square law detection), which fall under the headings of acquisition economy and, most generally, statistical regularity.

Acquisition economy means that fewer points are required to specify the Patterson function of a sample compared to its image. Although range in **k**-space (maximum gradient level and duration) must be sufficient for good spatial resolution for both diffraction and imaging, density of data points in **k**-space determines the spatial range, and this is much smaller for diffraction.

Statistical regularity refers to the aforementioned fact that statistical descriptions of systems often possess higher symmetry than the system itself. We have indicated that statistical characterization permits repeated signal acquisition for rearranging systems when statistical data are time-dependent but imaging data are not. Likewise, the rotational symmetry of Fig. 1, B and E, demonstrates angular symmetry, showing how 2-D features can be characterized by 1-D radial information. The lower density of required points allows a correspondingly reduced sample rate. The fact that diffraction information resides in an intrinsically narrower bandwidth than data for comparably resolved images is encouraging, and preliminary results indicate that Patterson functions are indeed "cleaner" than corresponding images. Signal averaging is quicker because the data are simpler, and extensions that would impractically lengthen imaging experiments become possible. For example, contrast-enhancing preparation sequences (*26*) may be applied to selectively weighted regions according to spectroscopic or mobility differences (*27*) between regions with different morphologies or composition. The statistical approach described here should eventually make NMR studies of small-scale inhomogeneities in plastics, ceramics, and structural materials a practical possibility.

REFERENCES AND NOTES

1. A. L. Patterson, *Phys. Rev.* **46**, 372 (1934).
2. P. Mansfield and P. K. Grannell, *J. Phys. C* **6**, L422 (1973).
3. ———, A. N. Garroway, D. C. Stalker, *Proc. Coll. Ampere (Krakow)* **1**, 16 (1973).
4. P. Mansfield, P. K. Grannell, A. A. Maudsley, *Proc. Coll. Ampere (Nottingham)* **2**, 431 (1974).
5. P. Mansfield and P. K. Grannell, *Phys. Rev. B* **12**, 3618 (1975).
6. P. Mansfield, *Contemp. Phys.* **17**, 553 (1976).
7. E. L. Hahn, *Phys. Rev.* **80**, 580 (1950).
8. H. Y. Carr and E. M. Purcell, *ibid.* **94**, 630 (1954).
9. E. O. Stejskal, *J. Chem. Phys.* **43**, 3597 (1965).
10. J. E. Tanner and E. O. Stejskal, *J. Phys. Chem.* **49**, 1768 (1968).
11. P. T. Callaghan, *Aust. J. Phys.* **37**, 359 (1984).
12. P. Stilbs, *Prog. Nucl. Magn. Reson. Spectrosc.* **19**, 1 (1987).
13. J. Kärger, H. Pfeiffer, W. Heink, *Adv. Magn. Reson.* **12**, 1 (1988).
14. J. Kärger and W. Heink, *J. Magn. Reson.* **51**, 1 (1983).
15. D. G. Cory and A. N. Garroway, *Magn. Res. Med.* **14**, 435 (1990).
16. L. H. Schwartz and J. B. Cohen, *Diffraction from Materials* (Springer-Verlag, Berlin, ed. 2, 1987).
17. P. T. Callaghan, D. McGowan, K. J. Packer, F. O. Zelaya, *J. Magn. Reson.* **90**, 177 (1990).
18. P. T. Callaghan, A. Coy, D. McGowan, K. J. Packer, F. O. Zelaya, *Nature* **351**, 467 (1991).
19. R. M. Cotts, *ibid.*, p. 443.
20. D. Twieg, *Med. Phys.* **10**, 610 (1983).
21. N. Wiener, *The Fourier Integral and Certain of Its Applications* (Dover, New York, 1958).
22. A. Khintchin, *Math. Ann.* **109**, 604 (1934).
23. S.-H. Chen, B. Chu, R. Nossal, Eds., *Scattering Techniques Applied to Supramolecular and Nonequilibrium Systems* (Plenum, New York, 1980).
24. W. A. Edelstein, J. M. S. Hutchison, G. Johnson, T. Redpath, *Phys. Med. Biol.* **25**, 751 (1980).
25. L. D. Landau and E. M. Lifshitz, *The Classical Theory of Fields*, M. Hammermesh, Transl. (Addison-Wesley, New York, 1962).
26. R. R. Ernst, G. Bodenhausen, A. Wokaun, *Principles of Nuclear Magnetic Resonance in One and Two Dimensions* (Oxford Univ. Press, Oxford, 1990).
27. D. B. Twieg, J. Katz, R. M. Peshock, *Magn. Reson. Med.* **5**, 32 (1987); G. C. Chingas, J. B. Miller, A. N. Garroway, *J. Magn. Reson.* **66**, 530 (1986).

28. We acknowledge the advice, encouragement, and support of A. Pines, in whose laboratory this work was carried out, and also acknowledge helpful discussions with M. Denn. Supported by the U.S. Department of Energy through the Director, Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division, under contract no. DE-AC03-76SF00098.

2 October 1991; accepted 19 December 1991

On the Probability of Matching DNA Fingerprints

NEIL J. RISCH* AND B. DEVLIN

Forensic scientists commonly assume that DNA fingerprint patterns are infrequent in the general population and that genotypes are independent across loci. To test these assumptions, the number of matching DNA patterns in two large databases from the Federal Bureau of Investigation (FBI) and from Lifecodes was determined. No deviation from independence across loci in either database was apparent. For the Lifecodes database, the probability of a three-locus match ranges from 1 in 6,233 in Caucasians to 1 in 119,889 in Blacks. When considering all trios of five loci in the FBI database, there was only a single match observed out of more than 7.6 million comparisons. If independence is assumed, the probability of a five-locus match ranged from 1.32×10^{-12} in Southeast Hispanics to 5.59×10^{-14} in Blacks, implying that the minimum number of possible patterns for each ethnic group is several orders of magnitude greater than their corresponding population sizes in the United States. The most common five-locus pattern can have a frequency no greater than about 10^{-6}. Hence, individual five-locus DNA profiles are extremely uncommon, if not unique.

VNTR (VARIABLE NUMBER TANDEM repeat) loci are used to generate the "DNA fingerprints" that have been presented as evidence in criminal and paternity cases. These loci are extremely polymorphic, having potentially hundreds of alleles at a single locus (*1*). Any particular genotype at a collection of such loci is deemed to be so rare that many forensic scientists believe the probability two unrelated individuals have matching genotypes across a set of loci to be extremely small. When many VNTR loci are tested (for example, up to five), the probability of a matching pattern occurring by chance has been reported in criminal cases to be extremely small, often on the order of 10^{-7} to 10^{-8} or even less, and sometimes the probability suggests less than one matching pattern in the total population of North America. Yet it is often argued that these probabilities are calculated in a conservative fashion, that is, the true probabilities are even smaller (*2, 3*). Others have argued, however, that the probabilities are invalid and are unrealistically small [(*4*) but see (*5*)].

In forensic cases, probability estimates are obtained by the multiplication rule. For multiplication to be valid, the events must be statistically independent. Statistical independence allows one to multiply allele frequencies within a locus to derive a single-locus genotype probability and to multiply genotype probabilities across loci to obtain a multilocus genotype probability. Statistical independence within a locus is referred to as Hardy-Weinberg equilibrium (HW), while statistical independence across loci is called linkage equilibrium (LE).

The assumption of independence, both within and across loci, has been challenged (*6*). For forensics, the reference population is divided into major ethnic components: for instance, Caucasians, Blacks, and Hispanics. Sometimes Hispanics are further subdivided by geography. The argument is put forth that none of these ethnic components is genetically homogeneous and that mating patterns are

N. J. Risch, Division of Biostatistics, Department of Epidemiology and Public Health, and Department of Genetics, Yale University, Post Office Box 3333, 60 College Street, New Haven, CT 06510.
B. Devlin, Division of Biostatistics, Department of Epidemiology and Public Health, Yale University, Post Office Box 3333, 60 College Street, New Haven, CT 06510.

*To whom correspondence should be addressed.

not truly random; within each racial group are subgroups, with random mating within subgroups and limited mating among subgroups (*6*). If there are differences in allele frequencies among subgroups, then the assumption of independence would not be strictly true. To the contrary, certain single and multilocus genotypes would occur in excess of expected values and others would be in deficit (*4*). Nevertheless, for deviations from HW and LE of practical importance to occur, the variation among subpopulations in allele frequencies must be large (*5, 7*). Even a small amount of intermarriage, referred to as gene flow, among subgroups seriously diminishes the allelic variation among subgroups, ameliorating correlations among alleles. Because of the extreme demographic and genetic conditions required to cause substantial violation of independence, most population geneticists accept independence as their operating assumption, especially for loci on different chromosomes.

Nonetheless, claims have been made for significant deviation from HW expectations (homozygote excess or Wahlund effect) for a number of VNTR loci, with such deviation attributed to population substructure (*6*). In all these cases, however, the apparent excess of homozygotes could be attributed to artifacts of the electrophoretic methods involved (*2, 7, 8*).

If population substructure were a serious problem, leading to significant departures from statistical independence, certain VNTR genotype patterns might occur significantly more often than independence would predict, and hence the probability that two unrelated individuals have a matching DNA pattern could be considerably higher than usually reported. In this report, we examine: (i) the statistical independence of genotypes matching across sets of loci (that is, LE) and (ii) the probability that two random, unrelated individuals have matching genotypes at a set of VNTR loci.

The FBI database includes the five loci D1S7, D2S44, D4S139, D10S7, and D17S79. From the Lifecodes database we analyzed the three loci D2S44, D14S13, and D17S79, which we analyzed previously for HW (*8*). Although the two laboratories analyze two of the same loci (D2S44 and D17S79), they use different restriction enzymes to create restriction fragments. The FBI uses Hae III, which results in smaller fragments than those of Lifecodes, which uses the enzyme Pst I.

VNTR polymorphism is based on fragment length variation. Fragments are measured on electrophoretic gels. Typically, the measurement error is greater than repeat unit size, making discrete alleles unresolvable. In forensic work, matches are usually determined visually, although measurement criteria are also employed. In particular, two fragment sizes are declared to match if their measured sizes are within a certain distance apart, with the distance based on measurement error (*9*). For the purposes of the analysis, we use a bound of 2.4% of the mean of the two fragment sizes. Specifically, two fragments of measured size x and y are called a match if

$$\frac{|x - y|}{0.5\,(x + y)} \leq 0.024$$

In words, this criterion is equivalent to the fragments being within about four measurement SD of each other (*10*).

To estimate match probabilities for individual loci, we evaluated all possible pairs of genotypes for each locus by the matching rule (Table 1, A and B). This is equivalent to asking for the probability that a random, innocent suspect and an evidentiary sample would be declared a match by chance.

For the FBI data, the minimum and maximum single-locus match probabilities always occur for D1S7 and D17S79, respectively. The large match probability for locus D17S79 is expected because this locus has several common alleles (*11*). Small match probability for the locus D1S7 is consistent with its lack of common alleles (*2*). The match probabilities for the other loci are uniformly small. Blacks have the lowest match probabilities for all loci, reflecting the fact that the Black population has the greatest gene diversity. Caucasians and Southeast and Southwest Hispanics have similar match probabilities.

For the Lifecodes data, the pattern is similar. The minimum and maximum single-locus match probabilities always occur for loci D14S13 and D17S79, respectively. As observed in the FBI data, Blacks have the smallest match probabilities. The match probabilities are higher for the Lifecodes data than for the FBI data, even for the same loci, because the FBI fragments are cut much shorter than the Lifecodes fragments. Measurement SD is proportional to fragment size.

Under the assumption of LE, the occurrence of genotypes at pairs of loci should be independent. Therefore, the probability that two individuals have matching genotypes at a pair of loci should be the product of the single-locus match probabilities.

To test for violations of pairwise independence, we used only phenotypes typed at both loci of interest. From these data, we constructed two-by-two tables, with match–no match at the first locus being the two rows and match–no match at the second locus being the two columns. The expected values for the cells are simply the products of

Table 1. (**A**) Single-locus matching results from FBI database. *N*, the sample size; NC, number of comparisons; O(M), observed number of matches; P(M), probability of matches. (**B**) Single-locus matching results from Lifecodes database. The FBI divides its database into four major groups: Blacks, Caucasians, Southeastern Hispanics (from Florida, primarily of Cuban or Caribbean origin or both), and Southwestern Hispanics (from Texas and California, primarily of Mexican and other Latin American origin). Lifecodes does not divide its Hispanic population.

Locus	*N*	NC	O(M)	P(M)	*N*	NC	O(M)	P(M)
				(**A**) *FBI database*				
	Blacks				Caucasians			
D1S7	360	64,620	71	0.0011	593	175,528	194	0.0011
D2S44	475	112,575	240	0.0021	790	311,655	1,116	0.0036
D4S139	447	99,681	186	0.0019	593	175,528	416	0.0024
D10S7	287	41,041	48	0.0012	428	91,378	213	0.0023
D17S79	549	150,426	1,640	0.0109	775	299,925	11,484	0.0383
	Southeast Hispanics				Southwest Hispanics			
D1S7	305	46,360	47	0.0010	288	41,328	67	0.0016
D2S44	300	44,850	137	0.0031	284	40,186	119	0.0030
D4S139	311	48,205	116	0.0024	265	34,980	132	0.0038
D10S7	230	26,335	64	0.0024	283	39,903	85	0.0021
D17S79	314	49,141	1,593	0.0324	293	42,778	1,467	0.0343
				(**B**) *Lifecodes database*				
	Blacks				Caucasians			
D2S44	1010	509,545	9,739	0.0191	3116	4,853,170	242,558	0.0500
D14S13	709	250,986	2,290	0.0091	2318	2,685,403	83,155	0.0310
D17S79	1007	506,521	23,562	0.0465	3104	4,815,856	480,558	0.0998
	Hispanics							
D2S44	403	81,003	4,127	0.0509				
D14S13	302	45,451	559	0.0123				
D17S79	405	81,810	4,662	0.0570				

the marginals divided by the total number of comparisons. The traditional test for independence is a chi-square test. We used the usual chi-square formula; in this case, however, the statistic does not have a chi-square distribution (*12*). To evaluate the test statistics, we used the method of bootstrapping to obtain the null distribution (*13*).

The pairwise match results are given in Table 2, A and B. For the FBI data, there were three tests with P values less than 0.05: D4S139 and D10S7 in Southeast Hispanics ($P = 0.032$), D1S7 and D4S139 in Southwest Hispanics ($P = 0.044$), and D4S139 and D17S79 in Southwest Hispanics ($P = 0.024$). However, given that a total of 40 tests was performed, these should not be considered significant. For all ten pairwise comparisons across loci combined, the observed versus expected number of matches under independence show good correspondence: 8 versus 6.1 (Blacks); 66 versus 69.7 (Caucasians); 14 versus 12.0 (Southeast Hispanics); and 16 versus 13.4 (Southwest Hispanics).

The pattern of results for the Lifecodes data is similar. No test statistic has a P value less than 0.05. Across the three pairs of loci, the observed versus expected number of matches shows close correspondence: 612 versus 583.3 (Blacks); 36,206 versus 35,981.4 (Caucasians); and 294 versus 283.3 (Hispanics).

The results in Table 1A would suggest that a three-locus match in the FBI data would be very infrequent in any of the populations examined, assuming independence. This expectation is met since there are no three-locus matches for any population tested. For the Lifecodes data, because of the higher probability of single-locus matches and the larger sample sizes, three-locus matches are expected and occur: for Blacks, two matches occurred versus 1.9 expected; for Caucasians, 412 occurred versus 400.8 expected; and for Hispanics, 2 occurred versus 1.5 expected (Table 2). Again, these results show close correspondence between the number of matches observed and the number expected under independence (*14*).

In conclusion, these analyses suggest no deviation from independence of match probabilities across loci for any of the populations or loci tested, supporting the assumption of LE across loci. These results should dispel recent objections to the statistical validity of DNA fingerprinting (*4*).

We also looked for matching genotypes across ethnic groups. For the FBI data, out of a total of 7,064,266 between-group comparisons for all pairs of five loci, there were 176 additional matches (a rate of 2.5×10^{-5}), compared to a within-group match frequency of 104 out of 2,701,834 comparisons, or 3.8×10^{-5}. For three-locus comparisons, there was one three-locus match observed (between a Caucasian and a Southeast Hispanic at loci D2S44, D4S139, and D17S79) out of 7,628,360 total comparisons for all ethnic groups combined. There were no four-locus matches.

For the Lifecodes data, with the loci taken as pairs, there was a total of 9,674,639 comparisons between ethnic groups, and 11,230 matches were observed, or an overall rate of 0.00116. This compares to a within-group match frequency total of 37,112 out of 11,202,501 comparisons, or 0.00331. For the three-locus comparison, there were 66 additional matches out of 2,430,032 comparisons between groups (or a rate of 2.7×10^{-5}); this compares to a frequency of 416 out of 2,878,879 (or 1.45×10^{-4}) within groups (mostly Caucasians). The reduction in the matching rate between groups compared to within groups is a function of differences in allele frequency distributions between groups. The modest reductions observed here, especially for the FBI data, are consistent with the substantial similarity between ethnic groups in allele frequency distributions (*1*).

The observed independence of matching among loci, both in the FBI and Lifecodes data sets, provides no support for claims of linkage disequilibrium within ethnic groups. Indeed, if linkage disequilibrium among loci does exist, it has little effect on the probability of two random individuals having matching genotypes.

For the FBI data, no three-locus matches were observed within any racial group, and only a single three-locus match was observed in the entire database out of a total of more than 7.6 million comparisons. When independence (multiplicability) of match probabilities across loci is assumed and single-locus match probabilities are used (Table 1A), the probability of a five-locus match for each of

Table 2. Analysis of independence of loci. (**A**) FBI database. (**B**) Lifecodes database. N, sample size; NC, number of comparisons; E(M), expected number of matches; O(M), observed number of matches; and Ts, test statistic.

Loci	N	NC	E(M)	O(M)	Ts	N	NC	E(M)	O(M)	Ts
					(**A**) *FBI database*					
	Blacks					Caucasians				
D1,D2	342	58,311	0.1	0	0.15	580	167,910	0.7	1	0.19
D1,D4	350	61,075	0.1	0	0.13	573	163,878	0.4	0	0.43
D1,D10	276	37,950	0.1	0	0.04	415	85,905	0.2	1	2.80
D1,D17	335	55,945	0.7	2	2.74	554	153,181	6.5	3	1.87
D2,D4	410	83,845	0.3	1	1.28	577	166,176	1.4	0	1.42
D2,D10	275	37,675	0.1	0	0.11	418	87,153	0.7	1	0.18
D2,D17	450	101,025	2.3	3	0.14	744	276,396	37.9	36	0.01
D4,D10	279	38,781	0.1	0	0.08	412	84,666	0.5	1	0.63
D4,D17	422	88,831	1.8	2	0.03	558	155,403	14.1	20	2.34
D10,D17	276	37,950	0.5	0	0.44	406	82,215	7.3	3	2.86
	Southeast Hispanics					Southwest Hispanics				
D1,D2	279	38,781	0.1	0	0.12	265	34,980	0.2	0	0.17
D1,D4	288	41,328	0.1	0	0.10	254	32,131	0.2	1	3.68*
D1,D10	218	23,653	0.1	0	0.06	260	33,670	0.1	0	0.12
D1,D17	292	42,486	1.4	3	1.71	271	36,585	2.0	1	0.47
D2,D4	285	40,470	0.3	0	0.28	248	30,628	0.3	0	0.40
D2,D10	208	21,528	0.2	0	0.17	267	35,511	0.2	0	0.22
D2,D17	295	43,365	4.3	4	0.02	267	35,511	3.6	4	0.03
D4,D10	216	23,220	0.1	1	5.08*	241	28,920	0.2	0	0.26
D4,D17	300	44,850	3.5	4	0.10	249	30,876	4.0	8	5.27*
D10,D17	220	24,090	1.9	2	0.03	265	34,980	2.6	2	0.09
					(**B**) *Lifecodes database*					
	Blacks					Caucasians				
D2,D14	703	246,753	43.0	47	0.22	2303	2,650,753	4,102.4	4,416	4.01
D2,D17	1000	499,500	441.1	454	0.24	3090	4,772,505	23,775.2	23,721	0.19
D14,D17	698	243,253	103.2	111	0.81	2292	2,625,486	8,103.8	8,069	0.37
D2,D14,D17	693	239,778	1.9	2	0.01	2280	2,598,060	400.8	412	0.31
	Hispanics									
D2,D14	299	43,071	27.0	24	0.52					
D2,D17	395	77,815	225.9	246	1.35					
D14,D17	295	43,365	30.4	24	1.56					
D2,D14,D17	287	41,041	1.5	2	0.01					

*$P < 0.05$.

the four racial groups is 5.59×10^{-14} for Blacks, 8.40×10^{-13} for Caucasians, 5.87×10^{-13} for Southeast Hispanics, and 1.32×10^{-12} for Southwest Hispanics. If one were to consider genotypes as discrete entities, the number of different genotypes must be at least as great as the inverse of the match probability. The minimum number of genotypes would occur only when the frequency of the different genotypes are equal; hence, the actual number of genotypes is likely to be greater than the inverse of the match probability. For the five loci used by the FBI, the minimum number of genotypes (or "genotype equivalents," assuming discrete genotypes) for the four populations (Blacks, Caucasians, Southeast Hispanics, and Southwest Hispanics, respectively), are: D1S7 (909, 901, 990, 617), D2S44 (469, 279, 327, 338), D4S139 (535, 422, 415, 265), D10S7 (855, 429, 412, 469), and D17S79 (92, 26, 31, 29). Therefore, considering all five loci together, the minimum number of possible genotypes is 1.79×10^{13} for Blacks, 1.19×10^{12} for Caucasians, 1.70×10^{12} for Southeast Hispanics, and 7.58×10^{11} for Southwest Hispanics. Hence, the number of possible genotypes, for each group, exceeds their total U.S. population size by several orders of magnitude.

If there are n possible genotypes in a population, and genotype i occurs with frequency p_i, then the probability two randomly selected people from the population have matching genotypes is

$$Q_m = \sum_i p_i^2$$

Hence, $Q_m \geq p_i^2$ for each i, and, in particular, $Q_m \geq p_k^2$ where genotype k is the most frequent. Hence, the frequency of the most common genotype can be no greater than $(Q_m)^{1/2}$. On this basis, the most common five-locus genotype can be no more frequent than 1 in 4,230,000 in Blacks, 1 in 1,090,000 in Caucasians, 1 in 1,310,000 in Southeast Hispanics, and 1 in 870,000 in Southwest Hispanics. Therefore, although matching five-locus genotype patterns may exist (that is, everyone is not automatically unique), the number of such matches is vanishingly small compared to the population size. Because the assumptions underlying this analysis are conservative (all of the match probability is assigned to a single genotype), all unrelated individuals may have unique five-locus patterns.

For forensic cases, it has been argued that to be conservative, the probability of a random person from the population matching an observed DNA profile should be reported as $1/N$, where N is the number of individuals in the database (*4*). Our calculations show that such an approach is unnecessarily conservative, because the most common five-locus genotype pattern occurs with a frequency no greater than about 1 in 1,000,000. In any particular case, the actual frequency of genotype patterns in the population matching the forensic sample is likely to be far smaller. Hence, there is no reason to discount the very small match probabilities often reported in criminal cases.

Another approach to determining the number of different genotype patterns, which does not rely on independence across loci, per se, is to use the distribution of observed genotypes to estimate the number of unobserved genotypes. This approach is similar to that of estimating the number of unseen species (*15*) and gives an estimate of 2.38×10^{11} possible genotypes (*16*).

Although we find the probability of a matching DNA profile between unrelated individuals to be vanishingly small, especially at five loci, related individuals, in particular identical twins and siblings, have a far greater probability of matching genotypes. For identical twins, the probability is 1.0, while for siblings it is $(0.25)^5$ or 0.001. Therefore, in the forensic setting, we conclude that an innocent suspect has little to fear from DNA evidence, unless he or she has an evil twin.

REFERENCES AND NOTES

1. I. Balazs, M. Baird, M. Clyne, E. Meade, *Am. J. Hum. Genet.* **44**, 182 (1989); J. Flint, A. J. Boyce, J. J. Martinson, J. B. Clegg, *Hum. Genet.* **83**, 257 (1989); J. R. Kidd, F. L. Black, K. M. Weiss, I. Balazs, K. K. Kidd, *Hum. Biol.* **63**, 775 (1991).
2. B. Budowle *et al.*, *Am. J. Hum. Genet.* **48**, 841 (1991).
3. B. Devlin, N. Risch, K. Roeder, *J. Am. Stat. Assoc.*, in press.
4. R. C. Lewontin and D. L. Hartl, *Science* **254**, 1745 (1991).
5. R. Chakraborty and K. K. Kidd, *ibid.*, p. 1735.
6. E. S. Lander, *Nature* **339**, 501 (1989); P. Green and E. S. Lander, *Science* **253**, 1038 (1991); J. Cohen, *Am. J. Hum. Genet.* **46**, 358 (1990); J. E. Cohen, M. Lynch, C. E. Taylor, *Science* **253**, 1037 (1991).
7. B. Devlin, N. Risch, K. Roeder, *Science* **253**, 1039 (1991).
8. ———, *ibid.* **249**, 1416 (1990).
9. Measurement error of fragment size is proportional to the size of the fragment, with the measurement error being very similar for both laboratories. We previously estimated the Lifecodes measurement SD to be $(0.00575)L$, where L is the length of the fragment in kilobases. Our estimate of the FBI measurement SD is $(0.00625)L$, on the basis of their report (*2*). We therefore employed the same matching criterion for both databases.
10. In determining whether two genotypes (pairs of fragments) match, we require both that the smaller fragments of the two pairs and that the larger fragments of the two pairs meet the match criteria described above. This criterion is somewhat more generous than the visual match evaluation employed by the FBI and Lifecodes. Measurement errors for bands on the same gel are correlated, which acts to conserve the distance between bands. Thus, the pattern of banding of two profiles is also considered, leading to fewer matches than with the simple size criteria.
11. B. Devlin, N. Risch, K. Roeder, *Am. J. Hum. Genet.* **48**, 662 (1991).
12. The distribution of matching patterns is not the usual multinomial required for the chi-square test of independence, and the test statistic has greater variance than a chi-square distribution.
13. For background information on bootstrapping, see B. Efron and R. Tibshirani [*Science* **253**, 390 (1991)]. We formed two-locus phenotypes by randomly and independently choosing one-locus phenotypes from the database for those loci. We generated n independent genotypes in this fashion, from which we calculated the test statistic. This procedure was performed 1000 times to form the null distribution.
14. These results argue against the notion (*4, 6*) that the ethnic groups are composed of very diverse subgroups. If genetically diverse subpopulations are mixed, disequilibrium should increase with the number of loci. Hence, the proportional increase of matching over that expected should be greater for two-locus matches than for single-locus matches, even greater for three loci versus two, and so on.
15. R. A. Fisher, A. S. Corbet, C. B. Williams, *J. Anim. Ecol.* **12**, 42 (1943).
16. The primary assumption (*15*) is that the number of each genotype (repeats) follows a Poisson distribution. If S is the total number of genotypes (species), N the sample size, and λ the Poisson parameter (mean), then $\lambda = N/S$, or $S = N/\lambda$. The parameter λ is estimated from the observed distribution of repeats. In our case, the number of repeats is extremely small, and only duplications were observed. Since the "0" repeat class is never observed, the mean repeat frequency $\mu = \lambda/(1 - e^{-\lambda})$. Since λ is quite small, an excellent approximation is obtained by $\mu = \lambda/(\lambda - 1/2\lambda^2) = (1 - 1/2\lambda)^{-1}$; hence, $\lambda = 2(1 - \mu^{-1})$. From the FBI data, evaluation of three-locus matching in all populations combined yielded a single match at loci D2S44, D4S139, and D17S79 out of a total of 1448 individuals, for a mean repeat frequency $\mu = 1448/1447 = 1.000691$, and $\lambda = 0.00138$. Hence, the number of genotypes is estimated to be 1,048,364. For the remaining two loci, D1S7 and D10S7, there were three observed matches out of 1169 total individuals, yielding a mean repeat frequency $\mu = 1169/1166 = 1.00257$, and $\lambda = 0.00513$. Hence, the number of genotypes is estimated to be $1169/0.00513 = 226{,}595$. Combining across the two sets of loci gives a total number of possible genotypes of $1{,}048{,}364 \times 226{,}595$ or 2.38×10^{11}. Another way to view this problem is by analogy with the classic birthday problem [W. Feller, *Introduction to Probability Theory and Its Applications* (Wiley, New York, 1968), vol. 1, p. 33]. Assuming all 365 days of the year are equally likely for a birthday, when only 23 individuals are assembled, the probability that at least two have the same birthday is greater than 50%. Out of all ten trios of loci we examined from the FBI data, only a single match was observed. Because no matches were observed for the other nine trios, the number of different three-locus patterns must be quite large. For example, for loci D1S7, D2S44, and D4S139, there were no matches observed out of 1400 individuals examined. Letting r be the number of individuals sampled, t the number of different three-locus patterns, and assuming the various patterns are equally frequent, the probability x of no matches is given by

$$x = \left(1 - \frac{1}{t}\right) \cdots \left(1 - \frac{r-1}{t}\right)$$

A good approximation for x can be obtained from the logarithm

$$\log x \approx -\frac{1}{t}\frac{r(r-1)}{2} - \frac{1}{t^2}\frac{r(r-1)(2r-1)}{12}$$

From this formula, we calculate that the probability of at least one match $(1 - x)$ out of 1400 individuals would be 95% if there were as few as 327,000 different patterns. Undoubtedly, the true number of patterns for these three loci is far greater, and at all five loci combined many orders of magnitude greater. For instance, returning to the problem of matching birthdays, suppose no two people out of a sample of 20 have matching birthdays. Then we can be 95% confident there are at least 70 days to a year, and 50% confident there are at least 280.

17. We thank the FBI and Lifecodes for supplying the data and K. DeSanctis for technical support. This work was supported by NIH grants HG00348 and CA45052 to N.J.R.

25 October 1991; accepted 13 January 1992