netic mechanism. A cell-lethal trait expressed very early in embryonic development would be undetectable or perhaps would cause a "biochemical pregnancy." Activation later in embryonic life would still cause male lethality but would be less likely to cause complete skewing of X inactivation in multiple tissues in the heterozygous female. In view of this delicate balance in timing, we feel that the genes in question are most likely to be transcribed early in fetal development and to impart a growth disadvantage rather than being cell lethal. The size of the deletion mutation, however, is less important to when the miscarriage occurs: size is simply being used as a surrogate to the assumed importance of the deletion region and gene(s) contained in that region. In the end, this is all an exercise in mental gymnastics, since the characterization of the causative gene(s) will enlighten us all as to the true mechanism.

ERIC HOFFMAN, ELENA PEGORARO, AND MARK LANASA
Department of Molecular Genetics and Biochemistry
University of Pittsburgh School of Medicine
Pittsburgh

References

Migeon BR, Haisley-Royster C (1998) Familial skewed X inactivation and X-linked mutations: unbalanced X inactivation is a powerful means to ascertain X-linked genes that affect cell proliferation. Am J Hum Genet 62:1555–1557 (in this issue)

Pegoraro E, Whitaker J, Mowery-Rushton P, Surti U, Lanasa M, Hoffman EP (1997) Familial skewed X inactivation: a molecular trait associated with high spontaneous-abortion rate maps to Xq28. Am J Hum Genet 61:160–170

Address for correspondence and reprints: Dr. Eric Hoffman, Department of Molecular Genetics and Biochemistry, University of Pittsburgh School of Medicine, E1240 Biomedical Science Tower, Pittsburgh, PA 15261. E-mail: eric@hoffman.mgen.pitt.edu

0002-9297/98/6206-0040$02.00

Am. J. Hum. Genet. 62:1558–1560, 1998

Difficulties in the Estimation of Ethnic Affiliation

To the Editor:

Although I disagree with their results, I am indebted to Shriver et al. (1997) for reawakening my attention to the interesting but tricky subject of the inference of ethnic origin by DNA typing.

They have taken the novel and daunting approach of culling through a vast catalogue of candidate DNA loci to find those which are particularly discriminating. They list a battery of 10 loci, obtained mostly through such a search, which they claim will be effective in determining whether an unknown stain is of African American (AA) or European American (EA) origin. Specifically, they predict that only "0.01% [of individuals will] show log likelihoods <3.0" favoring one origin over the other (Shriver et al. 1997, p. 962). If a prior probability of 50% is assumed for each alternative, this implies the posterior ability to make a correct guess at least 999 times in 1,000. Categorizing Americans as black or white by interviewing them probably does not achieve such a high level of reproducibility, so it seems natural to review with care the basis for such a claim.

I am concerned that the claim rests on serious flaws in statistical methodology. My reanalysis shows that the estimates of efficacy for race determination are significantly overstated because of bias in the algorithm for prediction of likelihood ratios. This is true even for the handful of loci from the literature the authors say that they were able to verify as useful. As for the majority of the recommended loci—those discovered by surveying the catalogue—there is an additional bias that is probably even more serious. I shall discuss a computer simulation that shows that the apparent good performance of the culled loci may be completely illusionary, explainable as mere sampling variation.

These concerns can be conveniently discussed and illustrated in the context of D7S657, the most highly rated of the loci found by the statistical survey. Figure 2 of the Shriver et al. article reveals enough information to allow a check of the calculations for this locus, calculations that assert a typical likelihood ratio of $r = 19$ ($\log_{10} r = 1.276$). I will argue that that number is inflated both by algorithmic errors and by sampling bias. A more realistic likelihood ratio estimation algorithm will reduce the value from 19 to ~8 ($\log_{10} r = 0.9$), and consideration of sampling bias will show that a value of 2.5 ($\log_{10} r = 0.4$) or even less is plausible and consistent with the reported results.

Let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ be the allele frequencies at some locus in populations A and B, for alleles 1, 2, . . . , respectively. Then for an allele whose true origin is A and for allele frequencies that are known,

$$\log_{10} r_{AB} = \sum a_i \log_{10}(a_i/b_i) \qquad (1)$$

is the expected value of the logarithm of the likelihood ratio that the origin is the reference population A rather than the target population B. The formulas in the article by Shriver et al. are equivalent, except that their notation refers to genotypes rather than to single alleles (which explains why their formula has factors of "$\frac{1}{2}$," whereas mine does not), and they formulate a statistic that is

symmetric with respect to A and B. Consequently, they define the "log-likelihood level," which in my notation is $\log_{10} r$ where $r = r_{AB} r_{BA}$. For r itself, the geometric expected value of the likelihood ratio, I use the phrase "typical" likelihood ratio.

In practice, it is necessary to deal with sample frequencies rather than with true frequencies. Therein lies the rub.

An obvious problem with (1) is possible division by zero when some allele is not observed in the target sample. The best strategy in this situation is probably the one that Shriver et al. (1997, p. 958) chose: "an allele not found in a sample is assumed to be the next allele to be observed"; that is, assume a minimum frequency of $b_i = 1/(2n + 1)$, where n is the number of people in the target sample. Apart from this proviso, the authors make computation (1) with sample frequencies $\hat{a}_i$, $\hat{b}_i$, as if they were population frequencies (the "hat" [^] crowning a variable indicates and emphasizes that it is a "statistic" computed from a population sample, and thus a mere estimate of the hatless "parameter"). If the only purpose is a rough comparison of loci, this approach could be accepted uncritically. However, since the likelihood ratio statistic is to be interpreted as what it is—as predicting the likelihood ratio performance that can be expected in practice—then it must be an unbiased estimate. For example, it must satisfy the validation criterion (Brenner 1997) that, if one compares two identical populations, then the likelihood ratio statistic should not imply that people will be more likely to come from one than from the other. Imagine collecting two n-person samples S and T from the same population A. The sample frequencies will, of course, by random chance, differ somewhat, so a likelihood ratio analysis of the samples, even a correct one, will sometimes suggest that A can be distinguished from A. However, when an average is taken over all samples S and T, there should be no expected distinction. The average log likelihood ratio should be zero. As an experiment, I posited a population with allele frequencies approximately like the D7S657 AA sample frequencies. Testing the Shriver et al. formulas on repeated computer samples S and T from this population, I found that $\log_{10}\hat{r}$ averaged ~0.11 ($\hat{r} = 1.3$) for distinguishing the simulated AA population from itself. In other words, there was a bias of ~30%.

A procedure that does seem to survive the validation test is the one used by Erikson and Svensmark (1994). They added one to the target sample count for every allele—not just for the unseen ones as do Shriver et al. When this method is applied to the D7S657 data, the likelihood ratio drops to $\hat{r} = 8$ ($\log_{10}\hat{r} = 0.9$).

The reader may wonder why the discrepancy between 19 and 8 is much larger than the 30% accounted for by the above validation test. It turns out that most of the discrepancy has a less subtle explanation. Those values that can be checked in Shriver et al.'s (1997) table 1 are consistent with misinterpreting n as the sample size in alleles rather than in people, suggesting a solecism in their computer implementation (a bug). Correcting the arithmetic would give $\log_{10}\hat{r} = 1.09$ (so $\hat{r} = 12$).

The most powerful locus in the Shriver et al. article is FY-null (Duffy blank), for which they give $\log_{10}\hat{r} = 1.858$ ($\hat{r} = 72$) for distinguishing AA and EA. The above bug is not an issue here since there were no zero-count alleles; nor is sampling bias an issue since FY-null is among the loci from the literature rather than from the statistical survey. Nonetheless, substituting the Erikson and Svensmark procedure reduces $\hat{r}$ to 39 ($\log_{10}\hat{r} = 1.59$), and that is probably a fairer guess of the efficacy of this locus based on the data given. It is still a very discriminating locus. In ~96% of the cases in which an unknown stain donor is African American, this locus alone will answer the question of ethnic origin. But a difficult question about allelic association: In estimating how many of the remaining 4% of such cases will be resolved by other loci, is it correct to use the overall AA allele frequencies, which after all come mostly from people who have FY-null?

It may be of interest to compare the r values discussed above with values for typical forensic loci, not intentionally selected for their ethnic-discrimination potential. Data from the Office of the Chief Medical Examiner of New York City (personal communication) comparing AA and EA samples ($n = 118$ and $n = 107$) in the tetrameric loci F13A, TH01, FES/FPS, and VWA, give values $\hat{r} = 2.56$, 1.56, 1.43, and 1.17 ($\log_{10}\hat{r} \leqslant 0.4$). With somewhat more care and difficulty, the same sort of evaluation can be made for RFLP loci. Six of them, in common use by United States law-enforcement and paternity laboratories for identification, average $r > 2$ ($\log_{10} r > 0.3$) per locus (Brenner 1997). Incidentally, whereas Shriver et al. (1997, p. 957) say in their paper that "most [DNA markers] offer little power to distinguish ethnicity," a handful of independent markers with $r = 2$ provide quite useful power (Brenner 1997), certainly better than "the best racial estimates [, which] are achieved" by bone and skull measurements, giving 80%–90% correct categorization as cited in Shriver et al.'s (1997 , p. 958) paper. Further, with a computationally more sophisticated approach (Evett et al. 1992; Brenner 1997), the RFLP efficacy increases to about $r = 3$ ($\log_{10} r = 0.5$) per locus. Combining a handful of loci of such power is sufficient to decide most cases with confidence, but the predicted distribution is such that 5%–10% of cases remain elusively ambiguous.

So compared with the standard repertoire of forensic loci, the claims of Shriver et al., even trimmed back a few orders of magnitude by the arguments that I have made above, would still be impressive. By my rough

estimates, the top 10 loci would still rigidly categorize Americans as black or white, to an implausible extent.

My remaining concern, the most vexing and possibly the most telling, is sampling bias. Of the 20 high-performing loci in Shriver et al.'s table 1 for discrimination between AA and EA, 17 were obtained by canvassing >1,000 loci. To be more precise, what they canvassed is >1,000 *pairs of samples* and sometimes rather small samples (e.g., $n = 21$ people). This suggests the possibility that most of the "high performers" are really ordinary performers with an atypically lucky sample.

How much can be explained by luck depends on the sampling distribution of the likelihood level statistic, $\hat{r}$, which I have investigated with a Monte Carlo computer experiment. Each experiment begins with 1,000 simulated loci whose AA and EA allele frequencies are assigned according to one or another of the New York data mentioned above, so $1.17 \leqslant r \leqslant 2.56$ ($0.08 \leqslant \log_{10} r \leqslant 0.4$) for each simulated locus. For each of these loci, a 21-person sample and a 22-person sample (mimicking the D7S657 sample sizes) are randomly selected according to the assumed frequencies, and the statistic $\hat{r}$ is computed from the two samples. To simplify the comparison with Shriver et al.'s table 1, I used the same (albeit incorrect, as per above discussion) formulas as were used for that table.

The 17 largest $\hat{r}$ values from such a 1,000-locus experiment are similar to the values for the 17 canvassed loci (out of 20 total) in the AA/EA column of table 1. The largest value is sometimes a little larger, sometimes a little smaller, than $\hat{r} = 19$ ($\log_{10}\hat{r} = 1.276$) of D7S657. The 17th largest simulated $\hat{r} \approx 5$ ($\log_{10}\hat{r} \approx 0.7$)—easily comparable to $\hat{r} = 3$ ($\log_{10}\hat{r} = 0.498$) in table 1. One might say that what the computer experiment screens is not nature but sampling variation. It lists loci with merely ordinary ethnic-discrimination power, but with extraordinary statistics. From among 1,000 loci, one could similarly find a set of 10 loci that differentiate the 9-year-old children from the 10-year-olds in the local playground. In the phrase of one of the referees of this letter, the process has the potential to create the appearance of signal where there is only noise.

Is the sieving procedure of Shriver et al. any different from the computer experiment? The bias problem would be mitigated if their sample sizes were mostly larger, or if some loci were screened twice. This may have been done to some extent; the description in the Shriver et al. paper is not explicit. Also, there is of course a tendency for the better loci to achieve a better score. But as I have shown, there is a strong countervailing tendency that the list of top scores will be dominated by scores that are particularly biased. Therefore, I do not believe that their conclusion—namely, that they have found "a set of genetic markers that would allow the confident determination of ethnicity" (Shriver et al. 1997, p. 962)—is likely to be correct.

CHARLES H. BRENNER

Electronic-Database Information

Brenner CH (1997) http://www.ccnet.com/~cbrenner/race.htm

References

Brenner CH (1997) Probable race of a stain donor. In: Proceedings from the Seventh International Symposium on Human Identification 1996. Promega, Madison, pp 48–52

Erikson B, Svensmark O (1994) DNA polymorphism in Greenland. Int J Legal Med 106:254–257

Evett IW, Pinchin R, Buffery C (1992) An investigation of the feasibility of inferring ethnic origin from DNA profiles. J Forensic Sci Soc 32:301–306

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60: 957–964

Address for correspondence and reprints: Dr. Charles H. Brenner, 2486 Hilgard Avenue, Berkeley, CA 94709. E-mail: cbrenner@ccnet.com

0002-9297/98/6206-0041$02.00

Am. J. Hum. Genet. 62:1560–1561, 1998

Reply to Brenner

To the Editor:

In response to the letter by Dr. Brenner (1998 [in this issue]), there are a number of issues open for discussion with regard to both our previously published article (Shriver et al. 1997) and, more generally, methods for estimation of biological ancestry. Dr. Brenner has identified some specific concerns with regard to our methods and results, which we address below. However, we remain confident of the main conclusions of our study: (1) the reliable estimation of ethnic affiliation by use of population-specific alleles (PSAs) is possible; and (2) many of the loci we identified will be useful markers for this effort.

We have examined the computer program that was used to calculate average single-locus log-likelihood levels and have found that Dr. Brenner is correct in his determination that alleles that were not observed were assigned a frequency of $1/(4n + 1)$, instead of $1/(2n + 1)$, where n is the number of individuals in the sample. The effect of this error was to inflate the average single-locus and multilocus log-likelihood estimates, to a small degree. Since the same program was used to screen all the allele-frequency data sets, it is reasonable to conclude