

Microsatellite Mutations and Inferences About Human Demography

Rusty Gonser,* Peter Donnelly,[†] George Nicholson[†] and Anna Di Rienzo*

*Department of Human Genetics, University of Chicago, Chicago, Illinois 60637 and [†]Department of Statistics, University of Oxford, Oxford OX1 3TG, United Kingdom

Manuscript received July 29, 1999

Accepted for publication November 29, 1999

ABSTRACT

Microsatellites have been widely used as tools for population studies. However, inference about population processes relies on the specification of mutation parameters that are largely unknown and likely to differ across loci. Here, we use data on somatic mutations to investigate the mutation process at 14 tetranucleotide repeats and carry out an advanced multilocus analysis of different demographic scenarios on worldwide population samples. We use a method based on less restrictive assumptions about the mutation process, which is more powerful to detect departures from the null hypothesis of constant population size than other methods previously applied to similar data sets. We detect a signal of population expansion in all samples examined, except for one African sample. As part of this analysis, we identify an "anomalous" locus whose extreme pattern of variation cannot be explained by variability in mutation size. Exaggerated mutation rate is proposed as a possible cause for its unusual variation pattern. We evaluate the effect of using it to infer population histories and show that inferences about demographic histories are markedly affected by its inclusion. In fact, exclusion of the anomalous locus reduces interlocus variability of statistics summarizing population variation and strengthens the evidence in favor of demographic growth.

INTEREST in the use of microsatellites as tools for the study of population processes followed soon after their discovery (LITT and LUTY 1989; WEBER and MAY 1989; BURKE 1991; PENA 1993; BOWCOCK *et al.* 1994; DI RIENZO *et al.* 1994). Population genetics theory can be used to relate aspects of the patterns of variation at microsatellite loci, in samples from a population, to the history of the population and the details of the mutation process generating variability. All attempts to infer population history from microsatellite data rely on assumptions, notably about the nature of the mutation processes at the loci involved (VALDES *et al.* 1993; DI RIENZO *et al.* 1994; ZHIVOTOVSKY and FELDMAN 1995; PRITCHARD and FELDMAN 1996; FELDMAN *et al.* 1997; REICH and GOLDSTEIN 1998). While the results of such analyses can depend critically on the assumptions made, there has been little experimental work aimed at observing directly the products of these mutation processes to provide an empirical foundation for population studies. Moreover, the validity of most methods or their power to detect population expansion relies on strong assumptions about the mutation mechanism: typically that both the mutation rate and the process that changes allele length are the same at each locus and/or that all changes to allele length are equally likely to involve the gain or loss of a single repeat unit (SHRIVER *et al.* 1993;

VALDES *et al.* 1993; KIMMEL *et al.* 1998; REICH and GOLDSTEIN 1998; REICH *et al.* 1999).

As a step toward an improved characterization of the mutation process on a locus-by-locus basis, we devised an approach based on the analysis of somatic microsatellite mutations in cancer patients, which allows the estimation of the distribution of mutation sizes for each locus (DI RIENZO *et al.* 1998). It transpires that one particular feature of this distribution, namely the average squared change in repeat number, or *mutation mean square*, plays a central role in the behavior of commonly used summaries of population variation such as the variance of the repeat number in a population sample. Our earlier study (DI RIENZO *et al.* 1998) suggested that locus-by-locus estimates of the mutation mean square from somatic mutations are informative for the germ-line mutation processes, and in addition, that these processes can differ across loci in important respects.

Here, we report on three results and a subsequent population analysis. First, we extend our previous findings to a broader range of human populations by applying the same approach to a second data set on 14 additional microsatellite loci. The results further validate the use of microsatellite instability as a means of characterizing the mutation process of microsatellites. Second, we investigate the variability of mutation rates across loci by using our locus-specific estimates of the distributions of mutation size. The purpose of this analysis is to identify loci that are "anomalous" in that their extreme pattern of population variation cannot be accounted for solely by mutation size variability. We iden-

Corresponding author: Anna Di Rienzo, Department of Human Genetics, University of Chicago, JFK Rm. 116, 924 E. 57th St., Chicago, IL 60637. E-mail: dirienzo@genetics.uchicago.edu

tify one such locus and conclude that its inclusion may compromise the inference about population histories. Third, we extend a result of SLATKIN (1995) to show that a linear relationship between the population variance and the mutation mean square would be expected under *any* demographic scenario for any generalized stepwise mutation model, even if there is a bias in the direction of the mutational change.

The amount of variation expected at a particular locus in a population increases with the variability (measured by the mutation mean square) of the mutation process. Interlocus variability in the mutation mean square is effectively another source of noise in an already very noisy system. Locus-specific estimates of the mutation mean square can be used to correct for this effect before combining information across loci. DI RIENZO *et al.* (1998) introduced a new statistic, called the normalized population variance (NPV), defined as the ratio of the population variance at a locus to (an estimate of) the mutation mean square at that locus. Inferences based on population data from a single locus are typically very imprecise, even if the genetic mechanisms at the locus are completely understood (*e.g.*, DONNELLY and TAVARÉ 1995). In making inferences about population history, it is thus desirable to use data from many unlinked loci. In the case of multilocus microsatellite analyses, the use of NPV values, rather than just the observed population variance, considerably improves the quality of the inference: tests of particular scenarios about population history will have greater power, and estimates of population parameters will be more precise. The better the estimates of the mutation mean square at each locus, the more marked this effect will be.

As a result, we are in a position to allow for most of the complexities of the mutation process at microsatellite loci and, thus, carry out an advanced multilocus analysis of different demographic scenarios. Because of our less restrictive assumptions about the mutation process, our method is more powerful to detect departures from the null hypothesis of constant population size than other methods that have been applied to similar datasets (KIMMEL *et al.* 1998; REICH and GOLDSTEIN 1998). Our analysis shows evidence for historical population expansions in all the populations examined with the exception of Africa in the second data set.

MATERIALS AND METHODS

Subjects: Study subjects included 219 patients with sporadic colorectal cancer diagnosed and treated at the Northwestern Memorial Hospital, Chicago, Illinois, as described in DI RIENZO *et al.* (1998). Sections of tumor and normal tissue were cut from paraffin-embedded tissue blocks and the DNA was extracted from the tissue sections as described in WRIGHT and MANOS (1990).

The Sardinia (Italy) population sample was randomly selected from previously described samples (DI RIENZO and WILSON 1991). The DNAs were extracted from placental tissue

TABLE 1
Microsatellite instability screening

Locus name	Patients with somatic mutations	Patients tested	Rate of instability (%)
D1S407	10	156	6.41
D1S399	26	194	13.40
D3S1537	11	176	6.25
D4S1530	20	146	13.70
D6S393	15	179	8.38
D8S384	16	204	7.84
D8S499	22	194	11.34
D10S516	39	189	20.63
D10S525	21	162	12.96
D10S526	18	81	22.22
D14S119	31	199	15.58
D17S919	17	191	8.90
D19S400	22	179	12.29
D20S161	19	201	9.45

of unrelated individuals selected from the general population. The remaining population data analyzed in this article were published in JORDE *et al.* (1995). We typed the 14 loci in Table 1 in the same Sardinian sample examined by DI RIENZO *et al.* (1998) to allow the direct comparison across data sets. We then calculated the population variance for these loci in Sardinia and in three major ethnic groups (Africans, Asians, and Europeans; second data set) on the basis of a published report (JORDE *et al.* 1995). Because each sample in the second data set was made up of several subpopulations, we analyzed the data by also using only the most numerous subpopulation sample for each ethnic group, *i.e.*, Sotho for Africa, Japanese for Asia, and French for Europe, to control for the effect of admixture on the pattern of microsatellite variation. In all the analyses performed in this article, the results for each subpopulation (not shown) were always consistent with those for the overall population sample.

Typing protocol: For both population and patient tissue samples, we used previously described typing protocols based on radioactively endlabeling one of the PCR primers (DI RIENZO *et al.* 1994, 1998). PCR conditions for each microsatellite locus were obtained through the Genome Database. Samples from the CEPH database that have been widely used as size markers were run on each gel to ensure consistent allele identification across gels. Every instance of instability was amplified and electrophoresed twice and classified as a somatic mutation only when a consistent pattern was obtained in at least two assays.

RESULTS

Microsatellite instability and patterns of somatic mutations: Fourteen out of the 30 tetranucleotide repeat loci described in JORDE *et al.* (1995) were typed in normal and tumor DNA extracted from surgical specimens of patients with sporadic colorectal cancer. A somatic mutation was determined to have occurred when one or more bands appeared in the tumor tissue in addition to those observed in the normal tissue of the same patient. The rate of microsatellite instability per locus, calculated as the proportion of loci with somatic muta-

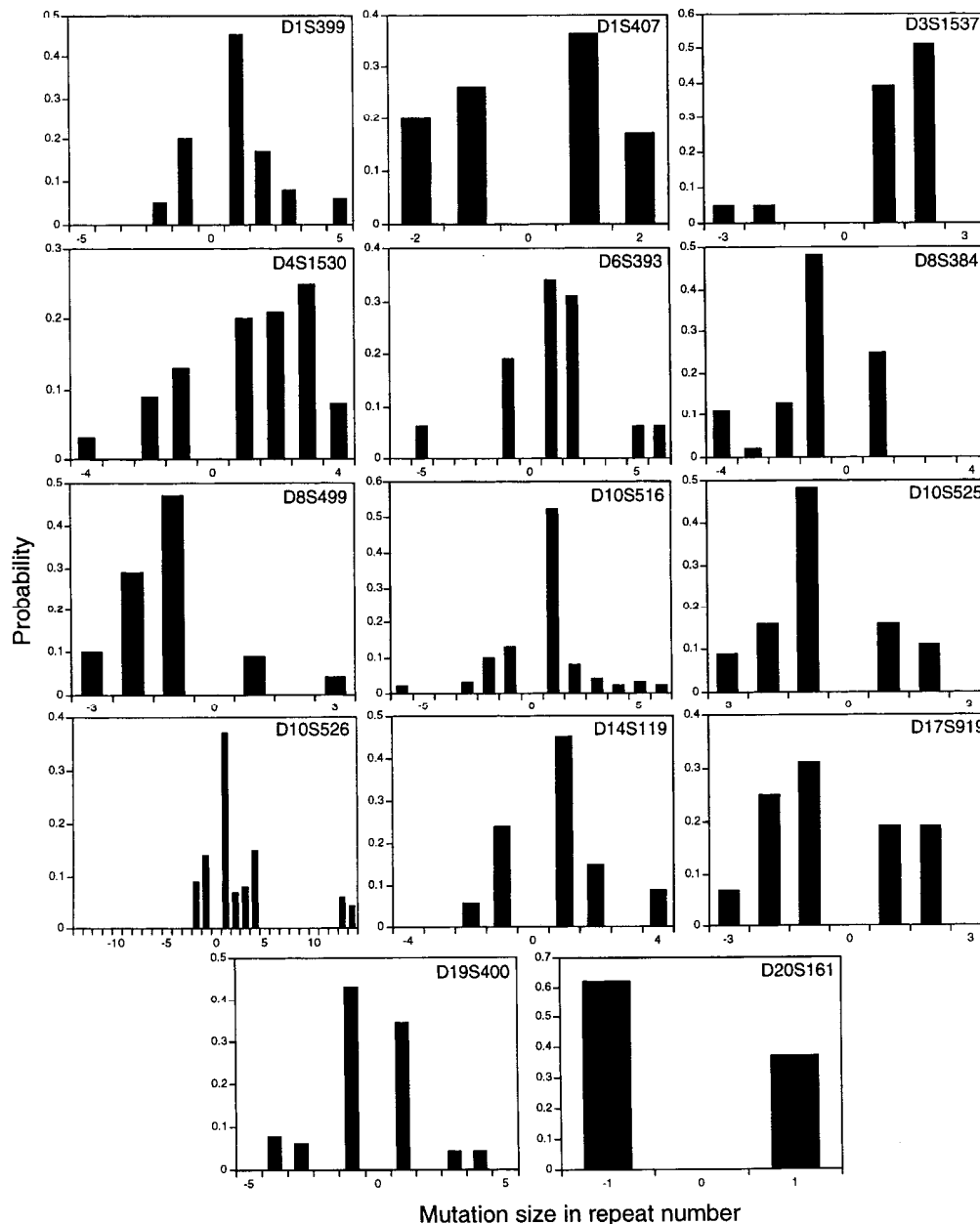


FIGURE 1.—Maximum-likelihood estimates of the mutation sizes at 14 microsatellite loci.

tions over the total number of patients tested, was 12.10% on average, in line with our previous estimates of the instability rate for tetranucleotide repeats (Table 1; DI RIENZO *et al.* 1998). To assess the pattern of somatic mutations for each locus, we estimated the distributions of mutation size based on the mutations observed in tumor tissue. Because the typing results do not allow one to determine unequivocally which allele was hit by a somatic mutation, we used the expectation maximization (EM) algorithm to produce maximum-likelihood estimates of the distributions of mutation sizes for each locus (DEMPSTER *et al.* 1977; DI RIENZO *et al.* 1998). These distributions, shown in Figure 1, show a moderate degree of interlocus variability and have on average a relatively narrow range of mutation sizes with most

mutations involving one or two repeat units. However, as in our previous analysis of somatic mutations, a minority of loci, such as D10S526, have a broader range of mutations asymmetrically distributed around the mean. To infer accurately population parameters, it is important to take this interlocus variability of mutation size into account. The estimated values of the mutation mean square can be found on the website of the Department of Human Genetics of the University of Chicago (<http://www.genes.uchicago.edu>).

For 8 out of 14 loci, the mean of the estimated distribution of mutation size was above zero, showing no evidence for a mutational bias toward an increase of repeat size.

Validating the use of somatic mutations for estimating

TABLE 2
Rank correlation between population variance (S^2) and mutation mean square (η_2)

Population	All loci		Without D10S244	
	ρ	<i>P</i> value	ρ	<i>P</i> value
First data set				
Luo (Africa)	0.815	0.0016	0.850	0.0015
Sardinia (Europe)	0.744	0.0040	0.868	0.0012
Kaingang (S. America)	0.662	0.0104	0.689	0.0099
Pooled sample	0.850	0.0010	0.911	0.0007
Second data set				
Africans	0.297	0.2847		
Asians	0.574	0.0386		
Europeans	0.710	0.0105		
Pooled sample	0.635	0.0220		
Sardinia (Europe)	0.736	0.0079		

germ-line mutation parameters: In APPENDIX A, we show that a linear relationship is expected between the variance of repeat number in a population sample and the mutation mean square for each microsatellite locus for *any* demographic scenario. This extends earlier results of ROE (1992), SLATKIN (1995), ZHIVOTOVSKY and FELDMAN (1995), KIMMEL and CHAKRABORTY (1996), PRITCHARD and FELDMAN (1996), CHAKRABORTY *et al.* (1997), and DI RIENZO *et al.* (1998). Therefore, regardless of the demographic history of the population, the population variance is expected to be linearly related to the mutation mean square estimated from somatic mutations in cancer patients, if the latter is similar to that underlying the observed population variability. These findings imply that the validation of our approach through the analysis of population variability does not depend on making the correct assumption about a particular demographic model.

The fit of the data to the general expectation of a linear relationship between population and mutation parameters can be tested by assessing the significance of the rank correlation between them. As shown in Table 2, all the population samples, with the exception of the African sample, have a significant rank correlation. A graphical representation of the relationship between the population variance and the mutation mean square is also shown in Figure 2.

We regard the significance for all populations other than the African sample as evidence that the somatic mutations in cancer are indeed informative for the mutation process generating population variation. Recall that if it were, we would expect a linear relationship between population variance and mutation mean square. The test based on rank correlation examines the null hypothesis of no association. Whether or not such a test will detect a linear relationship if one is present depends on the power of the test, which in turn will depend on the variability of the data around the linear relationship. This variability is considerable for our data, due to the

stochastic nature of the evolutionary process. If informative at all, the cancer data will be equally informative for all populations. In the light of the other results, we are thus inclined to view the lack of significance of the rank correlation for the African sample as a reflection of the low power of the test, rather than on the utility of the cancer data.

Bootstrap resampling was used to assess the sampling error in our estimates of mutation mean square. The bootstrap distributions are shown in Figure 3. We discuss below the consequences of this sampling variability for estimation of demographic parameters and testing of demographic scenarios.

Identifying “anomalous” loci: Equation A1 in APPENDIX A shows that at a given locus the expected population variance depends linearly on features of the mutation mechanism (the mutation mean square, η_2 , and the mutation rate, μ) and a feature of the population demographic history [the expected coalescence time for a pair of genes at the locus, $E(T_{12})$]. As noted above, the use of the NPV facilitates interlocus comparison by “correcting” for an estimate of mutation mean square at the locus. For example, microsatellites with large variability in mutation size, such as D10S526, are expected to show greater population variance, and allowance should be made for this effect when pooling results across loci. However, reliable empirical estimates of locus-specific mutation rates are not available and an analogous correction is not feasible. We now utilize the estimated distributions of mutation size to investigate the interlocus variability of mutation rate.

Write $\text{NPV}(l, p)$ for the NPV value at locus l in population p , and writing k_p for the number of loci tested in population p , define

$$\text{NPV}(+, p) = \frac{1}{k_p} \sum_{l=1}^{k_p} \text{NPV}(l, p),$$

the average NPV value across loci within population p .

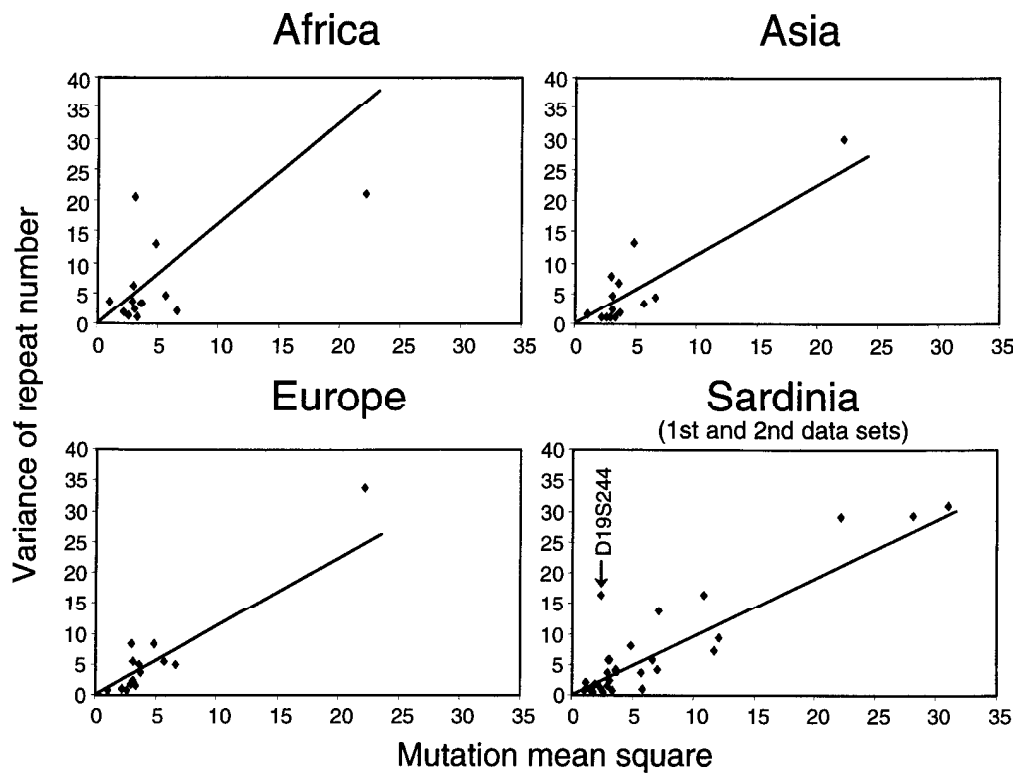


FIGURE 2.—Relationship between the mutation mean square (η_2) and the population variance (S^2). The data from the second data set are shown for the African, Asian, and European samples. The slope of the line was determined as the average NPV for each population sample. For the Sardinian sample, the NPV was averaged across loci for the first and second data sets without the outlier D19S244 (indicated by an arrow).

Now, writing $T_{12}^{(p)}$ for the mean pairwise coalescence time at an autosomal locus in population p , μ_l for the mutation rate at locus l , and $\bar{\mu}$ for the average mutation rate across these loci,

$$E(\text{NPV}(l, p)) = \mu_l T_{12}^{(p)} \quad \text{and} \quad E(\text{NPV}(+, p)) = \bar{\mu} T_{12}^{(p)},$$

so that

$$\gamma(l, p) \equiv \frac{\text{NPV}(l, p)}{\text{NPV}(+, p)}$$

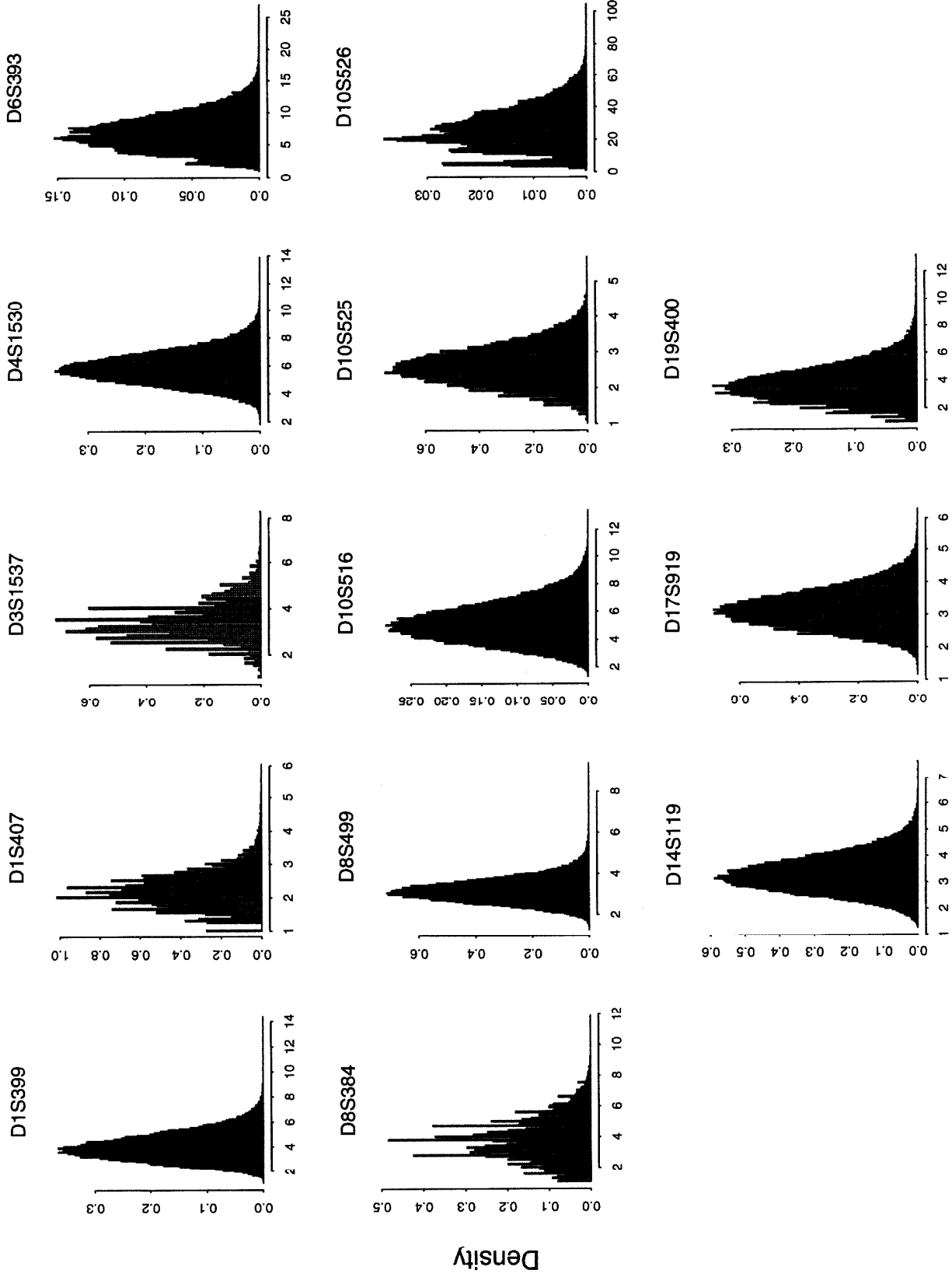
is a natural ratio estimator of $\mu_l/\bar{\mu}$ [the use of $\gamma(l, p)$ was suggested to us by M. STEPHENS, personal communication]. Figure 4 plots the values of $\gamma(l, p)$ for the data set of DI RIENZO *et al.* (1998; first data set) and the second data set reported here. Under selective neutrality and the generalized stepwise mutation model, the height of each bar in Figure 4 is an estimate of the mutation rate at the relevant locus relative to the average mutation rate across loci.

The most striking feature of Figure 4 is that the bar for locus D19S244 is much higher than that for any other locus in the corresponding samples. The difference is significant: a permutation test (GOOD 1994) of the null hypothesis of exchangeability of $\gamma(l, p)$ values across loci within populations has $P = 0.02$. The variability of the estimator $\gamma(l, p)$ will depend on the true demographic scenario. It could be substantial, especially in a population of constant size. We note, however, that the use of the permutation test, and hence our

conclusion that D19S244 is unusual, is valid regardless of the magnitude of such variability.

We propose three possible explanations for this observation. The first is that this locus has a substantially larger mutation rate than the other loci in the first data set. The second is that it has been affected by natural selection: if natural selection acted at or near the locus in all populations, it may systematically increase (balancing selection) or decrease the expected average coalescence time (background selection or a selective sweep; SMITH and HAIGH 1974; HUDSON and KAPLAN 1988; KAPLAN *et al.* 1988; CHARLESWORTH *et al.* 1995). The former case would tend to manifest itself as large $\gamma(l, p)$ values for that locus in all populations, and hence in a large bar in Figure 4, as for D19S244. (The latter forms of selection would have the opposite effect, leading to a small bar in Figure 4.) A third potential explanation is that the mutation mean square at that locus was substantially underestimated, hence inflating the NPV values in all populations.

A high rate of *de novo* mutations has been observed at the D19S244 locus in family studies (WEBER and WONG 1993). Further, a search of the Human Transcript Map database did not reveal any gene known to evolve under balancing selection in the neighborhood of this marker, and the number of mutations from which the mutation mean square was estimated was larger than for most other loci (29 instances of microsatellite instability were observed at D19S244). Thus, while we cannot be definitive, we propose that the increased $\gamma(l, p)$ value for the



Mutation Mean Square

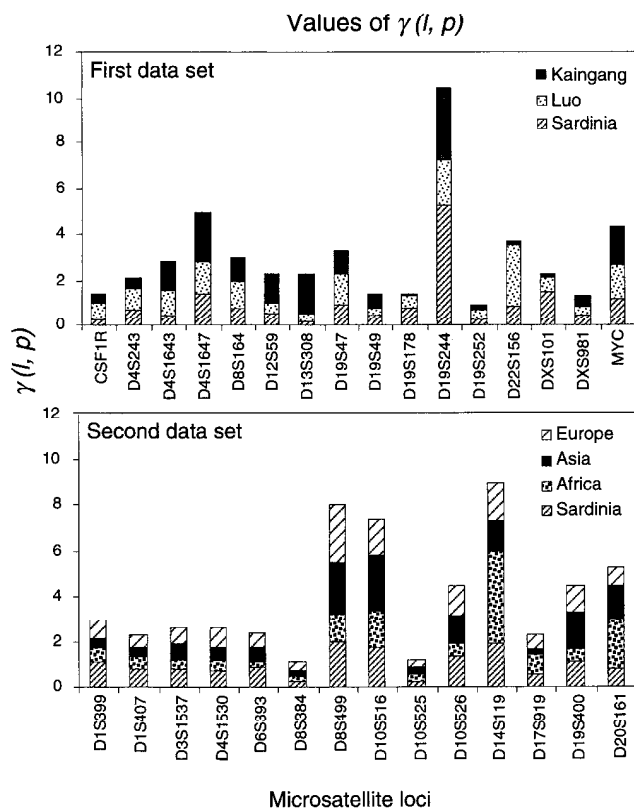


FIGURE 4.—Values of $\gamma(l, p)$ for the first and second data sets.

D19S244 locus reflects a substantially higher mutation rate for that locus. While the second data set (Figure 4) also suggests heterogeneity across loci, no locus appears to be markedly different from the others.

As shown in DI RIENZO *et al.* (1998; Equations A7 and A15), the precision of estimators of demographic parameters based on average NPV values is decreasing in the variance, across loci, of the mutation rate. So too is the power of tests of demographic scenarios based on interlocus variability of NPV values. Having identified D19S244 as an “anomalous” locus, it is thus most efficient to ignore this locus in subsequent statistical analyses. To evaluate the effect of this locus on population inferences, we compared the results from the first data set by including and excluding it from all the population samples.

Using the NPV to make inferences about human evolution:

Parameters of natural interest include the effective population size, in the case of a constant-sized population, or the timing or rate of changes in population size, under other demographic scenarios. It follows from Equation A1 in APPENDIX A that the average NPV value across loci within a population is an unbiased estimator of the product $\bar{\mu}E(T_{12})$, where $\bar{\mu}$ is the average mutation rate of the loci in the study, and $E(T_{12})$ is the expected coalescence time of a pair of genes at the locus. For a constant-sized population, the expected average coalescence time equals the effective number of chromosomes in the population, N , while for a population that was initially very small before undergoing a rapid and substantial growth in size T generations ago, it equals the time since the expansion (T) (e.g., DONNELLY and TAVARÉ 1995). For these respective demographic scenarios, our data can thus be used to provide estimates of N and T , respectively, under various assumptions about the average mutation rate for the loci. These are given in Table 3. The uncertainty in these estimates of population parameters results principally from stochastic variation in the evolutionary process and uncertainty in the estimation of the mutation mean square. The stochastic variation can be substantial (see DISCUSSION) and will depend on the nature of, and parameters in, the (unknown) demographic scenario. The effect of uncertainty resulting from the estimation of mutation mean square can be assessed through the bootstrap procedure described in Figure 3 legend. The distribution of the error in estimating the NPV at each locus may then be approximated (Figure 5). When the mean NPV is calculated across loci, the component of its standard error stemming from this source alone is $\sim 20\%$ of the estimated value (data not shown). For more realistic demographic scenarios, the relationship between the expected average coalescence time and parameters of interest is less clear, though it could of course be studied for particular scenarios of interest.

One scenario of interest in connection with human evolution would be a population of nontrivial size that grows rapidly to become quite large. Then the expected average coalescence time would be larger than the time since growth, so that it would provide an upper bound on that time (provided, as seems plausible for humans,

FIGURE 3.—Bootstrap estimates of the sampling error in our estimates of mutation mean square for each locus. For each locus, the following procedure was repeated 50,000 times: (1) A new data set was generated by choosing randomly one of the two normal alleles in each individual with a somatic mutation at the locus and then “mutating” it by adding a (positive or negative) mutation size chosen randomly according to the maximum-likelihood estimate of the mutation size distribution at that locus. (If this resulted in an allele of the same length as the other normal allele, the result was discarded and the procedure repeated for that pair.) The resulting bootstrap data set then had two normal alleles and one mutant allele for each individual. (2) For each bootstrap data set, the EM algorithm was used to calculate the maximum-likelihood estimate of mutation size distribution, and so of mutation mean square. A histogram is plotted of the 50,000 bootstrap values of the estimated mutation mean square for each locus. The procedure was vacuous for D20S161 because in the estimated distribution of mutation sizes mutations were always of size ± 1 , and so all bootstrap estimates of the mutation mean square were also 1.

TABLE 3
Estimates of population parameters based on the average NPV

Population sample	Aver (NPV)	Var (NPV)	<i>N</i>		<i>T</i> (kya)	
			$\mu = 10^{-3}$	$\mu = 10^{-4}$	$\mu = 10^{-3}$	$\mu = 10^{-4}$
Second data set—all loci						
Africa	1.632	2.975	1,632	16,320	33	330
Asia	1.123	0.710	1,123	11,230	23	225
Europe	1.107	0.432	1,107	11,070	22	221
Sardinia	0.979	0.286	979	9,790	20	196
Pooled	1.382	0.937	1,382	13,820	28	276
First data set—excluding D19S244						
Luo	1.248	0.742	1,248	12,480	25	254
Sardinia	0.901	0.287	901	9,010	19	189
Kaingang	0.579	0.205	579	5,790	12	117
Pooled	1.023	0.268	1,023	10,230	21	210

Estimates of *T* are based on a generation time of 20 years. Recall that *N* is the effective number of chromosomes in the population. The values of average NPV for the first data set are with the adjustments for X-linked loci as described in DI RIENZO *et al.* (1998). The values without the adjustments used to estimate *T* in the rapid expansion scenario are 1.272, 0.943, 0.586, and 1.048 for Luo, Sardinia, Kaingang, and pooled samples, respectively. kya, thousand years ago.

that the time since growth was less than the effective population size after growth).

The present data set also confirms our previous finding that the coalescence time of the overall pooled sample is very similar to the coalescence times of the individual populations. This strongly suggests that populations from different ethnic groups share a substantial portion of their genetic ancestry and is in agreement with previous studies indicating that a small proportion of human genetic diversity occurs between populations (LEWONTIN 1972; NEI 1987).

Testing demographic scenarios: Aside from estimating parameters, under the assumption that a particular demographic scenario applies, multilocus microsatellite data allow testing of demographic scenarios. It turns out that the expected variability, across loci, in statistics such as the population variance or the NPV changes considerably under different demographic scenarios. For example, under the null hypothesis of constant population size, relatively large values of the variance of NPV across loci would be expected, while under scenarios of recent population growth this variance will be smaller. DI RIENZO *et al.* (1995, 1998) exploited this fact in developing a method for inference about demographic history.

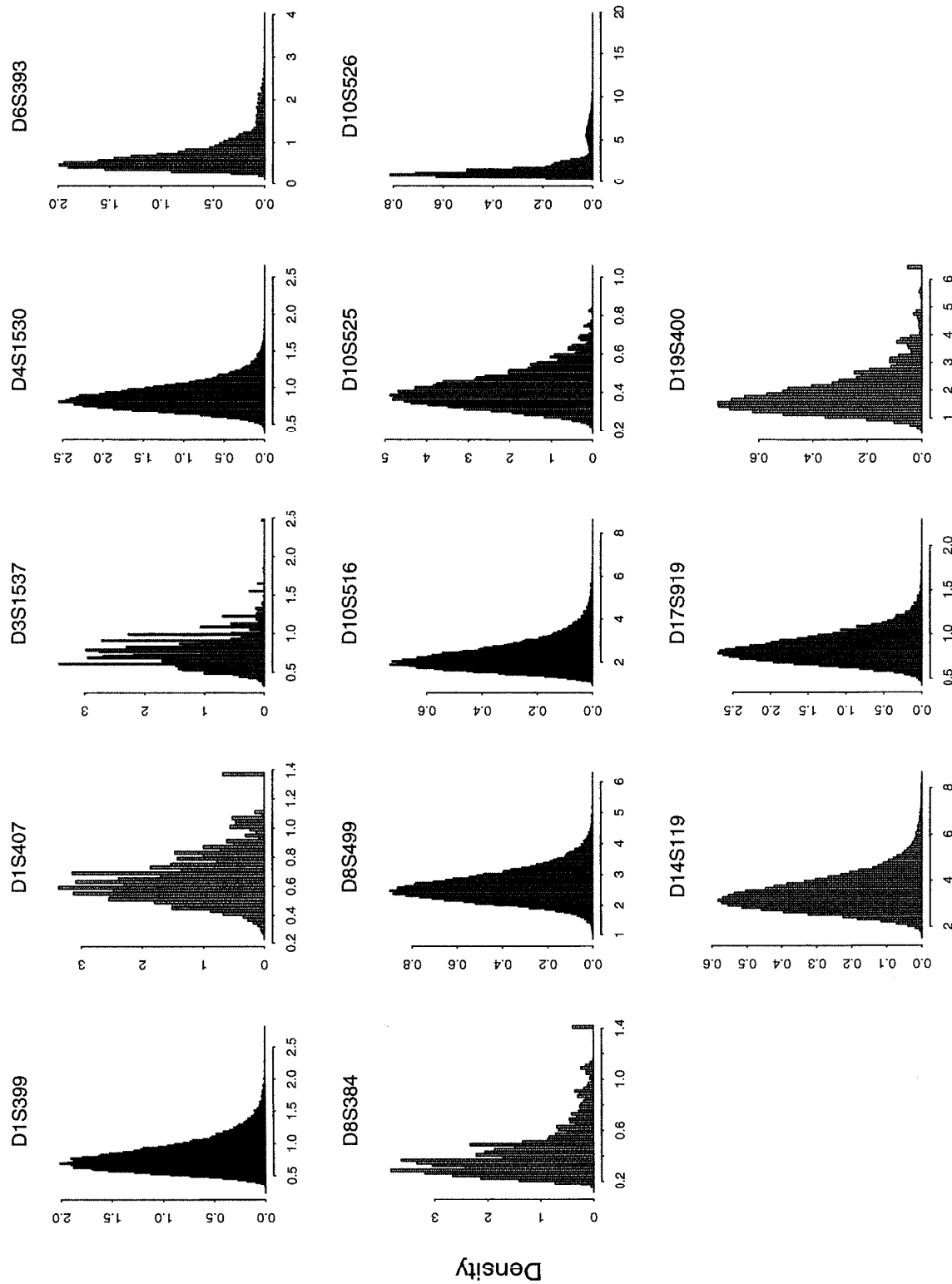
Here, we apply our method to the second data set. In the light of the earlier analysis suggesting an unusual status for the locus D19S244, we also reapply the method to the first data set, both with and without D19S244. The details of the hypothesis testing procedure are described in APPENDIX B. Table 4 gives the *P* values for three different test statistics, F_1 , F_2 , and g , defined in Equations B1, B2, and B3. The statistic g was effectively introduced by REICH and GOLDSTEIN (1998). (For ease

of comparison, we actually use the reciprocal of their statistic, but the testing procedure is equivalent.) This statistic uses only the population variance at each locus, without correcting for interlocus variability in the mutation process before combining information across loci. The statistics F_1 and F_2 each use NPV values at each locus, hence taking advantage of locus-specific information on the mutation mean square. Of the two, F_1 is much simpler, but F_2 leads to a slightly more powerful test.

Our test rejects the null hypothesis of constant population size for large values of the test statistics, corresponding to the data showing significantly less variability across loci than would be expected under this demographic scenario (see Table 4). Whatever the true demographic scenario, variation in mutation rates across loci will tend to increase the variation in population variances and NPVs across loci, thus decreasing the value of all three test statistics. While this means that it is conservative to calculate *P* values under the assumption of the same mutation rate for all loci, this can involve a considerable loss of power. We believe it is more helpful to calculate *P* values separately under a range of assumptions (see Table 4 for details) about the variation in mutation rate and have done so.

REICH *et al.* (1999) have shown that the procedure based on the statistic g is conservative under a range of other deviations from the simple assumptions usually made about the mutation mechanism. As described in APPENDIX B, we have conducted extensive simulations to establish that tests based on either F_1 or F_2 remain conservative if the mutation mechanism also varies across loci and (as for our data) the mutation mean square is estimated with error.

While the extent of variation in mutation rate across



Normalized Population Variance

FIGURE 5.—Effects of the sampling error on the calculation of the NPV at each locus in the pooled population sample. A histogram is plotted of 50,000 values of the NPV; each value corresponds to the sample variance of the allele length observed at a locus divided by a bootstrap value of the estimated mutation mean square for that locus.

TABLE 1
Estimated P values for the three F -statistic-based tests of the null hypothesis of constant population size for various levels of variability in μ

Data set	Population	Level of variability in μ											
		None			Moderate			Medium			High		
		g	F_1	F_2	g	F_1	F_2	g	F_1	F_2	g	F_1	F_2
First	Luo	0.81	0.16	0.14	0.51	0.06	0.05	0.32	0.03	0.02	0.12	0.00	0.00
	Sardinia	0.77	0.83	0.81	0.46	0.54	0.50	0.27	0.34	0.31	0.08	0.14	0.11
	Kaingang	0.81	0.49	0.40	0.51	0.23	0.18	0.32	0.10	0.07	0.12	0.01	0.01
	Pooled	0.70	0.25	0.21	0.39	0.10	0.08	0.21	0.04	0.03	0.05	0.00	0.00
Second	Africa	0.75	0.74	0.72	0.45	0.45	0.42	0.27	0.26	0.24	0.09	0.08	0.07
	Asia	0.91	0.34	0.29	0.66	0.15	0.12	0.50	0.07	0.05	0.31	0.01	0.00
	Europe	0.93	0.11	0.08	0.70	0.04	0.03	0.53	0.01	0.01	0.35	0.00	0.00
	Sardinia	0.91	0.06	0.04	0.67	0.02	0.01	0.49	0.01	0.00	0.31	0.00	0.00
	Pooled	0.84	0.25	0.22	0.57	0.11	0.09	0.38	0.04	0.03	0.17	0.00	0.00
First*	Luo	0.82	0.20	0.17	0.53	0.08	0.06	0.35	0.03	0.02	0.14	0.00	0.00
	Sardinia	0.83	0.07	0.05	0.53	0.02	0.02	0.35	0.01	0.00	0.14	0.00	0.00
	Kaingang	0.84	0.35	0.26	0.55	0.16	0.11	0.37	0.06	0.04	0.16	0.01	0.00
	Pooled	0.75	0.02	0.02	0.45	0.01	0.00	0.25	0.00	0.00	0.08	0.00	0.00

Significance levels of tests of the null hypothesis of constant population size based on three test statistics, g (effectively introduced by REICH and GOLDSTEIN 1998), which uses only population variance at each locus, and F_1 and F_2 , introduced in this article, based on NPV at each locus. The first and first* data sets refer to the data reported in DI RIENZO *et al.* (1998), including and excluding, respectively, the locus D19S244. The second data set refers to that reported in this article. Significance levels are given under a range of assumptions about the extent to which the mutation rate varies across loci: no variability, moderate variability (5% of loci having mutation rates 10^{-3} and 10^{-5} , respectively, with the remainder having rate 10^{-4}), medium variability (10% of loci having mutation rates 10^{-3} and 10^{-5} , respectively, with the remainder having rate 10^{-4}), and high variability (20% of loci having mutation rates 10^{-3} and 10^{-5} , respectively, with the remainder having rate 10^{-4}). Significance levels are based on 30,000 values simulated under the null hypothesis, with $N = 10,000$. (Changes in the value of N in the range 5,000 to 20,000 have little effect on the significance level; data not shown.) As described in APPENDIX B, tests based on F_1 and F_2 are shown to be conservative under variation in the mutation mechanism across loci and error in the estimation of mutation mean square.

microsatellite loci has not been well documented, either our “medium” or “high” variability scenarios may be most realistic (WEBER and WONG 1993; SEIELSTAD *et al.* 1999). In this case, excluding D19S244 from the analysis, all populations in the first data set show significant evidence for departure from the assumption of constant population size. The same is true for all populations, except Africa (in fact the Sotho) in the second data set. Under the scenario of high variability in mutation rate, the data are almost significant ($P = 0.07$).

All three test statistics lead to valid tests. The differences in P values on the same data are due to differences in their power to detect departures from the null hypothesis. The statistic F_2 is slightly more powerful than F_1 . We should expect the statistics based on NPV (F_1 and F_2) to be more powerful than one based simply on population variance (g), exactly because the normalization (division by mutation mean square) corrects for one source of variation before combining data across loci. Nonetheless, the difference in power between g and the two F statistics is striking, particularly if one were to use the procedure based on g without making allowance for variation in mutation rate.

DISCUSSION

Somatic and germ-line mutations: An understanding of microsatellite mutation patterns is central to their use for the accurate reconstruction of population processes. We have developed and validated an experimental approach to estimate the distribution of mutation sizes for each individual microsatellite locus. These distributions were estimated from somatic mutations observed in the tumor tissue of sporadic patients with colorectal cancer.

It is not known whether such mutations arise from the same events that produce variation in the normal population. Microsatellite instability in some cancer patients may reflect defects in mismatch repair; but, in other patients, it may be a consequence of the higher number of cell divisions that occurs in the tumor compared to the normal tissue. Nevertheless, in the absence of specific mechanistic or genetic information on the source of these mutations, it is still possible to test whether they reflect the mutation process in the general population by using population theory. Here, we demonstrate that under the generalized stepwise model with arbitrary distribution of mutation sizes, the relationship between the variance of repeat number at a

given locus in a population sample and the mutation mean square for the same locus is expected to be linear regardless of assumptions about the demographic history of the population (see APPENDIX A). Therefore, if the mutation mean square estimated in cancer patients parallels that of the “real” mutation process, one expects it to be linearly related to the variance of repeat number of different population samples. Three out of the four population samples examined in this article conform to this expectation. This observation extends our previous findings of a linear relationship between the population variance and the mutation mean square for an additional three population samples. Taken together, the results of these two studies indicate that the somatic mutations observed in sporadic colorectal cancer patients are a useful approach to the characterization of the mutation process of microsatellite loci on a locus-by-locus basis.

Even though most of the loci show a preponderance of short mutations, *i.e.*, gain or loss of one or two repeat units, our estimated distributions of mutation sizes (Figure 1) are relatively broad for a small subset of the loci examined. To investigate whether mismatch repair defects result in unusually large mutation sizes, we partitioned the patients into two groups. The first group includes patients with high levels of microsatellite instability (at least 20% of loci tested had somatic mutations). These patients are more likely to have mismatch repair defects, and this was recently confirmed by staining their tumor tissue with antibodies against MSH2 and MLH1 (A. DI RIENZO, K. HALLING and S. THIBODEAU, unpublished results; THIBODEAU *et al.* 1998). The second group includes patients with low levels of microsatellite instability; the somatic mutations observed in these patients may well be “normal” mutations probably reflecting the large number of cell divisions in tumor tissue. In this analysis we pooled classes of loci to maintain reasonable sample sizes. There were 123, 89, and 163 instances of somatic mutation at di-, tri-, and tetra-nucleotide repeats in the high instability group, respectively, and 47, 24, and 53 instances in the low instability group in the corresponding groups of loci. The preponderance of short mutations, with some larger mutations, was apparent in both the high and low microsatellite instability groups, and estimates of mutation mean squares were similar in each group (data not shown).

Furthermore, a similar broad range of mutation sizes (*e.g.*, from -12 to +11 repeat units) was observed in the largest survey reported to date of *de novo* mutations in family studies (1107 events over 952,962 parent-offspring transmissions; SEIELSTAD *et al.* 1999). In contrast, earlier, smaller studies observe predominantly one-step mutations (*e.g.*, WEBER and WONG 1993; BRINKMANN *et al.* 1998). These findings suggest that large samples are necessary to observe mutations of large amplitude and further support our inference of concordant patterns in somatic and germ-line mutations.

In addition to the study of germ-line *de novo* or somatic

mutations, another approach to understanding the mutation processes is to examine the variation at tightly linked microsatellite loci. For example, the analysis of multilocus haplotypes carrying the CCR5-Δ32 allele showed that 9.5% of the alleles at locus D3S4580, located 28 kb from CCR5, differ from the most common one by 4–10 repeat units. Detailed haplotype analysis revealed that this pattern cannot be easily explained by recombination and is more consistent with occasional large mutations (J. MARTINSON, personal communication; MARTINSON *et al.* 1998).

Overall, our finding in this and the preceding article of a significant rank correlation between population variance and mutation mean square estimated from the cancer data in five of the six population/loci pairs we have examined would seem extraordinarily unlikely if, in fact, the cancer data were uninformative for the germ-line processes. Further, the results of our analyses of human demography, utilizing the mutation mean squares estimated from the cancer data, are in broad agreement with those of analyses of other genetic systems.

Identifying anomalous loci: Here, we developed a method for identifying loci that are anomalous, either in the sense of having a different mutation rate from others in the study or because their evolution is not governed by the class of (neutral, generalized stepwise) models on which the analysis is based.

We identified one such locus in our studies, D19S244. In the light of independent evidence as to its unusually high mutation rate, we regard this as the most likely explanation for its status as an outlier (WEBER and WONG 1993). Whatever the reason for the anomaly, such loci should be excluded from population analyses. An important consequence of the ability to detect such loci is thus the potential for improved inferences as to population parameters and population history. In our study, excluding D19S244 led to clear differences in the population inferences obtained.

More generally, this method has the potential to detect loci at or near which natural selection has acted. Recall that balancing selection, respectively background selection or a selective sweep, acting near a locus will increase, respectively decrease, the observed $\gamma(l, p)$ value at that locus relative to others in the same population. In our analysis the effect of natural selection is confounded with a higher mutation rate at the locus. One way of distinguishing between the effects of selection and mutation rate changes would be to examine tightly linked microsatellite loci near the anomalous one. Selection should have an effect in the same direction on all such loci. If the original outlier results from an unusually high or low mutation rate, the effect should not extend to linked loci.

Inferences for population parameters and population history: Under general assumptions, the average NPV value across loci within a population is an unbiased estimator of $\bar{\mu}E(T_{12})$, the product of the average mutation rate for the loci and the mean pairwise coalescence

time. For the two simplified demographic scenarios of constant population size and sudden expansion from a small size, this leads naturally to estimators for the effective population size and the time since expansion, respectively. The estimates shown in Table 3 are in line with those obtained based on other studies suggesting an ancient expansion of the human population (HARPENDING *et al.* 1998). In this regard, it should be noted that methods that assume that all mutations involve only a single repeat unit will lead to an overestimate of the population parameters. Thus, in the presence of multi-step mutations, our estimates based on the NPV will result in comparatively lower, and more accurate, values.

Several points about this estimation procedure are noteworthy. The first is that recovery of time or population size estimates is dependent on assumptions about the average mutation rate for the loci used. Direct empirical evidence is scanty, yet a change by a factor of two, for example, in this average rate will change estimated times or sizes by a factor of one-half. This problem, effectively one of calibrating mutational events into numbers of generations, afflicts *all* estimates of such parameters from microsatellite data. Particularly in view of the current speculative nature of such calibrations, the actual value of such estimates should be interpreted with caution. We have presented point estimates with no attempt at assessing the precision of the estimates or, equivalently, of giving confidence intervals. In large part, this is because of the difficulty with the calibration just described. In addition, while the estimator is unbiased for the compound parameter $\bar{\mu}E(T_{12})$ rather generally, its sampling properties, and in particular its precision, will depend sensitively on the underlying demographic scenario. Finally, in our approach we have estimated the mutation mean square at each locus. This estimation also carries uncertainty, in the usual way through sampling, but in addition because we are only measuring a surrogate for the germ-line parameter. While we have used the bootstrap to quantify the former uncertainty, the latter is problematical. It thus does not seem straightforward to quantify the uncertainty in these kinds of estimates of population parameters. On the other hand, it is clear that the uncertainty is large, and in the absence of further relevant data any point estimate based on microsatellite data should be interpreted with great caution.

We performed significance tests of the null hypothesis of constant population size for both data sets. Taking the effective population size as 10,000 individuals, allowing medium variability in mutation rate across loci, and ignoring the locus D19S244 shown above to be anomalous, the null hypothesis would be rejected, in favor of scenarios involving population growth, for all populations in the first data set and for all but the African population in the second data set. If there were more variation in mutation rates across loci (our "high" variability scenario), then the African population in the

second data set also becomes highly suggestive of population expansion ($P = 0.07$).

The mutation process at microsatellite loci is clearly complex. Accordingly, we chose to use an analytical approach that takes into account most of these complexities. There are related recent approaches that use microsatellite data to estimate demographic histories (KIMMEL *et al.* 1998; REICH and GOLDSTEIN 1998). These methods typically make the restrictive assumption that all mutations involve a gain or loss of one repeat unit and do not allow for heterogeneities across loci in parameters of the mutation process. In the presence of such heterogeneities, and possible larger step sizes, estimates of parameters describing population history may be biased, and while tests for population expansion may be conservative, this will be at the cost of a loss of power to detect an expansion. Assuming, as seems plausible from the rank correlation results (Table 2), that our estimates of mutation mean square are related to the relevant germ-line processes, our approach to the reconstruction of demographic histories will have substantially more power to detect expansion. In particular, our analysis (see Table 4) has shown that a method that does not use locus-specific information on the mutation mean square is much less powerful than those developed here.

A signal of population expansion has been observed in virtually all major ethnic groups for mtDNA (HARPENDING *et al.* 1993; ROGERS 1995). However, based on microsatellite analyses, other authors have detected a significant signal only in nonoverlapping subsets of human populations (KIMMEL *et al.* 1998; REICH and GOLDSTEIN 1998). Our results are in broad agreement with those based on mtDNA and those obtained by Kimmel and colleagues on microsatellites: all but one of our signals result in unmistakable rejection of the null hypothesis of constant population size; in a direction consistent with expansion, only one (Luo) of the two African populations results in rejection of this null hypothesis. In this regard, it is interesting to note that the inclusion of a single "anomalous" locus in our data set, D19S244, renders the Sardinian population sample consistent with a constant population size without variability of mutation rates. Thus, failure to identify and allow for heterogeneity in the mutation process even at a single locus may dramatically affect the power to detect population expansion. With regard to microsatellite analyses, our methods provide more powerful tests in the presence of departures from the simple one-step mutation model or interlocus heterogeneity in either mutation rates or step size distributions. Thus, a more likely explanation for the discrepancies across microsatellite analyses is the different power of the statistical tests employed to detect population expansion. The fact that the null hypothesis of constant population size is not rejected by a particular test does not mean that the hypothesis is true. While some scenarios proposed for

expansion, say, in Africa but not in other regions (REICH and GOLDSTEIN 1998), are not without interest, our results suggest that they are not needed.

We thank D. Barch, G. K. Haines, and B. Sisk for help in sample collection; R. R. Hudson and C. Ober for comments on the manuscript; M. Stephens for helpful discussions; and L. Jorde for providing a file containing the original population data. This work was supported in part by grants from the National Science Foundation (SBR-9317266 to A.D. and DMS-9505129 to P.D.) and the American Cancer Society, Illinois Division, and Digestive Disease Research Center (DK-42086) to A.D. and a UK Engineering and Physical Sciences Research Council Advanced Fellowship (B/AF1255) to P.D.

LITERATURE CITED

- BOWCOCK, A. M., L. A. RUIZ, J. TOMFOHRDE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455–457.
- 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.
- BURKE, T. (EDITOR), 1991 *DNA Fingerprinting: Approaches and Applications*. Birkhäuser Verlag, Basel/Boston.
- CHAKRABORTY, R., M. KIMMEL, D. N. STIVERS, L. J. DAVISON and R. DEKA, 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. USA* **94**: 1041–1046.
- CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN, 1995 The pattern of neutral molecular variation under the background selection model. *Genetics* **141**: 1619–1632.
- DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum likelihood estimation from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* **39**: 1–38.
- DI RIENZO, A., and A. C. WILSON, 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **88**: 1597–1601.
- 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., C. TOOMAJIAN, B. SISK, K. HAINES, D. BARCH *et al.*, 1995 STRP variation in human populations and their patterns of somatic mutations in cancer patients. *Am. J. Hum. Genet.* **57** (Suppl.): A41.
- 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.
- DONNELLY, P., and S. TAVARÉ, 1995 Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29**: 401–421.
- FELDMAN, M. W., A. BERGMAN, D. D. POLLOCK and D. B. GOLDSTEIN, 1997 Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* **145**: 207–216.
- GOOD, P., 1994 *Permutation Tests*. Springer-Verlag, New York.
- HARPENDING, H. C., C. S. T. SHERRY, A. R. ROGERS and M. STONEKING, 1993 The genetic structure of ancient human populations. *Curr. Anthropol.* **34**: 483–496.
- HARPENDING, H. C., M. A. BATZER, M. GURVEN, L. B. JORDE, A. R. ROGERS *et al.*, 1998 Genetic traces of ancient demography. *Proc. Natl. Acad. Sci. USA* **95**: 1961–1967.
- HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. *Genetics* **120**: 831–840.
- JORDE, L. B., M. J. BAMSHAD, W. S. WATKINS, R. ZENGER, A. E. FRALEY *et al.*, 1995 Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* **57**: 523–538.
- KAPLAN, N. L., T. DARDEN and R. R. HUDSON, 1988 The coalescent process in models with selection. *Genetics* **120**: 819–829.
- KIMMEL, M., and R. CHAKRABORTY, 1996 Measures of variation at DNA repeat loci under a general stepwise mutation model. *Theor. Popul. Biol.* **50**: 345–367.
- KIMMEL, M., R. CHAKRABORTY, J. P. KING, M. BAMSHAD, W. S. WATKINS *et al.*, 1998 Signatures of population expansion in microsatellite repeat data. *Genetics* **148**: 1921–1930.
- LEWONTIN, R. C., 1972 The apportionment of human diversity. *Evol. Biol.* **6**: 381–398.
- LITT, M., and J. A. LUTY, 1989 A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**: 397–401.
- MARTINSON, J. J., E. C. LAWRENCE, M. J. ALEXANDER, M. SWEENEY and R. E. FERRELL, 1998 A population-based survey of STR allelic association at the chemokine receptor gene CCR5. *Am. J. Hum. Genet.* **63** (Suppl.): A216.
- NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- PENA, S. D. J. (Editor), 1993 *DNA Fingerprinting: State of the Science*. Birkhäuser Verlag, Basel/Boston.
- PRITCHARD, J. K., and M. W. FELDMAN, 1996 Statistics for microsatellite variation based on coalescence. *Theor. Popul. Biol.* **50**: 325–344.
- REICH, D., and D. GOLDSTEIN, 1998 Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. USA* **95**: 8119–8123.
- 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.
- ROE, A., 1992 Correlations and interactions in random walks and population genetics. Ph.D. Thesis, University of London.
- ROGERS, A. R., 1995 Genetic evidence for a Pleistocene population explosion. *Evolution* **49**: 608–615.
- SEIELSTAD, M., X. XU and X. XU, 1999 Direct observations of microsatellite mutations, p. 57 in *Human Evolution*, edited by L. L. CAVALLI-SFORZA, S. PÄÄBO and D. WALLACE. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- 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.
- SLATKIN, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**: 457–462.
- SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**: 23–35.
- THIBODEAU, S. N., A. J. FRENCH, J. M. CUNNINGHAM, D. TESTER, L. J. BURGART *et al.*, 1998 Microsatellite instability in colorectal cancer: different mutator phenotypes and the principal involvement of hMLH1. *Cancer Res.* **58**: 1713–1718.
- VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**: 737–749.
- WEBER, J. L., and P. E. MAY, 1989 Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**: 388–396.
- WEBER, J. L., and C. WONG, 1993 Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**: 1123–1128.
- WRIGHT, D. K., and M. MANOS, 1990 Sample preparation from paraffin-embedded tissues, pp. 153–158 in *PCR Protocols*, edited by M. A. INNIS, D. H. GELFAND and T. J. WHITE. Academic Press, San Diego.
- ZHIVOTOVSKY, L. A., and M. W. FELDMAN, 1995 Microsatellite variability and genetic distances. *Proc. Natl. Acad. Sci. USA* **92**: 11549–11552.

Communicating editor: N. TAKAHATA

APPENDIX A

Writing S^2 for the sample variance of allele length in a sample from a population at a particular locus, we show here that, for any demographic scenario,

$$E(S^2) = \eta_2 \mu E(T_{12}), \quad (A1)$$

where η_2 is the mutation mean square (the expected squared size of the change in allele length caused by mutation), μ is the mutation rate at the locus, and $E(T_{12})$

is the expected coalescence time (in generations) for a pair of genes at the locus. The effect of the demographic scenario enters through its effect on $E(T_{12})$. This extends a result of SLATKIN (1995) to an arbitrary generalized stepwise mutation process.

Note that for a population with constant effective size N chromosomes, $E(T_{12}) = N$, and for one that has expanded rapidly from a very small size T generations ago, $E(T_{12}) \approx T$. The general result (A1) thus reduces to known results [for example, DI RIENZO *et al.* (1998, Equation A2, and the equation below A11)] in these cases.

Throughout, we assume the generalized stepwise model for mutation, namely that neither the mutation rate nor the distribution of the change in allele length caused by mutation will depend on the length of the progenitor allele, and we assume selective neutrality. Aside from this, we allow arbitrary distribution of mutation sizes and, as we have noted, an arbitrary demographic scenario.

Recall from Equation A9 of DI RIENZO *et al.* (1998) that we can write

$$E(S^2) = 2\text{Var}(Y_1) - \frac{1}{2}\text{Var}(Y_1 + Y_2), \quad (\text{A2})$$

where Y_1 and Y_2 , respectively, are the differences between the lengths of two sampled copies of the locus and the length of their most recent common ancestor.

Now, write W_1 for the number of mutations on the lineage to the first sampled chromosome since its common ancestor with the second, and W_{12} for the total number of mutations along either lineage since their common ancestor. Conditional on T_{12} , the number of generations since this common ancestor, W_1 and W_{12} have binomial distributions with parameters (T_{12}, μ) and $(2T_{12}, \mu)$, respectively. In particular, conditional on T_{12} , the means of W_1 and W_{12} are μT_{12} and $2\mu T_{12}$, respectively, and their respective variances are $\mu T_{12}(1 - \mu)$ and $2\mu T_{12}(1 - \mu)$. Since μ is small, we approximate these conditional variances by μT_{12} and $2\mu T_{12}$, respectively.

As in DI RIENZO *et al.* (1998, equations preceding A10 and A11), we can write

$$\text{Var}(Y_1) = \sigma^2 E(W_1) + m^2 \text{Var}(W_1), \quad (\text{A3})$$

and

$$\text{Var}(Y_1 + Y_2) = \sigma^2 E(W_{12}) + m^2 \text{Var}(W_{12}), \quad (\text{A4})$$

where m and σ^2 are, respectively, the mean and variance of the distribution of mutation size.

Now,

$$E(W_1) = E(E(W_1|T_{12})) = E(\mu T_{12}) = \mu E(T_{12}). \quad (\text{A5})$$

Further,

$$\begin{aligned} \text{Var}(W_1) &= E(\text{Var}(W_1|T_{12})) + \text{Var}(E(W_1|T_{12})) \\ &= E(\mu T_{12}) + \text{Var}(\mu T_{12}) \\ &= \mu E(T_{12}) + \mu^2 \text{Var}(T_{12}). \end{aligned} \quad (\text{A6})$$

Analogously,

$$E(W_{12}) = 2\mu E(T_{12}) \quad \text{and} \quad \text{Var}(W_{12}) = 2\mu E(T_{12}) + 4\mu^2 \text{Var}(T_{12}). \quad (\text{A7})$$

The result (A1) now follows on substituting (A5) and (A6) into (A3) and (A7) into (A4), before substituting the resulting expressions into (A2). (Recall that $\eta_2 = m^2 + \sigma^2$.)

APPENDIX B

We describe here the details of the significance tests of demographic scenarios used in the article. Write S^2 and V , respectively, for the population variance and the normalized population variance, L for the number of loci in the data, and η_2 and η_4 , respectively, for the mutation mean square and the fourth moment of the distribution of mutation size. We use an overbar to denote the average of a quantity across loci in the data set, and Var to denote its variance across loci. Thus, for example, \bar{S}^2 and $\text{Var}(S^2)$ denote the average value and variance of S^2 across loci.

The three test statistics we consider are

$$F_1 \equiv \left[\frac{4}{3} (\bar{V})^2 \right] / \text{Var}(V), \quad (\text{B1})$$

$$F_2 \equiv \left[\frac{4}{3} (\bar{V})^2 + \left(\frac{1}{6} - \frac{2}{9L} \left(\frac{\bar{\eta}_4}{\eta_2^2} \right) \bar{V} \right) \right] / \text{Var}(V), \quad (\text{B2})$$

and

$$g \equiv \frac{(4/3)(\bar{S}^2)^2 + (1/6)\bar{S}^2}{\text{Var}(S^2)}. \quad (\text{B3})$$

The statistic g is the reciprocal of the g statistic introduced by REICH and GOLDSTEIN (1998), and so use of either statistic for testing is equivalent. To remove dependence of the numerator on $(\bar{\eta}_4/\eta_2^2)$, the statistic F_1 does not include a linear term in \bar{V} . The reason for the somewhat involved definition of F_2 is that we are aiming to rid the numerator of any dependence on $(\bar{\eta}_4/\eta_2^2)$, and under the assumption of constant population size,

$$E((\bar{V})^2) = N^2 \mu^2 + \frac{1}{L} \left(\frac{4}{3} N^2 \mu^2 + \frac{1}{6} N \mu (\bar{\eta}_4/\eta_2^2) \right).$$

For each statistic the null hypothesis is rejected for large values of the test statistic, corresponding to smaller variation across loci in the normalized or unnormalized

population variance than would be expected under the null hypothesis of constant population size.

For the values of L and n (the number of chromosomes in the sample) appropriate for each part of our data sets, we evaluated the null distribution of g by simulating 30,000 realizations of evolution with constant population size, $\theta \equiv 2N\mu = 4$ (where N is the number of chromosomes in the population), a simple stepwise mutation model, and no variation in either the mutation rate or the mutation mechanism across loci. For each of the scenarios in which there is variation across loci in the mutation rate, we found the null distribution of g by repeating these simulations, except that a population size of $N = 10,000$ was assumed, and in the simulations the mutation rate for each locus was chosen randomly according to

Moderate variability:

$$P(\mu = 10^{-5}) = P(\mu = 10^{-3}) = 0.05, \quad P(\mu = 10^{-4}) = 0.90$$

Medium variability:

$$P(\mu = 10^{-5}) = P(\mu = 10^{-3}) = 0.10, \quad P(\mu = 10^{-4}) = 0.80$$

High variability:

$$P(\mu = 10^{-5}) = P(\mu = 10^{-3}) = 0.20, \quad P(\mu = 10^{-4}) = 0.60.$$

(If variation in mutation rate across loci is assumed, the null distribution of g depends explicitly on N . With no such variation, it depends only on θ .)

In each case, the simulated null distribution for g (under the appropriate assumption about variability in μ) was also used as the null distribution for F_1 and F_2 . Significance levels for each of the three statistics were calculated as the percentage of times in the relevant simulation that the simulated value of g was larger than the observed value of the test statistic.

Our principal interest focuses on the use of the statistics F_1 and F_2 . To establish the validity of our procedure we carried out extensive simulations to check that the nominal P values calculated as just described were con-

servative, in understating the probability of a type I error, under more general assumptions about the processes involved.

First, we simulated from the null distribution of the F statistics under exactly the same assumptions as for g . Next, we weakened the assumption of simple stepwise mutation at each locus, simulating instead using the "two-phase" model introduced in DI RIENZO *et al.* (1994). Initially we assumed that all loci had the same two-phase distribution of mutation size, varying the parameters (notation as in DI RIENZO *et al.* 1994) in the range

$$p \in (0.0, 1.0), \quad \sigma_g^2 \in (0, 200), \quad \theta \in (0.01, 100).$$

For each of several sets of parameters spanning this range, 1000 data sets were simulated (with $L = 15$, $n = 100$).

Next, we introduced possible variation in the mutation mechanism across loci by choosing parameters for the two-phase model independently for each locus, with p being chosen uniformly over various ranges $((0,1)$, $(0.5, 1)$, $(0.8, 1)$, and $(0.9, 1)$) for each of which σ_g^2 was chosen uniformly over $(0, 50)$, $(0, 100)$, and $(0, 200)$. We also tried several discrete distributions on p and σ_g^2 in these ranges.

Finally, we allowed for uncertainty in the estimation of the mutation mean square at each locus by repeating the simulations described in the previous paragraph, but, in addition, using a value of η_2 for the locus that is chosen from a normal distribution, with mean given by the true value of η_2 (which is specified once the parameters for the mutation model are chosen) and variance $(\eta_2 - 1)^2/4$, independently for each locus. The choice of distribution for the sampling error in our estimates of mutation mean square is motivated by the bootstrap estimates of the sampling variability described in this article.

Each of these sets of simulations was performed under each of the assumptions about variation in mutation rate, with the nominal level of the test set at 0.05. On no occasion was the actual type I error >0.05 .