Statistical Issues in DNA Profiling

B.S. Weir and J.S. Buckleton*

Program in Statistical Genetics, Department of Statistics
North Carolina State University, Raleigh NC 27695-8203, USA

INTRODUCTION

DNA profiles consist of pairs of variants, or alleles, at one or more genetic loci. If the profile from an evidentiary sample does not match that from a suspect in a crime, then the suspect is excluded from having contributed to the sample. When the two profiles do match, however, then the suspect is not excluded from being a contributor. In this case, it may be necessary to attach some numerical weight to the evidence of a match, and statistical issues arise.

As discussed by Evett and Buckleton in this volume, and as described fully by Aitken (1995), the appropriate weighting of evidence is by means of a likelihood ratio. If E is the evidence of matching DNA profiles, and C and $\bar{C}$ are alternative explanations for that evidence, then the relative merits of the two explanations can be compared by the ratio of the probabilities of the evidence under each:

$$L = \frac{\Pr(E|C)}{\Pr(E|\bar{C})}$$

A court can be told that the evidence is L times more likely under C than under $\bar{C}$.

It is often the case that the prosecution explanation C fully explains the evidence, in which case $\Pr(E|C) = 1$ and L becomes the reciprocal of the probability of E under the defense explanation $\bar{C}$. A major exception to this simplification is when the evidentiary sample gives a profile that must have come from more than one person but the crime was committed by one person. For single stains, however, it should be straightforward to determine the probability of the evidence under $\bar{C}$ from the principles of population genetics. Translating a theoretical probability into a numerical estimate on the basis of samples of profiles requires statistical methods.

SINGLE STAINS

Suppose the profile A_iA_j is found for locus **A** from both an evidentiary sample and a suspect S. The circumstances of the crime imply that the sample was left by the perpetrator P, so that the evidence E can be written $S = A_iA_j, P = A_iA_j$. The two explanations are

Prosecution	C :	$S = P$
Defense	$\bar{C}$:	$S \neq P$

The likelihood ratio is

$$\begin{aligned} L &= \frac{\Pr(S = A_iA_j, P = A_iA_j | S = P)}{\Pr(S = A_iA_j, P = A_iA_j | S \neq P)} \\ &= \frac{\Pr(P = A_iA_j | S = A_iA_j, S = P)\Pr(S = A_iA_j | S = P)}{\Pr(P = A_iA_j | S = A_iA_j, S \neq P)\Pr(S = A_iA_j | S \neq P)} \\ &= \frac{1}{\Pr(P = A_iA_j | S = A_iA_j, S \neq P)} \end{aligned}$$

* ESR:Forensic, Mt Albert Science Centre
Private Bag 92-021, Auckland, NEW ZEALAND

This last expression is from a "suspect-anchored" viewpoint, and requires knowledge of the probability with which an unknown person P has the profile when it is known that a different person S has that profile. It has been the assumption that this probability does not depend on S since $\bar{C}$ says that S and P are assumed to be different people. This assumption is obviously false if S and P are related, meaning that they share common recent ancestors, and it is also false when S and P are related in an evolutionary sense. The latter dependence is more of a factor when S and P belong to the same population, particularly if that population is small.

Suspect and Perpetrator Independent

Evidently the expression

$$L = \frac{1}{\Pr(P = A_i A_j)}$$

is a very special case, even though it is the one in general use. The probability refers to people in some population of "potential perpetrators" defined by the circumstances of the crime. Although genetic frequencies are determined by evolutionary forces, and will therefore differ between ethnic groups, it is unlikely that the circumstances of the crime will describe the ethnicity of the perpetrator with any precision. It is equally unlikely that the ethnicity of the suspect is the reason that person is considered to be in the potential perpetrator population, and this population is certainly not defined by the suspect's ethnicity.

The simplest means for attaching a numerical value to $P_{ij} = \Pr(P = A_i A_j)$ is to sample people and determine the proportion $\tilde{P}_{ij}$ of the sample with this genotype. Debate over the appropriate population to sample can be avoided by sampling the entire population or by presenting estimates from several different ethnic groups. In either case, the estimate can differ from the true value. If the sample was a random one of size n, then it is helpful to report an upper confidence limit with the estimate and this is

$$\tilde{P}_{ij} + 1.645\sqrt{\tilde{P}_{ij}(1 - \tilde{P}_{ij})/n}$$

for a 95% limit *providing* $\tilde{P}_{ij}$ is not too small. This is likely to be the case for conventional blood groups or STR loci with only a few alleles.

For VNTR loci or other systems with many alleles, genotype frequencies are too small to allow confidence limits to be given by the binomial/normal theory of the previous paragraph. Indeed, samples of a few hundred people may not even contain representatives of the genotype in question. This is the point at which independence of the two alleles A_i and A_j is assumed and the genotype frequency estimated as the product of sample allele frequencies

$$P_{ij} \hat{=} \begin{cases} \tilde{p}_i^2 & A_i = A_j \\ 2\tilde{p}_i\tilde{p}_j & A_i \neq A_j \end{cases}$$

Methods for testing for independence have recently been reviewed by Maiste and Weir (1995). The very conditions under which use of the product rule is needed are those which invalidate normal-theory confidence limits, so the expressions of Chakraborty et al. (1993) are to be avoided. Instead, numerical resampling procedures such as the bootstrap (Efron and Tibshirani 1993) can be used.

In practice, DNA profiles involve several loci. Each particular genotype is very rare and product-rule estimates, along with bootstrap confidence limits, are necessary.

Suspect and Perpetrator Dependent

Calculating the likelihood ratio when the suspect and perpetrator are close relatives can be complex, although methods are in place (Weir 1994). This situation will not be treated here. Methods for pairs of people in the same population are not yet fully developed. Conditional genotypic frequencies of the form $\Pr(A\ A\ |A\ A\ )$ or $\Pr(A\ A\ |A\ A\ )$ are required. Direct estimates of these quantities are

under development but, as they refer to dependencies imposed by evolutionary forces, they cannot be estimated from a single population. The problem has been discussed extensively for the simpler case of pairs of allele frequencies $\Pr(A_i|A_i)$ in different individuals (e.g. Cockerham 1969).

It may be helpful to realize that

$$\Pr(A_i|A_i) \;=\; p_i + (1-p_i)\theta$$

where θ is often written as F_{ST}. This parameter, quantifying the correlation of alleles between individuals within populations also gives the variance component between populations. It is logically impossible to estimate this quantity from a single population.

There is an analogous parameterization for frequencies such as $\Pr(A_iA_i|A_iA_i)$, although three-gene and four-gene analogs of θ are needed (Cockerham 1971, Weir 1994). The expressions given recently by Balding and Nichols (1994) are true only for populations at equilibrium under the evolutionary forces of drift and infinite-alleles mutation. The relevance of these expressions for actual human populations has not yet been explored fully.

MIXED STAINS

There is currently some debate over the interpretation of mixed stains, even though the appropriate methodology was laid out by Evett et al. (1991). For the present discussion, it will be assumed that there are no dependencies between any of the principals in a crime: suspects, victims or perpetrators. The simplest case is where an evidentiary sample contains four alleles A_i, A_j, A_k, A_l and circumstances of the crime imply this sample has DNA from the victim V and one perpetrator P. If the victim has profile A_iA_j and a suspect S has A_kA_l then the suspect is not excluded from having contributed to the sample. The prosecution explanation C_1 is that the sample has DNA from V and S, whereas the defense explanation $\bar{C}_1$ is that it has DNA from V and some unknown perpetrator P.

The likelihood ratio for C versus $\bar{C}_1$ is

$$\begin{aligned} L &= \frac{\Pr(E|C_1)}{\Pr(E|\bar{C}_1)} \\ &= \frac{\Pr(V=A_iA_j, S=A_kA_l, P=A_kA_l|S=P)}{\Pr(V=A_iA_j, S=A_kA_l, P=A_kA_l|S\neq P)} \\ &= \frac{\Pr(S=A_kA_l, P=A_kA_l|S=P)}{\Pr(S=A_kA_l, P=A_kA_l|S\neq P)} \\ &= \frac{1}{\Pr(P=A_kA_l)} \end{aligned}$$

as in the single-stain case.

Another situation is where it is not known with certainty that the victim has contributed to the evidentiary sample, and the defense explanation $\bar{C}_2$ is that there are two unknown contributors $U1, U2$. Now the likelihood ratio is

$$\begin{aligned} L &= \frac{\Pr(E|C_1)}{\Pr(E|\bar{C}_2)} \\ &= \frac{\Pr(\text{evidence profile } A_iA_jA_kA_l|\text{contributors } S,V)}{\Pr(\text{evidence profile } A_iA_jA_kA_l|\text{contributors } U1,U2)} \\ &= \frac{1}{\Pr(\text{evidence profile } A_iA_jA_kA_l|\text{contributors } U1,U2)} \end{aligned}$$

The probability needed is that of finding the four distinct alleles A_i, A_j, A_k, A_l from two unknown people. Evidently these people must both be heterozygotes, but there are three possible heterozy-

gote pairs: $A_iA_j\&A_kA_l$, $A_iA_k\&A_jA_l$ and $A_iA_l\&A_jA_k$. Under the assumption of allelic independence, therefore,

$$L = \frac{1}{24p_ip_jp_kp_l}$$

A third situation may have a prosecution explanation C_2 of S and some unknown person $U1$ being the contributors, and then the likelihood ratio for C_2 versus $\bar{C}_2$ is

$$\begin{aligned} L &= \frac{\Pr(E|C_2)}{\Pr(E|\bar{C}_2)} \\ &= \frac{\Pr(\text{evidence profile } A_iA_jA_kA_l|\text{contributors } S, U1)}{\Pr(\text{evidence profile } A_iA_jA_kA_l|\text{contributors } U1, U2)} \\ &= \frac{\Pr(\text{profile } A_iA_j \text{ from } U1)}{\Pr(\text{profile } A_iA_jA_kA_l \text{ from } U1, U2)} \\ &= \frac{2p_ip_j}{24p_ip_jp_kp_l} \\ &= \frac{1}{12p_kp_l} \end{aligned}$$

The interpretation of mixed stains must therefore take into account the alternative explanations being offered. The calculations all hinge on the probabilities with which mixed stain profiles are found among two (or more) people. There is no logic in considering the probability with which a single person has a genotype included in the mixture (NRC 1992), or in working with probabilities with which single people would be excluded from a compound profile.

The number of contributors to a mixed stain may be dictated by circumstances of the crime, but in other cases the number may not be known with certainty. This could be the case when a location associated with a crime has several blood stains, some of which reveal six alleles at a locus and some of which reveal four alleles. Were the latter stains left by two contributors, or were there actually three and some alleles were not detected? Some of these complexities are present in the following example.

EXAMPLE

In the case of People of the State of California v. Orenthal James Simpson (Los Angeles County Case BA097211), evidence was presented for blood stains on the center console of a Bronco automobile. Three RFLP profiles were determined by the California Department of Justice DNA Berkeley Laboratory, and the band lengths are shown in Table 1, along with band lengths for two of the principals in the case, defendant OS and victim RG. At D4S139 and D5S110, evidentiary sample bands a, b match those of OS and bands c, d match those of RG. At D2S44, bands a, b match those of OS and bands a, c match those of RG. None of the 11 bands match those of the third principal, victim NB.

The fragment frequencies shown in Table 1 are from four FBI databases: AA=African American, CA=Caucasian, SE=Southeast Hispanic, and SW=Southwest Hispanic. Frequencies are assigned by the fixed-bin method described by Budowle et al. (1991). Consistency of genotype frequencies in these databases with the assumption of independence has been demonstrated by Maiste and Weir (1995).

The evidentiary stain could not have come from one person, although it could have come from two contributors, and indeed can be explained if OS and RG were the contributors. Other stains in the Bronco, however, contain alleles at other loci that match those of victim NB. One stain, involving locus DQα, was claimed by the defense to require a contributor other than the three principals, so the court ordered that statistics be provided for three and four contributors.

Table 1 RFLP profiles for Bronco Center Console.

		Fragment lengths			Frequency			
Locus	Allele	Sample	OS	RG	AA	CA	SE	SW
D2S44	a	2931	2925	3017	.0316	.0859	.0983	.0387
	b	1874	1877		.0842	.0827	.0750	.0898
	c	1684		1689	.0926	.1073	.1050	.1109
D4S139	a	8899	8915		.0770	.0951	.1013	.1264
	b	3281	3301		.0525	.0311	.0241	.0189
	c	7203		7192	.1094	.1911	.1672	.1830
	d	5683		5733	.0837	.1077	.1061	.1472
D5S110	a	11356	11355		1.0000	1.0000	1.0000	1.0000
	b	4777	4778		.0581	.0391	.0385	.0515
	c	5717		5722	.0765	.0274	.0367	.0455
	d	3015		3022	.0765	.0910	.0927	.0758

An example of the kinds of calculations needed is shown in Table 2. This table shows the genotypes, in terms of allelic names, of two possible contributors. The corresponding frequencies, with those for one contributor being taken from the AA database, and the other from the CA database, are also shown. For loci D4S139 and D5S110 the calculations are straightforward, and just lay out all six possible pairs of heterozygotes. For locus D2S44, care is needed.

Table 2 Details of two-contributor calculations.

AA Type	CA Type	Frequency			AA Type	CA Type	Freq.
		D2S44				D4S139	
		$\phi = 0.0$	$\phi = 0.1$	$\phi = 0.5$	a b	c d	.000333
a a / a d	b c	.000018	.000130	.000579	a c	b d	.000113
a b	a c	.000098	.000098	.000098	a d	b c	.000153
a b	b c	.000094	.000094	.000094	b c	a d	.000235
a b	c c / cd	.000061	.000175	.000632	b d	a c	.000319
a c	a b	.000083	.000083	.000083	c d	a b	.000108
a c	b b / c d	.000040	.000137	.000524		D5S110	
a c	b c	.000104	.000104	.000104	a b	c d	.000579
b b / b d	a c	.000131	.000441	.001683	a c	b d	.001089
b c	a a / a d	.000115	.000383	.001455	a d	b c	.000328
b c	a b	.000222	.000222	.000222	b c	a d	.001618
b c	a c	.000287	.000287	.000287	b d	a c	.000487
c c / c d	a b	.000122	.000385	.001437	c d	a b	.000915

The forensic scientist reported a possible fourth allele d at D2S44 close in size to allele a, as would be consistent with the explanation that OS and RG were the contributors. If it is not certain that a fourth allele is seen, however, there is the possibility that the RFLP technology simply did not detect the fourth allele d implied by two contributors. For this reason, the evidentiary profile a, b, c contributors include both aa, bc and ad, bc genotype pairs, with the d in the second pair not being seen. This unseen allele is assigned a frequency ϕ, and the value $\phi = 0$ corresponds to the case when it is known the evidentiary profile contains only three alleles at the locus. Some estimates for unseen RFLP allele frequencies have been reported, and have not been greater than 0.05 (Chakraborty et al. 1994). An extremely conservative upper bound, not at all supported by any data, would be $\phi = 0.5$. Each of these three values of ϕ was used in Table 2.

Reciprocals of frequencies for all combinations of racial groups, and for 2, 3 or 4 contributors, are shown in Table 3. Note that there is no need to double frequencies (or halve reciprocals) for pairs of different racial groups as these are conditional frequencies. A reciprocal for the pair AA,CA is the same as for CA,AA and for (AA,CA or CA,AA).

Table 3 Reciprocals of frequencies with which unknown contributors would have the evidentiary profile. (D2S44 unseen band frequency $\phi = 0.1$).

2 unknowns				4 unknowns				
AA	AA		114,237,467	AA	AA	AA	AA	2,746,756,219
AA	CA		62,197,505	AA	AA	AA	CA	1,228,857,678
AA	SE		60,855,431	AA	AA	AA	SE	1,278,168,818
AA	SW		68,058,586	AA	AA	AA	SW	1,287,152,804
CA	CA		67,204,161	AA	AA	CA	CA	631,696,172
CA	SE		66,357,214	AA	AA	CA	SE	661,957,388
CA	SW		58,133,830	AA	AA	CA	SW	615,392,828
SE	SE		68,882,220	AA	AA	SE	SE	698,478,400
SE	SW		60,114,492	AA	AA	SE	SW	642,501,663
SW	SW		74,741,492	AA	AA	SW	SW	646,894,041
				AA	CA	CA	CA	375,349,666
3 unknowns				AA	CA	CA	SE	393,922,618
AA	AA	AA	199,297,316	AA	CA	CA	SW	346,965,288
AA	AA	CA	94,627,491	AA	CA	SE	SE	416,665,128
AA	AA	SE	96,111,383	AA	CA	SE	SW	365,353,456
AA	AA	SW	101,405,846	AA	CA	SW	SW	342,048,695
AA	CA	CA	58,701,786	AA	SE	SE	SE	443,988,463
AA	CA	SE	60,198,945	AA	SE	SE	SW	387,829,353
AA	CA	SW	55,377,118	AA	SE	SW	SW	360,009,506
AA	SE	SE	62,636,897	AA	SW	SW	SW	365,721,013
AA	SE	SW	56,784,217	CA	CA	CA	CA	269,572,101
AA	SW	SW	60,333,497	CA	CA	CA	SE	281,441,469
CA	CA	CA	49,958,156	CA	CA	CA	SW	238,546,758
CA	CA	SE	51,055,589	CA	CA	SE	SE	297,186,005
CA	CA	SW	43,324,748	CA	CA	SE	SW	252,291,022
CA	SE	SE	53,300,230	CA	CA	SW	SW	226,565,281
CA	SE	SW	45,246,827	CA	SE	SE	SE	317,154,522
CA	SW	SW	43,308,048	CA	SE	SE	SW	269,724,555
SE	SE	SE	56,752,747	CA	SE	SW	SW	242,188,143
SE	SE	SW	48,248,000	CA	SW	SW	SW	233,397,498
SE	SW	SW	45,998,948	SE	SE	SE	SE	341,890,383
SW	SW	SW	52,037,162	SE	SE	SE	SW	291,370,357
				SE	SE	SW	SW	261,840,024
				SE	SW	SW	SW	251,670,721
				SW	SW	SW	SW	266,668,092

Simply presenting the frequencies with which unknown contributors have (only) the evidentiary profile between them does not provide the full forensic implication of the match between that profile and the profiles of OS and RG. As discussed above, the weight to be attached to this match is obtained as the ratio of the frequencies under alternative pairs of explanations, as shown in Table 4. In each case the probability for explanation C is divided by the probability for explanation $\bar{C}$. For example, the evidence is between 58 and 114 million times more likely to have arisen if OS and RG were the contributors than if two unknown people were the contributors, and the frequency of unseen bands at D2S44 is 0.1. The range of values reflects all possible combinations of racial groups for the unknown people.

Table 4 Likelihood ratios for interpreting evidence.

Explanation C	Explanation $\bar{C}$	Likelihood ratio		
		$\phi = 0.0$	$\phi = 0.1$	$\phi = 0.5$
Two Contributors				
OS+RG	OS+U1	65,000–150,000	42,000–96,000	17,000–38,000
OS+RG	RG+U1	38,000–73,000	23,000–52,000	9,400–23,000
OS+RG	U1+U2	100,000,000–220,000,000	58,000,000–114,000,000	21,000,000–38,000,000
OS+U1	U1+U2	720–20,000	630–1,700	560–1,400
RG+U1	U1+U2	1,000–5,800	1,100–4,800	910–4,000
Three contributors				
OS+RG+U1	OS+U1+U2	2,000–5,000	1,200–2,800	480–1,000
OS+RG+U1	RG+U1+U2	1,200–3,600	740–2,400	290–1,000
OS+RG+U1	U1+U2+U3	1,000,000–4,600,000	490,000-1,700,000	100,000–320,000
OS+U1+U2	U1+U2+U3	400–1,100	290–800	150–410
RG+U1+U2	U1+U2+U3	650–2,100	410–1,300	190–670
Four contributors				
OS+RG+U1+U2	OS+U1+U2+U3	880–2,200	500-1,100	190–430
OS+RG+U1+U2	RG+U1+U2+U3	550–1,500	320–1,000	120–430
OS+RG+U1+U2	U1+U2+U3+U4	260,000–1,300,000	100,000–460,000	18,000–71,000
OS+U1+U2+U3	U1+U2+U3+U4	110–680	87–440	42–200
RG+U1+U2+U3	U1+U2+U3+U4	410–1,100	230–690	97–290

U1,U2,U3,U4 are distinct unknown people.

CONCLUSION

The statistical issues arising in DNA profiling are currently centering on ways in which to accommodate population structure and to interpret mixtures. There is no problem with either provided likelihood ratios and the principles of population genetics are used. The care with which forensic samples are collected and analyzed must be matched by care in statistical analysis, and in explaining these analyses to a court.

ACKNOWLEDGEMENTS

This work was supported in part by NIH grant GM45344, and by award 95-IJ-CX-0007 from the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position of the U.S. Department of Justice. The cooperation of Mr. Gray Sims from the California Department of Justice is deeply appreciated.

REFERENCES

Aitken CGG (1995) Statistics and the Evaluation of Evidence for Forensic Scientists. Wiley, New York.

Balding DJ, Nichols RA (1993) DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Foren Sci Int 64:125–140.

Budowle B, Giusti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, Presley LA, Deadman HA, Monson KL (1991) Fixed-bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci, for use in forensic comparisons. Am J Hum Genet 48:841-855.

Chakraborty R, Zhong Y, Jin L, Budowle B (1994) Nondetectability of restriction fragments and independence of DNA fragment sizes within and between loci in RFLP typing of DNA. Am J Hum Genet 55:391–401.

Chakraborty R, Srinivasan MR, Daiger SP (1993) Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics. Am J Hum Genet 52:60-70

Cockerham CC, (1969) Variance of gene frequencies. Evolution 23:72-84

Cockerham CC (1971) Higher order probability functions of identity of alleles by descent. Genetics 69:235–246.

Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman and Hall, New York.

Evett IW, Buffery C, Wilott G, Stoney D (1991) A guide to interpreting single locus profiles of DNA mixtures in forensic cases. J Foren Sci Soc 31:41–47

Maiste PJ, Weir BS (1995) Comparison of tests for independence in the FBI RFLP databases. Genetica 96:125-138.

NRC (National Research Council) (1992) DNA Technology in Forensic Science. National Academy Press, Washington, DC

Weir BS (1994) Effects of inbreeding on forensic calculations. Ann Rev Genet 28:597–621.