13. D. M. Perkins, *U.S. Geol. Surv. Open-File Rep. 87-673* (1987), p. 428.
14. A. C. Johnston, *Earthquakes and Earthquake Engineering in the Eastern United States*, J. E. Beavers, Ed. (Ann Arbor Science, Ann Arbor, MI, 1981), pp. 161–181.
15. Catalogs used in this study are from M. S. Sibol and G. A. Bollinger, *Listing of Central and Eastern North American Earthquakes* (Virginia Polytechnical Inst. Seismological Observatory, Blacksburg, VA, 1989); *Electric Power Research Institute Catalog of Central and Eastern North American Earthquakes to 1985* (Electric Power Research Institute, Palo Alto, CA, 1985).
16. In the interval form of the frequency-magnitude equation, it is assumed that there is a finite or maximum magnitude cutoff for the analysis (we used m_b 7.5 to 8.0). As the cumulative form of the frequency-magnitude equation is unbounded, direct conversion of the interval results to the cumulative form for computing return times violates the maximum magnitude assumption.
17. Statistical tests of catalog completeness in the central and eastern United States indicate that intervals of complete reporting vary from 110 to 280 years for m_b 5, 125 to 300 years for m_b 6, and 160 to 300 years for m_b 7; O. W. Nuttli and R. B. Herrmann, *Misc. Pap. S-73-1* (U.S. Army Waterways Experiment Station, Vicksburg, MS, 1978); D. Veneziano and J. Van Dyck, *Seismic Hazard Methodology for Nuclear Facilities in the Eastern United States*, vol. 2 (Electric Power Research Institute, Palo Alto, CA, 1985).
18. The variability in magnitudes assigned to earthquakes in the catalogs consulted is greatest in the $m_b < 6$ range (which includes 85 to 93% of the total data set) whereas the number of events greater than m_b 6 is generally constant from catalog to catalog. By comparing different earthquake catalogs we have attempted to bracket at least some of the above variations in the smaller magnitude band.
19. Although a frequency–seismic moment regression would have dealt with the problem of instrumental magnitude saturation, we have used body-wave magnitudes (m_b) as the independent variable for our analysis to provide continuity with earlier studies. Table 1 includes seismic moment magnitude (**M**) equivalents of m_b values.
20. O. W. Nuttli, *Bull. Seismol. Soc. Am.* **73**, 519 (1983); T. C. Hanks and H. Kanamori, *J. Geophys. Res.* **84**, 2348 (1979).
21. B. Bender, *Bull. Seismol. Soc. Am.* **73**, 831 (1983); D. H. Weichert, *ibid.* **70**, 1337 (1980).
22. These consisted of five events: a m_b 5 foreshock to the 1897 m_b 5.8 Giles County, Virginia, earthquake, a m_b 5.5 aftershock of the 1940 m_b 5.5 New Hampshire event, and three of the 1811 to 1812 New Madrid earthquakes. The New Madrid earthquakes are determined to be dependent events as the probability of two independent earthquakes of $m_b \geq 7$ occurring within 2 months of the first event is $\sim 5 \times 10^{-8}$.
23. R. B. Herrmann, *Earthquake Notes* **48**, 47 (1977).
24. D. P. Russ, *Geol. Soc. Am. Bull.* **90**, 1013 (1979).
25. R. T. Saucier, *Geology* **17**, 103 (1989).
26. S. F. Obermeier, R. E. Weems, R. B. Jacobson, *U.S. Geol. Surv. Open-File Rep.* 87-504 (1987); P. Talwani and J. Cox, *Science* **229**, 379 (1985).
27. S. F. Obermeier, R. E. Weems, R. B. Jacobson, G. S. Gohn, "Earthquake hazards and the design of constructed facilities in the eastern United States," *Ann. N.Y. Acad. Sci.* **558**, 183 (1989).
28. A. J. Crone and K. V. Luza, *Bull. Geol. Soc. Am.* **102**, 1 (1990).
29. R. M. Hamilton and A. C. Johnston, *U.S. Geol. Surv. Circ. 1066* (1990).
30. O. W. Nuttli, in *Earthquakes and Earthquake Engineering in the Eastern United States*, J. E. Beavers Ed. (Ann Arbor Science, Ann Arbor, MI, 1981), pp. 25–51.
31. Damage area–magnitude comparison based on areas of MM VIII damage associated with the 1886 Charleston, South Carolina, and 1895 Charleston, Missouri, earthquakes. The minimum estimated MM VIII area for the m_b 6.7 (**M** 7.0) 1886 event is $\sim 3.8 \times 10^4$ km^2 [G. A. Bollinger, *U.S. Geol. Surv. Prof. Pap. 1028* (1977)]; the MM VIII area for the m_b 6.2 (**M** 6.4) 1895 event is $\sim 2 \times 10^4$ km^2 [I. N. Gupta and O. W. Nuttli, *Bull. Seismol. Soc. Am.* **66**, 743 (1976)]. On the basis of magnitude-intensity area relationships for California [T. R. Toppozada, *ibid.* **65**, 1223 (1975); J. F. Evernden, W. M. Kohler, G. D. Glow, *U.S. Geological Surv. Prof. Pap. 1223* (1981)], the equivalent magnitudes for California earthquakes with comparable damage areas are **M** 8.0 to 8.3.
32. The Poisson probability for a $M \geq 7$ earthquake in California is 0.85 for 30 years (11 events in 178 years). For southern California and the San Francisco Bay area, the 30-year Poisson probabilities are 0.57 and 0.54, respectively (five events in 178 years, and four events in 154 years, respectively). The 30-year time-dependent probabilities for **M** ≥ 7 earthquakes along the San Andreas fault system are 0.6 (*5*) in southern California and 0.67 for the San Francisco Bay area [Working Group on California Earthquake Probabilities, *U.S. Geol. Surv. Circular 1053* (1990)]. Probabilities for **M** ≥ 6 earthquakes in the eastern and central United States are shown in Table 1.
33. L. Seeber and J. G. Armbruster, in *The Geology of North America*, R. E. Sheridan and J. A. Grow, Eds. (Geological Society of America, Boulder, CO, 1988), vol. I-2, pp. 565–582.
34. We thank B. Bender, J. Dewey, K. Jacob, D. Perkins, M. Sibol, and R. Wheeler for stimulating discussion and advice.

30 May 1990; accepted 29 July 1990

No Excess of Homozygosity at Loci Used for DNA Fingerprinting

B. DEVLIN, NEIL RISCH,* KATHRYN ROEDER

Variable number of tandem repeat (VNTR) loci are extremely valuable for the forensic technique known as DNA fingerprinting because of their hypervariability. Nevertheless, the use of these loci in forensics has been controversial. One criticism of DNA fingerprinting is that the VNTR loci used for the "fingerprints" violate the assumption of Hardy-Weinberg equilibrium (H-W), making it difficult to calculate the probability of observing a genotype in the population. If one can assume H-W, the probability of observing the pair of alleles constituting an individual's genotype can be calculated by taking the product of the alleles' frequencies in the population and multiplying by two if the alleles are different. The evidence cited against assuming H-W is homozygote excess, which is presumed to be caused by an undetected mixture of two or more populations with limited interpopulational mating and distinct allele frequencies. For most VNTR loci, measurement error makes it impossible to test these claims by standard methods. The Lifecodes database of three VNTR loci used for forensics was used to show that the claimed excess of homozygotes is not necessarily real because many heterozygotes with similar allele sizes are misclassified as homozygotes. A simple test of H-W that takes such misclassifications into account was developed to test for an overall excess or dearth of heterozygotes in the sample (the complement of homozygote dearth or excess). The application of this test to the Lifecodes database revealed that there was no consistent evidence of violation of H-W for the Caucasian, black, or Hispanic populations.

THE DISCOVERY OF HYPERVARIABLE VNTR loci in human DNA in the early 1980s (*1*) was seen as a boon to a number of areas of scientific interest, in particular forensic science (*2, 3*). The loci are called hypervariable because, in any population, there are a very large number of alleles present at each locus (*3, 4*). Each allele of a VNTR locus is composed of a distinct sequence of base pairs, which one can detect indirectly by excising that region of the DNA with a restriction enzyme and estimating the length of the fragment by gel electrophoresis. Much of the allelic variation is generated by variation in the number of short, repeated sequences of base pairs linked in tandem in the core region of the locus (hence the acronym VNTR), which leads to fragment length variation on electrophoresis. It is the presence of a large number of alleles at these loci that renders the loci valuable as "fingerprints" (*5*), because it is quite likely that different individuals will have distinct genotypes at these loci. Nevertheless, the use of VNTR loci for forensics has led to controversy (*6, 7*). We focus on one of these controversies: whether H-W can be assumed for several VNTR loci used in forensics. Some researchers have asserted that H-W cannot be assumed because there is an excess of homozygotes at these loci (*6–8*).

H-W is an attribute of large, randomly mating populations. A population is in

B. Devlin, Division of Biostatistics, Department of Epidemiology and Public Health, Yale University, Post Office Box 3333, 60 College Street, New Haven, CT 06510.
N. Risch, Division of Biostatistics, Department of Epidemiology and Public Health, and Department of Human Genetics, Yale University, Post Office Box 3333, 60 College Street, New Haven, CT 06510.
K. Roeder, Department of Statistics, Yale University, Post Office Box 2179 Yale Station, New Haven, CT 06520.

*To whom correspondence should be addressed.

H-W for a particular locus if the probability of observing a genotype (a pair of alleles) in the population is the product of the probabilities of observing each of the pair of alleles in the population when the alleles are the same (for a homozygote) or twice this product when the alleles are distinct (for a heterozygote). Many forces could cause a population to deviate from the assumption of H-W: selection, inbreeding, phenotypic assortative mating, and population subdivision. For VNTR loci in human populations, it has been argued that population subdivision is the most important of these forces (*6, 8*). A population is considered subdivided if there are two or more groups within the population whose individuals experience limited intergroup mating. If the groups differ in their allele frequencies at a given locus and the sampling of the population ignores the substructure, an excess of homozygotes will be apparent in the sample even if the subpopulations are in H-W (*9, 10*). In this case, the violation of H-W is due to population subdivision, and the probability of observing a homozygote in the population is not the product of its alleles' frequencies. The degree of homozygote excess depends on the magnitude of allele frequency differences among such groups, as well as the admixture proportions.

Testing the assumption of H-W for a population is often straightforward. The usual tests involve examining the magnitude of the difference between the observed and expected number of each distinct genotype in a sample from the population (*11*). Testing H-W for VNTR loci is more complicated, however, because the length of a restriction fragment cannot be measured without substantial error. This error is often larger than the difference in the size of the alleles. Consequently, it is impossible to accurately assign discrete labels to alleles of a VNTR locus. Even if the discrete labels could be assigned, the large number of alleles and genotypes at VNTR loci seriously compromises the power of the usual tests. These difficulties have led to ad hoc methods of analysis of H-W (*6, 12*) and contradictory claims (*6, 13* versus *3, 12*). Here we have three objectives: (i) show that there is an apparent but not real excess of homozygotes at VNTR loci, making previous tests of H-W invalid; (ii) develop an appropriate method of testing H-W for VNTR loci; and (iii) demonstrate that there is no evidence that H-W is violated for three VNTR loci commonly used in forensics.

We analyze the data from three loci (D2S44, D14S13, and D17S79) generated by Lifecodes, Inc., for paternity testing and forensic inference. Each locus yields two restriction fragments per individual when cut with the restriction enzyme Pst I. There are three major sources of variation in the measured sizes of restriction fragments. The first source is intrinsic variation, due to differences in the number of tandem repeats, the size of the flanking region, or both. A second source of variation is measurement error, which can be large (*14*). The third source of variation results when an individual's pair of restriction fragments is visualized on x-ray film. Each fragment appears as a distinct band if the pair are substantially different in length (heterozygotes), but the bands are indistinguishable if the pair are the same length (homozygotes). For heterozygotes where the fragments are similar in size but not identical, however, only one band may be apparent because the distinct bands blur together or coalesce. These heterozygotes are indistinguishable from homozygotes; we call such a genotype a pseudohomozygote.

If coalescence were ignored, or if its consequences were not clearly understood, one might incorrectly infer that there is an excess of true homozygotes in the population. Indeed, the evidence usually cited to reject the assumption of H-W for VNTR loci is homozygote excess (*6–8*). Coalescence affects the kinds of phenotypes observed in the sample, resulting in an excess of apparent homozygotes and a dearth of "close heterozygotes" (heterozygotes with similar allele sizes). Therefore, if a substantial number of observations is affected, this should be readily detectable. Figure 1 demonstrates the presence of coalescence for a subset of the data for the D2S44 locus. Let x denote the length of the shorter allele and y denote the length of the longer allele. Note the relatively large number of observations on the line $y = x$; these are either homozygotes or pseudohomozygotes. Moreover, there are no heterozygotes in the interval adjacent to the line $y = x$. The difference between the fragment pair lengths of an observation in this interval would be small. Finally, the number of heterozygotes increases in the intervals of increasing distance from the line $y = x$. We call the absolute value of the fragment pair difference $\tau_j = |y_j - x_j|$, $j = 1, \ldots, n$ for the n individuals in the sample. If we plot τ_j against the mean fragment pair length $(x_j + y_j)/2$, coalescence is obvious in the data from all three loci (Fig. 2).

The difficulty of formulating a test of H-W for VNTR loci has led to some confusion and ad hoc methods of analysis. These ad hoc procedures (*6, 12*) treat the apparent homozygotes as if they were all true homozygotes and examine the magnitude of the difference between the observed and expected number of these homozygotes in the sample. These procedures have resulted in contradictory claims about H-W. It is illogical, however, to focus on the uncorrected number of apparent homozygotes in the sample because this number represents an indistinguishable mixture of true homozygotes and pseudohomozygotes. If these contaminated data are used without adjusting for coalescence, then any claim that there is no excess of homozygotes in the sample is false. Moreover, for the contaminated data, there will automatically be an excess of apparent homozygotes, making the claim of real homozygote excess irrelevant.

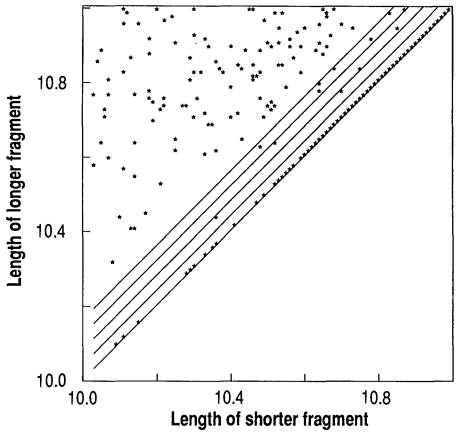

Fig. 1. Each point on the graph represents the phenotype of an individual. Points on the line $y = x$ represent either homozygotes or pseudohomozygotes. The data presented are a subset of the data from the D2S44 locus. The solid lines represent regions corresponding to adjacent intervals of $\tau(= |y - x|)$ of length 0.04 kb. Lengths are in kilobases.

We outline in the following paragraphs a simple test for an excess of homozygotes or heterozygotes formulated to use only the uncontaminated heterozygote data, which are unaffected by coalescence. Although the test uses a subset of the heterozygote data, it is informative nonetheless about homozygote excess because an overall excess of homozygotes must be accompanied by an overall dearth of heterozygotes, and vice versa. Moreover, most alternative hypotheses (such as assortative mating, inbreeding, and population subdivision) also affect the overall number of heterozygotes.

To determine the subset of the data that is uncontaminated, we first estimate the probability of coalescence $P(coal)$ of a pair of fragments as a function of the absolute value τ of the difference in their lengths. One can estimate $P(coal)$ by examining the observed versus expected number of heterozygotes found within adjacent intervals of τ. For instance, in Fig. 1, intervals of length 0.04 kilobase (kb) are depicted. The interval adjacent to the line $y = x$ is defined by $(0 < \tau \leq 0.04)$, and the interval adjacent to this first interval by $(0.04 < \tau \leq 0.08)$.

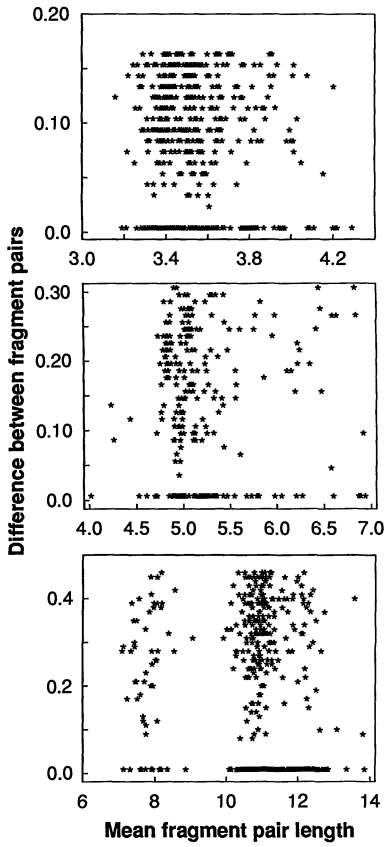

Fig. 2. The absolute value τ_j of the difference in fragment pair lengths (y axis) plotted against the mean of each fragment pair's length. The three figures present the data from the D17S79, D14S13, and D2S44 loci in descending order. A value of zero represents an apparent homozygote, and a value greater than zero represents a heterozygote. Note the missing observations for small values of $\tau > 0$. Lengths are in kilobases.

We index the adjacent intervals by k, for $k = 1, \ldots, q$, and define O_k to be the observed number of heterozygotes in interval k, H_k to be the expected number of heterozygotes in interval k, c_k to be the midpoint of interval k, and $2d$ to be the length of each interval (0.04 in the example above). One can obtain O_k by counting. If the cumulative distribution function F of fragment sizes were known, and if H-W could be assumed, then we could obtain the expected number H_k of heterozygotes in interval k by integrating over the distribution function. Specifically

$$H_k = n \int\int \mathbf{I}[(c_k - d) < |x - y| \leq (c_k + d)] \, dF(x)dF(y)$$

where $\mathbf{I}$ is 1 if the argument in brackets is satisfied and 0 otherwise. In words, H_k is the probability that two alleles fall in the kth interval of τ, assuming multiplicability of their probability densities (that is, H-W). Because F is unknown, we use the empirical distribution function to estimate it (*15*). Thus we estimate H_k by

Table 1. Logistic parameters and percentage of total variation explained by the model (R^2). Length is the average fragment length for the locus in kilobases.

Locus	β_0	β_1	R^2 (%)	Length
D2S44	−4.20	20.36	94	11.15
D17S79	−10.24	156.14	94	3.55
D14S13	−3.78	56.31	85	5.74

$$\frac{1}{4n} \sum_{i=1}^{2n} \sum_{j=1}^{2n} \mathbf{I}[(c_k - d) < |X_i - X_j| \leq (c_k + d)]$$

where X_i, $i = 1, \ldots, 2n$, denotes the $2n$ observed fragment lengths. Another way to view this calculation is that we obtain the expected number by creating all possible pairs from all $2n$ observations in the data and calculate n times the proportion of the $4n^2$ pairs that lie in the interval.

We then estimate *P(coal)* as a function of τ by fitting the points $[1 - (O_k/H_k)]$ versus c_k. Figure 3 suggests that the probability of coalescence *P(coal)* is a continuously decreasing sigmoidal function of τ. Therefore, we assume a logistic model with parameters β_0 and β_1 to relate *P(coal)* to τ (*16*). The fitted curve can be interpreted as the estimated values for *P(coal)* for the mean fragment pair length.

Given the fitted logistic curve, we can test for an excess or dearth of heterozygotes using only the data from heterozygotes for which coalescence is unlikely. We choose a bound b for τ such that the $P(coal|\tau > b)$ is small, say less than 0.01, and we then determine the observed (O) and expected (E) number of heterozygotes whose fragment pair length difference is greater than bound b. We determine O directly by counting, and we determine E from

$$E = \frac{1}{4n} \sum_{i=1}^{2n} \sum_{j=1}^{2n} \mathbf{I}(|X_i - X_j| > b)$$

Letting $p = E/n$, we formulate a test statistic

$$z = \frac{O - E}{\{n[p(1 - p) - 2v]\}^{1/2}}$$

where

$$v = \frac{1}{2n} \sum_{i=1}^{2n} \left[\frac{1}{2n} \sum_{j=1}^{2n} \mathbf{I}(|X_i - X_j| > b)\right]^2 - p^2$$

This statistic has approximately a standard normal distribution (*17*).

The logistic models fit the data for each of the three loci well (Table 1 and Fig. 3) (*18*). We chose a bound for *P(coal)* of 0.01, which translates into b = 0.434, 0.172, and 0.099 kb for the D2S44, D14S13, and D17S79 loci, respectively (*19*). The z statistic for each locus, reported by race, is given in Table 2. Clearly, there is no obvious violation of H-W. Without Bonferroni correction, there was a significant excess of homozygotes among Hispanics for locus D17S79 [but see (*19*)], but no excess at D2S44 or D14S13. Hence, these data provide little support for a consistent excess of homozygotes in any of the three populations tested, indicating that population subdivision has little impact on determining genotype frequencies. We do not claim, however, that our results demonstrate that there is no population subdivision. The black and especially the Hispanic populations may be heterogeneous, which would account for their slightly negative overall values in Table 2, but only large samples will be sufficient to definitively ascertain this heterogeneity. In addition, the assumption of H-W implies more than we have tested here; a general test would examine the expected and observed number of each genotype in the sample (though the problems noted earlier make the practicality of the general test question-

Table 2. Test of H-W by race and locus using only the uncontaminated heterozygote data. See text for calculation of bound, observed, and expected number of heterozygotes. A two-sided Bonferroni test with $\alpha = 0.05$ rejects the null hypothesis of H-W at $|z| \geq 2.8$. This test protects against spurious rejection of the null hypothesis due to multiple tests. Individual two-sided tests reject at $|z| \geq 1.96$. z^* is the value of the z-statistic when the bound is arbitrarily set at 0.0001 (that is, when the test is over all heterozygotes). N is the number of individuals in each sample.

N	Race	Observed	Expected	z	z^*
		D2S44			
1529	Caucasian	1165	1155.9	0.59	−84.5
693	Black	562	569.2	−0.76	−40.8
213	Hispanic	158	155.5	0.43	−30.1
		D17S79			
1399	Caucasian	1064	1047.9	1.06	−39.0
629	Black	519	516.9	0.24	−22.5
214	Hispanic	159	171.6	−2.25	−14.8
		D14S13			
701	Caucasian	579	575.2	0.43	−27.4
476	Black	421	432.4	−1.85	−19.7
122	Hispanic	105	108.3	−1.06	−11.9

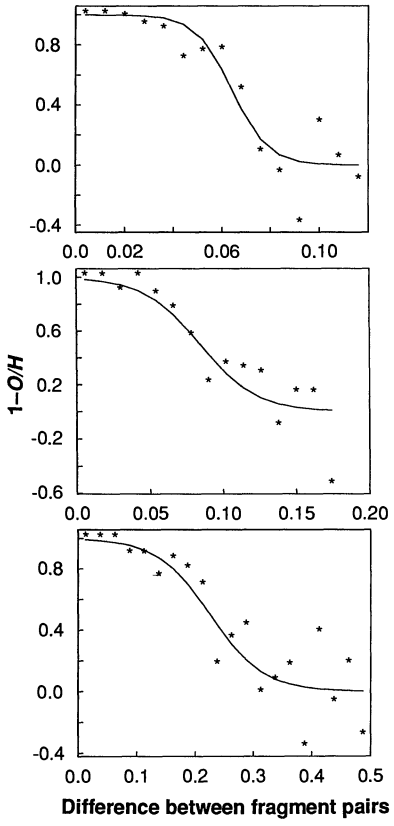

Fig. 3. The three figures present the data from the D17S79, D14S13, and D2S44 loci in descending order. The points in the figures represent the values of 1− the ratio of the observed (O) over expected (E) number of heterozygotes in each interval (y axis) versus the mean value of the interval (x axis). The curves are the fitted values of the logistic model. Lengths are in kilobases.

able). Despite these caveats, our test shows that the claim (*6*, p. 504) of "spectacular deviations from Hardy-Weinberg equilibrium" is totally unsupported. In fact, it would be surprising if the conclusion were otherwise, because conventional genetic markers such as blood group loci and enzyme loci rarely show deviations from H-W within human populations (*20*).

It is also clear why previous analyses have yielded evidence for homozygote excess. Ignoring coalescence by counting all apparent homozygotes as true homozygotes, we recalculated the z statistic (Table 2), calling it z^*. Ignoring coalescence results in enormously significant excesses of apparent homozygotes. Combining the data over races, there were 298, 113, and 329 apparent homozygotes in the samples for the D2S44, D14S13, and D17S79 loci, respectively, but only 9, 9, and 39 apparent homozygotes were expected. We caution that these values cannot be translated directly into the proportion of apparent homozygotes that were true homozygotes because of measurement error. A rigorous study (*21*) in which we estimated the allele frequencies of the D2S44 and D17S79 loci in the Caucasian population suggests that naively estimating the number of true homozygotes by setting the bound $b = 0.0001$ may underestimate the number of true homozygotes, and that one can obtain a superior estimate by choosing a bound equal to the size of the repeat of each locus and by calculating the expected number of fragment pairs that have the absolute value of the difference in their fragment lengths less than this bound. The expected value in this case should be composed predominantly of homozygotes and some close heterozygotes, giving an upper bound on the number of homozygotes. We obtained upper bound values of 51, 35, and 236 homozygotes or 17, 31, and 72% of the apparent homozygotes at the D2S44, D14S13, and D17S79 loci, respectively.

In forensic application, where the probability of a multilocus phenotype must be inferred, the usual approach is to assume multiplicability of allele frequencies, both within and across loci. The justification for multiplicability is random mating and H-W. It has been argued (*6, 8*) that a homozygote excess implies absence of H-W, nullifying the validity of multiplication across loci as well as within loci. Although the results presented here do not prove multiplicability across loci, they do prove that the arguments so far presented against it are incorrect.

REFERENCES AND NOTES

1. A. Wyman and R. White, *Proc. Natl. Acad. Sci. U.S.A.* 77, 6754 (1989).
2. R. Howlett, *Nature* **341**, 182 (1989); C. Norman, *Science* **246**, 1556 (1989).
3. I. Balazs, M. Baird, M. Clyne, E. Meade, *Am. J. Hum. Genet.* **44**, 182 (1989).
4. A. J. Jeffries *et al.*, *Nature* **314**, 67 (1985).
5. DNA fingerprinting can refer to either the band pattern of many restriction fragments from multiple unknown regions of the genome (*4*) or the patterns of restriction fragments from particular locations in the genome [Y. Nakamura *et al.*, *Science* **235**, 1616 (1987)]. We use the term only in the latter sense.
6. E. Lander, *Nature* **339**, 501 (1989).
7. S. Ford and W. C. Thompson, *The Sciences* **30**, 37 (1990); C. Norman, *Science* **246**, 1556 (1989).
8. J. Cohen, *Am. J. Hum. Genet.* **46**, 358 (1990).
9. S. Wahlund, *Hereditas* **11**, 65 (1928).
10. C. C. Li, *Ann. Hum. Genet.* **33**, 23 (1969).
11. See T. H. Emigh, *Biometrics* **36**, 627 (1980) for review; J. L. Hernandez and B. S. Weir, *ibid.* **45**, 53 (1989).
12. M. L. Balazs, M. L. Baird, K. McElfresh, J. Udey, in *Advances in Forensic Haemogenetics*, H. F. Polesky and W. R. Mayr, Eds. (Springer-Verlag, New York, 1990), vol. 3.
13. R. Lewin, *Science* **244**, 1033 (1989).
14. Lifecodes estimates their error to be $0.006L$, where L is the length of the fragment.
15. Using the empirical distribution function can be problematic when the number of alleles is small because the function is contaminated by coalescence. For a large number of alleles, however, the effect is negligible (B. Devlin, N. Risch, K. Roeder, unpublished data).
16. The form of the logistic model was

$$O_k/H_k = f(c_k,\beta_0,\beta_1) = \exp(\beta_0 + \beta_1 c_k)$$
$$[1 + \exp(\beta_0 + \beta_1 c_k]^{-1} \text{ and } P(coal) = 1 - f(\tau,\beta_0,\beta_1)$$

17. The observed count of heterozygotes in the uncontaminated region is $\Sigma_j \mathbf{I}(|X_j - Y_j| > b)$. Under H-W, $\mathbf{I}(|x - y| > b)$ is a Bernoulli random variable with probability of success estimated by p. Thus the variance of O is approximately $np(1 - p)$. To obtain the variance of E, note that $E \approx nW$ where

$$W = \frac{1}{2n(2n-1)} \sum_{i \neq j} \mathbf{I}(|X_i - X_j| > b)$$

Because the terms in the sum above are not independent, the variance cannot be calculated in a straightforward way. The U-statistic literature [R. J. Serfling, *Approximation Theorems of Mathematical Statistics* (Wiley, New York, 1980)] provides a convenient recipe for finding the approximate variance of statistics of the form of W. Applying these results we find that the variance of W is

$$2/n\{\int[\int \mathbf{I}(|x - y| > b)dF(x)]^2 dF(y) - p^2\}$$

Let v estimate the term in the braces; therefore the variance $VarE \approx 2nv$. Because O and E are calculated from indicators based on the sample fragments—$\mathbf{I}(|X_i - Y_i|)$ for O where X_i, Y_i are the paired fragments and $\mathbf{I}(|X_i - X_j|)$ for E where $X_1, \ldots, X_{2n}$ are the unpaired fragments—we need to account for the covariance between these terms. $Cov(O, E)$ is the sum of covariances across all pairings of indicators. For large n, $Cov(O, E)$ is dominated by the covariance between indicators for which one observation (X_i) matches; there are $8n^2 + o(n^2)$ of these indicators [where $o(n^2)$ denotes smaller order terms]. The variance of those terms where both observations (X_i and X_j) match can be ignored because there are only $o(n^2)$ of these terms; terms where no observations match have covariance zero. Those terms with one matching observation have covariance

$$Cov\,[\mathbf{I}(|X_1 - X_2| > b),\mathbf{I}(|X_1 - X_3| > b)] = \int[\mathbf{I}(|x - y| < b)dF(x)]^2 dF(y) - p^2$$

This happens to be the same term v as we observed in the variance calculation of the U-statistic. Therefore $Cov(O, E) \approx (1/4n)(8n^2 v) = 2nv$. Thus

$$Var(O - E) \approx np(1 - p) + 2nv - 4nv = np(1 - p) - 2nv$$

18. We fit the logistic model using only the Caucasian data set, which has the largest sample size. Gene frequencies among races may differ; hence the data sets cannot be combined over race. The results would be similar for blacks, however, and Hispanics were too few to analyze.
19. Reasonable changes in the width of the intervals produce very little change in the fit of the model or in the value of b. The effect on the z-statistic of choosing a bound greater than b is small, but choosing a bound less than b has substantial effects. All of our calculations ignored the fact that coalescence is also a function of mean fragment pair length; however, accounting for this fact leads to essentially the same results. Larger fragments are more likely to coalesce for a given difference in their fragment pair sizes than are smaller pairs because smaller fragments move further and separate more in the gel. To adjust for mean fragment pair length, let X_ℓ equal the mean value of the length of the fragments from locus ℓ (in our case, $\ell = 1, 2, 3$ for D2S44, D14S13, and D17S79) and ρ_ℓ equal the absolute value of the difference in fragment pair length such that $P(coal|\tau = \rho_\ell) = 0.5$ for locus ℓ, where $P(\ )$ is evaluated using the logistic model. If one assumes a simple linear relationship between fragment size and probability of coalescence, this relationship can be estimated as the slope α of ρ_ℓ against X_ℓ. To adjust the value of the bound b_ℓ for each mean fragment pair length $\bar{z}_j$, let the new bound be given by $\alpha(\bar{z}_j - \bar{X}_\ell) + b_\ell$. We estimate α to be 0.022. We then obtained the observed (O) and expected (E) number of heterozygotes in intervals with bounds defined by the above expression. A logistic model was fitted to the points $1 - O/H$ for each locus; the fit of these models was similar to that reported in Table 1 ($R^2 = 87$, 93, and 97% for the D2S44, D17S79, and D14S13 loci, respectively). The bound b changed noticeably only for the

D14S13 locus (b = 0.437, 0.104, and 0.203 for the D2S44, D17S79, and D14S13 loci, respectively). The z-statistics for the D2S44, D17S79, and D14S13 loci were 0.97, 0.99, and −0.14 for the Caucasian population, −0.72, 0.90, and −1.49 for the black population, and 0.09, −1.92, and −0.58 for the Hispanic population. Note that none of these z-statistics exceed the critical value for individual tests.

20. L. L. Cavalli-Sforza and W. F. Bodmer, *The Genetics of Human Populations* (Freeman, San Francisco, CA, 1971).
21. B. Devlin, N. Risch, K. Roeder, unpublished data.

22. We thank Lifecodes, Inc., for supplying the data, I. Balazs for discussion, and K. Kidd, R. Savage, and C. Schlichting for comments on a previous version of this manuscript. This work was supported by NIH grants GM39812 and CA45052 to N.R. and NSF grant DMS9001421 to K.R.

26 March 1990; accepted 22 June 1990

Detection of *Borrelia burgdorferi* DNA in Museum Specimens of *Ixodes dammini* Ticks

DAVID H. PERSING,* SAM R. TELFORD III, PAUL N. RYS, DEBORAH E. DODGE, THOMAS J. WHITE, STEPHEN E. MALAWISTA, ANDREW SPIELMAN

In order to investigate the potential for *Borrelia burgdorferi* infection before the recognition of Lyme disease as a clinical entity, the polymerase chain reaction (PCR) was used to examine museum specimens of *Ixodes dammini* (deer ticks) for the presence of spirochete-specific DNA sequences. One hundred and thirty-six archival tick specimens were obtained representing various continental U.S. locations; DNA sequences characteristic of modern day isolates of *B. burgdorferi* were detected in 13 1940s specimens from Montauk Point and Hither Hills, Long Island, New York. Five archival specimens of *Dermacentor variabilis* (dog tick) from the same collection and 118 *Ixodes* specimens from other endemic and nonendemic sites were negative. These data suggest that the appearance of the Lyme disease spirochete in suitable arthropod vectors preceded, by at least a generation, the formal recognition of this disease as a clinical entity in the United States.

ALTHOUGH SUBSETS OF THE DIverse clinical manifestations of Lyme disease were recorded in Europe early in this century (*1–4*), recognition of the disease as a distinct clinical entity did not occur until the mid-1970s (*5–7*). Today, Lyme disease is the most common tick-borne zoonosis in the United States, with more than 6000 human infections reported each year (*8*). Among the factors thought to contribute to the prevalence and spread of this disease are a burgeoning deer population, a constitutive rodent reservoir, and the indiscriminate feeding habits of the major tick vector (*9, 10*). Although some epidemiologic studies have suggested a relatively recent spread from earlier enzootic foci (*11*), it is not known how long *B. burgdorferi*, the spirochetal etiologic agent of Lyme disease (*12, 13*), has existed in its reservoir or vector populations.

The risk of a human aquiring Lyme disease is dependent on an interplay of microbial, environmental, and demographic factors. Ultimately, transmission is effected by nymphal ticks of the *Ixodes ricinus* complex (*14–18*). *Ixodes dammini* is the primary vector in enzootic areas of the northeastern United States, where its distribution correlates directly with that of cases of Lyme disease in humans and domestic animals (*14, 15*). The polymerase chain reaction (PCR) can be used (*19–21*) for the direct detection of *Borrelia* DNA and can detect spirochetes in field-collected specimens of *I. dammini* (*22*). The PCR-based method appears to be at least as sensitive as direct microscopic visualization after staining with a *Borrelia*-specific fluorescent antibody (DFA). PCR can also be applied to dried or alcohol-preserved specimens; the DFA method cannot be used on such specimens because of high levels of background fluorescence. This feature, coupled with success in the use of PCR to recover nucleic acids from archeological specimens (*23, 24*), prompted us to obtain and examine museum specimens of *I. dammini* to address the possible antiquity of Lyme disease in the United States.

We obtained 102 alcohol-preserved ticks from the Museum of Comparative Zoology, Cambridge, Massachusetts, archived between 1945 and 1951. These specimens were collected from Naushon Island and Martha's Vineyard in Massachusetts, and from various locations on Long Island, New York. Three additional groups of ticks from South Carolina and Florida were accessioned in the Museum of Comparative Zoology between 1925 and 1950 (*25*). Additional preserved specimens from other areas of the United States, dating from 1924 to 1950, were obtained from the Rocky Mountain Laboratories Acarine Collection (Smithsonian Institution). All specimens were stored in 70% alcohol and were cataloged according to the source and site of collection.

Our primary target for PCR detection of *Borrelia*-specific sequences was the gene encoding the major outer surface protein (OspA) of *B. burgdorferi* reference strain B31 (*26–28*). The *ospA*-specific primers (ospA2 and ospA4) were first used to test for the presence of *Borrelia* sequences in 15 specimens of nymphal *I. dammini* from Montauk Point, Long Island, New York, collected from a gray squirrel (*Sciuris carolinensis*) in the 1940s and in 5 contemporaneous nymphal ticks obtained from a vole (*Microtus pennsylvanicus*) on Naushon Island. Amplification products of the size expected for the *ospA* gene, 156 bp in length, were present in 3 of the 15 Montauk Point specimens by agarose gel electrophoresis (Fig. 1A). To verify the presence of *Borrelia*-specific DNA in these specimens, we used a second set of reagents designed to detect the flagellin gene (*fla*) of strain B31 (*29*). This primer pair produces a 200-bp genus-specific amplification product. *Borrelia* species of many types, including the relapsing fever agents *Borrelia hermsii* and *Borrelia recurrentis*, are detected with these primers, but not *Treponema*, *Leptospira*, or several exoflagellum-bearing organisms (*30*). The results of amplification of archival tick extracts with this primer pair are shown in Fig. 1B; the same three specimens previously identified with the *ospA* primer set also gave rise to an amplification product of the size expected for the *Borrelia fla* gene. Neither of the two target sequences was present at detectable concentrations in five specimens from Naushon Island dating to 1942. The identities of both the *ospA* and *fla* gene amplification products were confirmed by slot-blot hybridization with internal oligonucleotide probes (Fig. 2, a and b).

Analysis of the remaining specimens with the *ospA*- and *fla*-specific primer pairs indicated a relatively high prevalence of *Borrelia*-infected ticks from two eastern Long Island

D. H. Persing and P. N. Rys, Department of Laboratory Medicine, Yale University School of Medicine, New Haven, CT 06510.
S. R. Telford III and A. Spielman, Harvard School of Public Health, Boston, MA 02115.
D. E. Dodge and T. J. White, Roche Diagnostics Research, 1145 Atlantic Avenue, Alameda, CA 94501.
S. E. Malawista, Department of Internal Medicine, Yale University School of Medicine, New Haven, CT 06510.

*To whom correspondence should be addressed. Present address: Department of Laboratory Medicine and Pathology, and the Department of Internal Medicine, Mayo Foundation, Rochester, MN 55905.