J Forensic Sci, Sept. 2004, Vol. 49, No. 5
Paper ID JFS2003039
Available online at: www.astm.org

TECHNICAL NOTE

Bruce S. Weir,[1] *Ph.D.*

Matching and Partially-Matching DNA Profiles

ABSTRACT: The DNA profiles of two individuals can have 0, 1, or 2 pairs of alleles that are the same at each locus. These events may be called mismatches, partial matches or matches, respectively, and they have probabilities that depend on the population proportions of alleles as well as the population structure parameter theta. The observed and expected numbers of pairs of individuals with various numbers of matching or partially matching loci in FBI and Australian databases are found to be in good agreement provided theta is set equal to some small value greater than zero. The likelihood ratios for two individuals having a specified degree of relationship versus being unrelated also depend on the numbers of matching and partially matching loci, but even unrelated pairs of individuals can have likelihood ratios that support hypotheses of relatedness. Matching probabilities allow predictions to be made for the sizes of databases that are expected to contain a pair of individuals with high numbers of matching loci. It is very likely that two individuals with at least 9 matching loci among the 13 CODIS loci have already been typed.

KEYWORDS: forensic science, matching DNA profiles, population structure, relatedness, Australian DNA data, FBI DNA data

When the genotypes of two individuals are the same, the individuals are said to have matching profiles at that locus or to share two pairs of alleles identical in state (ibs). Forensic scientists have generally not been interested in the case where the individuals share only one pair of alleles ibs and so partially match at that locus. By contrast, the proportions of pairs of individuals that share zero, one or two pairs of ibs alleles are the key elements of "affected relatives" methods for linkage mapping of human disease genes (1). This note explores the ibs probabilities for pairs of individuals, with attention to the case where any two alleles in the population have a probability θ of being identical by descent (ibd). These probabilities are used to determine the expected number of profiles that match or partially match at various numbers of loci, and a comparison is made of these numbers to those observed in FBI and Australian forensic databases. This numerical work confirms the wisdom of incorporating the population structure parameter θ into match probability calculations. The probabilities allow a prediction of how large a database should be before a high number of matching loci can be expected. Although partially-matching profiles may suggest relatedness, the numerical results also show that care is needed if relatedness is to be inferred. Discussions on relatedness are better expressed in terms of matching and partially matching loci rather than the total number of shared alleles.

Matching Profiles

As the number of loci used for forensic profiling grows, the probability that a random person will have any specific profile will decrease. The forensic question of interest, however, is the probability that an untyped person has a profile given that it has aleady been seen. These matching probabilities can be expressed in terms of allelic frequencies and the population structure parameter θ (2) as shown in the Appendix. When $\theta = 0$, these reduce to the "product rule" result of assuming allelic independence. If two individuals are drawn at random from the population, the probability $P_2(\theta)$ that they match (i.e. have two alleles in common) is found by adding together the products of the probability of each possible genotype and the match probability of that genotype. An algebraic expression is shown in the Appendix.

The effect of θ on matching probabilities can be illustated with the CODIS data published by the FBI (3) for samples from US African American, Caucasian and Southwest Hispanics. Every complete 13-locus profile in each of these samples was compared with every other complete profile in the same sample. The number of profile pairs matching at each of the 13 loci is shown in Tables 1*a*–1*c*, along with the numbers expected from the value of $P_2(\theta)$. There is good overall agreement between the observed counts and the product rule result, although nearly half the time the observed counts are larger—meaning that the product rule is not conservative. Setting $\theta = 0.01$ does produce a conservative result at nearly every locus in all three samples.

Do the single-locus results offer a good guide to the matching probabilities for the whole profile? There are dependencies among matching probabilities even for unlinked loci (4), although these are not expected to be large for loci with low mutation rates. There is a slight tendency for the dependencies to rise with the number of loci (4), but the number of pairs of profiles in the FBI data (3) is too low to allow meaningful statements beyond three loci. A much larger set of nine-locus short tandem repeat (STR) profiles (Profiler Plus^{TM}) has been assembled by the Australian forensic agencies (5). This set represents people of various ethnic backgrounds, including Asian, Australian Aboriginal, Caucasian and Maori, and the ethnic composition of the set does not represent the ethnic composition of Australia. A conservative value of θ for this heterogeneous sample should certainly be conservative for a more homogeneous sample.

[1] Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC.

Received 8 Feb. 2003; and in revised form 28 Mar. 2004; accepted 3 April 2004; published 3 Aug. 2004.

TABLE 1*a*—*One-locus matches in FBI African American data (15,576 pairs of 13-locus profiles).*

Locus	Obs. No.	$\theta = 0.000$	$\theta = 0.001$	$\theta = 0.005$	$\theta = 0.010$	$\theta = 0.030$
D3S1358	1422	1467	1475	1507	1547	1712
vWA	1069	**982**	**989**	**1019**	**1056**	1211
FGA	488	515	521	546	578	712
D8S1179	1160	1270	1278	1311	1351	1519
D21S11	560	**534**	**540**	565	598	733
D18S51	429	456	462	486	516	643
D5S818	1826	**1748**	**1756**	**1789**	1830	2001
D13S317	2069	2123	2133	2176	2229	2442
D7S820	1278	**1252**	**1260**	1291	1331	1494
CSF1PO	1144	1254	1262	1294	1334	1498
TPOX	1346	1397	1405	1440	1483	1658
TH01	1724	**1702**	**1712**	1751	1800	1998
D16S539	1143	**1092**	**1099**	**1129**	1167	1324

Boldface when observed number is greater than expected number.

TABLE 1*b*—*One-locus matches in FBI Caucasian data (18,721 pairs of 13-locus profiles).*

Locus	Obs. No.	$\theta = 0.000$	$\theta = 0.001$	$\theta = 0.005$	$\theta = 0.010$	$\theta = 0.030$
D3S1358	1443	**1397**	**1406**	**1441**	1485	1669
vWA	1179	**1168**	**1177**	1212	1256	1440
FGA	679	**668**	**675**	705	743	903
D8S1179	1188	1256	1266	1305	1354	1555
D21S11	677	710	718	749	789	955
D18S51	509	530	537	564	599	749
D5S818	3054	**2960**	**2971**	**3012**	3065	3279
D13S317	1414	1588	1598	1639	1689	1897
D7S820	1170	1222	1231	1267	1312	1499
CSF1PO	2290	**2212**	**2222**	**2260**	2309	2509
TPOX	3860	**3646**	**3659**	**3712**	**3777**	4038
TH01	1393	1522	1531	1568	1614	1805
D16S539	1614	1658	1668	1708	1758	1963

Boldface when observed number is greater than expected number.

TABLE 1*c*—*One-locus matches in FBI Southwest Hispanic data (20,301 pairs of 13-locus profiles).*

Locus	Obs. No.	$\theta = 0.000$	$\theta = 0.001$	$\theta = 0.005$	$\theta = 0.010$	$\theta = 0.030$
D3S1358	2365	2439	2452	2501	2562	2811
vWA	1648	1751	1762	1806	1861	2088
FGA	535	560	568	597	635	796
D8S1179	1682	**1438**	**1448**	**1490**	1543	1760
D21S11	1084	1166	1177	1218	1271	1486
D18S51	608	**584**	**592**	622	660	825
D5S818	2355	2449	2461	2511	2573	2822
D13S317	995	1075	1084	1120	1166	1357
D7S820	1765	**1761**	1772	1814	1867	2084
CSF1PO	2720	2822	2833	2876	2930	3153
TPOX	4073	4244	4259	4316	4387	4669
TH01	1949	2008	2018	2060	2113	2330
D16S539	1789	**1768**	**1779**	1821	1874	2092

Boldface when observed number is greater than expected number.

Using these data has the advantages of a very large sample and of avoiding the issue of defining ethnicity. For three of the loci, the simple product rule match expectation is less than the observed number of matches and so is not conservative. The proportion of times the expected number is less than the observed number for each locus drops as θ increases, and is zero for $\theta = 0.005$. As the number of loci increases, the proportions of cases in which the product rule estimate of the multi-locus number of matches is less than the observed number are: $12/36 = 0.33$ for two loci, $42/84 = 0.50$ for three loci, $88/126 = 0.70$ for four loci and $92/126 = 0.73$ for five loci. The numbers of matches are too small to be meaningful for more than five loci.

It is necessary to put things in perspective. The observed and expected numbers of matches, from the Australian data, for all 126 combinations of five of the nine loci, are plotted in Fig. 1. The values shown in the figure for $\theta = 0$ are those for the product rule, and they assume that all 10 alleles in the five-locus profiles are independent. There is good overall fit of these to the observed numbers, with some sets of loci having more matches than expected and some having less. However, under-estimating low matching probabilities is prevented by using products over loci of the match probabilities with θ greater than zero, and Fig. 1 shows such values for $\theta = 0.005$ and $\theta = 0.01$. All of the values for $\theta = 0.01$ exceed the observed values, and they are very conservative for the higher

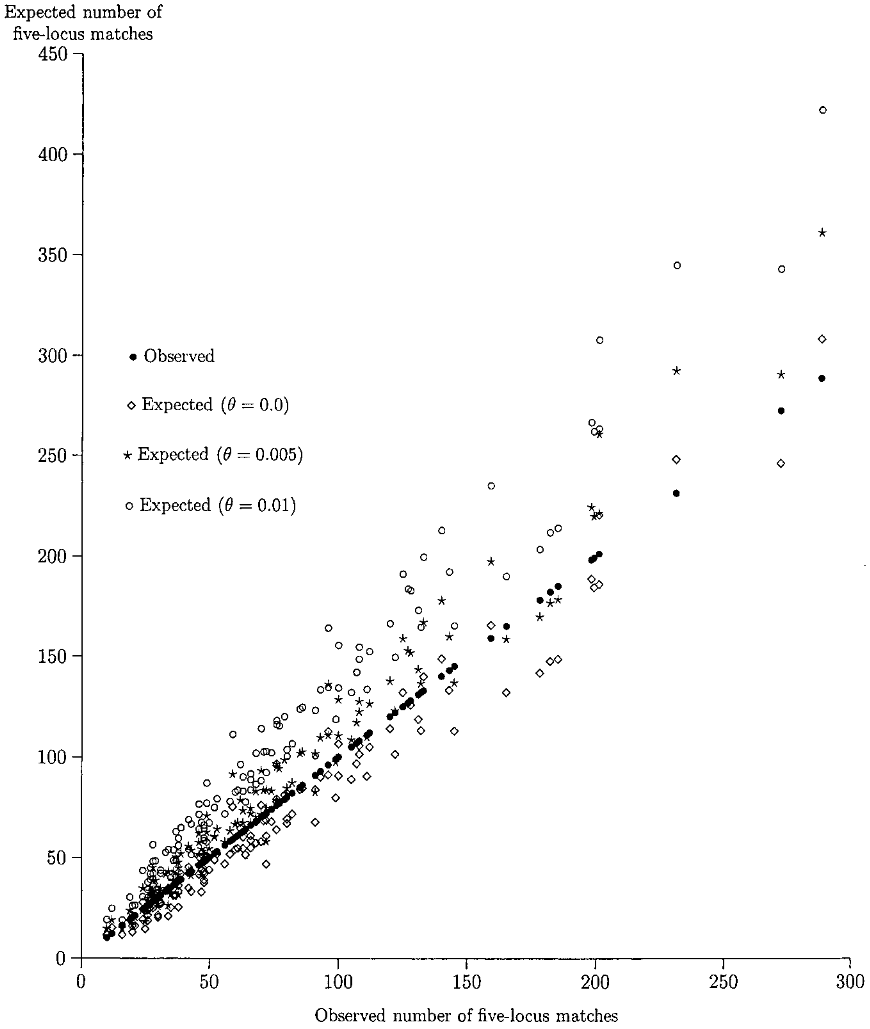

FIG. 1—*Observed and expected numbers of five-locus matches.*

TABLE 2—*One-locus matches in Australian data (109.039.528 pairs of 9-locus profiles).*

Locus	Obs. No.	θ = 0.000	θ = 0.001	θ = 0.005	θ = 0.010	θ = 0.030
D3S1358	9.091.486	9.154.812	9.208.646	9.425.305	9.699.114	10.827.586
vWA	6.931.377	6.973.970	7.025.350	7.232.343	7.494.403	8.579.544
FGA	3.377.511	**3.370.139**	3.410.904	3.576.069	3.787.241	4.683.813
D8S1179	5.558.642	5.615.844	5.666.469	5.870.460	6.128.816	7.199.803
D21S11	4.243.112	**4.214.155**	4.260.146	4.445.878	4.682.026	5.670.764
D18S51	2.854.192	2.876.360	2.914.837	3.070.981	3.271.167	4.126.894
D5S818	12.501.923	**12.345.349**	**12.404.405**	12.641.567	12.940.147	14.158.538
D13S317	7.783.307	7.842.196	7.896.707	8.116.011	8.392.982	9.532.738
D7S870	6.939.289	7.032.855	7.082.648	7.283.409	7.537.938	8.595.634

Boldface when observed number is greater than expected number.

matching probabilities. The observed numbers of matches are less for six or more loci and do not allow meaningful comparisons to be made but the trends for one to five loci suggest that it would be conservative to use an even larger value, say θ = 0.03, for nine-locus profiles. It is only because the Australian dataset is so large that these conclusions have been possible. It should be stressed that Table 2 and Fig. 1 are based on the combined data of Aboriginal, Asian, Cauucasian and Maori origin and they are affected by ethnic heterogeneity. It can be shown (unpublished results) that the dataset does not conform to the Hardy-Weinberg law, but Fig. 1 confirms that matching probabilities may be estimated in a conservative fashion by an appropriate "theta correction."

Partial Matches

At a single locus there are seven distinct pairings of individuals depending on allele sharing and whether or not the individuals are homozygous. General expressions for the probabilities of these seven cases have been given previously (6) and can be expressed in

TABLE 3—*Observed (o) and expected (e) numbers n^*_{xy} of matches and partial matches in Australian data.*

		n_{xy}									
x		$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$y = 7$	$y = 8$	$y = 9$
0	o	125.059	1.136.621	4.557.267	10.567.988	15.579.931	15.201.461	9.794.391	4.022.350	953.990	99.980
	e	106.387	1.012.655	4.231.719	10.189.442	15.578.703	15.682.188	10.392.445	4.371.272	1.058.818	112.516
1	o	155.283	1.233.623	4.246.000	8.288.485	10.005.378	7.664.890	3.636.565	976.872	114.164	
	e	139.135	1.149.315	4.103.359	8.269.178	10.286.150	8.085.981	3.922.172	1.073.131	126.790	
2	o	82.817	562.232	1.627.369	2.600.748	2.465.110	1.387.844	432.156	57.101		
	e	77.037	543.917	1.625.700	2.665.831	2.589.647	1.489.985	470.078	62.728		
3	o	24.370	140.382	334.303	419.197	291.803	107.937	16.651			
	e	23.745	140.360	341.353	437.082	310.712	116.255	17.885			
4	o	4.422	21.423	39.599	36.325	16.631	3.078				
	e	4.492	21.600	41.010	38.417	17.755	3.239				
5	o	559	1.973	2.778	1.713	400					
	e	540	2.028	2.816	1.715	386					
6	o	39	111	105	40						
	e	41	113	102	30						
7	o	0	8	5							
	e	2	3	2							
8	o	0	1								
	e	0	0								
9	o	0									
	e	0									

* x loci with two alleles matching, y loci with one allele matching.

TABLE 4—*Sample sizes for which specified matching is to be expected.*

	θ				
Population	0.000	0.001	0.005	0.010	0.030
Australian (all 9 loci)	660.000	640.000	540.000	450.000	230.000
US African-American (any 9 of 13 loci)	7.700	7.500	6.700	5.700	3.300
US Caucasian (any 9 of 13 loci)	6.400	6.200	5.500	4.800	2.800
US Southwest Hispanic (any 9 of 13 loci)	4.400	4.300	3.900	3.400	2.000
US African-American (all 13 loci)	43.000.000	41.000.000	34.000.000	27.000.000	11.000.000
US Caucasian (all 13 loci)	34.000.000	32.000.000	27.000.000	22.000.000	9.300.000
US Southwest Hispanic (all 13 loci)	21.000.000	20.000.000	17.000.000	13.000.000	5.900.000

terms of allele frequencies and the parameter θ (1). These probabilities are shown in the Appendix table and adding over all possible alleles leads to the probabilities $P_0(\theta)$, $P_1(\theta)$, $P_2(\theta)$ that two random individuals share zero or one, or two alleles ibs, and expressions for these are shown in the Appendix. The Appendix also shows how these probabilities lead to expressions for the numbers of loci for which two individuals either match, partially match (i.e., share only one allele ibs), or do not match.

If each individual in a sample is compared with each other individual, the number of loci (x) at which they match and the number of loci (y) at which they partially match can be found. The counts n_{xy} for matches and partial matches for the Australian data are shown in Table 3. For example, there were 13 distinct pairs of profiles that matched at 7 out of 9 loci: 8 of these pairs had partial matches at 1 of the remaining 2 loci, and the other 5 pairs had partial matches at both the remaining loci. In other words, $n_{71} = 8$, $n_{72} = 5$. Clearly, the number of matching loci decreases as the number of loci increases.

The corresponding numbers $e_{x,y}$ expected from the theory described in the Appendix are also shown in Table 3 for the case of $\theta = 0.001$. There is good overall agreement between observed and expected numbers (there is much less agreement if θ is zero). The expected values can also be used to predict the size of the Australian database when one pair of matching nine-locus profiles is expected to be found. These numbers decrease as θ increases, and are shown in Table 4 along with some predictions for US populations. The values in Table 4 refer to matching without specifying which particular profile it is that matches. The values do not give the size of the database in which a specific profile is expected to occur once.

Relatedness

The genotypes of a pair of individuals can be used to address the question of relatedness, and the likelihood ratio for the hypothesis that two individuals are related versus the hypothesis that they are unrelated is

$$\mathrm{LR}_{\mathrm{Rel.}} = P_0 + P_1 U + P_2 W \qquad (1)$$

The quantities P_0, P_1, P_2 are the probablities that individuals with the hypothesized relationship share 0, 1 or 2 pairs of alleles ibd (6). For full-sibs the values are $^1/_4$, $^1/_2$, $^1/_4$; for grandparent-grandchild or half-sibs or uncle-nephew the values are $^1/_2$, $^1/_2$, 0; for parent-child the values are 0, 1, 0; and for first cousins the values are $^3/_4$, $^1/_4$, 0. The quantities U, W are functions of frequencies of the shared alleles. If the two individuals have genotypes *ab* and *cd* at a locus, define u_1 for *ac*, u_2 for *ad*, u_3 for *bc* and u_4 for *bd*. If the two alleles of any of these pairs are ibs, then the u value is the reciprocal of the frequency of that allele. Otherwise $u = 0$. Then $U = (u_1 + u_2 + u_3 + u_4)/4$ and $W = (u_1 u_4 + u_2 u_3)/2$, and these two terms refer to partial matching and matching, respectively. The

TABLE 5—*Proportions of profile pairs for which the relatives likelihood ratio exceeds one in Australian data.*

x	Relationship	$y = 0$	$y = 1$	$y = 2$	$y = 3$	$y = 4$	$y = 5$	$y = 6$	$y = 7$	$y = 8$	$y = 9$
0	Full sibs	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.15	0.64
	First cousins	0.00	0.00	0.00	0.02	0.09	0.28	0.64	0.91	0.99	1.00
1	Full sibs	0.00	0.00	0.00	0.00	0.00	0.02	0.17	0.68	0.98	
	First cousins	0.00	0.01	0.06	0.19	0.52	0.85	0.98	0.99	1.00	
2	Full sibs	0.00	0.00	0.00	0.02	0.21	0.73	0.99	1.00		
	First cousins	0.03	0.13	0.40	0.77	0.96	0.99	1.00	1.00		
3	Full sibs	0.00	0.03	0.24	0.77	0.99	1.00	1.00			
	First cousins	0.29	0.66	0.93	0.99	1.00	1.00	1.00			
4	Full sibs	0.27	0.81	0.99	1.00	1.00	1.00				
	First cousins	0.87	0.99	1.00	1.00	1.00	1.00				
5	Full sibs	1.00	1.00	1.00	1.00	1.00					
	First cousins	1.00	1.00	1.00	1.00	1.00					
6	Full sibs	1.00	1.00	1.00	1.00						
	First cousins	1.00	1.00	1.00	1.00						
7	Full sibs	...	1.00	1.00							
	First cousins	...	1.00	1.00							
8	Full sibs	...	1.00								
	First cousins	...	1.00								
9	Full sibs	...									
	First cousins	...									

x loci with two alleles matching, y loci with one allele matching.

value of LR allows statements of the type "The probability of these two profiles if they came from relatives is LR times greater than the probability if they came from unrelated people."

The proportions of pairs of profiles in the Australian data for which the likelihood ratios are greater than one, supporting the hypothesis of relatedness, for two common relationships are shown in Table 5. Essentially, the same values were found for data simulated for unrelated individuals with the same allele frequencies—except that the simulated data had no eight-locus matches. Even if two individuals are not related, when they share one or two alleles at several loci there can be a substantial chance that likelihood ratios would support the hypothesis of them being related.

Discussion

As database sizes grow, the numbers of matching loci for any two profiles in the data also grows, and it is of interest to predict how much matching is to be expected by chance. The degree of matching depends on the relationship among the people for whom the profiles are determined, and account must be taken of the relationships caused by the shared evolutionary history of humans as well as those for members of the same family. The former can be conveniently summarized by a single parameter θ, and the numerical work presented here supports the practice of assigning a small non-zero value to θ. Questions about family relationships are best answered with matching and partially matching probabilities.

A high degree of allele sharing between pairs of profiles suggests relatedness, and the single instance of eight-locus matching in the Australian data was for a father and son. It is not known whether or not the seven-locus matches are for relatives, but several such matches were found in simulated data so that relatedness is by no means assured in those cases. There has recently been a discussion (8) on allele sharing in forensic profiles, but that discussion did not distinguish matching from partial matching. The need for both matching and partial matching data is illustrated by Eq 1 having separate terms for each. In Table 3 there are five instances where two profiles share eight of 18 alleles ($x, y = 0, 8$; 1, 6; 2, 4; 3, 2; 4, 0). These situations differ in the degree to which they favor the hypotheses of full sibs or first cousins (the proportions of times that the likelihood ratio for sibs is greater than that for cousins are 0.06, 0.09, 0.11, 0.15 and 0.21, respectively). Table 5, and the corresponding results for simulated data (not shown), shows that likelihood ratios favoring hypotheses of relatedness are expected even for unrelated pairs of individuals. Calculation of the probabilities of relatedness require prior probabilities as well as likelihood ratios.

A check of matching and partial matching among profiles in a database provides a useful diagnostic test. There were several instances of nine-locus matching profiles found initially in the combined Australian data. Subsequent investigation revealed that in each case the profiles were either from identical twins or from the same person typed by different agencies. There is no published explanation for the two pairs of matching profiles in the FBI Bahamian data (3). As offender databases grow, Table 3 illustrates that high degrees of matching are to be expected. It is very likely, for example, that there are already 9-locus matches within combined U.S. offender databases. The extent to which matching probabilities depend on the population structure parameter θ, as shown in all the numerical results in this note, points to the need for caution in basing "source attribution" arguments on the assumption of profile independence between individuals (i.e., assuming that θ is zero).

It should be stressed that the theory and results presented here are for averages over all possible profiles. The probabilities $P_0(\theta)$, $P_1(\theta)$, $P_2(\theta)$ do not refer to one specific profile. Matching of a suspect to a particular crime scene profile can constitute very strong evidence in favor of the hypothesis that the suspect is the source of the scene material.

Acknowledgments

This work was supported in part by NIH Grant GM 45344. Appreciation is extended to the Australian forensic scientists A. Bagdonavicius, B. Blair, C. Eckhoff, C. Pearman, P. Stringer, J. Sutton, J. West and L. Wynen who made their data available, and to Dr. Peta Stringer for coordinating that data collection. The published FBI data continue to be a valuable resource.

References

1. Liu W, Weir BS. Affected sib pair tests in inbred populations. Annals of Human Genetics. In press.
2. Balding DJ, Nichols RA. DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 1944;64:125–40.
3. Budowle B, Moretti TR. Genotype profiles for six population groups at the 13 CODIS short tandem repeat core loci and other PCR-based loci. Forensic Science Communications 1999. Available at: http://www.fbi.gov/programs/hq/lab/fsc/backissu/july1999/budowle.htm
4. Laurie C, Weir BS. Dependency effects in multi-locus match probabilities. Theor Pop Biol 2003;63:207–19.
5. Weir BS, Bagdonavicius A, Blair B, Eckhoff C, Pearman C, Stringer P, et al. Allele frequency data for Profiler Plus loci in Australia. J Forensic Sci 2004;49(5):1–3.
6. Evett IW, Weir BS. Interpreting DNA evidence. Sunderland, MA: Sinauer, 1998.
7. Brenner CH, Weir BS. Issues and strategies in the DNA identification of World Trade Center victims. Theor Pop Biol 2003;63:173–8.
8. Presciuttini S, Ciampini F, Alù M, Cerri N, Dobosz M, Domenici R. Allele sharing in first-degree and unrelated pairs of individuals in the Ge.F.I. AmpFℓSTR® and Profiler Plus™ database. Forensic Sci Int 2003;3488: 1–5.

Additional information and reprint requests:
Bruce S. Weir, Ph.D.
Bioinformatics Research Center
NC State University
Raleigh, NC 7695-7566

APPENDIX

Suppose that locus **A** has alleles A_i with population frequencies p_i. The matching probabilities for homozygotes A_iA_i and heterozygotes A_iA_j are

$$\Pr(A_iA_i \mid A_iA_i) = \frac{[2\theta + (1-\theta)p_i][3\theta + (1-\theta)p_i]}{(1+\theta)(1+2\theta)}$$

$$\Pr(A_iA_j \mid A_iA_j) = \frac{2[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]}{(1+\theta)(1+2\theta)}, \quad i \neq j$$

The Appendix Table shows the probabilities of all possible pairs of genotypes in terms of allele frequencies and θ. The probabilities that two individuals share 0, 1 or 2 alleles at a locus are found by adding the expressions in that Table over all possible alleles at the locus:

$$\begin{aligned} P_0(\theta) = \Bigg\{ & \theta^2(1-\theta)\left(1 - \sum_i p_i^2\right) + 2\theta(1-\theta)^2\left(1 - 2\sum_i p_i^2 \right. \\ & \left. + \sum_i p_i^3\right) + (1-\theta)^3\Bigg[1 - 4\sum_i p_i^2 + 4\sum_i p_i^3 \\ & + 2\left(\sum_i p_i^2\right)^2 - 3\sum_i p_i^4\Bigg]\Bigg\} \Big/ [(1+\theta)(1+2\theta)] \end{aligned}$$

$$\begin{aligned} P_1(\theta) = \Bigg\{ & 8\theta^2(1-\theta)\left(1 - \sum_i p_i^2\right) + 4\theta(1-\theta)^2\left(1 - \sum_i p_i^3\right) \\ & + 4(1-\theta)^3\Bigg[\sum_i p_i^2 - \sum_i p_i^3 - \left(\sum_i p_i^2\right)^2 \\ & + \sum_i p_i^4\Bigg]\Bigg\} \Big/ [(1+\theta)(1+2\theta)] \end{aligned}$$

$$\begin{aligned} P_2(\theta) = \Bigg\{ & 6\theta^3 + \theta^2(1-\theta)\left(2 + 9\sum_i p_i^2\right) \\ & + 2\theta(1-\theta)^2\left(2\sum_i p_i^2 + \sum_i p_i^3\right) \\ & + (1-\theta)^3\Bigg[2\left(\sum_i p_i^2\right)^2 - \sum_i p_i^4\Bigg]\Bigg\} \Big/ [(1+\theta)(1+2\theta)] \end{aligned}$$

If population structure or allelic dependence is ignored, $\theta = 0$, and the probabilities simplify:

$$P_0(0) = 1 - 4\sum_i p_i^2 + 4\sum_i p_i^3 + 2\left(\sum_i p_i^2\right)^2 - 3\sum_i p_i^4$$

$$P_1(0) = 4\sum_i p_i^2 - 4\sum_i p_i^3 - 4\left(\sum_i p_i^2\right)^2 + 4\sum_i p_i^4$$

$$P_2(0) = 2\left(\sum_i p_i^2\right)^2 - \sum_i p_i^4$$

If m loci are scored, then the ibs status of two individuals can be characterized by m_0, m_1, m_2, the numbers of loci at which they share zero, one or two pairs of alleles ibs respectively. It is convenient to index the loci by l, and to introduce indicator variables m_{l0}, m_{l1}, m_{l2} that are equal to one if the two individuals share zero, one or two alleles ibs respectively, and are zero otherwise. Then $\sum_i m_{li} = 1$, $m_i = \sum_l m_{li}$ and $\sum_i m_i = m$. If the loci are assumed to be independent, then the probabilities $P_{m_0,m_1,m_2}(\theta)$ of the allele-sharing status of two individuals are

$$P_{m_0,m_1,m_2}(\theta) = \sum_{m_{l0},m_{l1},m_{l2}} \prod_l P_{l0}(\theta)^{m_{l0}} P_{l1}(\theta)^{m_{l1}} P_{l2}(\theta)^{m_{l2}} \qquad (2)$$

where the sum is over all values of m_{li} such that $\sum_l m_{li} = m_i$ for $i = 0, 1, 2$. In the special case when each locus has the same set of allele frequencies, so that $P_{li}(\theta) = P_i(\theta)$ for $l = 1, 2, \ldots m$, this last result reduces to a multinomial expression:

$$P_{m_0,m_1,m_2}(\theta) = \binom{m}{m_0, m_1, m_2} P_0(\theta)^{m_0} P_1(\theta)^{m_1} P_2(\theta)^{m_2}$$

TABLE—*Joint genotypic probabilities.*

Genotypes	No. ibs Pairs	Probability
A_iA_i, A_iA_i	2	$p_i[3\theta + (1-\theta)p_i][2\theta + (1-\theta)p_i] \times [\theta + (1-\theta)p_i]/(1+\theta)(1+2\theta)$
A_iA_i, A_jA_j	0	$2(1-\theta)p_ip_j[\theta + (1-\theta)p_i][\theta + (1-\theta)p_i]/ (1+\theta)(1+2\theta)$
A_iA_i, A_iA_j	1	$4(1-\theta)p_ip_j[2\theta + (1-\theta)p_i][\theta + (1-\theta)p_i]/ (1+\theta)(1+2\theta)$
A_iA_i, A_jA_k	0	$4(1-\theta)^2p_ip_jp_k[\theta + (1-\theta)p_i]/(1+\theta)(1+2\theta)$
A_iA_j, A_iA_j	2	$4(1-\theta)p_ip_j[\theta + (1-\theta)p_i][\theta + (1-\theta)p_j]/ (1+\theta)(1+2\theta)$
A_iA_j, A_iA_k	1	$4(1-\theta)^2p_ip_jp_k[\theta + (1-\theta)p_i]/(1+\theta)(1+2\theta)$
A_iA_j, A_kA_l	0	$(1-\theta)^3p_ip_jp_kp_l/(1+\theta)(1+2\theta)$