A Microsatellite-Based Multilocus Screen for the Identification of Local Selective Sweeps

Christian Schlötterer[1]

Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, 1210 Wien, Austria

Manuscript received July 19, 2001
Accepted for publication November 5, 2001

ABSTRACT

With the availability of completely sequenced genomes, multilocus scans of natural variability have become a feasible approach for the identification of genomic regions subjected to natural and artificial selection. Here, I introduce a new multilocus test statistic, ln RV, which is based on the ratio of observed variances in repeat number at a set of microsatellite loci in two groups of populations. The distribution of ln RV values captures demographic history of the populations as well as variation in microsatellite mutation among loci. Given that microsatellite loci associated with a recent selective sweep differ from the remainder of the genome, they are expected to fall outside of the distribution of neutral ln RV values. The ln RV test statistic is applied to a data set of 94 loci typed in eight non-African and two African human populations.

IT is well understood that genetic change provides the basis for adaptation processes in natural and domesticated populations. Hence, the identification of those genetic changes causing a phenotype with an increased fitness has been of long-standing interest in biological sciences.

Three different approaches to identify targets of selection (and thus adaptation) have been pursued: (1) the candidate gene approach, (2) QTL mapping, and (3) the multilocus screen.

The candidate gene approach is based on an *a priori* knowledge about the function of a given gene. The ease of PCR amplification and DNA sequencing, combined with the availability of several test statistics to evaluate the statistical significance of the observed and expected patterns of DNA sequence variation (Otto 2000), has resulted in numerous studies using a candidate gene approach. Despite the unquestionable importance of these studies in understanding the partitioning of genetic variation in natural populations, this approach is limited to a small number of candidate genes. Hence, a screen for genes involved in adaptation is difficult to pursue as neither the traits nor their genetic basis are known.

QTL mapping (Lynch and Walsh 1998) is a more general approach. On the basis of the idea that many traits are of quantitative nature, QTL mapping aims to partition the phenotypic variance into a genotypic and environmental component. While this approach is becoming increasingly popular to identify genes contributing to a given trait, it suffers from the problem that the phenotypic trait of potential adaptive relevance must be known. Limited information is available, however, about the traits that are responsible for the adaptation of natural populations to their environment. Thus, QTL mapping has only limited potential for the identification of the genes that are involved in the adaptation process of natural populations.

The key for a multilocus screen is the idea that different forces act in characteristic ways on the genome. While genetic drift, migration, and inbreeding affect all loci to the same extent, selection is targeted to a few loci only. Hence, a locus, which shows a significantly different pattern from the remainder of the genome, is expected to reside in a genomic region that has been the target of selection. This idea was first used by Cavalli-Sforza (1966), who calculated *F* values over several human groups. Later, Lewontin and Krakauer (1973) proposed a formal test statistic to identify loci that deviate from a neutral pattern. This test statistic is based on the variance of the inbreeding coefficient *F*, which is proportional to the square of its mean value averaged across loci. This test has been subsequently criticized for several reasons. In particular, correlations in allele frequencies, which could be caused by stepping-stone migration and phylogenetic history, will inflate the variance in *F* relative to the expectations. Furthermore, skewed allele frequencies will also affect the Lewontin-Krakauer test (Nei and Maruyama 1975; Robertson 1975). Despite some recent improvements (Tsakas and Krimbas 1976; Bowcock *et al.* 1991; Beaumont and Nichols 1996; Vitalis *et al.* 2001) these test statistics have never been widely used to infer selection from multilocus data.

With the recent progress in genomics, various new markers, which are distributed over the genome at a high density, have become available. In combination with

[1] *Address for correspondence:* Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, Josef Baumann Gasse 1, 1210 Wien, Austria. E-mail: christian.schloetterer@vu-wien.ac.at

the (almost) complete genomic sequence of various organisms, multilocus screens should be reconsidered.

While the high density of available single nucleotide polymorphisms (SNPs) makes them the marker of choice for various studies, such as linkage disequilibrium mapping, they are biallelic markers with a limited information content of a single marker. Microsatellites, on the other hand, are less dense, but offer the advantage of a multiallelic marker, which is highly informative.

In this study I explore the potential of microsatellites to serve as a genetic marker for the identification of genomic regions that have been subject to selection. While microsatellites are unlikely to be the target of natural selection, linkage to a genomic region that has been the target of selection is expected to cause a deviation from neutral expectations. The spread of a novel beneficial mutation through a population results in a reduction of natural variability at the selected locus and flanking regions (Maynard Smith and Haigh 1974; Slatkin 1995a). The extent to which flanking sequences are affected by such a selective sweep depends largely on the strength of selection and the recombination rate. Hence, a microsatellite locus linked to a beneficial mutation is expected to have a reduction in variability below neutral expectations (Slatkin 1995a; Schlötterer *et al.* 1997; Pritchard and Feldman 1998; Wiehe 1998; Schlötterer and Wiehe 1999). Thus, a multilocus screen for genomic regions subjected to selection could take advantage of this reduction in variability.

This conceptionally simple approach is significantly hampered by the observed differences in variability among microsatellite loci. In neutrally evolving populations different coalescent times and variation in mutation rates are responsible for those differences. Hence, the goal of a multilocus test for selected genomic regions is the identification of those microsatellite loci that deviate from the neutrally evolving genome. While the variation in coalescent time can be estimated under certain assumptions of population history, the mutation rate of a microsatellite locus remains an unknown parameter. Here, I introduce a new test statistic, ln RV, which accounts for neutral variation in coalescent times and different microsatellite mutation rates. A microsatellite data set is used to evaluate the usefulness of the ln RV test statistic to identify genomic regions that differ between African and non-African human populations.

MATERIALS AND METHODS

Population samples: Microsatellite data from 10 different populations were analyzed. African populations were represented by Biaka Pygmies from the Central African Republic and the Mbuti Pygmies from northwestern Zaire. Non-African populations included a sample of unrelated Danish blood donors, a moslem community from Northern Israel, Han Chinese living in the United States, native Japanese from the Osaka area or visitors to Stanford or Yale, the Yakut from Siberia, the Nasioi from Melanesia, the Mayan from Mexico, and the Rondonian Surui from Brazil. More information about these populations is available at http://info.med.yale.edu/genetics/kkidd/pops.html.

Genetic markers used: Data from a total of 94 microsatellite loci were used. The loci are part of the ABI linkage panels 8–11 and 13–16 covering the chromosomes 5–11. All data were taken from the Kidd lab webpage: http://info.med.yale.edu/genetics/kkidd/abiinfo.html. GenBank searches were performed before March 2001.

Test of neutrality (ln RV test): Assuming the stepwise mutation model (Ohta and Kimura 1973), neutrality, and mutation drift equilibrium, the variance in repeat number (V) is a good estimator of microsatellite variability (Moran 1975; Goldstein *et al.* 1995; Slatkin 1995b):

$$E[V] = 4N_e\mu. \quad (1)$$

N_e is the effective number of diploid individuals and μ the microsatellite mutation rate. Given that microsatellite mutation rates differ substantially among loci (Di Rienzo *et al.* 1998; Harr *et al.* 1998), it is difficult to compare variances among loci directly. This problem can be circumvented by calculating the ratio of the variance in repeat number in two populations, which is independent of the mutation rate. It has to be noted that the expectation of RV is not identical to the ratio of the expectations of V_{Pop1} and V_{Pop2}. Computer simulations, however, indicate that over a reasonable range of parameters the two expectations are very similar (Table 1):

$$E[\mathrm{RV}] = E\left[\frac{V_{Pop1}}{V_{Pop2}}\right] \cong \frac{4N_{e_{Pop1}}\mu}{4N_{e_{Pop2}}\mu}. \quad (2)$$

A better approximation is provided by the delta method (Lynch and Walsh 1998):

$$E[\mathrm{RV}] \cong \frac{4N_{e_{Pop1}}\mu}{4N_{e_{Pop2}}\mu}\left(1 + \frac{V(V_{Pop2})}{(4N_{e_{Pop2}}\mu)^2}\right) \cong \frac{4N_{e_{Pop1}}\mu}{4N_{e_{Pop2}}\mu}\left(1 + \frac{1}{12}\right). \quad (3)$$

Higher-order approximations given in Lynch and Walsh (1998) are not included because of the large term $1/12$. Computer simulations show that (3) provides a better fit than (2) (Table 1).

Given the close fit of the approximation, it can be assumed that RV is independent of the mutation rate and all loci have approximately the same expectation for a comparison of two populations. Nevertheless, historic sampling causes variation in the coalescent times at the loci studied. Hence, a distribution of RV values is expected. To determine the shape of this distribution, I used computer simulations (see below) and found that for neutrally evolving microsatellite loci the ln RV values follow a Gaussian distribution.

Hence, it is possible to design a test statistic to identify individual microsatellite loci that deviate from neutral expectations. Assuming that most loci are evolving neutrally, the mean and standard deviation of the observed ln RV values could be used to describe the corresponding Gaussian distribution. Using the density function of the Gaussian distribution, it is possible to assign a P value to the ln RV value of each locus. The P values give the probability that a given ln RV value is consistent with the null hypothesis of a neutral evolution.

Test for normal distribution: Visual inspection of the distribution of ln RV values from computer simulations suggested that they are normally distributed. For a formal test, two different strategies were pursued. First, the nonparametric Kolmogorov-Smirnov test was used to evaluate the distribution of 1000 simulated ln RV values. Because the tail of the distribution is particularly important to define the significance level, I also constructed a "tail test." This test is based on two proper-

TABLE 1

Verification of approximations by computer simulation

Θ_1:	1	2	5	1	2	5
Θ_2:	10	10	10	1	2	5
$\ln(E[V1]/E[V2])^a$	−2.303	−1.609	−0.693	0	0	0
$\ln((E[V1]/E[V2])(1 + 1/12))^b$	−2.494	−1.744	−0.751	0	0	0
$E[\ln(V1/V2)]^c$	−2.660	−1.703	−0.740	0.021	0.007	0.003
Standard deviation of $\ln(E[V1]/E[V2])^c$	1.924	1.390	1.265	2.288	1.424	1.268

[a] Corresponds to Equation 2.
[b] Corresponds to Equation 3.
[c] Each simulation is based on 10,000 loci and 50 individuals.

ties of a normal distribution. First, the distribution is symmetrical with the same number of data points in the upper and lower tail. Second, 95% (99%) of the values of a standardized distribution are expected to fall within the interval between −1.96 (−2.58) and 1.96 (2.58). Hence, Fisher's exact test could be used to test whether or not the number of observations falling in the tail fits the expectations. I determined the significance from 1000 simulated ln RV values using the 1 and 5% tail. A distribution was considered to be normally distributed if the Kolmogorov-Smirnov test and the two-tail tests were not significant ($P < 0.05$).

Computer simulations: The coalescent process, which describes the genealogical history of chromosomes, provides a very simple approach to simulate population samples (Hudson 1990). I made the standard assumptions associated with the coalescent process including neutrality, constant population size, and panmixia.

If not stated otherwise, between 100 and 10,000 loci were simulated for two independent populations using the unbiased stepwise microsatellite mutation model (Ohta and Kimura 1973; Goldstein *et al.* 1995). For each simulated locus, the ln RV test statistic was calculated. When variation in microsatellite mutation rates was incorporated in the computer simulations, mutation rates varied by a factor 10 drawn from a uniform distribution. For these simulations the mean Θ-values are reported. For a restricted set of parameters, computer simulations were run with a two-phase mutation model of microsatellites (Di Rienzo *et al.* 1994). In addition to single repeat changes, a given fraction of microsatellite mutations was allowed to mutate by more than one repeat unit. The size change for such mutations was drawn from a uniform distribution ranging from 1 to a specified maximum.

The influence of demographic events, such as bottleneck and population expansion, was studied by a modification of the constant population size model. All demographic events affect the entire genome; therefore all loci were simulated using the same algorithm. A bottleneck was modeled as suggested by Hudson (1990). For the computer simulations of the population expansion, an instantaneous rise in population size was assumed.

To study the effect of admixture, I modified a recently proposed method (Pritchard *et al.* 2000) and simulated 100 chromosomes from three independent populations each. A set of randomly selected chromosomes was taken from population one and replaced the same number of chromosomes in population two. Rather than simulating two additional generations for the admixed populations, the ln RV test statistic was directly calculated for populations two and three.

The neutral coalescent simulations could be modified to study the properties of a single microsatellite locus, which is linked to a genomic region subjected to directional selection. I assumed an instantaneous selective sweep, which was simulated as a bottleneck occurring at the selected locus only. Hence, one locus in one of the two populations was simulated under the selection model, while all other loci were simulated under the constant population size model.

RESULTS

Verification of the test statistic: To explore the behavior of the ln RV test statistic computer simulations were performed under the following assumptions: neutrality, a constant population size, random mating, mutation drift equilibrium, no linkage of the microsatellite loci, and independence of the two populations. Using standard coalescent simulations (Hudson 1990), I obtained the variance in repeat number for a set of microsatellite loci. If not stated otherwise, computer simulations assumed the unbiased stepwise mutation model (Ohta and Kimura 1973; Goldstein *et al.* 1995).

Dependence on the mutation rate: Using a wide range of Θ-values (2–100) consistently resulted in a distribution of ln RV values, which was very similar to a Gaussian distribution (Figure 1, A and B). Based on 1000 loci and a sample of 100 chromosomes, no significant deviation from normality could be detected (Kolmogorov-Smirnov and tail test, $P > 0.1$). This observation is consistent with previous computer simulations (Goldstein *et al.* 1996; Pritchard and Feldman 1998) and empirical reports (Harr *et al.* 1998), which demonstrated that ln V generally approximates a normal distribution.

Microsatellite mutation rates vary by more than one order of magnitude (Di Rienzo *et al.* 1998; Harr *et al.* 1998). To account for this, microsatellite mutation rates were drawn from a uniform distribution, resulting in an up to 10-fold variation in mutation rate (Figure 1C). Simulations of 1000 loci for 100 chromosomes each did not result in statistically significant deviations from normality (Kolmogorov-Smirnov and tail test, $P > 0.1$). Finally, I also tested the influence of differences in population sizes among the groups compared (Figure 1D). The ratios of the effective population sizes were varied

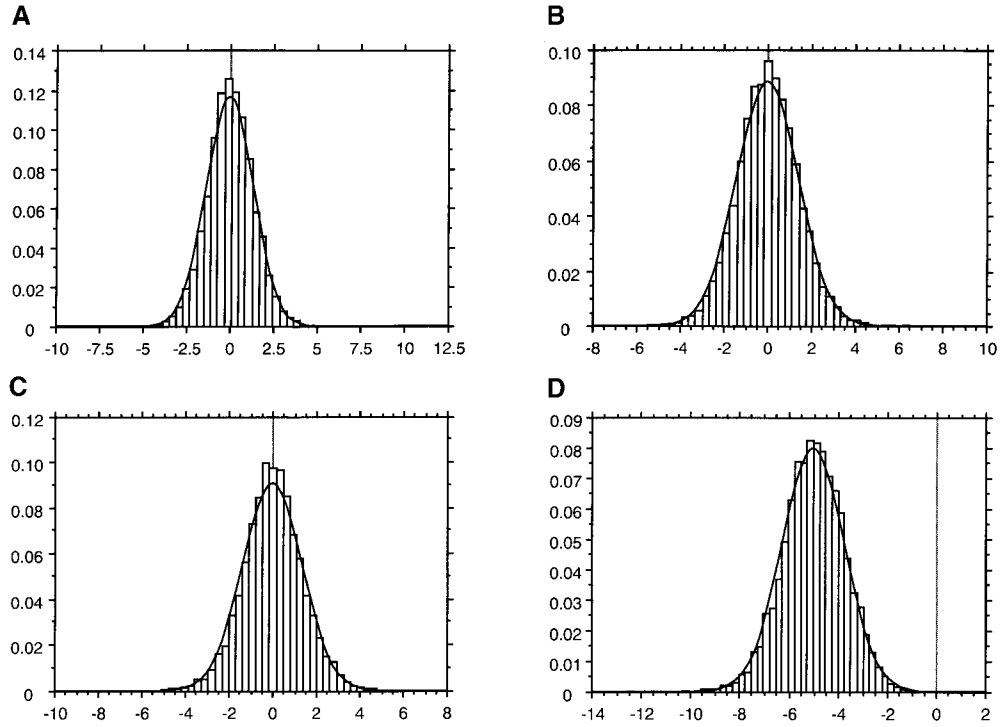

FIGURE 1.—Distribution of ln RV values as obtained from coalescent simulations of 10,000 independent microsatellite loci and 100 sampled chromosomes. The parameters used for the simulations are Θ-population1/Θ-population2/variation in mutation rate among loci (in percentages): (A) 5/5/0; (B) 500/500/0; (C) 5/5/1000; (D) 5/500/0. The variation in mutation rate was simulated on the basis of a uniform distribution.

(from 1:1 to 1:100) and no deviation from a normal distribution could be detected for 1000 simulated microsatellite loci (Kolmogorov-Smirnov and tail test, $P > 0.1$, 100 chromosomes). Hence, using a wide range of parameters, the ln RV test statistic can be approximated by a Gaussian distribution. This greatly facilitates the design of a statistical test to detect deviation from neutrality, as no *a priori* knowledge about the mutation rate or population size of the tested populations is required.

Deviation from stepwise mutation model: Inference from population data (DI RIENZO *et al.* 1994) and direct observations (WIERDL *et al.* 1997; BRINKMANN *et al.* 1998; HARR and SCHLÖTTERER 2000) indicated that microsatellite mutations are not confined to single repeat unit changes, but could also encompass larger gains and losses. To investigate whether such a modification of the mutation process affects the ln RV test statistic, I simulated 1000 microsatellite loci for two populations and a sample size of 100 chromosomes. No deviation from the normal distribution could be detected (Kolmogorov-Smirnov and tail test, $P > 0.3$, Table 2). The only notable difference was an increase in the variance of ln RV values with both a larger step size and a higher proportion of loci not evolving by single repeat unit changes (Table 2).

Power of the ln RV test statistic: To assess the power of the ln RV test statistic I simulated the variance in repeat number for 100 microsatellite loci of which one microsatellite locus was associated with a selective sweep. The rates of recombination between the selected site and the microsatellite as well as the strength of selection are two important parameters required for model selection. Given the large uncertainty for each of these parameters, I used the reduction in variability (r) at the marker locus as a compound parameter in the computer simulations. A strong reduction in variability could result from a large selection coefficient, tight linkage to the selected site, or both. Consistent with expectation, a more pronounced reduction in variability resulted in larger numbers of simulation runs with significant ($P < 0.05$) ln RV values (Table 3). Also, the mean ln RV of the selected locus was higher and had a lower variance when a large r was used. Hence, for a recent and strong reduction in variability, a large fraction of the selected loci will be identified by the ln RV statistic. Some differences could be detected between the simulations using different Θ-values (Table 3). In comparison to the large effect of r, the influence of Θ was found to be moderate.

Table 4 indicates that the power of the ln RV statistic

TABLE 2

Variance of the ln RV test statistic based on computer simulations of 1000 loci in two neutrally evolving populations under the two-phase microsatellite mutation model

	$K = 0$	$K = 0.2$	$K = 0.4$
$S = 5$	1.44*	1.98*	1.90*
$S = 10$		2.66*	1.99*

K is the probability of a mutation encompassing more than one repeat unit, and S is the upper boundary of size change by a single mutation event. $\Theta = 5$ (in both populations). *No deviation from normal distribution by Kolmogorov-Smirnov and tail tests ($P > 0.3$).

TABLE 3

Power of the ln RV test statistic in dependence of *r*

	$\overline{\Theta} = 3$			$\overline{\Theta} = 6$			$\overline{\Theta} = 30$		
	$\overline{\ln RV}$	Variance	FS	$\overline{\ln RV}$	Variance	FS	$\overline{\ln RV}$	Variance	FS
$r = 0.1$	−1.92	2.45	0.41	−2.03	1.86	0.46	−1.65	0.81	0.34
$r = 0.05$	−2.51	2.55	0.56	−2.48	1.95	0.63	−2.10	0.84	0.53
$r = 0.01$	−3.17	2.24	0.79	−3.24	1.83	0.86	−2.75	0.65	0.83

One locus was subjected to directional selection and 99 loci evolved neutrally. *r*, fraction to which variability was reduced due to linkage to a selected site; FS, fraction of significant ($P < 0.05$) simulations; $\overline{\ln RV}$, mean ln RV value of the selected locus over 1000 simulation runs; Variance, variance of ln RV at the selected locus over 1000 simulation runs with *t* (time elapsed since the selective sweep in $2N_e$) set to $0.02\,N_e$.

significantly decreases with the time elapsed since the selection (*t*) occurred. Only recent selective sweeps could be detected reliably. This observation is fully consistent with previous analytical results (Wiehe 1998). Similar to the simulations for which *r* was varied, Θ had only a moderate influence on the power of the *lnRV* test.

Influence of the sample size: Despite the continuous improvement in screening technologies, the analysis of large sample sizes is still an important hurdle in population genetics. Therefore, it is interesting to determine the influence of the number of sampled chromosomes on the ln RV test statistic. The power of the ln RV test statistic is dependent on the shape of the distribution of ln RV values. A larger variance in ln RV values requires a more extreme reduction in variability to obtain significance (Table 5). Therefore, I calculated the standard deviation of ln RV over 10,000 loci. Each ln RV value was simulated for different sample sizes (10–1000 chromosomes). Figure 2 clearly indicates that <30 chromosomes result in a large standard deviation, which will, in turn, result in a lower power of the ln RV test statistic. On the other hand, samples of >50 chromosomes will not significantly improve the test statistic. Hence, only a moderate number of individuals need to be typed to determine the significance level of the ln RV test statistic.

Influence of demographic events: Computer simulations were used to investigate the influence of common demographic events (population expansion, bottlenecks, and admixture) on the distribution of ln RV values in neutrally evolving populations. In all simulations one population was kept at a constant size, while for the other population either a change in size or admixture was simulated.

Despite that computer simulations covered quite radical population size changes, for most simulations the ln RV values were not found to deviate significantly from a normal distribution. The most notable exceptions were recent and strong bottlenecks in combination with a low Θ (Table 6). Under such extreme scenarios, microsatellite loci did not recover variability, resulting in an excess of loci with low variability in the bottlenecked population. Interestingly, a bottleneck occurring $0.1N_e$ generations ago resulted in a significant tail test ($P = 0.038$) for the population with a large Θ (Table 6). For expanding populations only the combination of large Θ-values with an older population expansion resulted in a significant deviation from a normal distribution (Table 7). This deviation is most likely the result of a large diversity generated in those samples, which was not adequately sampled with 100 chromosomes. When the sample size was increased to 200 chromosomes, no significant deviation from normality could be detected (data not shown). No significant deviation from a normal distribution was observed for various admixture pro-

TABLE 4

Power of the ln RV test statistic in dependence of *t*

	$\overline{\Theta} = 3$			$\overline{\Theta} = 6$			$\overline{\Theta} = 30$		
	$\overline{\ln RV}$	Variance	FS	$\overline{\ln RV}$	Variance	FS	$\overline{\ln RV}$	Variance	FS
$t = 0.1$	−1.46	1.31	0.26	−1.56	0.80	0.30	−1.51	0.49	0.26
$t = 0.05$	−1.93	1.41	0.41	−2.07	0.96	0.53	−1.89	0.53	0.42
$t = 0.01$	−3.17	2.24	0.79	−3.24	1.83	0.86	−2.75	0.65	0.83

One locus was subjected to directional selection and 99 loci evolved neutrally. *t*, time point at which selection occurred (in $2N_e$ generations); FS, fraction of significant ($P < 0.05$) simulations; $\overline{\ln RV}$, mean ln RV value of the selected locus over 1000 simulation runs; variance, variance of ln RV at the selected locus over 1000 simulation runs, with *r* (reduction in variability due to the selective sweep) set to 0.01.

TABLE 5

Power of the ln RV test statistic in relation to the variance of ln RV across loci

	Variance in ln RV		
r	2.7	1.7	1.5
0.01	0.76	0.84	0.90
0.1	0.35	0.42	0.48

The power is expressed as the fraction of simulations that identified a locus linked to a selective sweep as a significant ($P < 0.05$) outlier. r, fraction to which variability was reduced due to linkage to a selected site; the variance in ln RV was determined over 1000 replicate simulations of 99 microsatellite loci. One locus was assumed to be linked to a selected site. Different sample sizes in the computer simulations were used to obtain the differences in ln RV.

TABLE 6

Variance of ln RV when one population had passed through a bottleneck

	$\Theta = 3$	$\Theta = 6$	$\Theta = 30$
No bottleneck	1.56	1.58	1.97
$t = 0.1$	1.60	1.20	1.26**
$t = 0.05$	2.48***	1.32	1.30
$t = 0.01$	5.98***	2.02	1.48

A total of 10,000 microsatellite loci were simulated for two populations and r was set to 0.1. t, time (in $2N_e$) elapsed since the bottleneck. Significant deviations from normal distribution: $^{**}P < 0.05$, tail test; $^{***}P < 0.05$, Kolmogorov-Smirnov test and tail test.

portions as well as different effective population sizes of the source population (Table 8).

The power of the ln RV test statistic depends strongly on the behavior of the neutrally evolving loci. If the ln RV values have a broad distribution (large variance of ln RV), then the identification of a selected locus is more difficult. On the other hand, a very narrow distribution of ln RV values makes the identification of selected loci easier (Table 5). Given that the power of the ln RV test varies with the parameters used for the computer simulation (see above), a systematic power assessment is difficult. Therefore, I use the variance of the ln RV values as an indication for the power of the ln RV test under various demographic scenarios.

Table 6 indicates that population bottlenecks could have quite complex effects on the behavior of the ln RV test statistic. Simulations based on Θ-values of six produced a wider distribution of ln RV when one population recently ($t = 0.01$) went through a bottleneck. For those simulations, which are based on large Θ-values, a bottleneck resulted in a more narrow distribution of ln RV.

Population expansions were simulated using a wide range of times since expansion (t) and factors (r) by which the population size changed. Irrespective of the parameters used, population expansions always resulted in a smaller variance of ln RV values (Table 7).

Various proportions of admixture from a third population were simulated. As expected, admixture increased the variability in the admixed population, resulting in a shift of mean ln RV values (Table 8). This was also observed if immigrants from a population with a smaller effective population size replaced a large fraction (0.25) of the population (Table 8). The variance in ln RV values, however, was largely unaffected by admixture. Hence, the power of the ln RV test statistic is not significantly influenced by admixture.

Screening for adaptive mutations in the human genome: Data from mtDNA and microsatellites suggest that human populations left Africa about 100,000 years ago to colonize the rest of the world (Jorde *et al.* 1998). This migration challenged human populations in the form of a novel environment. Hence, a comparison of African and non-African populations could potentially

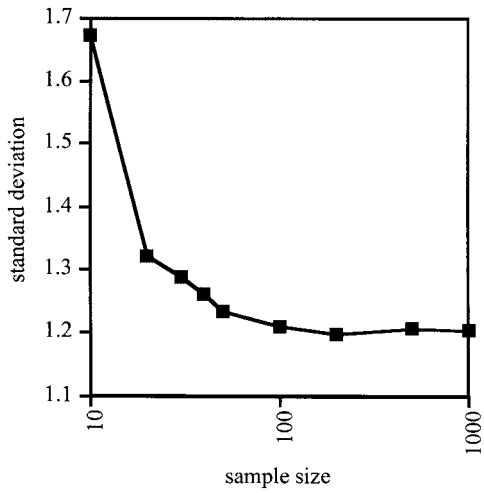

Figure 2.—Influence of the sample size (in chromosomes) on the standard deviation of the ln RV test statistic. Standard deviations were measured on 10,000 independently simulated microsatellite loci using the parameters 5/5/0 (see Figure 1 for further explanations).

TABLE 7

Variance of ln RV with one recently expanded population

	$\Theta = 3$	$\Theta = 6$	$\Theta = 30$
No expansion	1.56	1.58	1.97
$t = 0.1, f = 10$	0.97	0.95	1.36*
$t = 0.1, f = 100$	1.16	1.06	1.24*
$t = 0.01, f = 10$	1.33	1.29	1.95
$t = 0.01, f = 100$	0.91	0.91	1.34

A total of 10,000 microsatellite loci were simulated for two populations. t, time (in $2N_e$) since the expansion; f, factor by which the population expanded. Significant deviations from normal distribution: $^{*}P < 0.05$, Kolmogorov-Smirnov test.

TABLE 8

Mean ln RV values of computer simulations with one population experiencing admixture from a third population (variance of ln RV over 1000 loci)

	Proportion of admixture			
Θ	0	0.05	0.10	0.25
5	0.04*	0.22*	0.28*	0.45*
	(1.57)	(1.20)	(1.37)	(1.48)
20		0.47*	0.72*	1.08*
		(1.48)	(1.17)	(1.46)
2		0.11*	0.14*	0.18*
		(1.43)	(1.39)	(1.33)

Θ of the two test populations was set to 5, while Θ from the source population from which admixture occurred was varied. Simulations were based on 100 chromosomes per population. *$P > 0.3$, no deviation from normal distribution by Kolmogorov-Smirnov and tail tests.

identify genomic regions that were involved in adaptation processes in the two groups. Using the ln RV test statistic, it should be possible to identify some candidate regions bearing an adaptive mutation. In this report I used a data set consisting of 94 microsatellite loci, which were typed in 10 human populations, 2 African and 8 non-African. To apply the ln RV test statistic, I averaged the observed variances in repeat number in the non-African and African groups for each locus. The distribution of the ln RV values of the 94 microsatellite loci followed a Gaussian distribution (Kolmogorov-Smirnov test, $P = 0.94$). Out of the 94 loci analyzed, 4 loci had a ln RV value located outside the 95% confidence interval. Two loci had more variation in the non-African populations than expected by the level of variation detected in African populations (D10S249, $P = 0.002$; D6S305, $P = 0.023$). Microsatellite loci D6S462 ($P = 0.007$) and D10S197 ($P = 0.018$) had a reduced variability in non-African populations. Because the number of outliers is fully consistent with the neutral expectations, I evaluated the allele distribution of the two loci, which showed the strongest deviation from the remainder of the genome (D10S249 and D6S462). Figure 3 shows the allele distribution of both loci in the pooled African and non-African populations. Consistent with expectations under the selective sweep hypothesis, each locus showed a strongly peaked allele distribution in the population with reduced variability, while the other population had a scattered allele distribution.

Until now, populations were jointly analyzed as African or non-African groups. An alternative approach for the identification of loci that differ between African and non-African populations would be to make individual comparisons of all African against all non-African populations. Even under neutrality, 5% of the loci will be identified as significant outliers by the ln RV test. When

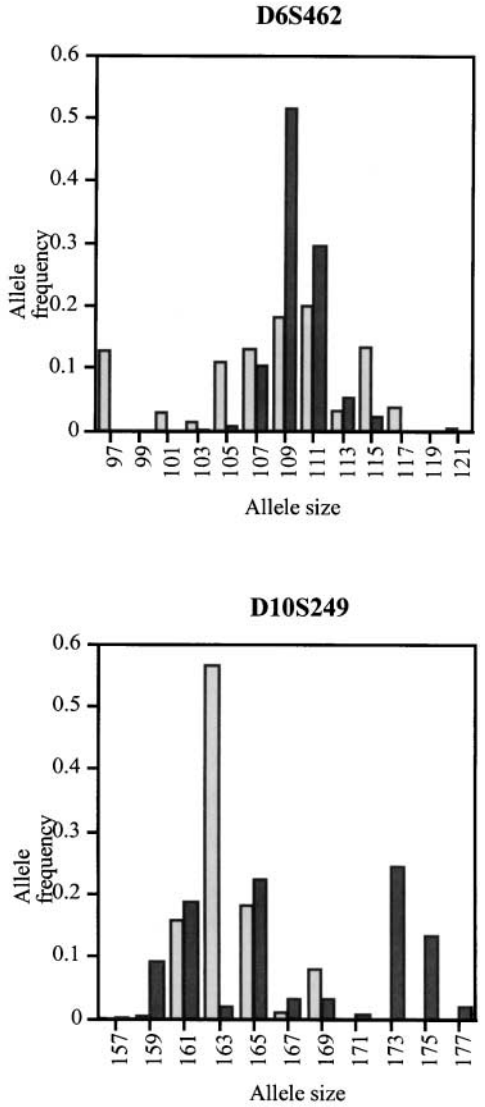

FIGURE 3.—Allele frequency distribution at the two microsatellite loci with the most extreme ln RV values in African and non-African populations. (□) Africa, (■) non-Africa.

multiple pairs of independent populations are compared, neutral outliers are expected to be confined to one comparison, but selected loci should be significant in all comparisons. Despite that neither African nor non-African populations are independent, I compared all African populations against each non-African population. In 16 pairwise comparisons of 94 microsatellite loci, 72 significant outliers were marked by the ln RV test. The probability of observing x significant ln RV tests for a given locus could be calculated by a binomial distribution. Loci D6S462 and D10S249 were significant in 9 and 16 comparisons, respectively, which would be extremely unlikely in 16 independent comparisons ($P < 0.0000001$). While the P values should not be taken at face value, given that the comparisons were not independent, Figure 4 clearly indicates that both loci are different from the remainder of the genome.

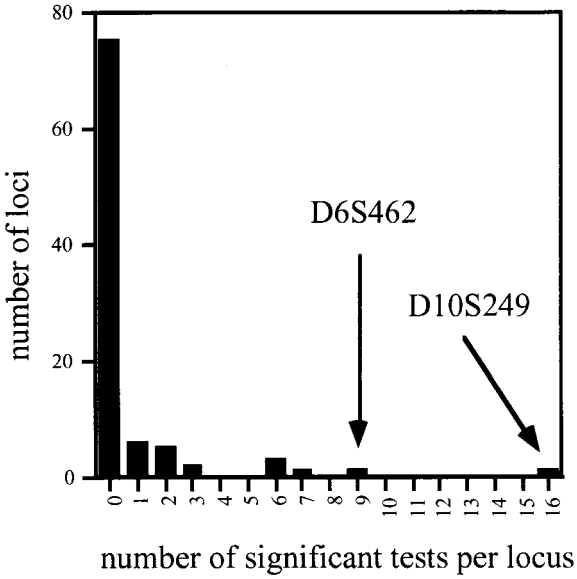

FIGURE 4.—Frequency distribution of the number of significant ($P < 0.05$) pairwise comparisons for all possible comparisons of African and non-African populations.

DISCUSSION

The interpretation of natural variability has been of long-standing interest in population genetics. Natural variability at a given locus is governed by various factors: mutation rate, effective population size, historic sampling, population demography, and selection. Any attempt to identify targets of selection in the genome is challenged by the need to account for the pattern expected under neutrality. In principle, each site in the genome may have its own specific neutral mutation rate. On the other hand, effective population size, demographic history, and historic sampling variation are shared across sites (at least for autosomes). Hence, it would be desirable to have a joint estimator of the parameters common to all loci and to adjust for differences in mutation rate.

The central variable of the new test statistic is ln RV. For every microsatellite locus analyzed, the ratio of the variance in repeat number is calculated for two populations. This ratio has the same expectation independent of the mutation rate of a given locus. Hence, ln RV values calculated for a number of microsatellite loci are independent of the mutation rate, but reflect population-specific parameters including effective population size and historic sampling. Computer simulations indicate that the distribution of ln RV follows to a very good approximation a Gaussian distribution. Thus, the mean and standard deviation summarize the neutral expectations of ln RV for a set of two populations.

Influence of demographic events on the ln RV test statistic: In contrast to selection, demographic events affect the entire genome. Hence, similar to the demographic model of a constant population size, selection at a genomic region may be detected by a deviation from the remainder of the genome. Computer simulations have been used to study population expansion, bottlenecks, and admixture. Two different aspects of the ln RV test were examined under those demographic scenarios: first, whether ln RV remains normally distributed, and second, the power of the ln RV statistic.

Distribution of ln RV: For some extreme demographic events, such as a recent and strong bottleneck, ln RV is no longer normally distributed (Table 6). This deviation is caused by a large number of microsatellite loci, which have lost almost all variability. It is obvious, however, that a data set containing a large number of loci with no or very little variability cannot be informative to infer a recent selective sweep at one or a few loci by the reduction in variability. Therefore, I do not consider this deviation as a major limitation for the application of the ln RV test. More serious is the deviation from normality that was observed for an old population expansion for highly variable loci (large Θ). While a larger sample size could solve this problem, these simulations indicated that it may be advisable to test the obtained ln RV values for normality before applying the ln RV test.

Overall, the distribution of ln RV values could be approximated by a normal distribution for most of the parameters of the demographic scenarios considered, suggesting that the ln RV test could also be applied for a wider range of demographic events than just constant population sizes.

In this article I did not consider the effect of population substructure within each of the two populations compared. While further computer simulations are required to determine influence of population structure on the distribution of ln RV values, it has to be noted that the effective population sizes can be determined for any hierarchical level of population structure (CHESSER *et al.* 1993). As under neutrality all autosomal loci have the same effective population size, the ln RV test statistic is most likely not affected by population substructure.

The independence of the ln RV test statistic for most of the demographic scenarios analyzed is in sharp contrast to many other statistical tests to identify selection, such as tests for linkage disequilibrium (DEPAULIS 1998; ANDOLFATTO *et al.* 1999; KOHN *et al.* 2000; VIEIRA and CHARLESWORTH 2000). These tests could be highly sensitive to admixture, which significantly complicates the identification of selected regions in the genome.

Power of the ln RV test: I estimated the power of the ln RV test for the three demographic scenarios considered by the variance of ln RV values. Given that for constant population sizes, the power of the ln RV statistic increased with a smaller variance of ln RV (Table 5), I assumed that this also applies to other demographic scenarios as long as ln RV follows a normal distribution. Exact power estimates, however, would require com-

puter simulations of the joint effects of selection and a given demographic event. While this may lead to slightly different power estimates, the overall picture is unlikely to be affected.

For all parameters evaluated population expansion resulted in a more narrow distribution of ln RV values (Table 7), suggesting a higher statistical power to detect local selective sweeps in growing populations. Admixture from a distantly related population (not included in the analysis) increases variability at all loci, resulting in a broader distribution of ln RV values (Table 8). Hence, admixture reduces the power of the ln RV test statistic. The effect of bottlenecks was less clear. For high Θ-values a bottleneck consistently resulted in a lower variance of ln RV when compared to a constant population size. Simulations based on intermediate Θ-values showed an increased variance of ln RV for very recent bottlenecks only (Table 6).

Limitations of the ln RV test statistic: Historic sampling is an important source of fluctuating variability in the genome. The ln RV test statistic uses the ratio of the observed variability at a given locus in two populations. Given that both populations are subjected to historic sampling, the ln RV test statistic has a considerable variability determined by historic sampling. The computer simulations assumed two completely unrelated populations, which maximizes the variation in ln RV due to historic sampling. Hence, less pronounced selective sweeps are very difficult to identify. This is reflected by the requirement of a strong recent reduction in variability to identify a selected locus with the ln RV test statistic. One possible approach to improve the power of the ln RV test statistic would be the comparison of two closely related populations. Closely related populations share a large fraction of their genealogy at each locus; hence, the variance of ln RV is expected to be significantly smaller than for distantly related populations. Thus, loci that have been exposed to a different selection regime in the two closely related populations should be easier to detect than in a comparison of two distantly related populations. Nevertheless, further simulation studies are required to verify the behavior of the ln RV test statistic for closely related populations or populations connected by gene flow.

The ln RV test statistic makes the important assumption of constant mutation rates across populations. This assumption could be easily violated given the strong correlation between microsatellite mutation rate and repeat number (Schlötterer *et al.* 1998; Harr and Schlötterer 2000). Furthermore, interruptions in the microsatellite repeat also reduce microsatellite mutation rates (Weber 1990). If the two populations differ in the average repeat number or an interruption in the repeat structure is more frequent in one population than in the other, this could result in a significant ln RV test statistic. Experimental evidence, such as sequencing of alleles, could provide further insight. Furthermore, a comparison of the allele distribution will also be informative (see below).

The ln RV test statistic uses the mean and standard deviation of the observed ln RV values to identify those loci that deviate from the remainder of the genome. The probability for each locus to deviate from the expectation can be directly inferred from the density function of the normal distribution. A larger number of loci results in a more accurate estimate of the mean and standard deviation, but also a larger number of loci with ln RV values located in the tail of the normal distribution. A generally accepted solution to this problem is the use of a smaller α-value, which reduces the number of false positives. Hence, the identification of loci subjected to selection becomes more difficult and the type 2 error increases. Computer simulations indicated that even an α-value of 0.05 results in a considerable type 2 error (Tables 3 and 4). A more practical approach is to use the ln RV test statistic as a first pass analysis for the identification of candidate regions. Whether or not an identified region has been subject to a selective sweep could then be further investigated by the analysis of flanking microsatellite loci, which are also expected to show the signature of a selective sweep (Wiehe 1998). Also, sequence analysis at the candidate locus could provide further evidence for a selective sweep when test statistics specific for sequence polymorphism are applied (Otto 2000). Additional evidence for a selective sweep could be obtained from those loci that have already recovered some variability after the sweep. The allele distribution of a locus that starts accumulating variation after the fixation of a single allele is strongly peaked. The more mutations occurred after the fixation, the broader the distribution becomes, until the random loss of alleles decays the single peaked distribution. Hence, after a sweep, the allele spectrum should be tighter than in nonsweep populations. Following the same rationale, multilocus test statistics based on the allele frequency distribution have been suggested to infer population size changes (Beaumont 1999; Reich *et al.* 1999; Garza and Williamson 2001). The power of an allele frequency distribution for a single locus, however, is generally weak and could only be used as confirmatory evidence.

Finally, the analysis of independent pairs of populations could serve as an additional tool for the verification of an identified deviation from neutral expectations. While outliers are expected to occur in a single comparison only, selected loci should be detected in most of the pairwise comparisons. Assuming independence among the populations the probability of observing multiple significant tests could be tested on the basis of the binomial distribution. In reality, however, populations are rarely independent, as they share a common history. Nevertheless, our analysis indicated that most of the significant ln RV tests between African and non-African populations occurred in a limited number of compari-

sons only. Out of 19 loci for which a significant ln RV test was recorded, 11 loci were found to be significant in 1 or 2 comparisons only. While it is impossible to rule out that some form of local selection has acted on those loci, the more likely explanation is that they are false positives. In any case, out of 16 tests the two candidate loci D6S462 and D10S249 had a significant ln RV value in 9 and 16 comparisons, respectively.

Much of the theory of the ln RV test is based on the variance in repeat number at mutation drift equilibrium. While the high mutation rate of microsatellites requires less time to reach mutation drift equilibrium, most natural populations are not expected to meet this condition. Computer simulations indicated that the normal distribution of ln RV seems to be quite robust to demographic events. Furthermore, no deviation from normality could be detected for the ln RV values of human populations. Further studies, in particular experimental ones with a large number of loci analyzed, will provide further insight into the behavior of the ln RV test statistic in natural populations.

Genomic regions associated with selective sweeps in human populations: The simple model of out-of-Africa-associated adaptive mutations would have predicted more loci with significantly reduced variability in non-African populations than in African populations. In the human data set of 94 microsatellite loci, however, the same numbers of outliers were observed on both sides of the ln RV distribution.

Locus D6S462 showed the strongest reduction in variability in the non-African populations, suggesting that this locus may have been linked to a genomic region that has swept in non-African populations. The approach to combining populations in African and non-African groups requires a consistently low level of variation across populations to result in a significant ln RV value. A separate analysis of the eight non-African populations against the pooled African populations indicated that each of the non-African populations had a reduced variability at locus D6S462 ($P < 0.065$, one-sided test). Given that the eight non-African populations covered a wide range of human diversity outside of Africa, the strong reduction in variability in the non-African populations is best explained by a selection event that coincided with the colonization. Further evidence for a recent selective sweep at D6S462 could be gleaned from the allele distribution. While the African population showed a scattered allele distribution, the non-African populations had a highly peaked allele distribution (Figure 3), a pattern that would be expected for an allele that has swept through the population and is starting to accumulate new mutations.

The recently published draft of the human genome (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001) could potentially indicate genes flanking the two candidate regions. While for D6S462 several flanking expressed sequence tags (ESTs) could be detected, a Genome Data Bank (GDB) search did not indicate known genes mapping to D6S462. Further analysis has to await the progress of the analysis of the human genome.

In contrast to the expectations for a selection event associated with the human habitat expansion out of Africa, the locus that deviates most from the remainder of the genome, D10S249, harbored a surplus of variability in non-African populations. Based on the allele spectrum at locus D10S249 (Figure 3), it is very likely that this locus has been subjected to a recent selective sweep. A BLAST search of the human subset of GenBank failed to identify locus D10S249 in the published draft of the human genome sequence. Thus, no information about flanking sequences is available. A GDB search indicated that locus D10S249 is located in the amplimer AFM207-wd12. The gene mapping closest to microsatellite D10S249 is called Severe Combined Immunodeficiency, Athabaskan type (SCIDA), a genomic region associated with both T-cell and B-cell immunity (Murphy and Stringer 1986). V(D)J recombination, which accounts for the diversity of T-cell receptor and immunoglobulin-encoding genes, is initiated by a specific double-strand break. The general DNA repair machinery is responsible for the resolution of this break. Previously, it was shown that an essential DNA repair/V(D)J recombination gene lies in the same region as SCIDA (Moshous *et al.* 2000). While it remains purely speculative until further proof (which will become feasible with the availability of the genomic sequence of this region), it is conceivable that a gene involved in immune defense is a potential target for adaptive mutations. Populations are constantly challenged by pathogen pressure and one way to counter this pressure is the acquisition of novel mutations to control pathogens.

Perspective: The introduced test statistic provides a means to search multilocus data to identify those loci that show a deviation from neutral expectations in one population (group). Given the inherent problem of a multilocus test statistic and the high type 2 error of the ln RV test statistic, it is obvious that loci identified as outliers by the ln RV test are no final proof of selection, but could serve as a starting point for subsequent studies.

I am particularly grateful to K. Kidd for making the data public on the web. Many thanks go to the C.S. lab, C. Haley, and R. R. Hudson for helpful discussions. R. Harding, B. Harr, M. van Staaden, and G. Muir provided comments on the manuscript. Special thanks to T. Wiehe for pointing out the glitches of the expectation of the ratio of two random variables. R. Bürger is acknowledged for his help in approximating the expectation of the ratio of two random variables. Three anonymous reviewers provided helpful comments, which improved the manuscript. W. Schlötterer helped with the C code. C.S. is supported by grants from the Fonds zur Förderung der wissenschaftlichen Forschung (FWF).

LITERATURE CITED

Andolfatto, P., J. D. Wall and M. Kreitman, 1999 Unusual haplotype structure at the proximal breakpoint of *In(2L)t* in a natural population of *Drosophila melanogaster*. Genetics **153:** 1297–1311.

Beaumont, M. A., 1999 Detecting population expansion and decline using microsatellites. Genetics **153:** 2013–2029.

Beaumont, M. A., and R. A. Nichols, 1996 Evaluating loci for use in genetic analysis of population structure. Proc. R. Soc. Lond. Ser. B **263:** 1619–1626.

Bowcock, A. M., J. R. Kidd, J. L. Mountain, J. M. Hebert, L. Carotenuto *et al.* 1991 Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. Proc. Natl. Acad. Sci. USA **88:** 839–843.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne and B. Rolf, 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. **62:** 1408–1415.

Cavalli-Sforza, L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. Ser. B **164:** 362–379.

Chesser, R. K., O. E. Rhodes, D. W. Sugg and A. Schnabel, 1993 Effective size for subdivided populations. Genetics **135:** 1221–1232.

Depaulis, F., 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. **15:** 1788–1790.

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.

Di Rienzo, A., P. Donnelly, C. Toomajian, B. Sisk, A. Hill *et al.* 1998 Heterogeneity of microsatellite mutations within and between loci and implications for human demographic histories. Genetics **148:** 1269–1284.

Garza, J. C., and E. G. Williamson, 2001 Detection of reduction in population size using data from microsatellite loci. Mol. Ecol. **10:** 305–318.

Goldstein, D. B., A. Ruiz Lineares, L. L. Cavalli-Sforza and M. W. Feldman 1995 An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463–471.

Goldstein, D. B., L. A. Zhivotovsky, K. Nayar, A. Ruiz Lineares, L. L. Cavalli-Sforza *et al.* 1996 Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol. Biol. Evol. **13:** 1213–1218.

Harr, B., and C. Schlötterer, 2000 Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics **155:** 1213–1220.

Harr, B., B. Zangerl, G. Brem and C. Schlötterer, 1998 Conservation of locus specific microsatellite variability across species: a comparison of two *Drosophila* sibling species *D. melanogaster* and *D. simulans*. Mol. Biol. Evol. **15:** 176–184.

Hudson, R. R., 1990 Gene geneologies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

International Human Genome Sequencing Consortium, 2001 Initial sequencing and analysis of the human genome. Nature **406:** 860–921.

Jorde, L. B., M. Bamsham and A. R. Rogers, 1998 Using mitochondrial and nuclear DNA markers to reconstruct human evolution. BioEssays **20:** 126–136.

Kohn, M. H., H. J. Pelz and R. K. Wayne, 2000 Natural selection mapping of the warfarin-resistance gene. Proc. Natl. Acad. Sci. USA **97:** 7911–7915.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175–195.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

Moran, P. A. P., 1975 Wandering distributions and electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

Moshous, D., L. Li, R. Chasseval, N. Philippe, N. Jabado *et al.*, 2000 A new gene involved in DNA double-strand break repair and V(D)J recombination is located on human chromosome 10p. Hum. Mol. Genet. **9:** 583–588.

Murphy, K. E., and J. R. Stringer, 1986 RecA independent recombination of poly[d(GT)-d(CA)] in pBR322. Nucleic Acids Res. **14:** 7325–7340.

Nei, M., and T. Maruyama, 1975 Letters to the editors: Lewontin-Krakauer test for neutral genes. Genetics **80:** 395.

Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

Otto, S. P., 2000 Detecting the form of selection from DNA sequence data. Trends Genet. **16:** 526–529.

Pritchard, J. K., and M. W. Feldman, 1998 A test for heterogeneity of microsatellite variation, pp. 47–56, in *Proceedings of the Trinational Workshop on Molecular Evolution*, edited by M. K. Uyenoyama and A. von Haeseler. Duke University Publications Group, Durham, NC.

Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945–959.

Reich, D. E., M. W. Feldman and D. B. Goldstein, 1999 Statistical properties of two tests that use multilocus data sets to detect population expansions. Mol. Biol. Evol. **16:** 453–466.

Robertson, A., 1975 Remarks on the Lewontin-Krakauer test. Genetics **80:** 396.

Schlötterer, C., and T. Wiehe, 1999 Microsatellites, a neutral marker to infer selective sweeps, pp. 238–248 in *Microsatellites—Evolution and Applications*, edited by D. Goldstein and C. Schlötterer. Oxford University Press, Oxford.

Schlötterer, C., C. Vogl and D. Tautz, 1997 Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. Genetics **146:** 309–320.

Schlötterer, C., R. Ritter, B. Harr and G. Brem, 1998 High mutation rates of a long microsatellite allele in *Drosophila melanogaster* provide evidence for allele-specific mutation rates. Mol. Biol. Evol. **15:** 1269–1274.

Slatkin, M., 1995a Hitchhiking and associative overdominance at a microsatellite locus. Mol. Biol. Evol. **12:** 473–480.

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

Tsakas, S., and C. B. Krimbas, 1976 Testing the heterogeneity of *F* values: a suggestion and a correction. Genetics **84:** 399–401.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural *et al.*, 2001 The sequence of the human genome. Science **291:** 1304–1351.

Vieira, J., and B. Charlesworth, 2000 Evidence for selection at the *fused* locus of *Drosophila virilis*. Genetics **155:** 1701–1709.

Vitalis, R., K. Dawson and P. Boursot, 2001 Interpretation of variation across marker loci as evidence of selection. Genetics **158:** 1811–1823.

Weber, J. L., 1990 Informativeness of human $(dC\text{-}dA)_n\cdot(dG\text{-}dT)_n$ polymorphisms. Genomics **7:** 524–530.

Wiehe, T., 1998 The effect of selective sweeps on the variance of the allele distribution of a linked multi-allele locus-hitchhiking of microsatellites. Theor. Popul. Biol. **53:** 272–283.

Wierdl, M., M. Dominska and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:** 769–779.

Communicating editor: J. Hey