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

et al. 1993; Lander and Kruglyak 1995), only the height of the peak (i.e., $\max_t X_t$) is used to decide whether there is sufficient evidence in favor of linkage. Recently, Terwilliger et al. (1997) published in the *Journal* a paper in which they challenged the claim that "there is no way to distinguish between small peaks that represent weak true positives and peaks of the same height arising from random fluctuations" (Lander and Kruglyak 1995, p. 244). By relying on some deeper results of the theory of stochastic processes, Terwilliger et al. (1997) showed that true positive peaks are expected to be longer than false-positive peaks. The purpose of this letter is to explain and support their result by presenting an extremely simplified situation in which elementary argumentation is sufficient to obtain the same conclusion.

Let M_1 and M_2 denote two different but linked marker loci separated by a genetic distance of $\theta \in (0,\frac{1}{2})$. The data consist of a single affected sib pair and only one parent. Both markers are assumed to be completely polymorphic. Thus, it is possible to decide whether, from this parent, the sib pair has inherited the same allele ($X_i = 1$) or has not inherited the same allele ($X_i = 0$), at marker locus i ($i = 1,2$). Let $p = P(X_1 = 1)$ denote the probability that the sib pair share an allele IBD at M_1. If M_1 is unlinked to the disease, then $p = \frac{1}{2}$; if M_1 is a disease locus, then $p > \frac{1}{2}$. In both cases, the conditional probabilities for the IBD score at the second marker locus, given the IBD score at the first marker locus, depend only on the genetic distance θ between M_1 and M_2 —that is, $P(X_2 = 1 \mid X_1 = 1) = P(X_2 = 0 \mid X_1 = 0) = \theta^2 + (1 - \theta)^2 =: \Psi$ and $P(X_2 = 1 \mid X_1 = 0) = P(X_2 = 0 \mid X_1 = 1) = 1 - \Psi$. The joint probability distribution of (X_1,X_2) is given in table 1.

Now the terms "peak" and "length of a peak" have to be defined. A peak occurs at marker locus i ($i \in \{1,2\}$) if $X_i = 1$. Given that there is a peak at marker locus i, the length of this peak is either 2 (if $X_{3-i} = 1$) or 1 (if $X_{3-i} = 0$). The following three conclusions are evident from table 1:

1. Given a peak at M_1, the length of this peak is 2, with probability Ψ. Thus, this probability does not depend on p.

2. However, given a peak at M_2, the length of this peak is 2, with probability $\Psi p/\{\Psi p + [(1 - \Psi)(1 - p)]\}$. Since $\Psi < 1$, this is a strictly increasing function in p, for $p \in [\frac{1}{2},1]$.

3. Given a peak "somewhere in the genome" (i.e., at M_1 and/or M_2), the length of this peak is 2, with probability $\Psi p/\{1 - [\Psi(1 - p)]\}$. Since the value for this expression is strictly increasing with p, this shows that true peaks ($p > \frac{1}{2}$) are expected to be longer than false-positive peaks ($p = \frac{1}{2}$).

Table 1

Joint Probability Distribution of (X_1,X_2)

l	m	$P(X_1 = l,X_2 = m)$
1	1	$\Psi \cdot p$
1	0	$(1 - \Psi) \cdot p$
0	1	$(1 - \Psi) \cdot (1 - p)$
0	0	$\Psi \cdot (1 - p)$

This example can be extended to a consideration of statistical testing. Let α denote an arbitrary but fixed value, with $\alpha \leqslant \Psi/2$. Since $P_{p=\frac{1}{2}}(\max_{t\in\{1,2\}} X_t = 1) = 1 - (\Psi/2)$, a (randomized) level α test based on the test statistic $\max_{t\in\{1,2\}} X_t$ is obtained by rejection of the null hypothesis of no linkage, with probability $\gamma_1 := \alpha/[1 - (\Psi/2)]$. The power of this test is $\alpha \cdot \{1 - [\Psi(1 - p)]\}/[1 - (\Psi/2)]$. Alternatively, a test based on the length of a peak can be constructed in the following way: for $p = \frac{1}{2}$, the probability is $\Psi/2$ that a peak of length 2 occurs. Thus, a randomized level α test based on the length of a peak is obtained by rejection of the null hypothesis, with probability $\gamma_2 := 2\alpha/\Psi$. The power of this test is $\alpha \cdot 2 \cdot p$. Since $2p > \{1 - [\Psi(1 - p)]\}/[1 - (\Psi/2)]$ for $p > \frac{1}{2}$, the second test, which is based on the length of a peak, is more powerful than the test based solely on the height of a peak. For a more realistic and relevant situation than the one considered in the present letter, it has to be determined how both height and length of a peak can be combined in a test for linkage. However, the observation described by Terwilliger et al. (1997) may prove very useful to increase the power for detection of disease-susceptibility loci by genomewide linkage scans.

MICHAEL KNAPP

Institute for Medical Statistics
University of Bonn
Bonn

References

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–251

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE (1997) True and false positive peaks in genomewide scans: application of length-biased sampling to linkage mapping. Am J Hum Genet 61:430–438

Address for correspondence and reprints: Dr. Michael Knapp, Institute for Medical Statistics, University of Bonn, Sigmund-Freud-Strasse 25, D-53105 Bonn, Germany. E-mail: umt70e@ibm.rhrz.uni-bonn.de

0002-9297/98/6206-0043$02.00