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

that the 40 loci with the highest log-likelihood levels, which we presented in tables 1 and 2 of our article (Shriver et al. 1997), are still good candidates for high performers among the loci tested.

Dr. Brenner is correct to recognize that our method for determining average single-locus log-likelihood ratios (LLRs) and multilocus ethnic-affiliation estimates is appropriate only when accurate allele-frequency data are available. We expect that, in the determination of biological ancestry, care will be taken to determine with precision the allele frequencies of potential contributing populations. If accurate allele frequencies are available (e.g., $n > 200$ individuals), no adjustment of the formula we presented will be needed. In cases for which frequency data are available only from small samples, the addition of one to the total allele count for each allele is a reasonable adjustment.

Dr. Brenner concludes that the differences in allele frequency that we observed between loci were largely due to bias resulting from small sample size. He bases this conclusion on a computer simulation in which he evidently resampled 1,000 × from frequency data on four short tandem-repeat identity markers. He then compared his results with the data in table 1 of our article (Shriver et al. 1997). We have two concerns with this approach. First, the 17 microsatellite PSAs that we presented in table 1 were culled from ~350 loci (1,000 loci/population combinations were tested in the work that we reported). Second, the range of variation in the frequency differential used in Dr. Brenner's model was very limited and, with only four loci (LLR of .08–.4), could not have reflected naturally observed levels of variation in the allele-frequency differential. We are well aware of the bias resulting from small sample sizes, which is why we presented a list of 20 loci in table 1 and not just the best 10. In fact, we stated, "It should be noted that the markers on this list need to be typed in larger samples from different parts of the country, both to have more accurate allele-frequency estimates and to identify the most efficient set for EAE [ethnic-affiliation estimation]" (Shriver et al. 1997, p. 963). Recently, we typed nine dimorphic autosomal PSAs in large samples from >20 ethnographically defined populations, including 12 African-American population samples, and indeed found these markers to be useful for the estimation of ethnic affiliation and admixture (Parra et al. 1997; E. J. Parra, A. Marcini, L. Jin, J. Akey, M. Batzer, R. Cooper, T. Forrester, et al., unpublished data). Overall and in view of Dr. Brenner's concerns, we still feel that this is a viable approach for the estimation of the biological ancestry of a person and that we have provided an important list of putative PSAs for this purpose.

Finally, in responding to Dr. Brenner's comments, we would like to suggest an alternative phrase that more accurately describes what is being estimated by means of the markers and methods that we, Dr. Brenner, and others have described. Ethnicity is a term that directly refers to the culture of a person or people and that encompasses their language, traditions, and national identity. Ethnicity is often related to biological ancestry but not always. In the United States, awkward terms that combine both ethnicity and biological ancestry are sometimes used—for example, "non-Hispanic whites," "black Hispanics," and "non-Hispanic blacks." Modern populations are highly complex, and the classification of genetic differences among individuals and populations is a potentially sensitive issue. We therefore propose and intend to use the term "estimation of biological ancestry," rather than "ethnic-affiliation estimation," to describe the methods that we have presented.

MARK D. SHRIVER,[1] MICHAEL W. SMITH,[2] AND LI JIN[3]
[1]*Department of Human Genetics, Allegheny University of the Health Sciences, Pittsburgh;*
[2]*National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, MD; and*
[3]*Human Genetics Centers, University of Texas, Houston*

References

Brenner CH (1998) Difficulties in the estimation of ethnic affiliation. Am J Hum Genet 62:1558–1560 (in this issue)

Parra E, Marcini A, Akey J, Ferrell RE, Shriver MD (1997) A systematic study of African-American admixture using population-specific alleles. Am J Hum Genet Suppl 61:A17

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. Mark D. Shriver, Department of Human Genetics, Allegheny University of the Health Sciences, 3290 William Pitt Way, Building B4, Room 125, Pittsburgh, PA 15212-4772. E-mail: mshriver@phg.auhs.edu

0002-9297/98/6206-0042$02.00

Am. J. Hum. Genet. 62:1561–1562, 1998

Discriminating between True and False-Positive Peaks in a Genomewide Linkage Scan, by Use of the Peak Length

To the Editor:

A standard method to map disease-susceptibility loci consists of collecting n affected sib pairs and their parents, genotyping them for a dense set of genetic markers, and counting, at each marker locus t, the number, X_t, of parental alleles shared identical by descent (IBD). According to current statistical practice (e.g., see Feingold