Heterogeneity of Microsatellite Mutations Within and Between Loci, and Implications for Human Demographic Histories

Anna Di Rienzo,*,†† Peter Donnelly,† Chris Toomajian,* Bronwyn Sisk,* Adrian Hill,‡ Maria Luiza Petzl-Erler,§ G. Ken Haines and David H. Barch††**

**Department of Anthropology, Northwestern University, Evanston, Illinois 60208, †Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom, ‡Wellcome Trust Center for Human Genetics and Institute of Molecular Medicine, University of Oxford, Oxford, OX3 9DU United Kingdom, §Department of Genetics, Federal University of Paraná, 81531-970 Curitiba, Brazil, **Department of Pathology, Northwestern University Medical School, Chicago, Illinois 60611 and ††Lurie Cancer Center, Northwestern University Medical School, Chicago, Illinois 60611*

Manuscript received July 16, 1997
Accepted for publication November 20, 1997

ABSTRACT

Microsatellites have been widely used to reconstruct human evolution. However, the efficient use of these markers relies on information regarding the process producing the observed variation. Here, we present a novel approach to the locus-by-locus characterization of this process. By analyzing somatic mutations in cancer patients, we estimated the distributions of mutation size for each of 20 loci. The same loci were then typed in three ethnically diverse population samples. The generalized stepwise mutation model was used to test the predicted relationship between population and mutation parameters under two demographic scenarios: constant population size and rapid expansion. The agreement between the observed and expected relationship between population and mutation parameters, even when the latter are estimated in cancer patients, confirms that somatic mutations may be useful for investigating the process underlying population variation. Estimated distributions of mutation size differ substantially amongst loci, and mutations of more than one repeat unit are common. A new statistic, the normalized population variance, is introduced for multilocus estimation of demographic parameters, and for testing demographic scenarios. The observed population variation is not consistent with a constant population size. Time estimates of the putative population expansion are in agreement with those obtained by other methods.

MICROSATELLITES, also called simple tandem repeat polymorphisms (STRPs), are an abundant group of repetitive DNA sequences with repeat units up to six nucleotides long. Their high level of variation in the number of repeat units is thought to reflect their high mutation rate (Tautz and Schlotterer 1994). Microsatellites have become important in genetics because they are involved in the phenomenon of repeat expansion and instability and, as excellent markers for genetic analysis, they have potential for the study of population processes.

Instability of simple sequences is responsible for a number of genetic diseases. There are two main types of instability. The first involves a single locus at a time and is characterized by expansion of the repeat and increased mutation rate at the germline and often somatic level (Ashley and Warren 1995). In most cases linkage disequilibrium studies have shown that only a subset of the chromosomes carrying a normal size allele are susceptible to expansion, raising the questions of whether and how the mutational processes producing the normal and the expanded alleles are mechanistically related (Imbert *et al.* 1993; MacDonald *et al.* 1992; Neville *et al.* 1994; Richards *et al.* 1992; Tanaka *et al.* 1996). The second type of instability is observed at the somatic level in cancer tissue; it is not associated with repeat expansion and involves many microsatellite loci per individual genome (Aaltonen *et al.* 1993; Thibodeau *et al.* 1993). The vast majority of hereditary colon cancer patients and 11–28% of sporadic colon cancer patients are characterized by microsatellite instability (Honchel *et al.* 1995), *i.e.*, the accumulation of somatic mutations in their cancer tissue. In hereditary colon cancer patients, this phenomenon, also called mutator phenotype, is due to defects in mismatch repair. In contrast, only about 43% of sporadic colon cancers with microsatellite instability have mutations in mismatch repair genes (Liu *et al.* 1995). Thus, the molecular bases of microsatellite instability in these patients are not completely understood; it is possible that other genes controlling genome stability may be altered in sporadic cancers with microsatellite instability. Mismatch repair corrects mispairs and insertion/deletion loops that occur spontaneously during DNA metabolic processes such as replication and recombination (Fishel and Wilson 1997; Kolodner 1996). This is of particular interest in the case of STRP loci because they are normally prone to forming such structures through slipped

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

strand mispairing (Levinson and Gutman 1987). Thus, somatic mutations of microsatellites in cancer may generally result from defects in replication fidelity. It may be hypothesized that the main difference between normal and cancer cells with microsatellite instability is that the latter, lacking the ability to correct mutations, simply show higher mutation rates, but do not differ in the mutation mechanism.

The high degree of polymorphism, with alleles that can be typed by automated assays, makes microsatellites promising candidates as tools for the study of population processes. However, certain assumptions about the mode of producing length variation are needed to relate the variation among populations to population genetic processes. The mechanism of slipped strand mispairing leads to a process generating gain or loss of repeat units. The most important implication is that alleles with the same repeat number may not be identical by descent because several mutations may have produced the same allele more than once. The stepwise mutation model, which assumes mutational changes of one repeat unit, is potentially suitable for describing microsatellite variation in populations (Shriver *et al.* 1993; Valdes *et al.* 1993). However, there may be substantial variation in mutation size among alleles as well as across different loci. In particular, the analysis of population patterns has led to the proposal that microsatellites mutate most frequently by gain or loss of one repeat unit, but more rare mutations of larger amplitude also occur (Di Rienzo *et al.* 1994). Such mutation events, even if rare, may complicate the interpretation of population patterns of variation. Several worldwide surveys of STRP variation have been carried out in human populations (Bowcock *et al.* 1994; Deka *et al.* 1995; Goldstein *et al.* 1995; Jorde *et al.* 1997). Some of these studies concluded that genetic distances based on the infinite allele model "perform" better than those based on the stepwise mutation model with regard to the reconstruction of human evolution (Jorde *et al.* 1995; Perez-Lezaun *et al.* 1997). These data taken together raise the possibility that microsatellite loci may vary in mutation processes and that the strict formulation of the stepwise mutation model may not be directly applicable to human population variation.

In this paper, we characterized the process generating variation in human populations at 20 microsatellite loci, on a locus-by-locus basis, to investigate possible interlocus heterogeneity of mutation patterns. This was done by analyzing the somatic mutations detected in a large sample of colon cancer patients. To determine whether the observed somatic mutations in cancer indeed reflect the "real" mutation process, the same microsatellites were typed in three population samples from different ethnic backgrounds. This approach allowed us to determine that the patterns of mutation in cancer are consistent with the pattern of population variation, implying that somatic mutations in cancer cells are useful for estimating the parameters of the "real" mutation process (Di Rienzo *et al.* 1995). We also developed new multilocus methods for the estimation of demographic parameters and for testing different demographic scenarios. The application of these methods to our present dataset suggests that the observed pattern of microsatellite variation is not consistent with a history of constant population size.

MATERIALS AND METHODS

Patient material: Study subjects were identified through a search of the Northwestern Memorial Hospital Tumor Registry, Chicago, IL. All patients with adenocarcinoma of the colon who had their primary diagnosis and primary resection of their colon carcinoma at Northwestern Memorial Hospital between January 1, 1984 and December 31, 1988 (219 cases) were identified. The original slides for each patient were reviewed by a pathologist (G.K.H.) and sections of tumor and normal tissue were cut from paraffin-embedded tissue blocks. The presence of tumor or normal tissue was confirmed by an additional hematoxylin and eosin stained slide. Additional sections were then cut from each block and the normal and tumor portions of the sections separated. DNA was then extracted from the tissue sections as described in Wright and Manos (1990).

Population samples: Sardinia (Italy) and Kaingang (Brazil) population samples were randomly selected from previously described samples in Di Rienzo and Wilson (1991) and (Belich *et al.* 1992), respectively. The DNAs were extracted either from placental tissue or peripheral blood of unrelated individuals selected from the general population. A Luo sample consists of unrelated healthy male blood donors collected in Saradidi, Kenya. Peripheral blood samples were obtained and the DNA was extracted according to a standard phenol-chloroform protocol. Each population sample consisted of 46 individuals.

Typing protocol: For both population and patient tissue samples, we used a typing protocol based on radioactively endlabeling one of the PCR primers as described in Di Rienzo *et al.* (1994). Population samples were amplified by using the following cycling profile: initial denaturation at 94° 2 min; 9 cycles of 94° 5 sec, 92° 30 sec, 55° 1 min, 72° 1 min; 31 cycles of 90° 40 sec, 55° 1 min, 72° 1 min; final extension at 72° 5 min. The patient tissue DNA samples were amplified with the same cycling profile by doubling the duration of each step except for the final extension. The tetranucleotide repeat at locus D19S244 was amplified using a radioactive incorporation protocol described in Weber and Wong (1993). PCR products were separated on 6% denaturing polyacrylamide gels and exposed to autoradiography film. Samples from the Centre d'Etudes du Polymorphisme Humain (CEPH) (Paris) database that have been widely used as size markers were run on each gel to ensure consistent allele identification across gels. Each gel was scored independently by two readers to check for consistency. In addition, 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.

Estimating mutation sizes: Owing to the high degree of heterozygosity of microsatellites, it was not possible to determine unequivocally which of the two alleles in the normal tissue mutates to produce the additional band(s) observed in the tumor DNA. We describe here four methods for estimating the distribution of mutation sizes at a locus from tumor data.

Measure length in repeat copy number and suppose l mutant alleles are observed at a particular locus. Write $Y_{i,1}$, and

$Y_{i,2}$, $i = 1, 2,\ldots,l$, for the two possible mutation sizes for each of the l somatic mutations observed, with these values being negative if mutation reduces allele length and positive otherwise. Each of the methods used effectively assigns a probability, $p_{i,1}$, that for the ith mutant, the mutation was of size $Y_{i,1}$. Probability $p_{i,2} \equiv 1 - p_{i,1}$ is then defined to be the probability that the mutation was of size $Y_{i,2}$. Having chosen values for the $p_{i,j}$, the underlying distribution of mutation size is estimated to give probability

$$\gamma(k) = \frac{1}{n} \sum_{i,j:Y_{ij}=k} p_{i,j} \qquad (1)$$

to changes of value k. That is, to estimate the probability of a change of a particular size k, add up the probabilities assigned to each of the n possible observed changes that take the value k and divide by n.

The simplest method assumes that the mutation was definitely of the smaller of the two possible (absolute) values. That is, it puts $p_{i,1} = 1$ if $|Y_{i,1}| < |Y_{i,2}|$, $p_{i,1} = 1/2$ if $|Y_{i,1}| = |Y_{i,2}|$, and $p_{i,1} = 0$ if $|Y_{i,1}| > |Y_{i,2}|$.

Two natural Bayesian methods make specific prior assumptions about the relationship between the probabilities of certain changes and the size changes involved. One such assumes *a priori* that the probabilities of certain changes decrease inversely with the (absolute) length of the change. Write $f_{i,j} = 1/|Y_{i,j}|$. Then put

$$p_{i,1} = \frac{f_{i,1}}{f_{i,1} + f_{i,2}}. \qquad (2)$$

If it is assumed *a priori* that probabilities decrease inversely with the square of the length change, one simply uses (2) with $f_{i,j}$ redefined as $1/Y_{i,j}^2$.

Each of the three methods just described tries to encapsulate the belief that of the two possible (absolute) changes, the smaller one is thought more likely than the larger one. The first method always adopts the smaller change, the third has a strong bias toward the smaller (absolute) change, and the second has a less marked bias toward the smaller (absolute) change. A comparison between the results of each of these methods and maximum likelihood estimation is given in Figure 1.

The maximum likelihood estimate of the underlying mutation size distribution cannot be found directly. Instead, we used the expectation maximization (EM) algorithm to approximate the maximum likelihood estimate (Dempster *et al.* 1977). The algorithm repeatedly updates estimates of the $\gamma(k)$ in a two stage process. First, we put all the $p_{i,j} = 1/2$, and define $\gamma(k)$ as in (1). Then repeat sequentially each of the two steps below until respective values of the γ's agree very closely in successive iterations. (1) Calculate the $p_{i,j}$ is from (2) with $f_{i,j} = \gamma(Y_{i,j})$, for the current values of the γ's. (2) Using the current values of the $p_{i,j}$, calculate new values for the γ's from (1).

Having estimated the mutation size distribution, we estimate its mean square by

$$\eta_2 = \sum_k k^2 \gamma(k). \qquad (3)$$

Some of the literature relates population statistics to the variance of the mutation size distribution. Such analyses typically assume that the mean of the mutation size distribution is 0, in which case the variance and mean square of the distribution are identical. In appendix i, we extend the theoretical results to general mutation size distributions. In this more general setting, the mean square of the distribution, rather than its variance, plays the central role.

Finally, we note that in connection with the constant population size scenario described below, some numerical adjustment is needed for X-linked loci, because of the differing number of chromosomes in the population. Perhaps the easiest correction is simply to redefine the mutation mean square for these loci, multiplying it by a factor of 3/4 for X-linked loci. With this change, the results in appendix i all apply, where N is now the effective number of chromosomes at an autosomal locus. We have made this adjustment in the presentation of our data. (A related adjustment would also be needed for data on Y-linked loci.)

RESULTS

Microsatellite instability screening: To observe an adequate number of somatic mutation events at each microsatellite locus, we used surgical specimens from patients with sporadic colon cancer as a source of paired normal and tumor DNA. Normal and tumor DNA patterns at 22 autosomal and two X-linked microsatellites were compared in each patient to detect somatic mutations (Table 1). A somatic mutation was considered to have occurred when there were one or more bands in the tumor tissue that were absent in the normal tissue (Figure 2). The fact that the same bands are seen in the normal and tumor tissue in the presence of an extra band can be explained either by the presence of some normal tissue in the tumor tissue sample or by the fact that the somatic mutation occurred after the onset of tumor growth. The average rate of instability per patient was 9.13%, with a slightly higher rate in tetranucleotide repeats compared to the di- and trinucleotide repeats.

Patterns of somatic mutations: Heterogeneity of mutation sizes may account for many of the inter-locus differences in population patterns of variability as well as the propensity to produce extreme size alleles, such as those leading to expanded triplets. Therefore, we examined the somatic mutation data to estimate the distribution of mutation sizes for each STRP. Because the mutation patterns described above (*e.g.*, Figure 2) do not allow unequivocal determination of mutation size, we used the EM algorithm, as described in materials and methods, to produce maximum likelihood estimates of the distribution of mutation sizes in the colon cancer patients. The distributions, shown in Figure 3, vary greatly in shape and range, supporting the idea that microsatellites may differ significantly in mutation patterns.

In our estimation we ignored the possibility of "double" (or higher) mutations, in which an observed mutant allele has undergone more than one mutation event, via unobserved intermediate alleles, from the progenitor normal allele. The growing tumor represents an exponentially expanding population of cells, and standard population genetics theory suggests that it is unlikely that mutant alleles that arise during growth will subsequently be lost to the population. Additional, empirical, support for our assumption comes from the fact that most individuals had at most one observable mutation.

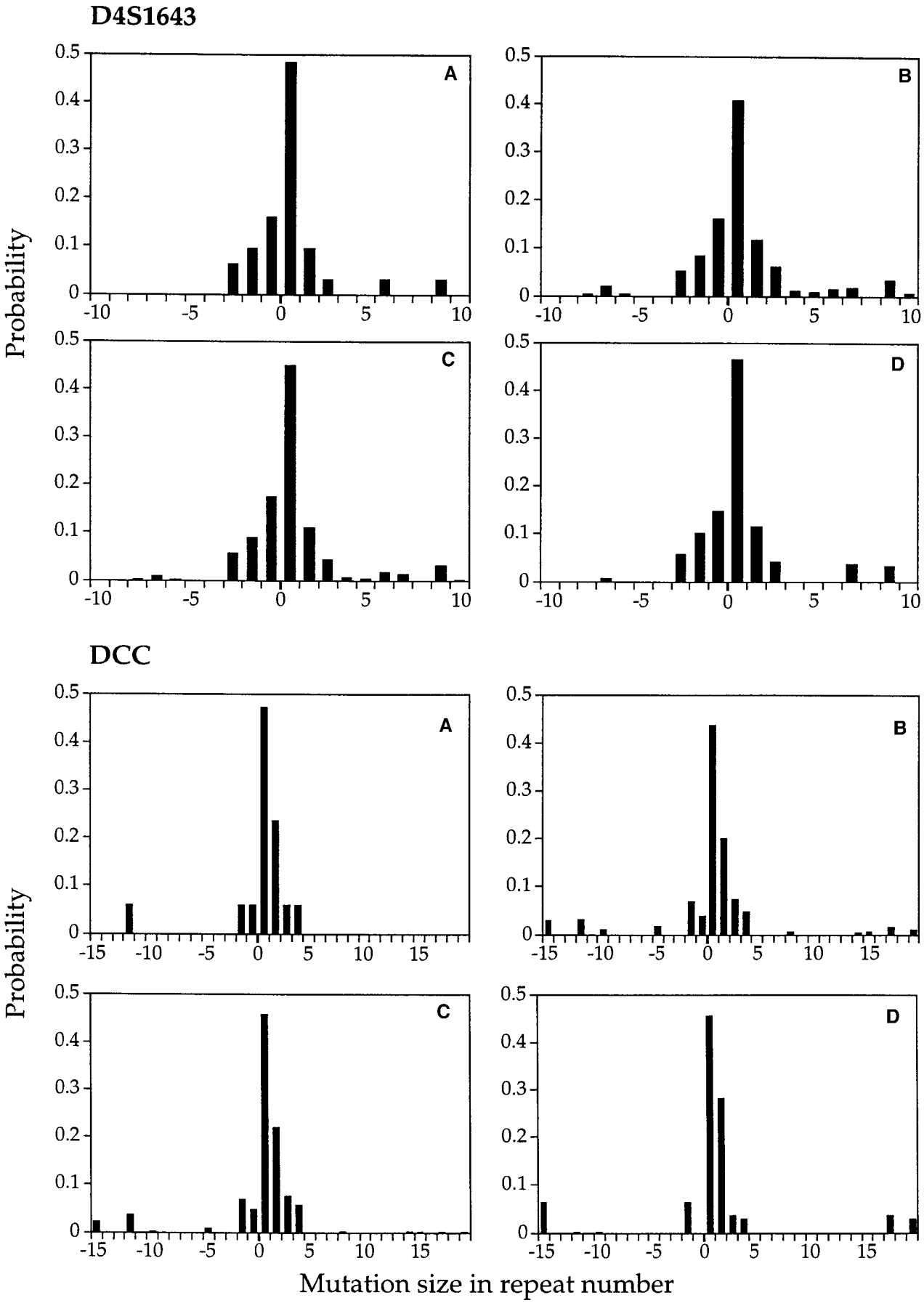
D4S1643
A
B
C
D
DCC
A
B
C
D
Probability
Probability
Mutation size in repeat number

TABLE 1

Screening for microsatellite instability at 24 loci

Locus	Repeat type	Chr	GDB no.	No. of patients	
				Unstable	Tested
DCC	DINUC	18	162735	15	202
D8S164	DINUC	8	160831	24	212
D19S49	DINUC	19	158604	21	208
D19S178	DINUC	19	160842	13	218
D22S156	DINUC	22	158863	9	218
D12S59	DINUC	12	160892	21	220
D19S47	DINUC	19	158602	23	218
MYC	DINUC	8	156494	15	217
CRYGB	TRINUC	2	155416	6	206
D13S308	TRINUC	13	210978	22	221
D5S373	TRINUC	5	162873	19	218
D5S556	TRINUC	5	209058	10	208
D6S366	TRINUC	6	162758	18	216
DM	TRINUC	19	523959	11	214
DRPLA	TRINUC	12	303908	20	208
DXS101	TRINUC	X	207649	27	214
CSF1R	TETNUC	5	156913	23	219
D4S243	TETNUC	4	161149	29	213
DXS981	TETNUC	X	162916	23	216
D4S1643	TETNUC	4	689648	29	218
D4S1647	TETNUC	4	691158	29	219
D19S244	TETNUC	19	162356	29	202
D19S252	TETNUC	19	162366	18	208
TH	TETNUC	11	225009	3	211

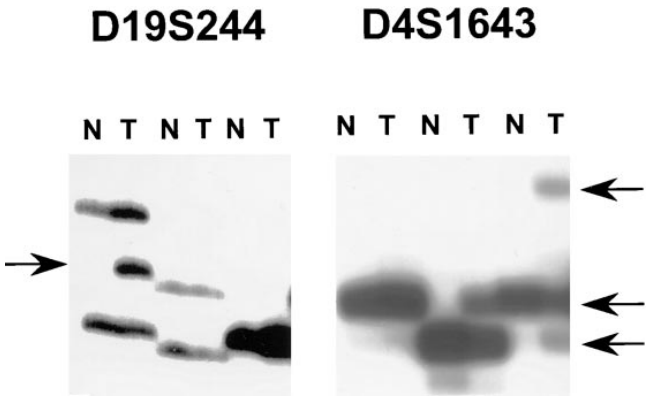

Figure 2.—Detection of somatic mutations in colon cancer patients at loci D19S244 and D4S1643. Arrows point to the bands observed in the tumor tissue lane (T) and absent from the normal tissue lane (N) which are the result of somatic mutations.

Some care should be exercised in interpreting the estimated distributions shown in Figure 3. For the distributions that show a wide range of possible mutation sizes (*e.g.*, D8S164), the sample sizes do not allow precise inferences about the underlying distribution. For example, the apparently "ragged" estimated distributions may arise even if the underlying distribution is much smoother. However, the number of observed mutations for these loci is comparable with those for loci showing a small range of mutation sizes (compare, for example, D8S164 and DXS981 each with 24 and 23 patients with somatic mutations) (see Table 1). Note that in estimating an unknown distribution, the method of maximum likelihood will tend to produce "clumped" estimates.

In addition to the maximum likelihood method, we implemented three natural Bayesian methods for estimating the underlying distribution of mutation sizes. These correspond respectively to the prior assumption that: (1) the shorter of the two possible mutation sizes occurred; (2) the probability of a given mutation size is inversely proportional to that size, and (3) the probability of a given mutation size is inversely proportional to the square of that size. Broadly, these methods gave estimates similar to those obtained by maximum likelihood. Figure 1 illustrates the four estimated distributions for two representative loci. Our subsequent analysis only uses the mean square of the distribution estimated by the maximum likelihood method, *i.e.*, the average squared mutation size, and neither this quantity nor our subsequent conclusions are sensitive to the estimation method.

It was previously proposed that microsatellites evolve directionally with a bias toward increase of repeat number (Rubinsztein *et al.* 1995). We examined this hypothesis by calculating the average mutation size in the estimated distribution for each locus. However, the numbers of loci with an average above and below zero were approximately equal, suggesting that any bias is either too weak to be detected in this analysis or it does not occur at the mutational level.

Population data: The pattern of population variation for a genetic locus depends on the features of the mutation process of the locus and the demography of the population examined. To test whether the somatic mutations in cancer tissue follow the same rules as the "real" mutation process underlying the population variation, we typed 16 of the same microsatellite loci in three population samples from different ethnic backgrounds: the Sardinian population from Europe, the Luos from Kenya, and the Kaingang from Brazil. These populations were chosen to maximize, within the human evolutionary timescale, the amount of independent evolution. Hence, the alleles observed in these populations may well share only a small part of their mutational

Figure 1.—Comparison of four methods to estimate distributions of mutation size for loci D4S1643 and DCC. (A) The shorter of the two possible mutation sizes is assumed to have occurred; (B) the probability of a given mutation size is inversely proportional to that size; (C) the probability of a given mutation size is inversely proportional to the square of that size; (D) maximum likelihood estimate through the EM algorithm.

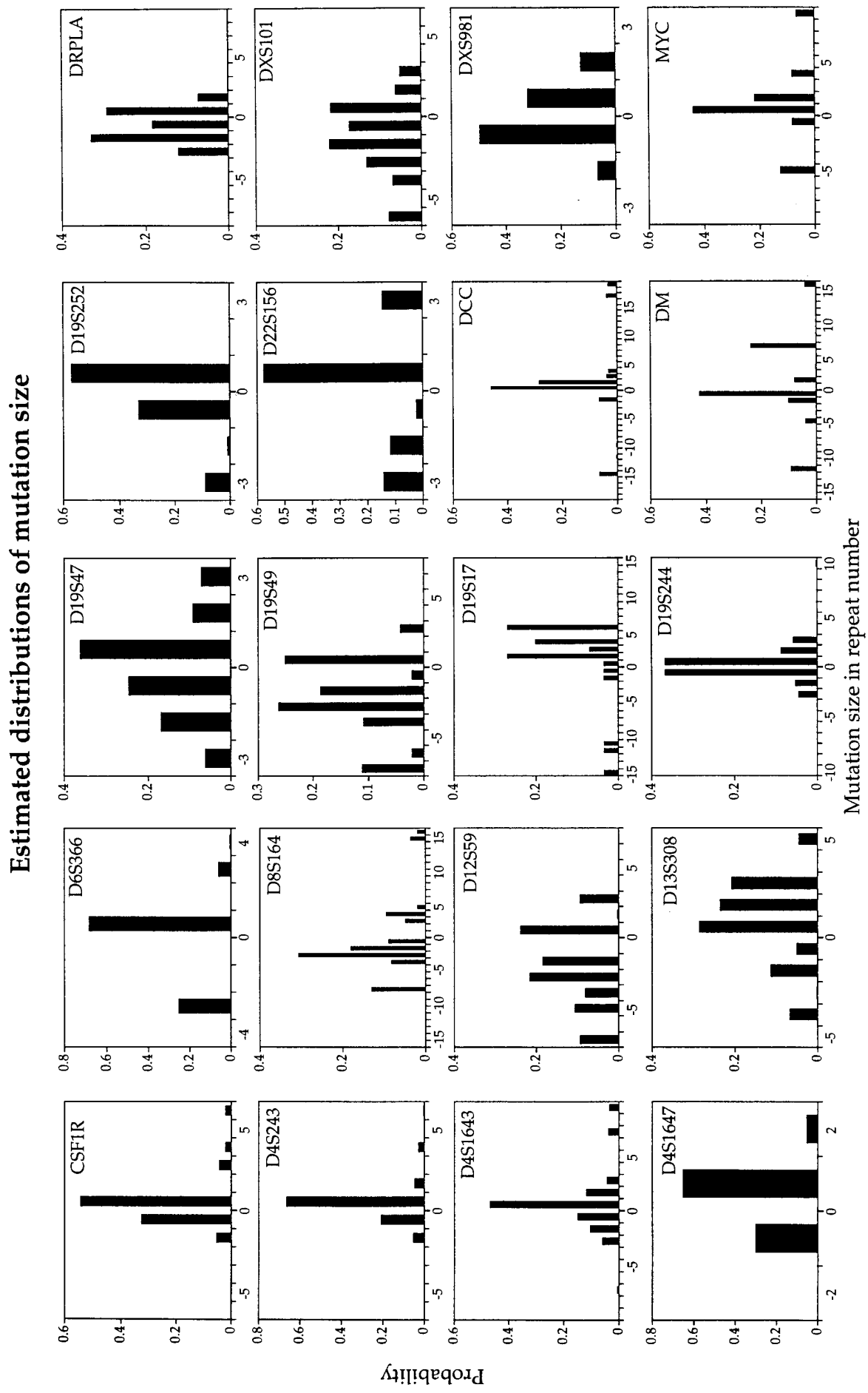

Figure 3.—Maximum likelihood estimates of the distribution of mutation size for 20 microsatellite loci.

history. Compared to the other populations, the Kaingang showed a lower degree of genetic heterogeneity, a higher homozygosity, and a lower number of alleles (Table 2). These results are in agreement with other surveys of genetic variation in Amerindian populations, and suggest a smaller effective population size, probably because of a bottleneck that occurred during the colonization of the Americas (Cavalli-Sforza *et al.* 1994). Also, in agreement with other surveys of microsatellite variation, the Sub-Saharan African population showed a level of heterozygosity that is marginally higher than that observed for the European population (Bowcock *et al.* 1994; Deka *et al.* 1994; Jorde *et al.* 1995; Jorde *et al.* 1997; Perez-Lezaun *et al.* 1997).

Expected relationship between mutation and population parameters: The generalized stepwise mutation model offers a broad theoretical framework for studying the process generating length variation at microsatellite loci (Di Rienzo *et al.* 1994; Ohta and Kimura 1973; Shriver *et al.* 1993; Valdes *et al.* 1993). It assumes that mutations for all alleles happen at the same rate and that when a mutation occurs the distribution of the mutation size does not depend on the length of the progenitor allele. Our use of the term "generalized" in conjunction with this model refers to the fact that we allow mutations of any integer number of repeat units rather than restricting attention to mutation sizes of ± 1 repeat unit. In other words, our model does not entail any constraints on allowable mutation sizes and allows possible biases toward increase or decrease of allele size.

We used population genetics theory to relate aspects of the mutation size distribution to patterns of variation expected in population samples (Di Rienzo *et al.* 1995). However, as with other forms of genetic data (Di Rienzo and Wilson 1991; Marjoram and Donnelly 1994; Rogers and Harpending 1992; Slatkin and Hudson 1991), these expected patterns are sensitive to the demographic history of the population. We considered two scenarios for demographic histories: constant population size and rapid demographic expansion. Such theoretical scenarios are to be considered extreme cases unlikely to reflect accurately the demographic history of any real human population. Nevertheless, we tested them as scenarios that have been widely used in the literature.

One summary of population variation that is convenient for assessing the consequences of various demographic assumptions is the variance of repeat number in a population sample, denoted here by S^2. By this we mean, for a particular locus, the sample variance of the collection of repeat copy numbers in the chromosomes sampled from the population. That is, if X_1, X_2,...,X_n denote the number of repeats in each of the n chromosomes sampled, and $\overline{X}$ denotes their average,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

TABLE 2

Summary of population data

	Luo (Africa)	Sardinia (Europe)	Kaingang (S. America)
Heterozygosity	0.96	0.95	0.89
Number of alleles	9.64	7.88	4.95
Variance of repeat number	8.30	7.49	5.26

Population genetics theory will be used to relate this to the mean square of the distribution of the mutation size, which we denote here by η_2. As described in materials and methods, we estimate the value of η_2 for each locus from the estimated distributions shown in Figure 3.

(a) Constant population size scenario: In this case the population in question is panmictic and has maintained a constant effective size of N chromosomes throughout its history. We assumed a generalized stepwise mutation model with arbitrary distributions of mutation sizes; in particular, we allow for asymmetric distributions as well as biases towards increase or decrease of repeat number. In appendix i, we show that under this scenario of demography and mutation, population theory predicts the following relationship

$$E(S^2) = N\mu\eta_2 \qquad (4)$$

where μ is the mutation probability per gene per generation.

In Equation 4, the expectation E pertains to averages over realizations of evolutionary and sampling processes. In a particular realization of evolution the actual pattern of variation observed in a sample for a particular microsatellite locus is the result of various chance processes. One of these is sampling: a different sample from the same population would not produce exactly the same pattern. The second and more important source of randomness stems from the chance events intrinsic to the evolutionary history of the population: unlinked loci within the same population and subject to identical mutation processes will give rise, because of chance events, to different patterns of variation. Thus, the actual value of S^2 observed in a population sample at a given locus is not expected to be given exactly by Equation 4, but to deviate from it because of the above chance effects.

One particular consequence of Equation 4 pertains to values of S^2 at different loci within a population. If we can assume that the mutation rate is approximately the same across these loci, and plot the observed values of S^2 against the mutation mean square, η_2, for each locus, the points should lie around a straight line. The slope of the line is proportional to the product of population size and mutation rate.

(b) Rapid population growth scenario: We considered a

scenario, based on mtDNA evidence, that the human population underwent rapid demographic growth in the past (Di Rienzo and Wilson 1991; Rogers and Harpending 1992; Slatkin and Hudson 1991). For comparative purposes, we make the same assumptions here about the mutation process as for the constant population size scenario.

Under this combined scenario of demography and mutation process, for a given microsatellite locus we derive in appendix i the following approximate relationship

$$E(S^2) = T\mu\eta_2 \qquad (5)$$

where T is the number of generations since the rapid expansion occurred. The interpretation of Equation 5 is similar to that of Equation 4 and analogous considerations about the chance variability apply. In particular, under the assumption of an approximately constant mutation rate across loci, a plot of the observed values of S^2, one for each locus, against the mutation mean square η_2 for that locus, should again give points lying around a straight line. For this scenario, the slope of the line is proportional to the product of the time since the population expansion, and the mutation rate.

Observed relationship between mutation and population parameters: Plots of the variance of repeat number against the mean square of mutation size estimated from the somatic mutations observed in cancer patients are shown in Figure 4 for each of the three populations in the study. If the constant population size scenario is valid, all loci within a population will share a common value of N (with suitable adjustment, as described in materials and methods, for X-linked loci), while under the rapid expansion scenario all loci within a population will share the same value of T. Neither of these values need necessarily be shared between populations; thus each population is shown separately.

As shown in Figure 4, in each case the points showed clustering around a straight line. This linear relationship observed consistently in each of the three population samples strongly suggests that the estimated mutation parameters based on the somatic mutations in colon cancer are good approximations of the "real" mutation parameters, namely those characterizing the process underlying the population variation (Di Rienzo *et al.* 1995). To assess the strength of the association between mutation mean square and population variance at each locus, we calculated Spearman's rank correlation coefficient. The calculated values of this correlation coefficient for the Sardinian, Luo, Kaingang, and pooled samples were 0.74, 0.81, 0.66, and 0.85, respectively. The related nonparametric test rejects the null hypothesis of no association with a one-sided P value less than 0.01 in each case. This empirical finding further supports our contention that the distributions of mutation sizes (Figure 3) may be a reliable reconstruction of the mutation process of microsatellites.

Normalized population variance: The above theoretical framework leads to the introduction of a new statistic for each locus, the normalized population variance (NPV), which is defined here as the ratio between the population variance, *i.e.*, the variance of repeat number at a locus in a population sample, and the (estimated) mutation mean square at the same locus. Thus, if we use S_j^2 and η_{2j}, respectively, to denote population variance and mutation mean square at locus j, the normalized population variance, V_j, for that locus is defined by

$$V_j = S_j^2/\eta_{2j}.$$

Intuitively, a locus with a large mutation mean square, and hence large variability in mutation size, is expected to show greater population variance, and allowance should be made for this effect when combining information across loci. The normalized population variance provides the "correct" measure for comparison across loci. Its use is particularly appropriate in light of the apparent heterogeneity of mutation processes across loci (Figure 3).

The theoretical expectations for the demographic scenarios outlined above can thus be expressed in terms of the normalized population variance:

$$E(V) = N\mu \text{ and } E(V) = T\mu$$

for the constant population size and the rapid expansion scenarios, respectively. Thus, the average value of the normalized population variance across the loci studied provides a natural estimate of $N\bar{\mu}$, in the constant population size scenario, and $T\bar{\mu}$ in the rapid expansion scenario, where, $\bar{\mu}$ is the average mutation rate across the loci considered. By using estimates of the (average) mutation rate, this offers a convenient way to estimate the parameters N and T. Table 3 shows such estimates for the population samples examined. It should be noted that our approach has the advantage that the estimates of underlying population parameters are obtained by combining independent loci, thus reducing the substantial variability inherent in single-locus estimates. In addition, the estimation allows and corrects for observed interlocus variability of mutation processes because it uses the mean square mutation size estimated individually for each locus rather than making restrictive and possibly unrealistic assumptions about the mutation mechanism, such as that gain or loss of only one repeat is possible. In essence, locus-specific information about the mutation mean square allows us to calibrate, and hence to combine efficiently, population variability at the different loci studied.

Discriminating between different demographic scenarios: While both demographic scenarios described above predict that the expected variance of repeat number in the population sample and the mean square of mutation size will be related linearly, there are major differences in other aspects of the predicted relationship between these two variables. As described in appen-

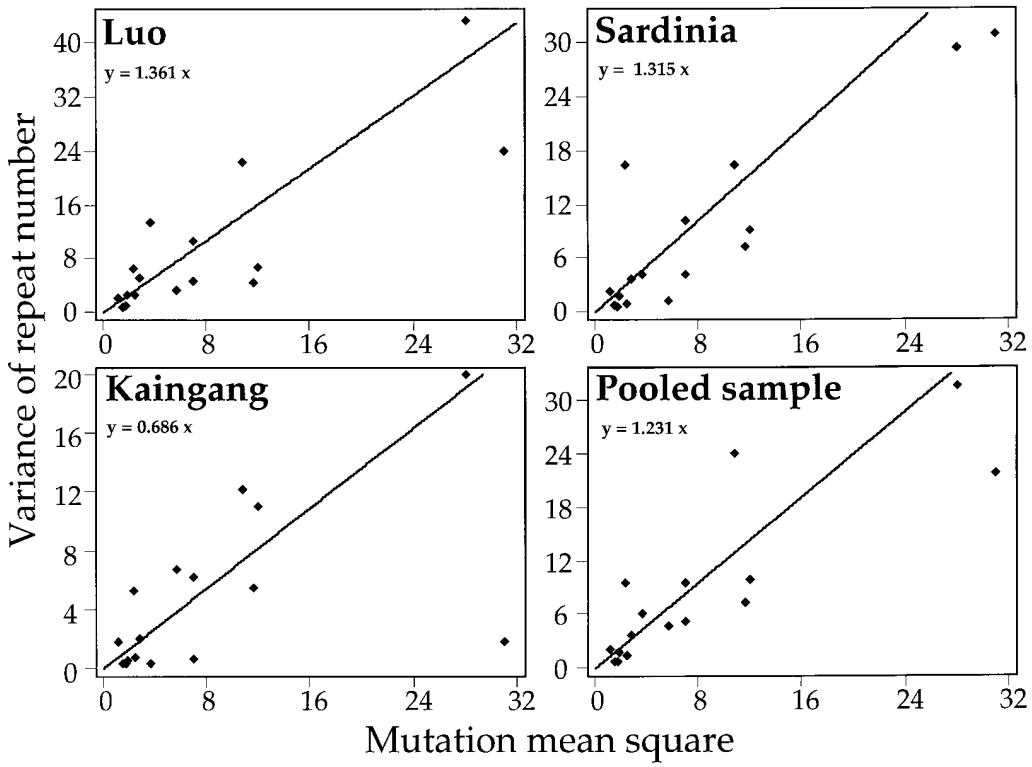

Figure 4.—Relationship between mutation mean square estimated in cancer patients and variance of repeat number in three human populations and in the pooled sample. The slope of the fitted line is given by the average normalized population variance across loci (without making the adjustments for X-linked loci). Since the predicted variance around the line in both demographic scenarios increases with increasing mutation mean square, it is not appropriate to use ordinary (least squares) regression techniques in fitting the line. The approach we use is equivalent to weighted least squares with weights equal to the inverse of the square of the mutation mean square value at each point, with the added condition that the fitted line passes through the origin.

dix i, one can quantify the amount of variability around the line expected under each scenario, *i.e.*, the extent to which the points are scattered around the line. Equivalently, one can assess the variability in measured values of the normalized population variance across loci. As shown in Table 3, much more scatter is expected under the constant population size scenario than in the case of a rapid expansion; loosely speaking, evolutionary variability (genetic drift) will be more pronounced in the constant population size scenario. Sampling variability is effectively the same in both scenarios.

It should be noted that in our data there are sources of variability additional to those incorporated in the model because we have had to use estimates of the mutation mean square. Thus, if the model were valid we should expect to observe more than the predicted variability. In addition, if mutation rates are not constant across loci, the variability around the line is expected to be even larger (see Equations A6 and A14 in appendix i). The average normalized population variances and their observed variances are presented in Table 3.

We can use population genetics theory to estimate a lower bound on the variance across loci of the normalized population variance for each of the two demographic scenarios. In addition, for each scenario of demography, we considered three models of mutation rate: (1) constant rate across loci, (2) high, and (3) moderate variability of rate across loci. A range of values of T and N were used to calculate the expected variance of the normalized population variance under the constant population size and the rapid expansion scenarios, respectively. The expected values of variance of the normalized population variance are shown in Table 3.

Even under the assumption of constant mutation rate across loci, we observe markedly less variability in normalized population variance than predicted under the constant population size scenario for the Luo, Kaingang, and pooled samples, and slightly less than predicted for the Sardinian sample. Allowing for even moderate variability in mutation rates, the observed variability is substantially less than predicted under the constant population size scenario for all samples. While a formal significance test does not seem feasible in this context, we interpret this as evidence against the assumption that the populations have evolved with roughly constant size. For a detailed discussion see appendix ii. Conversely, the level of observed variability is consistent with the rapid growth scenario if interlocus variability of mutation rate is incorporated into the model.

TABLE 3

Normalized population variance (NPV) and its variance

A.

Sample	Average (NPV)[a]	T (kya)[b]	N[b]	Observed Var. (NPV)
Luo	1.337	52,600–526,000	2,674–26,740	0.851
Sardinia	1.275	51,000–510,000	2,550–25,500	2.464
Kaingang	0.680	27,440–274,400	1,360–13,600	0.361
Pool	1.207	49,240–492,400	2,414–24,140	0.807

B.

	Predicted Var (NPV)[c]									
			Variable μ							
			High variability[d]				Moderate variability[e]			
	Constant μ		Constant *N*		Rapid growth		Constant *N*		Rapid growth	
			N		T(kya)		N		T(kya)	
Sample	Consant *N*	Rapid growth	5,000	10,000	50	250	5,000	10,000	50	250
Luo	3.10	0.09	11.12	35.16	0.97	22.09	5.39	12.26	0.34	6.37
Sardinia	2.85	0.08	10.87	34.91	0.96	22.08	5.14	12.01	0.33	6.36
Kaingang	0.98	0.04	9.00	33.04	0.92	22.04	3.27	10.14	0.29	6.32
Pool	2.59	0.08	10.60	34.64	0.96	22.08	4.88	11.75	0.33	6.36

[a] The values presented in the Table are with the adjustments for *X*-linked loci described in materials and methods. The corresponding values without the adjustments used to estimate *T* in the rapid expansion scenario are 1.361, 1.315, 0.686, and 1.231.

[b] Estimates of *T* and *N* are based on $\mu = 0.0005$–0.00005 and generation time of 20 years. Recall that *N* is twice the effective number of individuals in the population.

[c] The predicted variance of the normalized population variance was calculated from Equations A6 and A14 in appendix i, with $T\bar{\mu}$ and $N\bar{\mu}$ estimated by Average (NPV) and η_4 and η_2 estimated from the mutation size distributions shown in Figure 3.

[d] $\mu = 0.001$ for 20% of the loci; $\mu = 0.0001$ for 60% of the loci; and $\mu = 0.00001$ for 20% of the loci.

[e] $\mu = 0.001$ for 5% of the loci; $\mu = 0.0001$ for 90% of the loci; and $\mu = 0.00001$ for 5% of the loci.

Analysis of the pooled population sample: All three populations showed a pattern consistent with rapid growth and the estimates of *T* for Sardinia and Luo were reasonably similar. Therefore, we pooled the microsatellite typing data for the three populations and calculated the variance of repeat number in the pooled sample. We then plotted the mutation mean square against the pooled population variance, as shown in Figure 4. As with the single population analysis, we observe a linear relationship and a low variability of the normalized population variance. Again, the average across loci of the normalized population variance can be used to obtain an estimate of *T*, which in this context represents the coalescence time for the species under the assumption of rapid population growth. Interestingly, the estimate of the time since the rapid population growth obtained on the pooled sample is very similar to those for the Sardinian and Luo samples (see Table 3).

DISCUSSION

It is not easy to obtain experimental evidence to describe the mutation process of microsatellites on a locus-by-locus basis. Here, we present a novel experimental and theoretical approach to this problem that uses the mutator phenotypes of cancer patients with microsatellite instability and the theoretical framework of the generalized stepwise mutation model. By analyzing the somatic mutations occurring in the tumor tissue of such patients, we estimated the distribution of mutation sizes and the mean square mutation size for each locus independently. The variance of repeat number in a population sample is related to the mean square of the size of the mutations that produce the observed population variation. By using our estimated mean square mutation size in tumor tissue, we observed the expected relationship supporting the idea that the mutation sizes in cancer cells are similar to those occurring in the germline. Further analysis of the relationship between mean square mutation size and variance of repeat number in human population samples allowed testing of different demographic scenarios of human evolution.

Somatic mutations in cancer and in the germline of the general population: Our integrated analysis of microsatellite instability and population variation demanded testing of the idea that mutation patterns in cancer do not differ substantially from those in normal

germline cells. In agreement with the predictions of population genetic theory, a significant correlation between the variance of repeat number and the mutation mean square is evident even though the latter was obtained through the analysis of somatic mutations in cancer. The predicted relationship between the population and mutation parameters is observed consistently in three samples from distantly related populations, which represent partially independent realizations of evolution.

Differences between the mutations observed in cancer patients and those underlying population variation might be expected if the mechanisms maintaining genome stability recognized and fixed certain mutation types more efficiently than others, *e.g.*, replication slippage events involving a larger number of repeats were more likely to be recognized. One such mechanism is mismatch repair, but others may exist in sporadic cancer patients (daCosta *et al.* 1995).

Information on mutation patterns resulting from defects in genome stability is available only for mismatch repair mutants in yeast. In this organism, the rate and pattern of mutations were compared in wild-type and mismatch repair mutant strains, including *msh2* mutants (the human homolog of *msh2* is the most common mutant in hereditary colon cancer) (Sia *et al.* 1997; Wierdl *et al.* 1997). The rate of repeat mutations in mismatch repair mutants relative to that in the wild-type strain decreases gradually, going from a mono-nucleotide repeat to a 20-nucleotide repeat. In addition, mismatch repair mutants showed a higher frequency of short (± 1 repeat unit) mutations relative to the wild type. This suggested that the mismatch repair binding apparatus in yeast recognizes and fixes short mutations more efficiently than large ones.

The significant proportion of relatively large mutations in our sporadic cancer patients as well as a higher (rather than lower) rate of instability at tetranucleotide compared to di- and tri-nucleotide repeats contrasts with the above scenario of mismatch repair mutants in yeast. This difference could be due to the presence of mutations in genes other than those involved in mismatch repair. In fact, mutations in mismatch repair genes could be identified only in approximately 43% of sporadic cancers with microsatellite instability (Liu *et al.* 1995). Alternatively, the mismatch repair genes of yeast and human have significantly diverged with regard to mutation size specificity. Possible differences between yeast and human with regard to the mutation specificity of mismatch repair proteins may be seen also in colon cancer cell lines known to harbor mismatch repair defects. Unlike the yeast mismatch repair mutants, single cell clones from these cell lines show a higher rate of instability in tetranucleotide compared to dinucleotide repeats (G. Ybazeta, B. Scaglione-Sewell, and A. Di Rienzo, unpublished results). The possible divergence between the yeast and human mismatch repair systems could be reflected in the higher relative abundance of microsatellites in the human genome compared to yeast.

Our finding of a significant correlation between the mutation parameters estimated in cancer and the population parameters in samples from three ethnically diverse human populations suggests that the mutation process in our sporadic cancer patients is closely related to the process that generates population variation at human microsatellites.

Mutation sizes have a wide spectrum: Previous surveys of microsatellite mutations have consisted of the observation of mutation events in family studies. Over all loci, the majority of mutations (78% in Weber and Wong 1993) changed allele size by a single repeat unit. However, due to the small number of observations at any given locus, rare mutation types may have been missed. One interesting feature of the estimated distributions of mutation sizes in Figure 3 is that while most mutations involve one or two repeat units, many loci also show a substantial proportion of large length changes. Also of note is the heterogeneity across loci: only a proportion of the loci show large changes and/or possible asymmetries. The molecular causes of this heterogeneity are not clear. Based on findings on minisatellite mutations (Monckton *et al.* 1994), it might be postulated that flanking regions may play a role in shaping the mutation process, a suggestion that is also consistent with the founder effect observed at several unstable triplet loci (Imbert *et al.* 1993; MacDonald *et al.* 1992; Neville *et al.* 1994; Richards *et al.* 1992; Tanaka *et al.* 1996).

Estimating population parameters from multilocus microsatellite data: Most estimators of population parameters in the literature are of necessity based on single-locus data. It is well understood that because of the extent of evolutionary variability, estimates based on such data have limited precision (Donnelly and Tavarè 1995; Hudson 1990, 1992; Tajima 1989).

Here, we have shown how to exploit locus-specific information on microsatellite mutation processes to combine data from distinct loci in estimating population parameters. Central to our approach is the introduction of what we have called normalized population variance, *i.e.*, the population variance at a locus divided by the estimated mean square mutation size. Our theoretical model explicitly allows for differences across loci in the distribution of mutation sizes (in particular, we do not assume that changes only involve one repeat unit) and we have shown that the use of the normalized population variance properly allows for this interlocus variability.

Data from different loci can then be combined using the normalized population variances. For the two demographic scenarios that were considered, the average of these normalized population variances across loci leads to a natural estimator for the effective population size, or time since the rapid expansion, respectively. The ability to use data from multiple loci has considerable advantages for estimating population parameters or for

the assessment of evolutionary hypotheses. Unlinked loci effectively provide independent sources of information on population parameters, so that as in classical statistics the variance of our estimators decreases as the inverse of the number of loci involved. In addition, the combination of information across loci acts to reduce the effect of special features (such as selection) at particular loci, another problem that may affect single-locus analyses.

Theoretical novelties of this approach include the derivation of the mean and variance of population variance under extreme population growth, for a general stepwise mutation model. (In particular, there are no restrictions on mutation size, and the distribution of mutation size is allowed to have a bias toward either increase or decrease, and to be asymmetric.) We have also derived the mean of the population variance, for this level of generality in the distribution of mutation size, in the constant population size scenario.

We also note that information on the mean square (or variance if the mean is assumed to be 0) of mutation size at different loci could be used to improve the use of recently introduced measures of genetic distance for microsatellite loci [see for example Goldstein *et al.* (1995) and references therein]. Corrections for interlocus heterogeneity analogous to the normalized population variance would allow more efficient combination of data across loci, and hence increase the precision of the resulting distances.

Discriminating between different demographic scenarios: In addition to the estimation problem, we have developed a multilocus method for discriminating among competing demographic scenarios. The central idea is that while both scenarios should result in a linear relationship between certain statistics, they make very different predictions of the amount of variability to be expected across loci.

The finding of relatively low variability in the normalized population variance is not consistent with the scenario of constant population size. It should be noted that our results are biased toward higher variability because of the estimation of the mutation mean square. This bias further supports the rejection of the constant population size scenario. However, the observed variability in the normalized population variance is higher than expected under the rapid expansion scenario, if the mutation rate is assumed to be constant across loci. This may be due either to the estimation of the mutation mean square or to a violation of the assumption of lack of interlocus variability of mutation rate; these explanations are not mutually exclusive.

It is important to ask whether departures from the assumptions of the mutation model, rather than failure of the constant population size scenario, may explain the observed low variability in the normalized population variance. One such possibility is the presence of constraints on microsatellite allele size. It has been proposed that microsatellite variation between species is subject to size constraints; more specifically, that the expected difference in allele sizes between human and chimpanzee is larger than observed (Bowcock *et al.* 1994; Garza *et al.* 1995). The interpretation of these data with regard to intraspecific variation is problematic. In fact, when different human populations are compared the interpopulation differences in average allele sizes are smaller than between human and chimpanzee. This may suggest that, even if allele size is constrained, the evolutionary time scale for human populations may not have been long enough for such constraints to come into effect. If constraints on allele size exist, it is unclear whether they would act at the mutation or selection level. It is difficult to develop theoretical expectations for a mutation model with constrained allele size since there are many different ways in which such constraints may operate.

Mutation rates which depend on the length of the progenitor allele would also violate the assumptions of the generalized stepwise model. The most obvious possibility is presumably for mutation rates to increase with allele length (Chung *et al.* 1993; Weber 1990; Zhang *et al.* 1994). This would induce differences between loci, with those with larger average allele length tending to have higher mutation rates. In itself, this should have the same effect as differences between loci in mutation rate. It would increase the variability in normalized population variance between loci, in contrast to the observed underdispersion under the constant population size scenario. This effect would violate modeling assumptions only if the dependence of mutation rate on allele size is strong enough to induce reasonable differences in mutation rate for different alleles at the same locus. While a formal analysis of this situation is not straightforward, one might expect it to increase variability over the constant mutation rate scenario, again in contrast to the data.

Implications for human evolution: The hypothesis of a rapid population growth during the evolutionary history of human populations has been proposed on the basis of mtDNA data showing a star-shaped genealogy and an approximately Poisson distribution of pairwise sequence differences (Di Rienzo and Wilson 1991; Rogers and Harpending 1992). Theoretical analyses of demographic patterns revealed that only specific scenarios of population expansion, namely rapid growth from a very small initial population size, would produce the patterns of mtDNA sequence variation mentioned above (Marjoram and Donnelly 1994; Slatkin and Hudson 1991). The mtDNA data could also be explained by a selective sweep acting on an advantageous mutation, without the need for invoking population growth. (For further discussion, see Di Rienzo and Wilson 1991; Excoffier 1990; Marjoram and Donnelly 1997).

Our findings are difficult to reconcile with a constant

population size scenario, but are consistent with a rapid population growth and are in agreement with the above proposal. The multilocus analysis performed here allows us to rule out the possibility that our findings are the results of selection.

Under the assumption of extreme population growth, our analysis allows for multilocus estimation of the time since the expansion event occurred. These estimates, for the three populations separately, and for the pooled population sample are given in Table 3. The estimates of T obtained for the pooled sample can be interpreted as the time since the most recent common ancestor of humans. They are not inconsistent with analogous estimates obtained for mtDNA and *Y* chromosome loci (Vigilant *et al.* 1991; Hasegawa *et al.* 1993; Hammer 1995; Tavar *et al.* 1997).

This work was supported in part by grants from the National Science Foundation (SBR-9317266 to A.D.R. and DMS-9505129 to P.D.) and the American Cancer Society, Illinois Division, to A.D. and a UK EPRSC Advanced Fellowship (B/AF1255) to P.D. We thank M. Nordborg, R. Thomas, and G. Wichmann for technical help, D. J. Balding, D. R. Cox, and P. McCullagh for helpful discussions, and M. Kreitman, C. Ober, and A. Turkewitz for critical reading of the manuscript.

LITERATURE CITED

Aaltonen, L. A., P. Peltomaki, F. S. Leach, P. Sistonen, L. Pylkkanen *et al.*, 1993 Clues to the pathogenesis of familial colorectal cancer. Science **260:** 812–816.

Ashley, C. T., and S. T. Warren, 1995 Trinucleotide repeat expansion and human disease. Annu. Rev. Genet. **29:** 703–728.

Belich, M. P., J. A. Madrigal, W. H. Hildebrand, J. Zemmour, R. C. Williams *et al.*, 1992 Unusual HLA-B alleles in two tribes of Brazilian Indians. Nature **357:** 326–329.

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.

Box, G. E. P., 1953 Non-normality and tests on variances. Biometrika **40:** 318–334.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza, 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Chung, M., L. P. W. Ranum, L. A. Duvick, A. Servadio, H. Y. Zoghbi *et al.*, 1993 Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nat. Genet. **5:** 254–258.

daCosta, L. T., B. Liu, W. S. El-Deiry, S. R. Hamilton, K. W. Kinzler *et al.*, 1995 Polymerase delta variants in RER colorectal tumors. Nat. Genet. **9:** 10–11.

Deka, R., S. DeCroo, L. Jin, S. T. McGarvey, F. Rothhammer *et al.*, 1994 Population genetic characteristics of the D1S80 locus in seven human populations. Hum. Genet. **94:** 252–258.

Deka, R., L. Jin, M. D. Shriver, L. M. Yu, S. DeCroo *et al.*, 1995 Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. Am. J. Hum. Genet. **56:** 461–474.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood estimation from incomplete data via the EM algorithm. J. Roy. Stat. Soc. B **39:** 1–38.

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., and A. C. Wilson, 1991 Branching pattern in the evolutionary tree for human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **88:** 1597–1601.

Donnelly, P., and S. Tavarè, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

Excoffier, L., 1990 Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J. Mol. Evol. **30:** 125–139.

Fishel, R., and T. Wilson, 1997 MutS homologs in mammalian cells. Curr. Opin. Genet. Dev. **7:** 105–113.

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

Goldstein, D. B., L. A. Ruiz, L. L. Cavalli-Sforza and M. W. Feldman, 1995 Genetic absolute dating based on microsatellites and the origin of modern humans. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

Hammer, M. F., 1995 A recent common ancestry for human Y chromosomes. Nature **378:** 376–380.

Hasegawa, M., A. Di Rienzo, T. D. Kocher and A. C. Wilson, 1993 Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. **37:** 347–354.

Honchel, R., K. C. Halling and S. N. Thibodeau, 1995 Genomic instability in neoplasia. Sem. Cell Biol. **6:** 45–52.

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

Hudson, R. R., 1992 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by T. N. C. A. G. Sinauer Associates, Sunderland, MA.

Imbert, G., C. Kretz, K. Johnson and J. L. Mandel, 1993 Origin of the expansion mutation in myotonic dystrophy. Nat. Genet. **4:** 72–76.

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.

Jorde, L. B., A. R. Rogers, M. Bamshad, W. S. Watkins, P. Krakowiak *et al.*, 1997 Microsatellite diversity and the demographic history of modern humans. Proc. Natl. Acad. Sci. USA **94:** 3100–3103.

Kimmel, M., and R. Chakraborty, 1996 Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor. Pop. Biol. **50:** 345–367.

Kolodner, R., 1996 Biochemistry and genetics of eukaryotic mismatch repair. Genes Dev. **10:** 1433–1442.

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

Liu, B., N. C. Nicolaides, S. Markowitz, J. K. Willson, R. E. Parsons *et al.*, 1995 Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. Nat. Genet. **9:** 48–55.

MacDonald, M. E., A. Novelletto, C. Lin, D. Tagle, G. Barnes *et al.*, 1992 The Huntington's disease candidate region exhibits many different haplotypes. Nat. Genet. **1:** 99–103.

Marjoram, P., and P. Donnelly, 1994 Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics **136:** 673–683.

Marjoram, P., and P. Donnelly, 1997 Human demography and the time since mitochondrial Eve, pp. 107–131 in *Progress in Population Genetics and Human Evolution*, edited by P. D. S. Tavare. Springer-Verlag, New York.

McCullagh, P., 1987 *Tensor Methods in Statistics*. Chapman & Hall, London.

Monckton, D. G., R. Neumann, T. Guram, N. Fretwell, K. Tamaki *et al.*, 1994 Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. Nat. Genet. **8:** 162–170.

Neville, C. E., M. S. Mahadevan, J. M. Barcelo and R. G. Korneluk, 1994 High resolution genetic analysis suggests one ancestral predisposing haplotype for the origin of the myotonic dystrophy mutation. Hum. Mol. Genet. **3:** 45–51.

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

Perez-Lezaun, A., F. Calafell, E. Mateu, D. Comas, R. Ruiz-Pacheco *et al.*, 1997 Microsatellite variation and the differentiation of modern humans. Hum. Genet. **99:** 1–7.

Pritchard, J. K., and M. W. Feldman, 1996 Statistics for microsatellite variation based on coalescence. Theor. Pop. Biol. **50:** 325–344.
Richards, R. I., K. Holman, K. Friend, E. Kremer, D. Hillen *et al.*, 1992 Evidence of founder chromosome in fragile X syndrome. Nat. Genet. **1:** 257–260.
Roe, A., 1992 Correlations and interactions in random walks and population genetics, pp. 137–203. University of London, London.
Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.
Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain *et al.*, 1995 Microsatellite evolution—evidence for directionality and variation in rate between species. Nat. Genet. **10:** 337–343.
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.
Sia, E. A., R. J. Kokoska, M. Dominska, P. Greenwell and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol. Cell. Biol. **17:** 2851–2858.
Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Tanaka, F., M. Doyu, Y. Ito, M. Matsumoto, T. Mitsuma *et al.*, 1996 Founder effect in spinal and bulbar muscular atrophy (SBMA). Hum. Mol. Genet. **5:** 1253–1257.
Tautz, D., and C. Schlotterer, 1994 Simple sequences. Curr. Opin. Genet. Dev. **4:** 832–837.
Tavaré, S., D. J. Balding, R. C. Griffiths and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145:** 505–518.
Thibodeau, S. N., G. Bren and D. Schaid, 1993 Microsatellite instability in cancer of the proximal colon. Science **260:** 816–819.
Valdes, A. M., M. Slatkin and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.
Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes and A. C. Wilson, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.
Weber, J. L., 1990 Informativeness of human (dC-dA)n(dG-dT) polymorphisms. Genomics **7:** 524–530.
Weber, J. L., and C. Wong, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. **2:** 1123–1128.
Wierdl, M., M. Dominska and T. D. Petes, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics **146:** 769–779.
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.
Zhang, L., E. P. Lee flang, J. Yu and N. Arnheim, 1994 Studying human mutations by sperm typing: instability of CAG trinucleotide repeats in the human androgen receptor gene. Nat. Genet. **7:** 531–535.
Zhivotovsky, L. A., and M. W. Feldman, 1995 Microsatellite variability and genetic distances. Proc. Natl. Acad. Sci. USA **92:** 11549–11552.

Communicating editor: M. Slatkin

APPENDIX I

Theory for rapid expansion: Focus attention on a particular locus within a population. It is natural to derive results via genealogical methods. In the case of a large constant size population, the genealogy is well described by the coalescent. The effect on this genealogical tree of changes in population size is well understood, at least under plausible assumptions on the dynamics of the process governing change in the population size (Slatkin and Hudson 1991; Rogers and Harpending 1992; Donnelly and Tavarè 1995). Loosely speaking, the consequence of a rapid expansion in population size from an initially small value is to make the genealogical tree effectively star-shaped. To facilitate the analysis here, we will approximate the setting of rapid expansion by assuming that under these conditions the tree is actually star-shaped, that is, that all lineages trace back and coalesce exactly at the common ancestor of the sample. Write T for number of generations since this common ancestor, that is, the depth of the tree. It is well known (*e.g.*, Donnelly and Tavarè 1995) that the effects of mutation act independently on different branches of the genealogical tree.

Write A for the number of repeats in the common ancestor. Then write M_i for the change from the common ancestor along the lineage to the ith chromosome sampled, so that X_i, the observed repeat number in the ith chromosome, is given by

$$X_i = A + M_i.$$

Then, writing $\overline{X}$ for the sample average: $\overline{X} = n^{-1}(X_1 + X_2 + \cdots + X_n)$, and analogously for $\overline{M}$,

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(M_i - \overline{M})^2 \\ &\equiv S_M^2. \end{aligned}$$

Now, the random variables M_i are independent and identically distributed. Write W for the number of mutation events along the lineage leading from the common ancestor to the ith chromosome. The random variable W has a binomial distribution with parameters T and μ, the mutation probability at the locus in question. Since T is large and μ small, we will assume that in fact W has a Poisson distribution with mean $T\mu$. If F denotes the distribution governing mutation sizes, we can write

$$M_i = \sum_{j=1}^{W} Z_j,$$

where the Z_j are independent and identically distributed with distribution F: Z_j denotes the change in length caused by the jth mutation.

It turns out to be convenient to calculate the cumulants (see *e.g.*, McCullagh 1987) of the M_i. Now, writing $\phi(t)$ for the moment generating function of the distribution F,

$$\begin{aligned} \mathrm{E}(e^{tM_i}) &= \mathrm{E}(e^{tM_i}|W) \\ &= \mathrm{E}\{[\phi(t)]^W\} \\ &= e^{-T\mu[1-\phi(t)]}, \end{aligned}$$

using the assumption that W has a Poisson distribution with mean $T\mu$. It follows that the cumulant generating function of M_i is $T\mu\phi(t) - T\mu$, so that, writing κ_r for the rth cumulant of M_i and η_r for the rth moment of F, about the origin: $\eta = \mathrm{E}(Z_j^r)$

$$\kappa_r = T\mu\mathrm{E}(Z_j^r) = T\mu\eta_r. \qquad \text{(A1)}$$

Standard results for independent and identically distributed random variables (McCullagh 1987) then give

$$\mathrm{E}(S^2) = \kappa_2 = T\mu\eta_2, \tag{A2}$$

and

$$\begin{aligned}\mathrm{Var}(S^2) &= \frac{\kappa_4}{n} + \frac{2\kappa_2^2}{n-1} \\ &= \frac{T\mu\eta_4}{n} + \frac{2T^2\mu^2\eta_2^2}{n},\end{aligned} \tag{A3}$$

ignoring terms of order n^{-2}. Similar formulas are given in Kimmel and Chakraborty (1996), although there appears to be a misprint in the first term of their Equation 23.

Thus, writing V for the normalized population variance (NPV),

$$\mathrm{E}(V) = T\mu \text{ and } \mathrm{Var}(V) = \frac{T\mu}{n}\left(\frac{\eta_4}{\eta_2^2} + 2T\mu\right). \tag{A4}$$

If the mutation rate varies across loci, then

$$\mathrm{E}(V) = T\bar{\mu} \tag{A5}$$

and

$$\begin{aligned}\mathrm{Var}(V) &= \mathrm{E}[\mathrm{Var}(V|\mu)] + \mathrm{Var}[\mathrm{E}(V|\mu)] \\ &= \frac{T\bar{\mu}}{n}\left(\frac{\eta_4}{\eta_2^2} + 2T\bar{\mu}\right) + T^2\mathrm{Var}(\mu)\left(1 + \frac{2}{n}\right),\end{aligned} \tag{A6}$$

where $\bar{\mu}$ and $\mathrm{Var}(\mu)$ are the average and variance, respectively, of the mutation rate. With NPV values for L different loci, their average $\bar{V}$ provides a natural estimate of $T\bar{\mu}$. Further, the sampling variance of the estimator is

$$\mathrm{Var}(\bar{V}) = \frac{T\bar{\mu}}{Ln}\left(\frac{\eta_4}{\eta_2^2} + 2T\bar{\mu}\right) + \frac{T^2\mathrm{Var}(\mu)}{L}\left(1 + \frac{2}{n}\right), \tag{A7}$$

Note the decrease in this sampling variance as a function of the number of loci, and for loci with the same value of μ, [and hence $\mathrm{Var}(\mu) = 0$] the decrease also as a function of the sample size n (assumed here to be the same at all loci, though this is easily relaxed).

Theory for constant population size: In the context of constant large population size, genealogy is well described by the coalescent. Write X_i, M_i, and S^2, as above, for the number of repeats in the ith sampled chromosome, the change in repeat number in that chromosome since the common ancestor of the sample, and the population variance, respectively. Then, as above, $S^2 = S_M^2$, so that

$$\begin{aligned}\mathrm{E}(S^2) &= \frac{1}{n-1}\left[\mathrm{E}\left(\sum_{i=1}^{n} M_i^2\right) - n\mathrm{E}(\bar{M}^2)\right] \\ &= \frac{1}{n-1}\left[n\mathrm{E}(M_1^2) - n^{-1}\mathrm{E}\left(\sum_{i=1}^{n} M_i^2 + \sum_{i\neq j} M_iM_j\right)\right] \\ &= \mathrm{E}(M_1^2) - \mathrm{E}(M_1M_2).\end{aligned} \tag{A8}$$

Now, consider the lineages leading to the first two chromosomes in the sample. It may be that the common ancestor of these two chromosomes is the common ancestor of the sample, or it may be that their common ancestor occurred more recently than the common ancestor of the entire sample, in which case the two chromosomes will have shared some ancestry subsequent to the sample's common ancestor. Write Y_c for the change in repeat number along this shared lineage, with $Y_c = 0$ if there is no shared lineage, and Y_1 and Y_2 for the respective changes in repeat number since the common ancestor of the two chromosomes. Then

$$M_1 = Y_c + Y_1 \text{ and } M_2 = Y_c + Y_2.$$

Some algebra then gives

$$\begin{aligned}\mathrm{E}(M_1^2) - \mathrm{E}(M_1M_2) &= \mathrm{E}(Y_1^2) - \mathrm{E}(Y_1Y_2) \\ &= \mathrm{Var}(Y_1) - \mathrm{Cov}(Y_1,Y_2) \\ &= 2\mathrm{Var}(Y_1) - \frac{1}{2}\mathrm{Var}(Y_1 + Y_2).\end{aligned} \tag{A9}$$

Now, in the coalescent, the number of mutations, W_1, on the lineage to the first chromosome since its common ancestor with the second has a geometric distribution with mean $\theta/2$, where $\theta \equiv 2N\mu$ is the usual scaled mutation parameter (recall that N is the number of chromosomes, rather than individuals, in the population). The total number of mutations, W_{12}, along either lineage since their common ancestor is also geometric, with mean θ.

With $Z_1, Z_2, \ldots$ denoting independent random variables with mean m, variance σ^2 and distribution F, we can write $Y_1 = \sum_{i=1}^{W_1} Z_i$. Thus

$$\begin{aligned}\mathrm{Var}(Y_1) &= \mathrm{E}[\mathrm{Var}(Y_1|W_1)] + \mathrm{Var}[\mathrm{E}(Y_1|W_1)] \\ &= \sigma^2\mathrm{E}(W_1) + m^2\mathrm{Var}(W_1) \\ &= \sigma^2\frac{\theta}{2} + m^2\frac{\theta(2+\theta)}{4}.\end{aligned} \tag{A10}$$

Similarly, $Y_1 + Y_2 = \sum_{i=1}^{W_{12}} Z_i$, so that

$$\begin{aligned}\mathrm{Var}(Y_1 + Y_2) &= \mathrm{E}[\mathrm{Var}(Y_1 + Y_2|W_{12})] \\ &\quad + \mathrm{Var}[\mathrm{E}(Y_1 + Y_2|W_{12})] \\ &= \sigma^2\mathrm{E}(W_{12}) + m^2\mathrm{Var}(W_{12}) \\ &= \sigma^2\theta + m^2\theta(1+\theta).\end{aligned} \tag{A11}$$

Finally, on substituting (A10) and (A11) into (A9), and then (A8) we have,

$$\begin{aligned}\mathrm{E}(S^2) &= (\sigma^2 + m^2)\frac{\theta}{2} \\ &= \eta_2\frac{\theta}{2}.\end{aligned}$$

This result has recently been derived independently, by related methods, in Kimmel and Chakraborty (1996). In the case of symmetric F, it was first derived by Roe (1992). For other special cases see Valdes *et al.* (1993), Zhivotovsky and Feldman (1995), and Pritchard and Feldman (1996).

Recall that

$$\mathrm{Var}(S^2) \approx \frac{1}{3}\eta_2^2\theta^2 + \frac{1}{12}\eta_4\theta,$$

a result due to Roe (1992) in the case of symmetric

F and recently extended to general *F* through a nice symmetrization argument by Kimmel and Chakraborty (1996). [For special cases see Zhivotovsky and Feldman (1995) where $m = 0$, and Pritchard and Feldman (1996) when mutation is symmetric and changes always involve only 1 repeat unit.] Thus, with *V* denoting NPV,

$$\mathrm{E}(V) = N\mu \text{ and } \mathrm{Var}(V) = \frac{4}{3}N^2\mu^2 + 0.16\frac{\eta_4}{\eta_2^2}N\mu. \tag{A12}$$

If the mutation rate varies across loci, then

$$\mathrm{E}(V) = N\bar{\mu} \tag{A13}$$

and

$$\begin{aligned}\mathrm{Var}(V) &= \mathrm{E}[\mathrm{Var}(V|\mu)] + \mathrm{Var}[\mathrm{E}(V|\mu)] \\ &= \frac{4}{3}N^2[\mathrm{Var}(\mu) + \bar{\mu}^2] + 0.16\frac{\eta_4}{\eta_2^2}N\bar{\mu} + N^2\mathrm{Var}(\mu) \\ &= \frac{7}{3}N^2\mathrm{Var}(\mu) + \frac{4}{3}N^2\bar{\mu}^2 + 0.16\frac{\eta_4}{\eta_2^2}N\bar{\mu}, \end{aligned} \tag{A14}$$

where $\bar{\mu}$ and $\mathrm{Var}(\mu)$ are the average and variance respectively of the the mutation rate. With data on NPV values at *L* different loci, then under this demographic scenario, their average $\bar{V}$ provides a natural estimator of $N\bar{\mu}$. The variance of this estimator is then

$$\mathrm{Var}(\bar{V}) = L^{-1}\left[\frac{7}{3}N^2\mathrm{Var}(\mu) + \frac{4}{3}N^2\bar{\mu}^2 + 0.16\frac{\eta_4}{\eta_2^2}N\bar{\mu}\right]. \tag{A15}$$

Note the decrease in this sampling variance with *L*, the number of loci. For loci with the same mutation rate μ, this sampling variance reduces to

$$\mathrm{Var}(\bar{V}) = L^{-1}\left(\frac{4}{3}N^2\mu^2 + 0.16\frac{\eta_4}{\eta_2^2}N\bar{\mu}\right). \tag{A16}$$

In addition to variation in mutation rate, the mutation mechanism may also vary across loci, as suggested for example by Figure 3 of the paper. In this case, A6, A7, A14, A15, and A16 still apply, with η_4/η_2^2 replaced by the average of this quantity across loci.

APPENDIX II

A formal significance test of the null hypothesis of constant population size and generalized stepwise mutation is complicated by several factors. One of these is the nuisance parameters, *N*, and the variation of mutation rates across loci. Another is the fact that only moments of some observables, rather than their full distribution, are known under the null hypothesis. The following informal analysis may nonetheless be helpful in assessing the strength of the evidence in the data against the null hypothesis.

For definiteness, consider the case of constant *N*, with $N = 10{,}000$ (recalling that *N* is twice the effective number of individuals in the population) and moderate variability in mutation rate, as described in Table 3. (It is straightforward, if desired, to mimic the analysis below for different parameter values, including the possibility of these differing between populations.) If the NPV values were normally distributed under the null hypothesis, and in addition, the null hypothesis exactly specified the predicted variance in NPV values, then the statistic

$$G^2 = \frac{(n-1)\text{Observed var}(NPV)}{\text{Predicted var}(NPV)}$$

has a chi-squared distribution with $n - 1$ degrees of freedom (d.f.), where *n* is the number of loci for which NPV values are available. Here, $n = 16$. Using the predicted variances from the appropriate column of Table 3, the calculated values of G^2 for the Luo, Sardinian, Kaingang, and Pooled populations are 1.04, 3.08, 0.53, and 1.03, respectively. All of these would be significant with $P < 0.001$ in a two-sided chi-squared test with 15 d.f. One approach for nonnormally distributed data (Box 1953) is to adjust the degrees of freedom appropriately. Box's method recalculates the degrees of freedom as

$$(n-1)/\left(1 + \frac{1}{2}\gamma_2\right)$$

where γ_2 is the standardized fourth cumulant (see, *e.g.*, McCullagh 1987) of the distribution from which the data are drawn.

Unfortunately we do not know the standardized fourth cumulant for the distribution of NPV values. A sensitivity analysis shows that our data would show a significant departure from the null hypothesis ($P = 0.05$, two-sided) unless γ_2 took the values 3, 1, 4, 3 for the Luo, Sardinian, Kaingang, and Pooled population, respectively. Except for the Sardinian, these would represent rather extreme departures from normality.

In fact, our null hypothesis does not specify the predicted variance of NPV values. We have used estimates of relevant parameters in the formula (A14). A natural statistic to consider is thus

$$F = \frac{\text{Predicted var}(NPV)}{\text{Observed Var}(NPV)}.$$

Under appropriate assumptions, this would have an *F* distribution, whose d.f. would reflect the precision of the estimates in numerator and denominator. An adjustment to the d.f. for the denominator can be made exactly as above for the chi-squared test. It is far from clear how to assess the appropriate degrees of freedom for the numerator; however, sensitivity analyses then show that the calculated statistic would be significant ($P = 0.05$, two-sided) for the Luo, Kaingang, and Pooled populations, for most pairs of d.f. values, and for many such pairs for the Sardinian population.

The above analyses are not intended as definitive. Our aim is to assist an assessment of whether the lower than predicted variation in NPV values is compatible with the null hypothesis. Our conclusion is that it is unlikely to be.