Am. J. Hum. Genet. 41:948–956, 1987

Brief Communication

Neutrality Tests of Highly Polymorphic Restriction-Fragment-Length Polymorphisms

ANDREW G. CLARK

Department of Biology, Pennsylvania State University, University Park

SUMMARY

The allele frequency data of Baird et al. were tested using Ewens-Watterson sampling theory for goodness of fit to the infinite-alleles model of neutral evolution. Although probes of both the *HRAS-1* and *D14S1* loci identify highly diverse restriction-fragment-length polymorphisms, the observed values of gene identity (F) and the common allele frequency (C) are not significantly different from the neutral expectation. Allele frequency distributions show a tendency toward a deficit in diversity for *HRAS-1* and a slight excess diversity for *D14S1*. The direction of these departures is consistent with potential selective effects of the *Harvey-ras* oncogene and hitchhiking of the *D14S1* locus to closely linked immunoglobulin genes. Direct χ^2-tests of goodness of fit of the observed and expected allele frequency distributions reveal significant departures in the caucasoid and Hispanic *HRAS-1* distributions but not in any of the other tests.

INTRODUCTION

Baird et al. (1986) recently showed that the *HRAS-1* and *D14S1* loci exhibit extensive sequence variation in humans, and they emphasize the utility of these markers for paternity testing and forensic genetics. Such genetically diverse loci are also very useful for testing hypotheses concerning the evolutionary

Received January 23, 1987; revision received May 5, 1987.

Address for correspondence and reprints: Dr. Andrew G. Clark, Department of Biology, 208 Mueller Laboratory, Pennsylvania State University, University Park, PA 16802; phone (814) 863-3891.

history of the locus. With suitable sampling, the degree of population mixing through migration and gene flow can be determined (Wright 1978). The goodness of fit to expected allele frequency distributions can also reveal whether specific evolutionary models acceptably fit the data. In addition to the evolutionary insights that such tests might reveal, paternity testing and forensic genetics require that the population genetics of the loci be well characterized. As Baird et al. (1986) are careful to verify, the loci exhibit proper Mendelian segregation and satisfy the Hardy-Weinberg proportions, but the extraordinary allelic diversity raises suspicions concerning the stability (mutation rate) of these regions.

One particular model that provides a test of the past effects of natural selection is the infinite-alleles model. In a finite population with a constant rate of mutation to an infinite series of novel alleles, there will be a steady-state configuration of allele frequencies representing a balance between the mutational gain of new alleles and the stochastic loss of alleles by drift. A sample from a population may deviate by having an excessive frequency of the most common allele (and many very rare alleles) or by having an excessively even distribution. Sampling theory developed by Ewens (1972, 1979) and Watterson (1978*a*, 1978*b*) enables one to test the goodness of fit of observed and expected homozygosity.

The genetic variation at the *HRAS-1* and *D14S1* loci are particularly interesting because of their sequence structure and opportunity for hitchhiking (Thomson 1977). The *HRAS-1* locus itself may reveal evidence for selection due to the function of the *Harvey-ras* oncogene (Capon et al. 1983). This locus is also tightly linked to the insulin gene on chromosome 11. There appears to be a tandemly repeated 28-bp sequence within *HRAS-1* (deMartinville et al. 1983; Gerald and Grzeschik 1984), suggesting the possibility for unequal crossing-over as a means for generating diversity. The *D14S1* gene also has a tandemly repeated sequence (Wyman and White 1980). *D14S1* polymorphism may be affected by its linkage to the heavy-chain g1 immunoglobulin (Balazs et al. 1982), a genetic region whose diversity may well be partly accounted for by balancing selection, as in the HLA glycoprotein loci (Klitz et al. 1986).

MATERIAL AND METHODS

As described in Baird et al. (1986), DNA samples of ~700 people from three ethnic groups in the New York area were analyzed by means of Southern blotting. *Eco*RI digests were probed with pAW101 (Wyman et al. 1984), which recognizes the *D14S1* locus, and *Taq*I digests were probed with pLM0.8, which recognizes the *HRAS-1* locus. For the analyses presented here, data were read from figures 2 and 4 of Baird et al. (1986), using a Numonics 2200 digitizing tablet and a simple microcomputer routine called RULER. The *HRAS-1* data unambiguously identify each variant by its fragment length, but the resolution of the gels was not sufficient to identify unambiguously every variant of the *D14S1* locus. Baird et al. (1986) calculated that the error in size measurement was 0.6% of the fragment length. For the largest fragments (~25 kb), this means that the 95% confidence interval for fragment size is ~ ±300 bp. In an attempt to account for this uncertainty in allele identification, the *D14S1* data were pooled in several ways. Two new allele frequency distributions were constructed by pooling adjacent 200-bp units in both of the two phases. Three more distributions were constructed via the three possible 300-bp poolings. Statistical analyses described below were run on all five poolings as well as on the raw *D14S1* data.

The expected allele frequency distribution under the infinite-alleles model is determined by the population size and the mutation rate. The sampling theory of Ewens (1972, 1979) shows that the number of genes sampled (N) and the count of observed alleles (k) in the sample are sufficient statistics to determine the expected allele frequency distribution. Watterson (1978*a*, 1978*b*) showed that a useful test statistic can be constructed by comparing the observed and expected values of $F = \Sigma x_i^2$, where x_i is the frequency of allele i. F is known as the gene identity and is the Hardy-Weinberg expected homozygosity. By means of the algorithm of Stewart (Fuerst et al. 1977), 1,000 populations were generated for each set of N and k. Each of these simulations gives a realization of the vector of allele frequencies, so empirical distributions of F, the common allele frequency (C) and the count of singletons (S) are easily obtained. The average allele frequency distribution is obtained by pooling and averaging the ranked allele frequencies. Statistical tests of F, S, and C were performed by sorting the empirical distributions and determining where in the empirical distribution the observed values fell. These tests are not independent of each other but serve to illustrate different modes of departure of the observed distribution from the neutral expectation.

RESULTS

By means of the method of Nei (1978), the statistics of population subdivision were calculated for *HRAS-1* ($\bar{H}_S = 0.697$; $G_{ST} = 0.011$) and for the unpooled *D14S1* data ($\bar{H}_S = 0.978$; $G_{ST} = 0.003$). The values of G_{ST} appear lower than is typical for blood proteins (Nei and Roychoudhury 1982), but this is not surprising in light of the high diversity of these loci.

Table 1 gives the results of the Ewens-Watterson tests for the *HRAS-1* and *D14S1* restriction-fragment-length polymorphisms (RFLPs), stratified by ethnic group. Empirical tests for fits of F, S, and C, are given. As described above, several poolings of the *D14S1* data were tested. In all cases, the homozygosity test (F) showed acceptable fit of the infinite-alleles model to the data. Similarly, C was within acceptable limits of the model prediction, but the *D14S1* locus had an excess of singletons in caucasoids. This result must be interpreted with caution, owing to the inability to identify unambiguously all alleles at this locus.

The simulation procedure generates expected allele frequency distributions, and these are presented along with the observed distributions in figures 1 (*HRAS-1*) and 2 (*D14S1*). In the case of *HRAS-1*, there is a consistent tendency for the observed common allele frequency to exceed the expectation, and there is a deficit of alleles of intermediate frequency. In agreement with the data in table 1, the *HRAS-1* data has less diversity than expected under the infinite-alleles model. The *D14S1* data show the opposite pattern, with a common allele

TABLE 1

Goodness of Fit Tests to the Infinite-Alleles Model

	Parameter								
	F			*S*			*C*		
Locus Data and Ethnic Group (*N*), *k*	Observed	Expected	P^a	Observed	Expected	P^a	Observed	Expected	P^a
HRAS-1:									
Blacks (602), 18	0.249	0.218	.261	3	3.30	.434	0.448	0.360	.206
Caucasians (490), 14	0.379	0.269	.127	4	2.51	.078	0.592	0.418	.102
Hispanics (308), 16	0.319	0.220	.114	2	3.38	.701	0.545	0.367	.089
D14S1 unpooled:									
Blacks (544), 80	0.029	0.037	.886	22	24.37	.677	0.070	0.102	.927
Caucasians (494), 99	0.020	0.026	.924	27	34.56	.953*	0.055	0.080	.937
Hispanics (182), 70	0.022	0.024	.519	37	33.48	.110	0.065	0.080	.665
D14S1 pooled by groups of 200 bp:									
Blacks (544), 49	0.057	0.070	.769	15	12.58	.158	0.108	0.165	.901
Caucasians (494), 65	0.041	0.047	.666	13	19.12	.949	0.104	0.124	.671
Hispanics (182), 46	0.043	0.047	.587	18	17.56	.354	0.098	0.128	.760
D14S1 pooled by groups of 300 bp:									
Blacks (544), 36	0.085	0.103	.664	10	8.38	.177	0.153	0.217	.826
Caucasians (494), 46	0.053	0.073	.902	6	11.89	.974*	0.107	0.169	.925
Hispanics (182), 32	0.070	0.081	.614	8	10.35	.770	0.153	0.185	.681

[a] The one-tailed probability of obtaining a value greater than the observed value by chance, drawing from a population that fits the infinite-alleles model. Other possible poolings of *D14S1* gave very similar results.

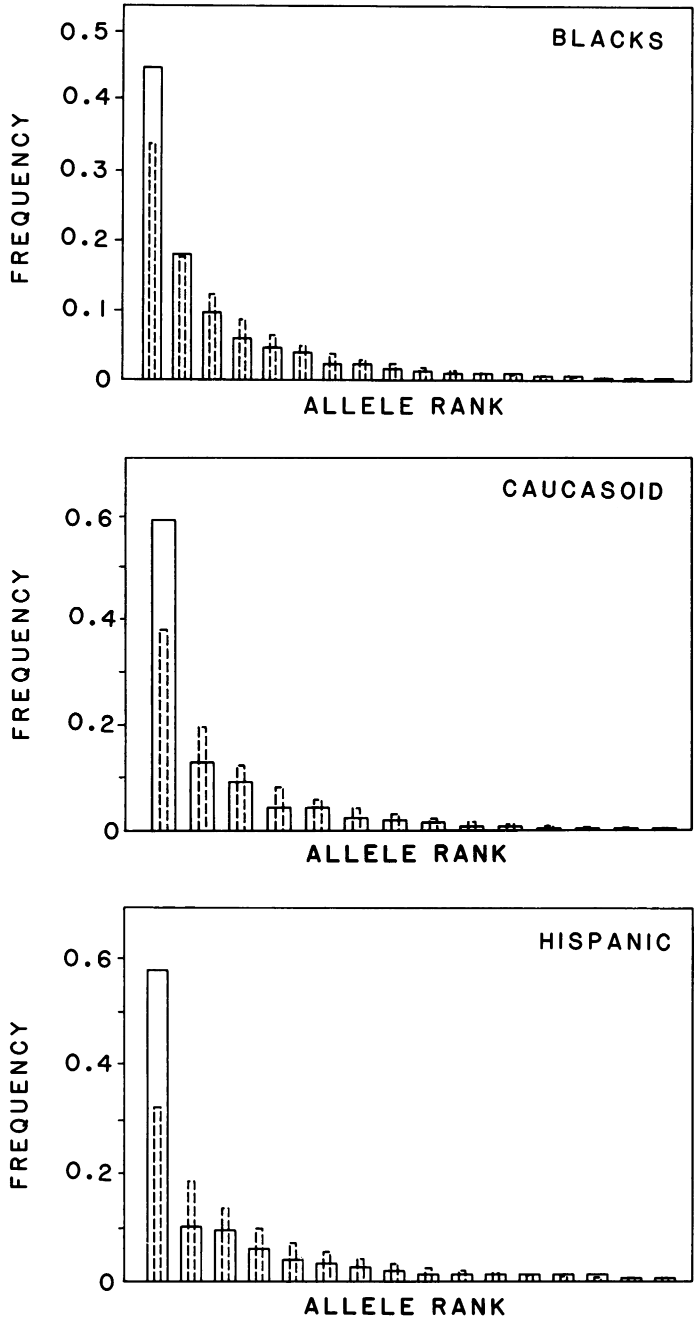

FIG. 1.—Observed (solid lines) and expected (dashed lines) allele frequency distributions of the *HRAS-1* locus, identified by Southern blotting with the pLM0.8 probe and *Taq*I digests. Observed data are from Baird et al. (1986), and expected distributions have been generated using Ewens's (1972) sampling theory.

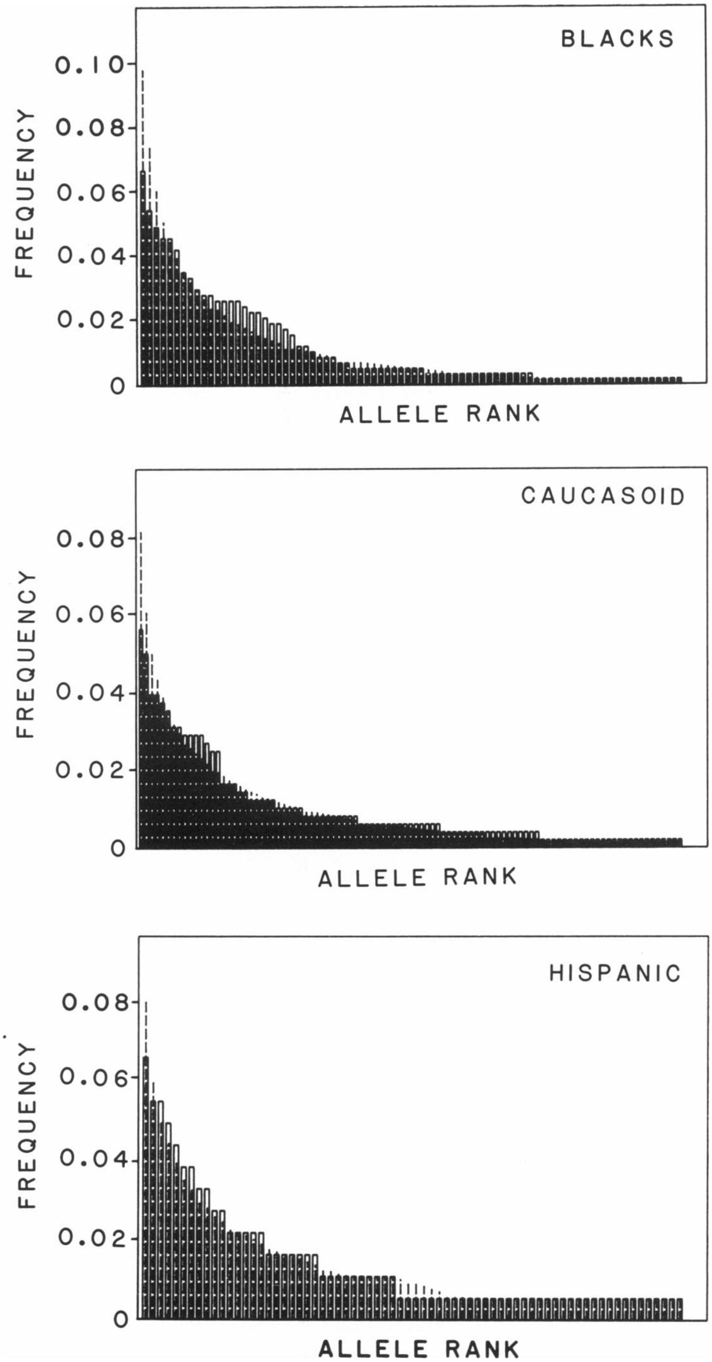

FIG. 2.—Observed (solid lines) and expected (dashed lines) allele frequency distributions identified at the *D14S1* locus with use of the pAW101 probe and *Eco*RI digests.

frequency lower than expected and an excess frequency of intermediate alleles, but the departures are more subtle for this locus. The slight excess observed diversity in the *D14S1* RFLPs is seen regardless of the way in which fragments are pooled (table 1).

The goodness of fit of the observed distributions to those expected in figures 1 and 2 was also tested by means of simple *G* tests, pooling alleles to keep the minimum expected allele frequency >5 (Sokal and Rohlf 1981). For the black, Caucasian, and Hispanic data on *HRAS-1* alleles, the *G*-statistics were 21.10 (13 df; not significant), 45.03 (10 df; $P < .001$), and 40.66 (11 df; $P < .001$), respectively. For the *D14S1* locus, the respective *G*-statistics were 27.46 (45 df), 13.83 (51 df), and 9.52 (27 df), none of which were significant.

DISCUSSION

The allelic variation defined by Southern analysis of the *HRAS-1* and *D14S1* loci in a sample of humans shows an acceptable fit to the expected neutral distribution given by the Watterson test. This test collapses the information from the observed and expected distributions into a single statistic measuring the level of homozygosity. When the observed and expected distributions of allele frequencies were plotted (figs. 1, 2), departures in the direction of a deficit of diversity at the *HRAS-1* locus and a slight excess evenness at the *D14S1* locus were seen. These departures are consistent with the idea that purifying selection is operating on the *HRAS-1* locus and that weak balancing selection is acting at the *D14S1* locus. The goodness of fit of the observed and expected distributions, tested by means of *G*-statistics, verified the significance of the *HRAS-1* departure. The tests do not, however, constitute proof of any particular evolutionary force. The results are best interpreted in the context of other applications of Ewens-Watterson sampling theory.

For most genetic variation identified by means of protein electrophoresis, there is an acceptable fit to neutrality (Nei 1987). Sequential protein electrophoresis successfully characterizes the richness of variation at loci such as *Est-5* and *Xdh* in *Drosophila pseudoobscura* and yields allele frequency distributions that depart from neutrality owing to an excess gene identity (Keith 1983; Keith et al. 1985). This observation, in concert with the remarkable similarity of allele frequency distributions among two widely separated populations, led to the conclusion that purifying selection caused the departure. The most striking departures from the predicted allele frequency distribution have been found in the *HLA* loci (Klitz et al. 1986). The values of *F* for the five loci that encode membrane glycoproteins were significantly below the expected values, whereas the closely linked complement loci showed either no significant departure or an excess of *F*. The direction of departure for the former class of genes is consistent with other evidence of balancing selection in the *HLA* region (Hedrick and Thomson 1983; Hedrick et al. 1986).

The mitochondrial genome in humans also shows extensive variation, and human mitochondrial DNA data have recently been tested for goodness of fit to the infinite-alleles model (Whittam et al. 1986). Most of these tests conformed with neutrality, whereas the departures that were seen were in the direction of an excess in *F*. As discussed by Whittam et al. (1986), reasons for the departures may include purifying selection, nonequilibrium of mutation and drift, population expansion, and hidden population subdivision. The latter two of these should affect all loci, whereas different loci may show different patterns of selection. The fact that departures have been found in both directions (excess and deficit of diversity) suggests that, with respect to mutation and drift, human populations may be sufficiently close to steady-state that the Ewens-Watterson test is applicable.

Genes with repeated structure may be sites of unequal crossing-over, and this mode of generating new alleles may violate assumptions of the infinite-alleles model. In particular, the model assumes (1) that any allele has the same probability of mutating and (2) that, if it mutates, it does so to a novel allele. Depending on copy number or arrangement, repeated genes may vary in their chances of undergoing unequal crossing-over, and crossover events that change copy number may result in alleles that already exist. Despite these hypothetical complications, the two loci investigated in the present study show fairly good fits to the expected distributions, and the departures in homozygosity that were seen are in opposite directions.

Although the Ewens-Watterson test failed to reject an acceptable fit to the infinite-alleles model on the basis of observed and expected *F* and *C*, other aspects of the distributions were revealing. *S* was found to be significantly in excess at the *D14S1* locus in Caucasians, and plots of the distributions (figs. 1, 2) revealed consistent departures. The direct tests of goodness of fit of the entire distribution by means of a *G*-statistic may have advantages for such genetically diverse loci but should be interpreted with caution. The test is based on the assumption of multinomial sampling of alleles, which would only be strictly true if the number of alleles in the sample were fixed. The degree of departure from multinomial sampling—and hence the validity of the test—remain questionable, but the heuristic value and the utility of plotting observed and expected distributions of allele frequencies (as in figs. 1, 2) should be apparent. The Watterson test, which compares observed and expected values of *F*, may be particularly weak in cases of high genetic diversity, since both observed and expected values of *F* become very small; yet both are constrained to be >0. A disadvantage of using the entire distribution for testing goodness of fit is that no test statistic can be easily put into a table. This makes the procedure more computationally intensive, but not prohibitively so. The power and robustness of this test are the subjects of a numerical study currently in progress.

ACKNOWLEDGMENTS

I thank Dr. Thomas Ryan, Director of Minitab, Inc., for generous donation of computer time, and Drs. Warren Ewens, Phil Hedrick, and Thomas Whittam for critical discussion. This work was supported by National Institutes of Health grants GM36496 and HD00743.

REFERENCES

Baird, M., I. Balazs, A. Giusti, L. Miyazaki, L. Nicholas, K. Wexler, E. Kanter, J. Glassberg, F. Allen, P. Rubinstein, and L. Sussman. 1986. Allele frequency distribution of two highly polymorphic DNA sequences in three ethnic groups and its application to the determination of paternity. Am. J. Hum. Genet. **39**:489–501.

Balazs, I., M. Purrello, P. Rubinstein, B. Alhadeff, and M. Siniscalco. 1982. Highly polymorphic DNA site *D14S1* maps to the region of Burkitt lymphoma translocation and is closely linked to the heavy chain g1 immunoglobulin locus. Proc. Natl. Acad. Sci. USA **79**:7395–7399.

Capon, D. J., E. Y. Chen, A. D. Levinson, P. H. Seeburg, and D. V. Goeddel. 1983. Complete nucleotide sequence of the T24 human bladder carcinoma oncogene and its normal homologue. Nature **302**:33–37.

de Martinville, B., J. Giacolone, U. Franke, C. Shih, and R. A. Weinberg. 1983. Oncogene from human EJ bladder carcinoma is located on the short arm of chromosome 11. Science **219**:498–501.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3**:87–112.

———. 1979. Mathematical population genetics: biomathematics. Springer, New York.

Fuerst, P. A., R. Chakraborty, and M. Nei. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. Genetics **86**:455–483.

Gerald, P. S., and K. H. Grzeschik. 1984. Report of the Committee on the Genetic Constitution of Chromosomes 10, 11 and 12. Cytogenet. Cell Genet. **37**:103–126.

Hedrick, P. W., and G. Thomson. 1983. Evidence for balancing selection at HLA. Genetics **104**:449–456.

Hedrick, P. W., G. Thomson, and W. Klitz. 1986. Evolutionary genetics: HLA as an exemplary system. Pp. 583–606 *in* S. Karlin and E. Nevo, eds. Evolutionary processes and theory. Academic Press, New York.

Keith, T. P. 1983. Frequency distributions of esterase-5 alleles in two populations of *Drosophila pseudoobscura*. Genetics **105**:153–155.

Keith, T. P., L. D. Brooks, R. C. Lewontin, J. C. Martinez-Cruzado, and D. L. Rigby. 1985. Nearly identical allelic distributions of xanthine dehydrogenase in two populations of *Drosophila pseudoobscura*. Mol. Biol. Evol. **2**:206–216.

Klitz, W., G. Thomson, and M. P. Baur. 1986. Contrasting evolutionary histories among tightly linked HLA loci. Am. J. Hum. Genet. **39**:340–349.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89**:583–590.

———. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Nei, M., and A. K. Roychoudhury. 1982. Genetic relationship and evolution of human races. Evol. Biol. **14**:1–59.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry. W. H. Freeman, San Francisco.

Thomson, G. 1977. The effect of a selected locus on linked neutral loci. Genetics **85**:753–788.

Watterson, G. A. 1978*a*. An analysis of multi-allelic data. Genetics **88**:171–179.

———. 1978*b*. The homozygosity test of neutrality. Genetics **88**:405–417.

Whittam, T. S., A. G. Clark, M. Stoneking, R. L. Cann, and A. C. Wilson. 1986. Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. Proc. Natl. Acad. Sci. USA **83**:9611–9615.

Wright, S. 1978. Evolution and the genetics of populations. Vol. **4**. Variability within and among natural populations. University of Chicago Press, Chicago.

Wyman, A., and R. White. 1980. A highly polymorphic locus in human DNA. Proc. Natl. Acad. Sci. USA **77**:6754–6758.

Wyman, A., L. Wolfe, and D. Botstein. 1984. Isolation of the highly polymorphic locus *D14S1* and other lethal/unstable human sequences in a new cloning host. Am. J. Hum. Genet. **36**:159S.