An Evaluation of Evolutionary Constraints on Microsatellite Loci Using Null Alleles

Tovi Lehmann,*,† William A. Hawley*,‡ and Frank H. Collins*,†

**Division of Parasitic Diseases, Centers for Disease Control and Prevention, Chamblee, Georgia 30341, †Department of Biology, Emory University, Atlanta, Georgia 30322 and ‡Kenya Medical Research Institute, Clinical Research Centre, Nairobi, Kenya*

Manuscript received May 13, 1996
Accepted for publication August 9, 1996

ABSTRACT

A test to evaluate constraints on the evolution of single microsatellite loci is described. The test assumes that microsatellite alleles that share the same flanking sequence constitute a series of alleles with a common descent that is distinct from alleles with a mutation in the flanking sequence. Thus two or more different series of alleles at a given locus represent the outcomes of different evolutionary processes. The higher rate of mutations within the repeat region (10^{-3} or 10^{-4}) compared with that of insertion/deletion or point mutations in adjacent flanking regions (10^{-9}) or with that of recombination between the repeat and the point mutation (10^{-6} for sequences 100 bp long) provides the rationale for this assumption. Using a two-phase, stepwise mutation model we simulated the evolution of a number of independent series of alleles and constructed the distributions of two similarity indices between pairs of these allele series. Applying this approach to empirical data from locus AG2H46 of *Anopheles gambiae* resulted in a significant excess of similarity between the main and the null series, indicating that constraints affect allele distribution in this locus. Practical considerations of the test are discussed.

MICROSATELLITE loci have been described as "ideal" markers to measure population-level phenomena (*e.g.*, population structure) due to their high polymorphism, codominance, abundant presence throughout the genome, and relative ease in scoring (*e.g.*, BOWCOCK *et al.* 1994; BUCHANAN *et al.* 1994; SCRIBNER *et al.* 1994; ESTOUP *et al.* 1995; LANZARO *et al.* 1995). The high polymorphism of microsatellite loci results from high mutation rates, estimated to range from 10^{-2} to 10^{-5} locus/gamete/generation (DALLAS 1992; EDWARDS *et al.* 1992; WEBER and WONG 1993), with most estimates being between 10^{-3} and 10^{-4}. Replication slippage events are considered to be the main process producing insertion/deletion (indel) mutations of 1, or infrequently, several repeat units (LEVINSON and GUTMAN 1987; WEBER and WONG 1993). However, the forces that shape allele composition in these loci are poorly understood (*e.g.*, SLATKIN 1995), and some evidence suggests that these forces include biased mutation rate (GARZA *et al.* 1995) and/or selection acting on allele size (EPPLEN *et al.* 1993). If such forces strongly affect allele composition at these loci, interpopulation differentiation will be underestimated and gene flow will be overestimated to an unknown extent. Different microsatellite loci probably experience different intensities of such constraints, ranging from negligible to strong. Ideally, a test would permit evaluation of constraints acting on a single locus. A possible test is described below using null alleles. This test is also applied to our data at locus AG2H46 of the principal malaria vector in Africa, the mosquito *Anopheles gambiae.*

Corresponding author: Tovi Lehmann, Division of Parasitic Diseases, Centers for Disease Control and Prevention, Mailstop F22, 4770 Buford Highway, Chamblee, GA 30341.
E-mail: lbt2@ciddpd2.em.cdc.gov

Null alleles in microsatellite loci cannot be visualized on the gel due to insufficient PCR product resulting from a mutation(s) in the flanking sequence that is complementary with one of the oligonucleotide primers. They can be detected when PCR is done with alternative primers. "Null" alleles represent a common complication in the interpretation of microsatellite genotype data, resulting in a reduced level of observed heterozygosity (CALLEN *et al.* 1993). The incidence of null alleles was 30% (seven of 23 loci) in one study of human microsatellites (CALLEN *et al.* 1993). Our data on microsatellite loci in the mosquito *A. gambiae* also suggest high incidence of null alleles. At one of the five loci examined (AG2H46), 2/3 of the homozygotes were found to be heterozygotes for a null allele when tested with an alternative set of primers (LEHMANN *et al.* 1996 and this report). Null alleles indicate the presence of polymorphism in the sequence flanking the repeat region. The test below exploits the presence of polymorphism regardless if the polymorphic site(s) occurs in the primer annealing sequence (resulting in null alleles) or closer to the repeat region.

A TEST FOR THE EFFECT OF CONSTRAINTS ON MICROSATELLITE EVOLUTION

Rationale and assumptions: Null alleles most commonly arise from point mutations in the sequence flanking the repeat region, but indels have been re-

ported also (CALLEN *et al.* 1993). A null allele series at a given microsatellite locus is defined as all alleles that share the same flanking sequence including the mutation. The distinction between null and "not null" (the main allele series) is arbitrary. Therefore, we defined the main series as the most common series, *i.e.*, the flanking sequence haplotype represented by the largest proportion of chromosomes.

The rate of indel mutations in the repeat regions of microsatellite loci is two to four scale orders higher than the rates of both point mutations in a specific base pair (10^{-9}, LI *et al.* 1985; LI and GRAUR 1991:69–73; LEWIN 1994:106) and recombination events within ~100-bp regions embracing the repeat region and the flanking sequence containing the mutation up to the mutation (10^{-6}, HILLIKER *et al.* 1991; LEWIN 1994:128). (Microsatellite alleles are typically <200 bp in total length.) Therefore, a new null allele that emerges by a mutation in the flanking region evolves into a null series primarily by the accumulation of indel mutations in the repeat region. Furthermore, due to the substantial difference between rates of indel mutations within the repeat sequence and recombination, a series of null alleles evolves essentially independently from other series of the locus and represent the result of a separate evolutionary process. Only if the effective population size of the species is $>10^5$ can recombination between different series of alleles homogenize the allele distributions in the series to an extent where this independence will not hold. Otherwise, excessive similarity (more than expected by chance) in allele size distributions among such independent allele series (from the same locus) attests to the existence of evolutionary constraints on that locus. It is also assumed that the slight differences in the flanking sequence between series do not affect the mutation process in the repeat region.

The test: A general test to assess the excess of similarity between any number of null series (having at least one null series and one main series) is based on simulation results. The simulation generates a large number ($n = 25$ or more) of independent series that allow multiple pairwise comparisons ($n*(n-1)/2$) that are used to build distributions of similarity indices. The indices we used were (1) the absolute pairwise difference between the mean repeat number and (2) the value of a G statistic measuring the homogeneity of allele distributions in a contingency table (without significance test) calculated for each series pair. While the absolute difference between means reflects the difference in the central location of the distributions, the G statistic incorporates information on the overall similarity between the two distributions. Because the resulting distributions of similarity indices depend on parameters whose point estimates are usually unknown, one needs to estimate conservative (exaggerating the similarity among the series) values from the range of plausible values in every case. A conservative test ensures greater validity of a finding indicating excess similarity between the observed series in comparison with expectations based on simulations. The parameters required for the simulation include effective population size (N_e), mutation rate (μ), and minimum time of evolution of the series. The empirical data available from the main and null series provide guidance to evaluate the range of plausible values of these parameters. Thus, if the heterozygosity, number of alleles, and allele range of the null series is similar to that of the main series, the minimum evolutionary time of the series compared can be set to $4N_e$ generations (SHRIVER *et al.* 1993). If these measures of the null series are significantly smaller or larger than those of the main series, the minimum evolutionary time of the series compared can be set to the (average) minimum time needed to attain the lowest values of heterozygosity, number of alleles, and allele size range of that series, which is estimated by the simulation. Because the heterozygosity estimated from empirical data depends on N_e and mutation rate (assuming neutrality), one needs to estimate only one to derive the other (NEI 1987, and see below).

MATERIALS AND METHODS

Simulation: Assuming neutrality (no constraints on allele size), we simulated the effects of genetic drift and mutation on the distribution of allele size in a microsatellite locus. We used a two-phase stepwise mutation model (SMM) where most mutations are single step mutations but larger steps also occur. The simulation process was based on SHRIVER *et al.* (1993) and DI RIENZO *et al.* (1994) with modifications as described below. A population of 1000 individuals (2000 alleles), all fixed to one allele, was set at the first generation. Each allele was associated with a reproductive value randomly drawn from a Poisson distribution with a mean of 1. Alleles were represented in the next generation according to their reproductive value, such that alleles with a reproductive value of 0 were eliminated while those having a value equal to 1, 2, or other number were represented accordingly in the next generation. Each allele was also associated with a mutation index randomly drawn from a uniform distribution (0, 1) to determine if the allele mutates. If the mutation index was lower than the mutation rate, a mutation took place. The range of the mutation index culminating in a mutation event (0 to μ) was divided such that a value within 90% of this range resulted in a deletion or insertion of one repeat unit, while mutations involving two, three, four, five, and six repeat units occurred each in 2% likelihood. Insertion mutations occurred when the mutation index multiplied by 10^8 was an even number, while deletion mutations occurred otherwise. This symmetrical mutation process assumed no boundaries on allele size (but see below). Population size could fluctuate between 750 and 1250 individuals, but when it passed any boundary, the mean of the Poisson distribution from which reproductive values were drawn was changed to 1.334 or 0.8, respectively, and returned to one as soon as the population size backtracked. Simulation continued for 4000 ($4N_e$) generations, by which time a steady state in heterozygosity, allele size range, number of modes, and number of alleles was reached (SHRIVER *et al.* 1993). The distribution of allele frequency of each generation was stored in another computer file.

To consider the effect of sampling from populations, we randomly sampled 100 alleles of each simulation's last genera-

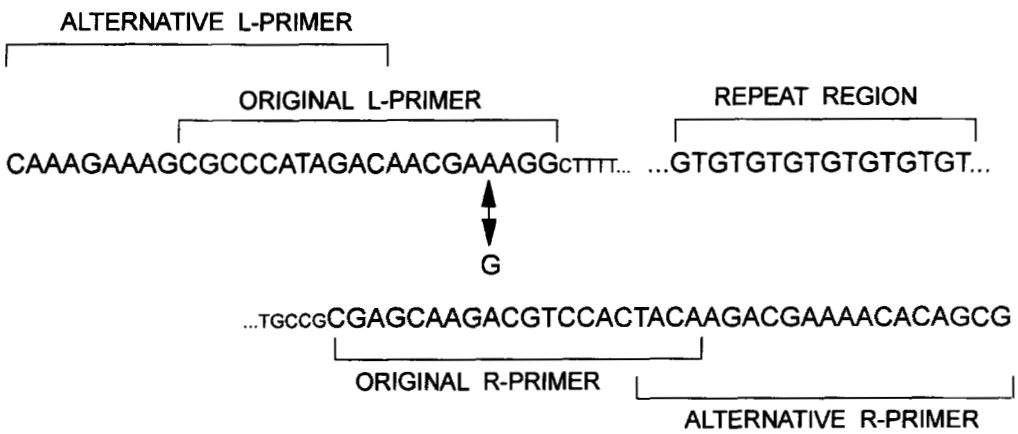

FIGURE 1.—Flanking regions of the GT strand near the primer annealing sites of locus AG2H46 showing the transition creating the null allele series.

tion and used these samples (one per simulation) to calculate the distributions of our similarity indices. Sample size can be set to the actual size of the null allele series when this test is employed. To avoid inflating the G statistic, rare alleles were pooled such that no expected frequency was smaller than 2 and no more than 20% were smaller than 5.

Because heterozygosity of microsatellite loci is usually over 0.7, the expected value of the product $\mu*N_e$ ranges from 1 to 10 for a stepwise mutation model (see also VALDES *et al.* 1993). Thus, we simulated the compositions of allele sizes for mutation rates of 10^{-3} and 10^{-2} with N_e of 1000 individuals. The simulations provided an example of how to perform and interpret such a test and also provided a basis to compare the power of this test using the different similarity indices.

Empirical data: Repeated failure to produce PCR products with primers for locus AG2H46 from *A. gambiae* specimens that were successfully scored for other four loci and an unusually high deficiency of heterozygotes suggested the presence of null alleles in that locus. Using alternative primers (Figure 1), these specimens were successfully scored. Mosquito collection, extraction of DNA from individual specimens, PCR conditions, visualization and scoring of allele length were described in detail previously (LEHMANN *et al.* 1996), thus only cloning and sequencing of microsatellite alleles is described below. Specimens collected in Asembo Bay (western Kenya) during June–July 1994 were included in the present study. The sampling sites were located <10 km apart from each other. Three specimens that repeatedly failed to produce PCR product with original primers but were successfully scored with alternative primers were selected. The PCR products were ligated into a linearized pCRII vector provided in the TA cloning kit (Invitrogen) and transformed into competent cells, from which seven clones per individual were sequenced using Sequenase version 2 (United States Biochemicals). Sequences labeled with ^{35}S were read from autoradiographs after 24 hr exposure time.

RESULTS

Simulation: Simulations generated allele compositions with average heterozygosity coinciding with expectations of the one-step SMM (Figure 2) based on $H = 1 - 1/(1 + 8*N_e*\mu)^{1/2}$ (NEI 1987). The occurrence of mutations involving two and up to six repeat units in 10% of the mutation events had no obvious impact on heterozygosity. The expected values of heterozygosity based on the infinite allele model (IAM) were 0.98 and 0.80 for N_e of 1000 and mutation rates of 10^{-2} and 10^{-3}, respectively. These IAM predictions were clearly higher than the values obtained by most simulations after leveling off (Figure 2). The better fit with SMM predictions is expected because either single and multiple step mutations result in independent events in which the same mutant allele size was created. This "convergence" in allele size is an important characteristic of microsatellite evolution. That the heterozygosity of the simulations based on the two-phase mutation process converged on the predicted value based on the one step SMM formulation allows the derivation of one of the two parameters, N_e or μ, given the empirical heterozygosity of the series and an estimate of one parameter.

The distributions of the indices of similarity between independent allele series allow the determination of whether two or more empirically determined allele series are more similar to each other than would be expected by chance for independent series (Figure 3). According to these distributions, an absolute difference in the mean repeat number smaller than 1.3 for a situation with $N_e \simeq 1000$ and $\mu \simeq 0.01$, or smaller than 0.3 for a situation with $N_e \simeq 1000$ and $\mu \simeq 0.001$ are significant at the 95% confidence level. Likewise, a G statistic smaller than 88 for the first case, and smaller than 65.5 for the other case will result in the same decision. It is noteworthy that these critical values are conservative because they assume that both series have originated from alleles of the same size and the minimal time of evolution of the series is used.

Demonstration of null allele series: Rescoring all homozygotes using the alternative primers (Figure 1) in addition to 13 individuals that had produced no PCR products with the original primers, we found 85 null alleles. All 20 clones representing six alleles (from three individuals that could only be PCR amplified using the alternative primers) had a G instead of an A at the fourth site from the 3′ end of the left primer (Figure 1). This site was located 49 bases from the first GT repeat in the core motif. Thus, a null primer was designed that was identical to the original primer but had a G instead of an A in the fourth site from the primer 3′ end (Figure 1). Only five

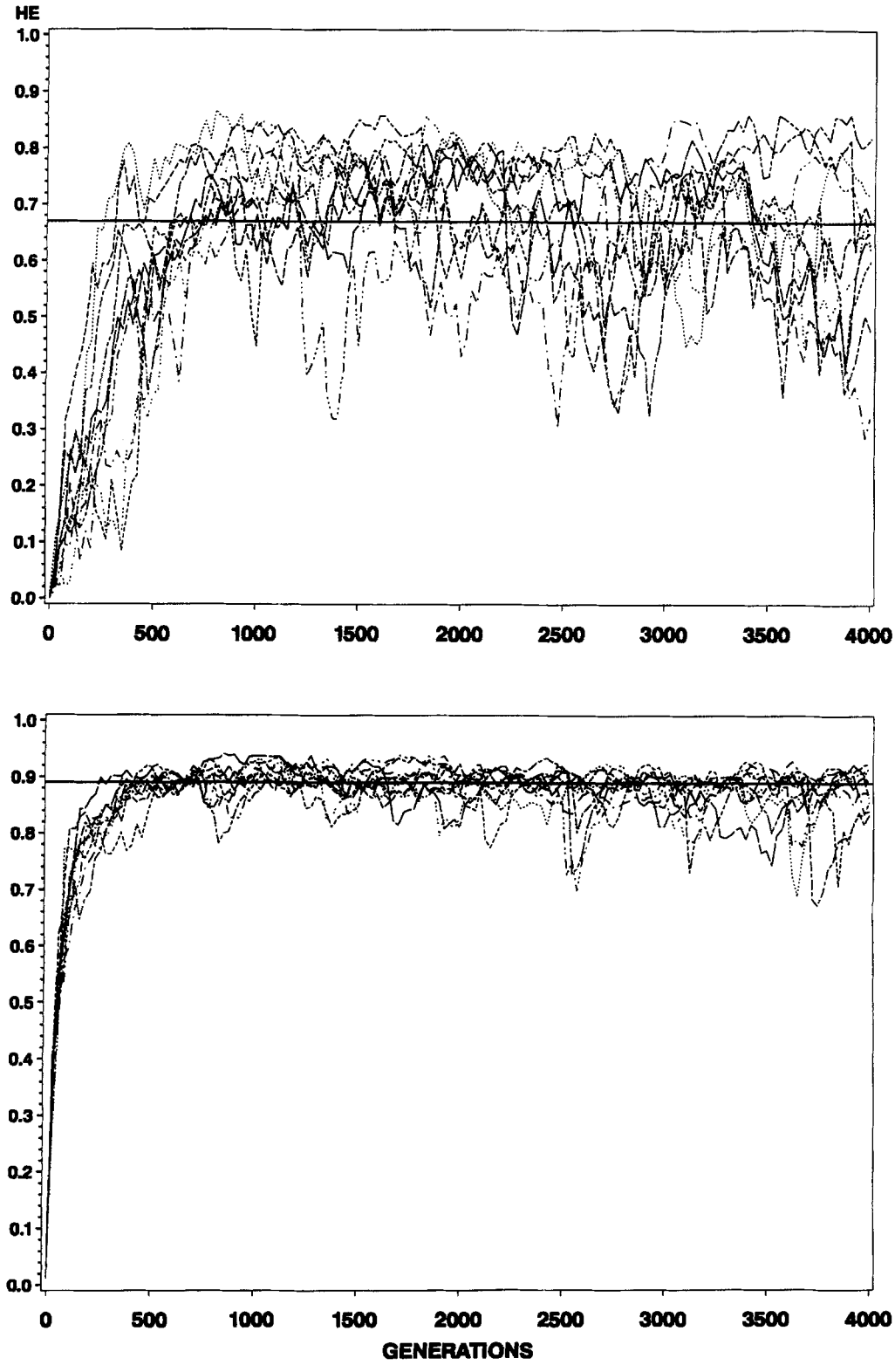

FIGURE 2.—Convergence of heterozygosity on predicted values based on SMM for effective population size of 1000 individuals and mutation rate of 10^{-3} (top) and mutation rate of 10^{-2} (bottom). Only 25th generations of 10 simulations were plotted.

of 85 alleles were not PCR amplified with the null primer (and with the original R primer) and thus were excluded from the null series. Using the null primer we verified that the 80 null alleles constitute an homogenous series with regard to the A-G transition.

Applying the test to locus AG2H46 of *A. gambiae*: The allele distributions of the null (as defined above) and the main series (the alleles amplified by the original primers) were remarkably similar (Figure 4, Table 1). The overlap in allele size range was nearly complete (the only unique allele was 130, which was represented by one copy in the main series, Figure 4). The difference in the mean size of the series was 1.55 and the *G* statistic was 20.5 (Table 1).

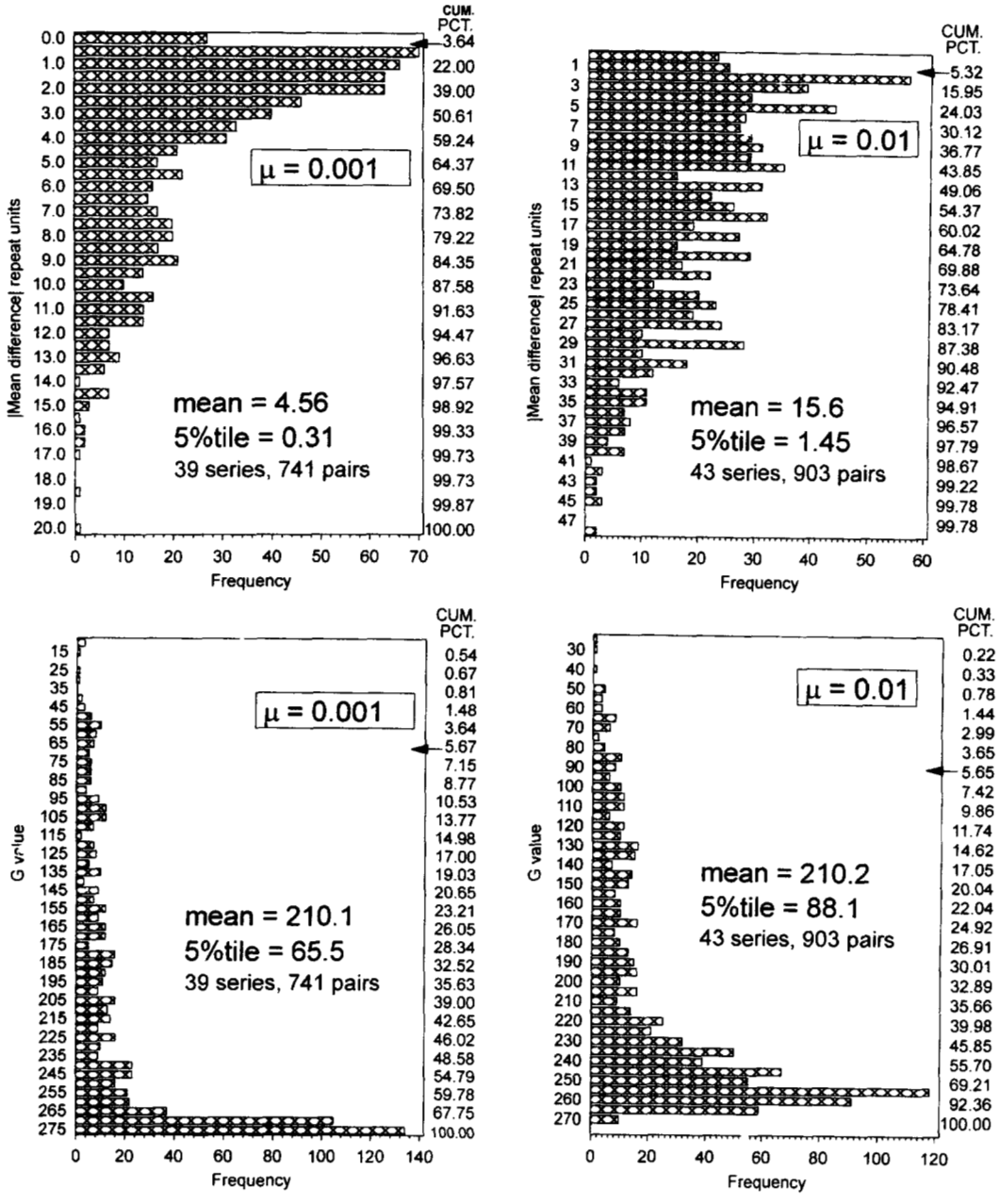

FIGURE 3.—Distribution of absolute mean difference between all possible pairs of simulations (top panels) and *G* statistic of pairwise contingency tables of all possible pairs of simulations (bottom panels). Only the 4000th generation was used for each simulation. Mutation rate of each distribution is shown in frame. From each distribution only 50 individuals (100 chromosomes) were randomly sampled and the resulting allele distribution was used. Arrows point to the 5 percentile of the distribution, where the range of smaller values represent excess similarity at the 95% confidence level for two-allele series (one main and the other null). The cumulative frequency values are shown above bars.

To evaluate whether the similarity between the null and the main allele series indicated the action of constraints on this locus, two simulation processes were employed. The first (unconstrained) followed the process described above, and the second (semiconstrained) simulation process was modified to represent a more realistic situation. Thus, the allele generating a null series was not the same every time but was selected from the alleles in the main series such that the chances of an allele to mutate into a null series was determined by its frequency. The actual number of repeats in the main series was used instead of an arbitrary number and when an allele with one repeat units was created its mutation rate was set to zero, representing a lower boundary of allele size. Both simulation processes used the same effective population size (N_e) and the same mutation rate. The N_e was estimated for *A. arabiensis* to be 2000 individuals (TAYLOR *et al.* 1993). *A. arabiensis* is a sibling species of *A. gambiae* with a very similar biology. To provide a conservative test we doubled this estimate and used a value of $N_e = 4000$ in our simulations. Given N_e, and the expected heterozygosity at the main series (Table 1), we calculated a microsatellite mutation rate of 0.00132 (see above), which is within an acceptable range for microsatellite loci (see above). Finally, we sampled 80 alleles from each distribution before calculating the similarity indices. To provide a deeper insight into the divergence of allele series through time, we compared allele distributions generated within $4N_e$, $2N_e$, N_e, and $N_e/2$ generations.

The simulations based on the unconstrained model followed the course described above (data not shown),

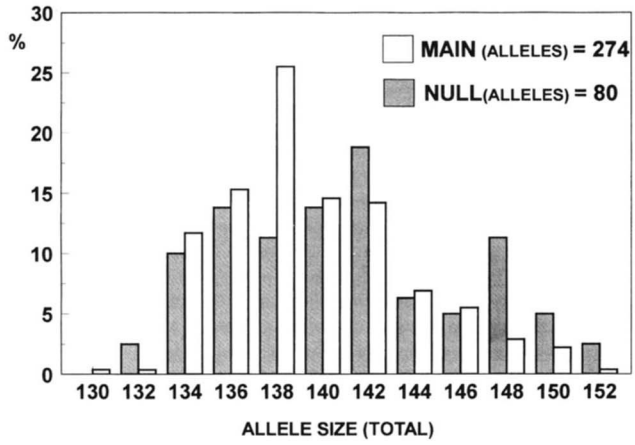

FIGURE 4.—Allele distribution in the main and the null series (allele size of 124 nucleotides is equivalent to one repeat unit).

whereas those based on the semiconstrained model differ in their course. Some of the series whose mean repeat number drifted toward 1 had relatively high frequencies of that allele that was not prone to mutations. Apparently, these series followed a process of a loss of a microsatellite locus. In such series, heterozygosity declined temporarily below 0.6 (*e.g.,* Figure 5). However, during the $4N_e$ generations none of the 27 allele series became fixed on the one repeat allele. The expected value of heterozygosity was attained by most series in the 2000th generation ($N_e/2$). All series were included in calculation of the pairwise similarity indices.

The observed G value (20.5) was found to represent significant excessive similarity by both simulation processes throughout all the time sections (Table 2). The absolute difference in the mean number of repeats was always lower than the mean expected value but the difference was not significant ($P = 0.12$ for the $4N_e$ time section, Table 2). These results provided strong evidence for constraints acting on locus AG2H46 in *A. gambiae.*

DISCUSSION

A simple, although computationally intensive test to assess the effect of constraints on the evolution of microsatellite loci is described. This is a locus-specific test that allows discriminating loci affected by high intensity of constraints once empirical data on (at least one) null allele series per locus are available. The validity of the test depends on correct estimates of the minimum time of evolution of the "youngest" series and either N_e or the mutation rate. Because point estimates of these parameters usually are not available, the use of conservative estimates (that maximize the expected similarity *i.e.,* low mutation rate, large population size, short minimum time of evolution) is recommended. Application of the test to empirical data indicated that locus AG2H46 in *A. gambiae* is significantly affected by constraints and thus provided support for the statistical power of this test. Although we doubled the only available estimate of N_e (TAYLOR *et al.* 1993) to attain a more conservative test, the validity of our results obviously depends on this being a reasonable approximation. Until verification of this value, our conclusion should be regarded as preliminary. However, a visual inspection of allele distributions in the main and the null series (Figure 4), viewed on the basis of the theory described above, lends support for the action of constraints on this locus without a formal test.

TABLE 1

Comparisons of null and main allele series

	Null ($N^a = 80$)	Main ($N^a = 274$)
Expected heterozygosity	0.88	0.85
No. of alleles	11	12
% most common allele	18.8	25.5
Range repeat number	10	11
No. of modes	3	1
Mean allele size	141.05	139.50
G test of homogeneity (rare alleles pooled)	d.f. = 9, $G = 20.50$, $P < 0.015$	

[a] No. of chromosomes.

SHRIVER *et al.* (1993) suggested that microsatellite loci attain steady state in heterozygosity, allele size range, number of modes and number of alleles within $4N_e$ generations. We suspect $4N_e$ generations may be an overestimate and suggest use of simulations for guidance about the time period and to evaluate the results at several time points (*e.g.,* Table 2). The method is supposedly quite robust with respect to errors in the values of N_e and μ [as long as $N_e \leq 100{,}000$ (see below)] because they are constrained by the heterozygosity of the series, which can be reliably estimated based on the empirical data. Thus, the expected degree of similarity among independent series at a certain time period is a function of the product of both parameters, which is derived directly from the heterozygosity.

The lack of a detailed understanding of the mutation process(es) in microsatellite loci impedes modeling of a fully "realistic" simulation. Including a lower bound for allele size and the actual locus repeat numbers and allele distribution in the "semiconstrained" model was an attempt to improve the realism of the simulation. That both models produced agreeable test results was encouraging. Nevertheless, a realistic simulation process and comparing results obtained by "neutral" process with alternative models specifying different type and intensities of constraints will provide more insight into microsatellite evolution and evaluation of the power of this test.

The validity of the results depends on negligible "leaking" between the different series due to recombi-

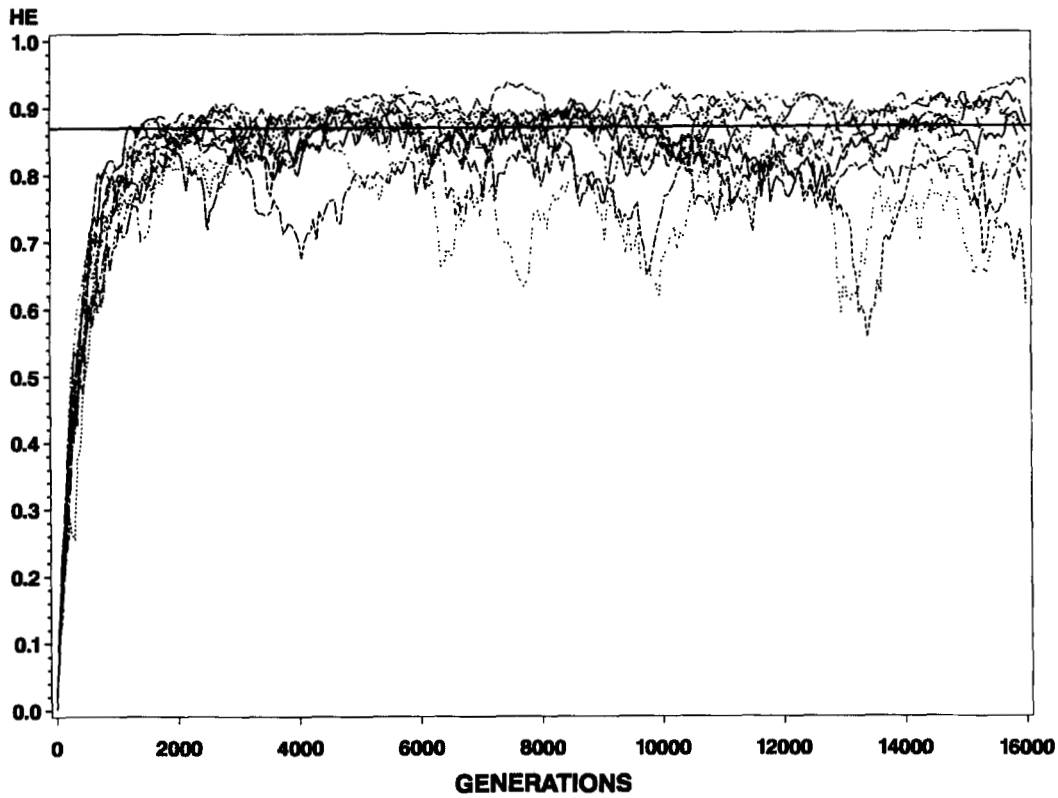

FIGURE 5.—The change in expected heterozygosity through time in the semi-constrained simulations in relation to the predicted value based on SMM (see text for more details). Only 50th generations of 10 simulations were plotted.

nation between the mutation in the flanking sequence (creating the null allele) and the distal end of the repeat region. A recombination within a sequence of 100 bp long is expected in a rate of 10^{-6} for *Drosophila melanogaster* (HILLIKER *et al.* 1991) and slightly lower for a mammal (mouse, human, LEWIN 1994:128). The recombination rate (c) acts on allele series as the migration rate acts on separate populations: as a force that determines whether they evolve together or independently. The critical value of $Nm = 1$ is considered as the "threshold" value to discern these situations (see SLATKIN 1987 for a discussion on this limit for Nm). We recommended therefore, that N_e be smaller than 10^5 to achieve $Nc < 0.1$. In cases where the effective population size is $>10^5$, the test may be used only with loci known to map in regions of the genome where recombination rates have been shown to be low enough such that $Nc < 0.1$. Loci mapped to the Y chromosome (in organisms with XY system) may be particularly useful because recombination can be ignored. The contribution of recombination to the evolution of an allele series can be also minimized by selecting mutations responsible for the null series as close as possible to the repeat region itself. Locus AG2H46 is mapped in subdivision 7A, located at the tip of chromosome *II*, where recombination is probably lower than the estimate cited above. Thus, the test results obtained for this locus may prove robust even if the N_e of *A. gambiae* is slightly $>10^5$.

The test applies to simple repeat loci that have only one type of repeat unit and cannot be applied to loci in which two or more different repeat units mutate simultaneously without serious risk of compounding errors of the point estimates used in the simulation.

The repeated, unintentional discovery of null alleles attests to their relative abundance. Their presence has been indicated in many independent studies (PHILLIPS *et al.* 1991; WEBER *et al.* 1991; CALLEN *et al.* 1993; BOWCOCK *et al.* 1994; LANZARO *et al.* 1995; LEHMANN *et al.* 1996). In most cases, however, the identification of the mutation involved was not attempted. When null alleles are discovered by the failure of an original set of primers to amplify all possible alleles, the null alleles revealed by a new set of primers may include more than one null series. It is necessary to confirm that a set of nulls identified by such a procedure actually represents a single series to use this test. Not all null alleles need to be sequenced, however, PCR amplification with an oligonucleotide primer that includes at its 3′ end the base(s) complimentary to the mutated base(s) defining the null allele may be useful for determining which of the alleles found to be null with an original set of primers actually belong to a series defined by this mutation.

When several series are available for a given locus, additional tests can be used. Obviously, the power of the test increases as more null series are available. In that case, a one sample test (t-test or nonparametric equivalent *e.g.*, Wilcoxon signed-rank test) of the hypothesis that the mean similarity index of the empirical series is equal to that of the simulated series is probably of higher statistical power than testing the individual value. If constraints confine the distributions of allele size into a narrow range by "pressing" the margins of

TABLE 2

Comparing unconstrained with semiconstrained simulation processes in different time sections (see text for details)

Time section (generations)	Unconstrained simulation process	Semiconstrained simulation process
$N_e/2$		
No. of series; no. of pairs	34; 561	35; 595
Mean difference[a]	1.29 (0.088)	2.65 (0.20)
Mean G value	80.9 (25.4)	102.0 (36.9)
N_e		
No. of series; no. of pairs	32; 496	33; 528
Mean difference[a]	3.35 (0.23)	3.24 (0.325)
Mean G value	86.6 (41.6)	114.2 (58.7)
$2N_e$		
No. of series; no. of pairs	32; 496	30; 435
Mean difference[a]	7.50 (0.61)	5.29 (0.45)
Mean G value	98.5 (45.1)	107.9 (48.07)
$4N_e$		
No. of series; no. of pairs	26; 325	27; 351
Mean difference[a]	16.67 (1.46)	8.44 (0.85)
Mean G value	101.1 (56.5)	119.1 (47.1)

Fifth percentiles shown in parentheses.
[a] Absolute value of mean difference.

the distributions toward their center, then a correlation is predicted between the deviation of the mean allele size of each null series from the grand mean across all series and each series' skeweness toward the grand mean. A positive and significant correlation (between the skewness of each null distribution toward the center of all distributions pooled and the absolute distance of each series' mean from the grand mean) provides evidence that constraints exist. A regression coefficient measuring the increase in the skewness (toward the center) in relation to the deviation of series' mean from the grand mean may be used to quantify the intensity of constraints. The power of this test, however, depends on the variation among the means of the different series. If all means differ only slightly from each other, then a nonsignificant correlation is meaningless and the previous test may be used. Nevertheless, a finding that the means of independent series do not differ much from each other is itself an indication of constraints on allele size.

The evaluation of the effect of constraints on microsatellite loci is invaluable for correct interpretation of population genetics data. Empirical data on the relationship between different null allele series or between null and main allele series at a given locus will become available as the study of microsatellite polymorphism continues. With a relatively small additional effort to produce data on the homogeneity of null allele series, it seems possible to evaluate the intensity of constraints acting on a locus specific basis. We hope that our results, indicating measurable constraints acting on locus AG2H46 of *A. gambiae*, the first locus to be tested, will not be found as representing most microsatellite loci.

We thank BRUCE LEVIN and his colleagues from Emory University, NORA J. BESANSKY, DIANE M. HAMM, ALLAN HIGHTOWER, MICHEL TIBAYRENC, ANANIAS ESCALENTES, and CATHRINE WALTON from Centers for Disease Control and Prevention, and WILLIAM C. BLACK IV from Colorado State University for useful suggestions and help. We also thank anonymous reviewers for insightful suggestions that were incorporated into the manuscript. This work was supported by a fellowship to T.L. from the National Center for Infectious Diseases American Society for Microbiology Postdoctoral Research Associates Program and a grant to F.H.C. from the John D. and Catherine T. MacArthur Foundation.

LITERATURE CITED

BOWCOCK, A. M., A. RUIZ-LINARES, J. TOMFOHRDE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

BUCHANAN, F. C., L. J. ADAMS, R. P. LITTLEJOHN, J. F. MADDOX and A. M. CRAWFORD, 1994 Determination of evolutionary relationships among sheep breeds using microsatellites. Genomics **22:** 397–403.

CALLEN, D. F., A. D. THOMPOSN, Y. SHEN, H. A. PHILLIPS, R. I. RICHARDS, *et al.*, 1993 Incidence and origin of "null" alleles in the (AC)n microsatellite markers. Am. J. Hum. Genet. **52:** 922–927.

CHOVNICK, A., W. GELBART and M. MCCARRON, 1977 Organization of the Rosy locus in *Drosophila melanogaster*. Cell **11:** 1–10.

DALLAS, J. F., 1992 Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. Mammal. Genome **3:** 452–456.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

EDWARDS, A., H. A. HAMMOND, L. JIN, T. CASKEY and R. CHACRABORTY, 1992 Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics **12:** 241–253.

EPPLEN, C., G. MELMER, I. SIEDLACZCK, F.-W. SCHWAIGER, W. MAUELER *et al.*, 1993 On the essence of "meaningless" simple repetitive DNA in eukaryote genomes, pp. 29–46 in *DNA Fingerprinting: State of the Science*, edited by S. D. J. PENA, R. CHAKRABORTY, J. T. EPPLEN and A. J. JEFFREYS. Birkhauser Verlag, Basel.

ESTOUP, A., L. GARNERY, M. SOLIGNAC and J. CORNUET 1995 Microsatellite variation in Honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. Genetics **140:** 679–695.

FITZSIMMONS, N. N., C. MORITZ and S. S. MOORE, 1995 Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. Mol. Biol. Evol. **12:** 432–440.

GARZA, J. C., M. SLATKIN and N. B. FREIMER, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12:** 594–603.

HILLIKER, A. J., S. H. CLARK and A. CHOVNICK, 1991 The effect of DNA sequence polymorphisms on intragenic recombination in the *rosy* locus of *Drosophila melanogaster*. Genetics **129:** 779–781.

LANZARO, G. C., Y. T. TOURE, L. ZHENG, F. C. KAFATOS and K. D. VERNICK, 1995 Microsatellite DNA and isozyme variability in a West African population of *Anopheles gambiae*. Ins. Mol. Biol. **4:** 105–112.

LEHMANN, T., W. A. HAWLEY, L. KAMAU, D. FONTENILLE, F. SIMARD *et al.*, 1996 Genetic differentiation of *Anopheles gambiae* from East and West Africa: comparison of microsatellite and allozyme loci. Heredity **77:** 192–200.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

LEWIN, B., 1994 *Genes IV,* Ed. 4. John Willey & Sons, New York.

LI, W.-H., and D. GRAUR, 1991 *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA.

LI, W.-H., C.-C. LUO and C.-I. WU, 1985 Evolution of DNA sequences, pp. 1–94 in *Molecular Evolutionary Genetics,* edited by R. J. MACINTYRE. Plenum Press, New York.

NEI, M. 1987 *Molecular Evolutionary Genetics.* Columbia University Press. New York.

PHILLIPS, H. A., V. J. HYLAND, K. HOMAN, D. F. CALLEN, R. I. RICHARDS *et al.*, 1991 Dinucleotide repeat polymorphism at D16S287. Nucleic Acids Res. **19:** 6664.

SCRIBNER, K. T., J. W. ARNTZEN and T. BURKE, 1994 Comparative analysis of intra- and interpopulation genetic diversity in *Bufo bufo,* using allozyme, single-locus microsatelllite, minisatellite and multilocus minisatellite data. Mol. Biol. Evol. **11:** 737–748.

SLATKIN, M., 1987 Gene flow and the geographic structure of natural populations. Science **236:** 787–792.

SLATKIN, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

SHRIVER, M. D., L. JIN, R. CHAKRABORTY and E. BOERWINKLE, 1993 VNTR Allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies and microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

WEBER, J. L., and C. WONG, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.

WEBER, J. L., M. H. POLYMEROPOULUS, P. E. MAY, A. E. KWITEK, H. XIAO *et al.*, 1991 Mapping of human chromosome 5 microsatellite DNA polymorphisms. Genomics **11:** 695–700.

Communicating editor: W.-H. LI