am. J. Hum. Genet. 51:992–997, 1992b.

[50] ZAYKIN, D., L. ZHIVOTOVSKY AND B.S. WEIR, *Exact tests for association between alleles at arbitrary numbers of loci*, Genetica 96:169–178, 1995.