Am. J. Hum. Genet. 51:992–997, 1992

Independence of VNTR Alleles Defined as Floating Bins

B. S. Weir

Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh

Summary

Data bases of VNTR fragments determined for Caucasians and blacks by Cellmark Diagnostics and Lifecodes Corporation are analyzed for independence of variants within and between loci. Floating bins are constructed around specific fragment lengths and are used to define discrete genotypes. Simple χ^2 test statistics for independence of bins within and between loci are described and applied to large sets of randomly generated four-locus profiles. The proportions of significant test statistics were about as expected under the hypotheses of independence, suggesting the absence of both Hardy-Weinberg and linkage disequilibrium. In any particular forensic application, however, these tests need to be performed on the fragments in question.

Introduction

One of the central issues in the forensic uses of VNTR-array data concerns the estimation of population frequencies of VNTR profiles that are found to match between crime-scene material and a suspect. The profiles generally consist of pairs of VNTR fragment lengths at three or four loci, and frequencies are obtained as products of the frequencies of individual fragment lengths. Use of this product rule requires that the individual lengths, within and between loci, be independent. One approach to demonstrating this independence with data collected by the Federal Bureau of Investigation (FBI) was given by Weir (1992), and here is given a demonstration for data collected by Cellmark Diagnostics (CD) and by Lifecodes Corporation (LC).

Forensic calculations are based on data bases consisting of estimated fragment lengths for VNTR profiles collected on people characterized by racial group. The CD information comes from blood-bank material, and the LC information comes from paternity tests (L. Forman and I. Balazs, personal communication). The data bases and loci analyzed for this paper are described in table 1.

Because of measurement error, estimated VNTR fragment lengths are essentially continuous variables, and continuous analyses have been presented (e.g., see Devlin et al. 1991; Berry et al. 1992). Current forensic practice in the United States, however, discretizes the lengths into "bins." The FBI uses a set of "fixed bins" that are defined in advance of any particular criminal investigation (Budowle et al. 1991), and this means that analyses can be performed on data bases of discrete genotypes, or "binotypes" (Weir 1992). Both CD and LC use "floating" bins, and there has not been a published demonstration of independence at this level for these data bases. The rationale for the floating bins is the same for both CD and LC, although the details differ. CD recognizes an ability to measure migration distances on an electrophoretic gel, to the nearest millimeter, and translates this uncertainty to a window of lengths around an estimated fragment length. LC works directly with precisions of estimated lengths and constructs a window of $\pm 1.8\%$ around estimates. Both companies have performed experiments to validate these rules (L. Forman and I. Balazs, personal communization).

A result of the floating-bin approach is that bins are defined anew for each profile, so that demonstrations of independence of binned fragment lengths cannot be done in advance of a particular criminal investigation as in the fixed-bin approach. It is a simple matter to

Received February 12, 1992; revision received July 23, 1992.

Address for correspondence and reprints: B. S. Weir, Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203.

© 1992 by The American Society of Human Genetics. All rights reserved.

0002-9297/92/5105-0008$02.00

Table 1

Sample Sizes and Ranges of Fragment Sizes in LC and CD Data Bases

Data Base and Locus	Caucasian			Black		
	n[a]	Lower Limit (kb)	Upper Limit (kb)	n[a]	Lower Limit (kb)	Upper Limit (kb)
LC:						
D2S44	2,249	6.87	20.94	831	6.73	19.52
D14S13	1,564	2.15	20.46	542	2.09	17.71
D17S79	2,238	2.06	5.59	827	2.42	5.80
D18S27	1,086	3.83	8.77	411	3.92	9.72
CD:						
D1S7	262	2.05	21.71	240	2.05	22.21
D7S21	264	3.46	14.75	238	2.05	10.91
D7S22	325	2.05	19.82	200	2.05	12.22
D12S11	294	3.21	10.43	223	2.80	19.93

[a] Number of individuals.

perform tests for independence, however, and the results for such tests are now given.

One-Locus Tests

When there is only one distinguishable fragment at a locus, whether because of coalescence of neighboring bands on a gel, or because of short fragments migrating off the end of the gel, or because of true homozygosity, it is the practice to estimate the genotype frequency by doubling the bin frequency for this fragment. This estimates an upper bound for the genotype frequency and avoids having to assume the product rule. It is therefore sufficient to confine attention to the case of locus X having distinguishable fragments falling in bins X_1 and X_2. When these two bins are different, the frequencies of all fragments in a data base having lengths in these two bins are $\tilde{p}_1$, $\tilde{p}_2$, and the frequency all fragments outside the two bins is $1-\tilde{p}_1-\tilde{p}_2$. This third collection of fragments may be termed "bin X_3." The estimated frequency of X_1 X_2 is $2\tilde{p}_1\tilde{p}_2$, providing that the two bins have independent frequencies, and independence can be tested very simply with the χ^2 test statistic given by Hernández and Weir (1989). Using the product rule instead of using the observed frequency of X_1X_2 pairs in the data base generally gives estimates with smaller SDs and avoids problems with X_1X_2 pairs not appearing in the data base.

To perform the test, three disequilibrium coefficients are determined, one for each heterozygote among the three bins X_1, X_2, and X_3:

$$\tilde{D}_{12} = \tilde{p}_1\tilde{p}_2 - \tilde{P}_{12}/2$$
$$\tilde{D}_{13} = \tilde{p}_1\tilde{p}_3 - \tilde{P}_{13}/2$$
$$\tilde{D}_{23} = \tilde{p}_2\tilde{p}_3 - \tilde{P}_{23}/2\,,$$

where $\tilde{P}_{ij}$ is the data-base frequency of heterozygotes X_iX_j and $\tilde{p}_3 = 1-\tilde{p}_1-\tilde{p}_2$. The test statistic is calculated as

$$X^2 = \frac{2n\tilde{D}_{12}^2}{\tilde{p}_1\tilde{p}_2[(1-\tilde{p}_1)(1-\tilde{p}_2)+\tilde{p}_1\tilde{p}_2]+(\tilde{p}_1^2\tilde{D}_{23}+\tilde{p}_2^2\tilde{D}_{13})}. \quad (1)$$

When the X_1X_2 heterozygote has a frequency consistent with Hardy-Weinberg equilibrium, this statistic has a χ^2 distribution with 1 df.

If a data base represents a population in which the product rule (i.e., Hardy-Weinberg hypothesis) holds at a locus, then the proportion of genotypes causing rejection of the Hardy-Weinberg hypothesis should be 5% if a 5% significance level is used. If the population does not have Hardy-Weinberg frequencies, then the proportion of rejections should be higher. Recall that the test is to be applied to VNTR profiles in criminal investigations and that these profiles are found independently of the data base. Accordingly, a set of 10,000 profiles was created by selecting fragment lengths independently and uniformly over the range of values present in the data base. At any locus, only those profiles with nonempty bins were used. The test for these profiles is conducted with data-base genotype frequencies (not simulated frequencies), as it is the

Table 2

Empirical Powers for Hardy-Weinberg Tests in LC and CD Data Bases

Data Base and Locus	Power for Hardy-Weinberg Tests		
	Caucasian Sample	Black Sample	Combined Sample
LC:			
D2S44	.07	.05	.08
D14S13	.04	.03	.02
D17S79	.05	.02	.07
D18S27	.05	.04	.04
CD:			
D1S7.....	.04	.02	.03
D7S21	.08	.05	.08
D7S22.....	.02	.03	.02
D12S11	.02	.07	.13

data base that is being tested for consistency with Hardy-Weinberg equilibrium. In table 2, the proportions of tests indicating significant departures from Hardy-Weinberg frequencies are shown. These empirical powers are close to the 5% significance level, suggesting that there is Hardy-Weinberg equilibrium in the corresponding populations.

A reviewer of an early version of this paper suggested that the set of random profiles be constructed by drawing fragments from the empirical distribution of fragment lengths found in the data base. This is a reasonable suggestion, and it leads to the same conclusions for the CD data base. For the LC data base, however, this sampling scheme runs into the problem of coalescence of neighboring bands that has been discussed by Devlin et al. (1990). For locus D17S79, in particular, most of the fragments drawn by sampling from the empirical distribution function lie in the interval 3.14–4.04 kb, and many pairs have lengths with a difference that falls within the range at which they would coalesce on a gel. Devlin et al. (1990) estimate this range as 0.099 kb for D17S79. Omitting fragment pairs whose lengths are closer together than the coalescence limit as given by Devlin et al. (1990) or as estimated by their methods gives empirical powers, for the Hardy-Weinberg tests, that are again close to those shown in table 2. This is consistent with the implicit suggestion of Devlin et al. (1990) that Hardy-Weinberg equilibrium will be found when very close pairs of fragments are ignored, and it is consistent with the approach of this paper—i.e., testing only heterozygotes. The trouble with this sampling strategy is that it uses the somewhat ad-hoc coalescence limits, which themselves depend on an assumption of Hardy-Weinberg equilibrium.

It should be stressed that table 2 is merely an attempt to describe the Hardy-Weinberg situation for the various loci in the CD and LC data bases. These tests are not *global* in the sense described by Weir (1992). In practice, tests should be performed on the particular pairs of fragments found in matching profiles. Only these *local* tests are relevant in the forensic setting.

Two-Locus Tests

The very low frequencies reported for VNTR profiles arise when several loci are used. When data are available from four loci, the product of the frequencies of all eight bins is taken, under the assumption that these frequencies are independent. A test of independence of pairs of alleles, one at each of two loci, in the case where phase is unknown was given by Weir (1979), and this linkage-disequilibrium test was applied to the CD and LC data bases. The empirical powers in table 3 result from 10,000 random four-locus profiles constructed by drawing sets of eight fragments independently and uniformly from their observed ranges. The tests use the data-base two-locus genotypic frequencies. As with the one-locus case, the empirical powers are close to the 5% significance level used.

To perform the test for bins X_i and Y_j at two loci, it is necessary to find the counts of the nine two-locus genotypic classes for all combinations of bins X_i, X_I, Y_j, and Y_J, where X_I is not-X_i and Y_J is not-Y_j. When these classes set out in a two-way array, the nine counts may be written as in table 4. Frequencies for the two bins of interest are estimated as

$$\tilde{p}_i = [2(n_1 + n_2 + n_3) + (n_4 + n_5 + n_6)]/2n$$
$$\tilde{p}_j = [2(n_1 + n_4 + n_7) + (n_2 + n_5 + n_8)]/2n ,$$

where n is the number of individuals in the data base. A coefficient of linkage disequilibrium, for genes on the same or different gametes, is estimated as

$$\tilde{\Delta}_{ij} = \frac{(2n_1 + n_2 + n_4 + n_5/2)}{n} - 2\tilde{p}_i\tilde{p}_j .$$

The presence of disequilibrium is tested for with the statistic

Table 3

Empirical Powers for Linkage-Disequilibrium Tests in LC and CD

Data Base, Loci, and Sample	Alleles i,j[a] 1,1	1,2	2,1	2,2
LC:				
D2S44 and D14S13:				
Caucasian	.05	.05	.04	.04
Black	.05	.05	.05	.05
Combined	.05	.05	.06	.04
D2S44 and D17S79:				
Caucasian	.05	.06	.05	.06
Black	.04	.05	.03	.04
Combined	.10	.09	.08	.08
D2S44 and D18S27:				
Caucasian	.05	.05	.04	.04
Black	.04	.06	.05	.06
Combined	.07	.07	.06	.06
D14S13 and D17S79:				
Caucasian	.04	.05	.04	.04
Black	.04	.06	.05	.05
Combined	.06	.06	.04	.05
D14S13 and D18S27:				
Caucasian	.04	.04	.04	.04
Black	.04	.05	.04	.04
Combined	.05	.05	.05	.05
D17S79 and D18S27:				
Caucasian	.04	.04	.04	.04
Black	.04	.04	.04	.06
Combined	.08	.07	.07	.07
CD:				
D1S7 and D7S21:				
Caucasian	.07	.07	.05	.05
Black	.05	.06	.05	.05
Combined	.06	.06	.05	.06
D1S7 and D7S22:				
Caucasian	.06	.06	.08	.08
Black	.04	.04	.04	.05
Combined	.07	.06	.08	.07
D1S7 and D12S11:				
Caucasian	.04	.04	.03	.03
Black	.05	.05	.06	.06
Combined	.04	.04	.05	.04
D7S21 and D7S22:				
Caucasian	.04	.04	.04	.04
Black	.05	.05	.05	.06
Combined	.05	.05	.04	.04
D7S21 and D12S11:				
Caucasian	.04	.04	.04	.04
Black	.04	.05	.04	.05
Combined	.06	.07	.07	.07
D7S22 and D12S11:				
Caucasian	.05	.05	.05	.05
Black	.05	.05	.04	.04
Combined	.06	.06	.05	.05

[a] Allele i at first locus and allele j at second locus.

Table 4

Two-Locus Genotypic Classes for All Combinations of X_i, X_I, Y_j, and Y_J

	Y_jY_j	Y_jY_J	Y_JY_J
X_iX_i	n_1	n_2	n_3
X_iX_I	n_4	n_5	n_6
X_IX_I	n_7	n_8	n_9

$$X^2 = \frac{n\tilde{\Delta}_{ij}^2}{[\tilde{p}_i(1-\tilde{p}_i)+\tilde{D}_i][\tilde{p}_j(1-\tilde{p}_j)+\tilde{D}_j]}$$

where the one-locus disequilibria are estimated as

$$\tilde{D}_i = [(n_1+n_2+n_3)/n] - \tilde{p}_i^2$$
$$\tilde{D}_j = [(n_1+n_4+n_7)/n] - \tilde{p}_j^2 .$$

When there is no disequilibrium, this test statistic has a χ^2 distribution with 1 df.

The testing procedure described here can be applied to all four combinations of pairs of bins between two loci.

The demonstration of independence of *pairs* of bins, within and between loci, suggests independence of *all* bins in a VNTR profile. For a four-locus profile with eight distinct and nonempty bins in the data base, the profile frequency in the population represented by the data base can be estimated at 16 times the product of the eight bin frequencies. When a locus has distinct fragments but the corresponding bins overlap, a larger bin can be defined as the union of the two, and the frequency of the one-locus genotype can be estimated as the square of the frequency of this larger bin.

If a test indicates that a pair of fragments do not have independent bin frequencies, then the product rule should not be applied to that pair. The actual pair frequency in the data base could be used, or an upper bound on the pair frequency could be obtained as the smaller of the two bin frequencies. If a fragment is found for which the bin is empty in the data base, then an upper bound on that genotype frequency can be constructed from standard multinomial theory (Nelson 1978). With $100(1-\alpha)\%$ confidence, the population frequency of a class that has an observed frequency of zero in a sample of size n can be said to be $<1-\alpha^{1/n}$.

Power of Tests

The tests for independence of binned frequencies are 1 df χ^2 and may have low power. In other words, in the population there may be disequilibrium that is not being detected by the tests. While it is important to characterize power, it is more relevant to ask what effect a low power could have on estimated multilocus profile frequencies.

For the Hardy-Weinberg test, only underestimation of heterozygote frequencies is of concern, since overestimation would be favorable to the suspect whose VNTR profile matches that of the crime-scene material. If it can be assumed that departures from Hardy-Weinberg equilibrium are small, as suggested by the small bin frequencies, then appeal can be made to the noncentral χ^2 distribution, to relate the power of a test to its significance level.

Specifically, the largest departure, from Hardy-Weinberg equilibrium, not expected to be detected with a power of 90% when a 5% significance level is used can be found. Such a value is found by solving for D_{12} when X^2 in equation (1) is set to 10.51. The bin frequencies and other disequilibria are taken as their observed values.

When this largest departure and the observed bin frequencies were used to construct heterozygote frequencies, four-locus heterozygote VNTR-profile frequencies were computed for 500 random profiles. This is an extremely conservative calculation, since it assumes that all four loci show heterozygote excess, whereas these VNTR systems tend to show single-band excesses (e.g., see Devlin et al. 1990). In each of the 500 cases, the worst-case frequency was greater than the frequency calculated under the assumption of independent bin frequencies, but the discrepancy was least for the largest frequencies (those most favorable to a suspect). For the LD Caucasian data base, for example, the expected and worst-case frequencies are shown in figure 1. Similar results hold for the other data bases.

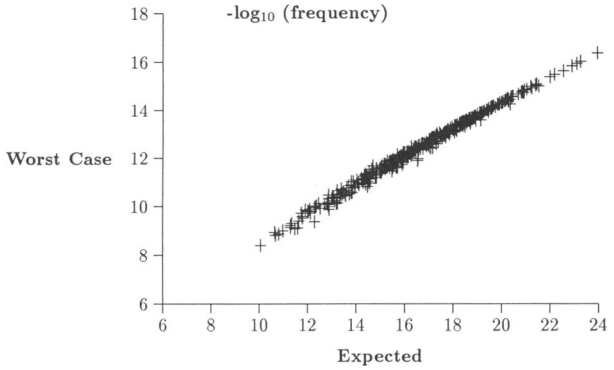

Figure 1 Maximum effect of undetected Hardy-Weinberg disequilibrium in LC Caucasian data base.

Population Substructuring

Estimates of the frequency of a particular VNTR profile in a population are valid only if the correct bin frequencies are used. One of the objections that has been raised (e.g., see Lewontin and Hartl 1991) is that the target population may be composed of a number of subpopulations, each with different bin frequencies. Frequencies calculated from a composite data base may not apply to a particular subpopulation, and such differences may lead to departures from Hardy-Weinberg or linkage equilibrium.

For the CD and LD data bases, a response to this criticism follows the argument given, by Weir (1992) and Chakraborty and Kidd (1991), for fixed-bin data bases. A composite CD or LC data base was constructed by amalgamating the Caucasian and black data bases in each case. Five hundred VNTR profiles were constructed by choosing two fragments at each of four loci, independently and uniformly, from the ranges in the composite data bases. Only those cases with eight nonempty bins were used. On the basis of bootstrap confidence intervals (e.g., see Weir 1992) it was found that many of the bins had different frequencies in the black versus the Caucasian data bases, but the columns denoted "Combined" in tables 2 and 3 show that the composite data bases exhibit little evidence of departures from Hardy-Weinberg or linkage equilibrium. If a data base is used that actually contains samples from two populations with the same degree of genetic divergence as is exhibited by U.S. blacks and Caucasians, this fact will not often be recognized by tests for independence of bin frequencies.

What is the effect of different bin frequencies? Population substructuring is important to the extent that it affects an estimated profile frequency. For the 500 random VNTR profiles generated for the comparison of Caucasian and black data bases, frequencies were estimated, by the product rule, for each data base separately. The results are shown in figure 2. The estimated frequency of a random four-locus VNTR profile is low, regardless of which data base is used. There are

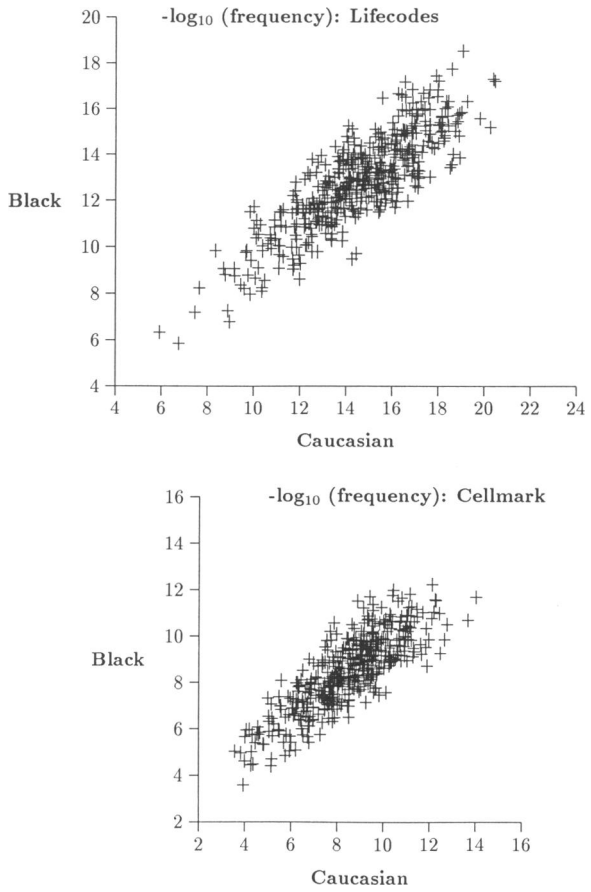

Figure 2 Values of $-\log_{10}$ (frequency) calculated with LC and CD data bases.

few values $>10^{-5}$ for CD or $>10^{-8}$ for LC (the difference between these numbers reflects the larger size of the LC data base). There is overall similarity between the estimates calculated from the black and Caucasian data bases.

Discussion

It would be desirable to determine whether data bases used with a floating-bin methodology have frequencies consistent with independence within and between loci. Since the bins are not defined in an absolute sense, such determinations cannot be made. Tests need to be conducted after the bins are defined around the particular fragments in question. By randomly generating many profiles, however, this study has demonstrated that as many significant tests are found as are expected under the hypothesis of independence. Furthermore, it has been demonstrated that whatever levels of dependence do exist are unlikely to have a meaningful impact on forensic calculations.

Acknowledgments

This investigation was supported in part by NIH grant GM45344. I received generous cooperation from Dr. Lisa Forman of Cellmark Diagnostics and Dr. Ivan Balazs of Lifecodes Corporation, and I had helpful discussions with Dr. Bruce Budowle of the FBI.

References

Berry DA, Evett IW, Pinchin R (1992) Statistical inference in crime investigations using deoxyribonucleic acid profiling. Appl Stat 41:499–531

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

Chakraborty R, Kidd KK (1991) The utility of DNA typing in forensic work. Science 254:1735–1739

Devlin B, Risch N, Roeder K (1990) No excess of homozygosity at loci used for DNA fingerprinting. Science 249: 1416–1420

——— (1991) Estimation of allele frequencies for VNTR loci. Am J Hum Genet 48:662–676

Hernández JL, Weir BS (1989) A disequilibrium approach to Hardy-Weinberg testing. Biometrics 45:53–70

Lewontin RC, Hartl DL (1991) Population genetics in forensic DNA typing. Science 254:1745–1750

Nelson SL (1978) Nomograph for samples having zero defectives. J Quality Tech 10:42–43

Weir BS (1979) Inferences about linkage disequilibrium. Biometrics 25:235–254

——— (1992) Independence of VNTR alleles defined as fixed bins. Genetics 130:873–887