A Power Analysis of Microsatellite-Based Statistics for Inferring Past Population Growth

J. Patrick King,* Marek Kimmel,* and Ranajit Chakraborty†

*Department of Statistics, Rice University; and †Human Genetics Center, University of Texas Health Science Center, Houston, Texas

We present results concerning the power to detect past population growth using three microsatellite-based statistics available in the current literature: (1) that based on between-locus variability, (2) that based on the shape of allele size distribution, and (3) that based on the imbalance between variance and heterozygosity at a locus. The analysis is based on the single-step stepwise mutation model. The power of the statistics is evaluated for constant, as well as variable, mutation rates across loci. The latter case is important, since it is a standard procedure to pool data collected at a number of loci, and mutation rates at microsatellite loci are known to be different. Our analysis indicates that the statistic based on the imbalance between allele size variance and heterozygosity at a locus has the highest power for detection of population growth, particularly when mutation rates vary across loci.

Introduction

The classical Wright-Fisher model of genetic drift is based on several assumptions, including discrete non-overlapping generations, random mating, and constant effective population size (Ewens 1979, p. 16). These assumptions may deviate from reality. It is known, for example, that many species have not maintained a constant population size throughout their existence. An important case is that of our own species, which underwent rapid numerical expansions in both the early and the later stages of its existence. Therefore, it seems important to develop methods to make inferences concerning past population demography. Indeed, such methods were developed and various inferences were made based on them (e.g., Rogers and Harpending 1992; Griffiths and Tavaré 1994; Polanski, Kimmel, and Chakraborty 1998; Relethford 1998).

More recently, several statistics were developed for detection of past population growth based on short tandem repeat DNA sequences, also referred to as microsatellite loci (Di Rienzo et al. 1998; Kimmel et al. 1998; Reich and Goldstein 1998; Reich, Feldman, and Goldstein 1999). Some of their properties were investigated in the original communications. However, a systematic comparison has yet to be carried out. Microsatellites are an important source of information about the phylogenetic relationships among human populations (Bowcock et al. 1994; Deka et al. 1995; Mountain and Cavalli-Sforza 1997). Also, they were found to be of crucial importance in gene mapping (Dib et al. 1996) and estimation of individuals’ relatedness and identity (Chakraborty and Jin 1993). The estimates vary, but there exist at least several tens of thousands of these loci throughout the human genome. Most of them are considered neutral, but some are associated with coding sequences, and some are known to cause disease in individuals carrying highly expanded alleles (reviewed by Rubinsztein 1999).

The purpose of this paper is to present results concerning the power to detect past population growth using three microsatellite-based statistics available in the current literature: (1) the interlocus g statistic (Reich and Goldstein 1998), (2) the within-locus k statistic (Reich and Goldstein 1998), and (3) the imbalance index β (Kimmel et al. 1998). The analysis was based on the single-step stepwise mutation model (SSMM). The power of the statistics was evaluated under both constant and variable mutation rates. The latter case is important since it is a standard procedure for pooling data collected at a number of microsatellite loci, and mutation rates at these loci are known to be variable (Chakraborty et al. 1997; American Association of Blood Banks 1998; Brinkmann et al. 1998; Deka et al. 1999a). These power studies are preceded by a brief summary of relevant theoretical concepts and definitions of the indices. The effects of variations in the mutation process, particularly the presence of multistep mutations, on the power of the statistics are also presented.

Microsatellite Evolution Under Demographic Change

Microsatellite loci are tandem repeat loci with repeat motifs of 2–6 nucleotides in length. The most frequent mutations in microsatellites are considered to be single-step stepwise mutations, i.e., changes of the allele size by 1 or −1 repeat unit. Estimates vary, but the mutation rate seems to be on the order of \( v \approx 10^{-4} \) per generation (Weber and Wong 1993), with the value depending on the length of the repeat motif (Chakraborty et al. 1997) and on the genomic location of the repeat locus (Deka et al. 1999a). Due to this relatively high mutation rate, neutral microsatellite loci are informative about the short-term evolutionary history of sampled chromosomes.

The evolution of a microsatellite locus is shaped by demography, as well as by stepwise mutation. This is because genetic drift (the loss of alleles through random sampling of genotypes of new individuals from the gamete pool) acts with strength inversely proportional to the
effective population size. This effect determines the distribution of the branch lengths in the genealogy of the sampled locus. For a sample of \( n \) individuals, the genealogy may be partitioned according to the times \( T_k \), \( k = 2, 3, \ldots, n \), where \( T_k \) denotes the time for which the sample represents \( k \) distinct lineages (see fig. 1). In the case of constant population size, it is known that the coalescence times \( T_k \) are independent, exponentially distributed random variables, each with parameter \((1/2N)k\) in the population.

Under constant population size, the most ancient coalescence times tend to be long relative to branches of the tree associated with more recent bifurcations. For this reason, coalescent trees in constant populations frequently exhibit two or three clusters at the tips of the tree connected with the root by a few long branches. Mutations accumulate on these long branches, accounting for much of the allelic variation observed in the sample. Figure 2a shows such a genealogy for a sample of \( n = 10 \) alleles. Under the SSMM, the resulting allele length distributions are characterized by clusters of similar alleles separated by troughs of low frequency. Figure 3a shows three simulated allele size distributions based on samples of size \( n = 40 \) from the same constant population scenario used to generate figure 2a.

When the population size varies over time, the above description of the coalescence process must be modified. The times to coalescence are no longer exponentially distributed. Intuitively, the constant coalescence intensity \( 1/(2N) \) is replaced by the time-dependent coalescence intensity \( 1/[2N(t)] \), where \( N(t) \) gives the population size \( t \) generations in the past. Given that there are \( k \) lineages represented in the sample at time \( t \), the distribution of the time to coalescence is given by

\[
P[T_k \geq \tau] = \exp \left\{ -\frac{1}{2} \int_{\tau}^{\infty} \frac{ds}{2N(s)} \right\}, \quad k = 2, \ldots, n.
\]

As a result, the distributions of the coalescence times \( T_k \) in a population of variable size will be distorted relative to their counterparts in a constant population.

The growth pattern of interest here is rapid expansion. Consider, for example, a population originally of size \( N_0 \) which undergoes a stepwise expansion \( t_{e} \) generations in the past to its current size \( N \), where \( N \gg N_0 \). Looking backward in time, this demographic expansion corresponds to a sudden increase in the coalescence intensity from \( 1/(2N) \) to \( 1/(2N_0) \). The effect of this change

Fig. 1.—Genealogy for a sample of \( n = 5 \) chromosomes from a population of constant size.

Fig. 2.—Simulated genealogies for samples of size \( n = 10 \) from (a) a population of constant size \( N = 20,000 \) and (b) a population which experienced a stepwise expansion \( t_e = 8,000 \) generations in the past (indicated by dashed line) from size \( N_0 = 200 \) to size \( N = 20,000 \). In both panels, mutation rate \( v = 5 \times 10^{-4} \), and x’s indicate mutation events.
on the genealogy of a sample of \( n \) chromosomes depends on the time since the expansion. If the expansion event is very ancient, even the coalescence times closest to the root will reflect the current population size \( 2N \).

However, if the growth is sufficiently recent, the lineages in the genealogy at the time of expansion will be subject to the pre-expansion coalescence intensity \( 1/(2N_0) \), and expected coalescence times for these lineages will be much shorter. With high probability, then, the most recent common ancestor of the sample is found close to time \( t_e \).

Figure 2b depicts a genealogy sampled from a stepwise expansion scenario. The mutation rate and final population size are the same as those used to simulate the tree in figure 2a, so that coalescence times and the number of mutations in the two scenarios are identically distributed over the time period \( t = 0 \) to \( t = t_e \). In figure 2b, the lineages which remain at time \( t_e \) coalesce quickly, and mutation has little time to act on the deepest branches of the genealogy. Relative to the constant-population scenario, observed allelic variation depends more on mutation events which occur near the tips of the tree, and observed allele sizes correspond more closely to independent, identically distributed random variables. These features are apparent in figure 3b, which shows three simulated allele size distributions based on samples of size \( n = 40 \) from the expansion scenario just discussed. In comparison with figure 3a, these distributions are less variable and more clearly unimodal.

The three tests for demographic expansion discussed in this paper represent three attempts to detect these qualitative and quantitative differences between samples of microsatellite alleles from constant and recently expanded populations.

### Indices of Demographic Expansion

**Mutation-Drift Equilibrium and the Imbalance Index**

Consider a sample of \( n \) haploid individuals (or chromosomes) from a population of size \( 2N \) typed at a single microsatellite locus evolving under the SSMM. Two parameters indicative of past demography are genetic variance \( V(t) = E[(X_i(t) - X_j(t))^2] \) and homozygosity \( P_0(t) = P[X_i(t) = X_j(t) = 0] \). Here, \( X_i(t) \) and \( X_j(t) \) are the sizes of two alleles sampled at time \( t \) from the population. Unbiased estimators for these two parameters are

\[
\hat{V} = \frac{1}{n(n-1)} \sum_{i \neq j} (X_i - X_j)^2 = \frac{2}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]

(2)

\[
\hat{P}_0 = \left( n \sum_{k \in K} p_k \bar{x} - 1 \right) / (n-1),
\]

(3)

where \( K \) is the set of allele sizes represented in the sample and \( p_k \) is the relative frequency of alleles of size \( k \) (Kimmel and Chakraborty 1996; Kimmel et al. 1998).

Under a generalized stepwise mutation model (i.e., a model in which \( U \), the allele size change in a single mutation event, is an arbitrary integer-valued random variable), mutation-drift equilibrium is mathematically predicted (Kimmel and Chakraborty 1996). Under the SSMM (\( U \) restricted to \( \geq 1 \)), both \( V(t) \) and \( P_0(t) \) converge to limits which depend only on the composite parameter \( \theta \):
Fig. 4.—Power to detect (a) 2-fold, (b) 10-fold, and (c) 100-fold stepwise growth as a function of time since expansion. Curves 1 and 2 are for tests based on and , respectively; g refers to the interlocus g test; and k indicates the within-locus k test. Pre-expansion population size \( N_0 \) = 2,500, mutation rate \( v \) = 5 \times 10^{-4}.

The imbalance index (Kimmel et al. 1998) is based on comparison of two estimators of \( \theta \) derived from these limiting values:

\[
\hat{\theta}_V = \bar{V} \quad \text{and} \quad \hat{\theta}_{P_o} = (1/P_o^2 - 1)/2,
\]

called the (allele size) variance estimator of \( \theta \) and the homozygosity estimator of \( \theta \), respectively. At mutation-drift equilibrium,

\[
\frac{E(\hat{\theta}_V)}{E(\hat{\theta}_{P_o})} \approx \frac{V(\infty)}{(1/P_o^2 - 1)/2} = 1,
\]

which leads to a parametric definition of an index \( \beta(t) \), given by

\[
\beta(t) = \frac{V(t)}{(1/P_o(t)/2 - 1)/2}.
\]

Deviation of \( \beta(t) \) from 1 at a microsatellite locus is a signature of disequilibrium or departure from the SSMM. In particular, populations which have recently expanded exhibit values of \( \beta(t) \) less than 1. The stepwise expansion scenario discussed above illustrates this effect. After expansion, both \( V(t) \) and the heterozygosity \( 1 - P_{o}(t) \) increase as the population approaches a new mutation-drift equilibrium. If the expansion is sufficiently recent, typical sample genealogies resemble the one in figure 2b, with mutation events largely restricted to the postexpansion epoch and distributed among relatively many lineages. Such mutations are effectively placed to generate heterozygosity, but not allele size variance, since the latter quantity is more sensitive to mutations which distinguish the oldest lineages. As a result, heterozygosity approaches its new limit value faster than genetic variance, causing the index \( \beta(t) \) to deviate downward from 1.

As is the case with any ratio-type index, \( \beta(t) \) can be estimated in several ways. Here, we consider two estimators of \( \ln \beta(t) \) which yield normal-like distributions. Let \( L \) be the number of microsatellite loci in the sample. The first estimator is the log ratio of means, used by Kimmel et al. (1998) and given by

\[
\ln \beta_1 = \ln \hat{\theta}_V - \ln \hat{\theta}_{P_o}
\]

\[
= \ln \bar{V} - \ln [(\bar{P}_o^2 - 1)/2],
\]

where \( \bar{V} \) and \( \bar{P}_o \) are averages of the \( L \) single-locus estimates given by equations (5). The second estimator is based on the mean of log ratios, given by

\[
\ln \beta_2 = \frac{1}{L} \sum_{i=1}^{L} (\ln \hat{\theta}_V)_i - (\ln \hat{\theta}_{P_o})_i,
\]

where \( i \) indexes loci. We note that this statistic depends on the joint distribution of \( \hat{\theta}_V \) and \( \hat{\theta}_{P_o} \), while \( \ln \beta_1 \) depends solely on their marginal distributions.

Between-Locus Variability and the \( g \) Statistic

Introduced by Reich and Goldstein (1998) (also see Di Rienzo et al. 1998), the \( g \) statistic is defined by the following ratio

\[
g = \frac{\text{Observed Var}(\hat{\theta}_V)}{\text{Predicted Var}(\hat{\theta}_V)}.
\]

The predicted variance of \( \hat{\theta}_V = \bar{V} \) is based on the mutation-drift equilibrium expression derived by Roe (1992), Zhivotovsky and Feldman (1995), and, in a more general setting, by Kimmel and Chakraborty (1996):
The test statistic is

\[ n = \frac{2}{3} \theta^2 + \frac{1}{3} \theta. \]  

(10)

For a sample of \( L \) loci, this quantity can be estimated by \( \frac{1}{\hat{\theta}^2} \sum_i (\hat{\theta} - \hat{V})_i \), where \( \hat{\theta} \) is the average of the \( L \) single-locus estimates \( \hat{\theta}_i = \hat{V}_i \). The observed variance of \( \hat{\theta} \) is simply the sample variance of these same single-locus estimates. The statistical properties of \( \hat{g} \) were studied extensively by Reich, Feldman, and Goldstein (1999).

The behavior of the \( g \) statistic can be illustrated by considering the genealogical effect of expansion. Mutations in the deepest branches of a genealogy strongly influence the allele size variance estimated by \( \hat{\theta} \). In a population of constant size, the high interlocus variability of these branch lengths inflates the variance of \( \hat{\theta} \) across loci. In expanded populations, however, mutation has little opportunity to act on the oldest lineages. As a result, \( \text{Var}(\hat{\theta}) \) is decreased (compare fig. 3a and b), and we expect \( g < 1 \).

Shape of the Allele Size Distribution and the Within-Locus \( k \) Test

Developed by Reich and Goldstein (1998), the within-locus \( k \) test is based on the observation that allele size distributions in expanding populations tend to be unimodal, more peaked than the ragged distributions associated with constant populations (compare fig. 3a and b). The test statistic is

\[ k = 2.5 \hat{\sigma}^4 + 0.14 \hat{V} - 0.95/n - \hat{\gamma}^4. \]  

(11)

where \( n \) is the number of chromosomes in the sample, \( \hat{\sigma}^4 \) is an unbiased estimator of \( \text{Var}(\hat{V})/4 \), and \( \hat{\gamma}^4 \) is an unbiased estimator of the fourth central moment of the allele size distribution. Derivations of these estimators are provided by Reich, Feldman, and Goldstein (1999). The coefficients were determined via simulation so that samples of size \( n \approx 10 \) from a locus evolving in a population of constant size (\( \theta > 1 \)) yield \( k > 0 \) with probability \( p \in [0.515, 0.550] \). When \( L \) loci are sampled, a conservative test for expansion compares the number of loci for which \( k \) exceeds 0 to cutoffs from the binomial distribution with parameters \( L \) and \( p = 0.515 \).

Simulation Studies

The power of a statistical test is the probability with which it detects a specified alternative to the null hypothesis. Here, the null hypothesis is long-term constancy of the population size, which leads to mutation-drift equilibrium. The alternatives are various scenarios of past population expansion, including different times and scales for the expansion event, as well as different values of \( \theta = 4Nv \), treated as a nuisance parameter. We also consider the power to detect expansion when the mutation rate is variable. Our focus on power as a criterion of comparison is justified, since all four tests are simulation-based. Type I error can be controlled by including suspected nuisance variation (such as mutation rate variability) in the simulations.

In the simulation studies which follow, a single replicate consists of a sample of \( n = 40 \) individuals, each typed at \( L = 30 \) unlinked microsatellite loci. A replicate corresponds to a data set of modest size, and all four tests would perform at least as well with more data. Reich, Feldman, and Goldstein (1999) have studied the effect of increasing both \( n \) and \( L \) on the \( g \) and \( k \) statistics. For each replicate, we used the coalescent to generate 30 independent genealogies and applied single-step mutations according to a Poisson process. (The rate of the mutation process was either fixed at \( v = 5 \times 10^{-4} \) or sampled from a lognormal distribution with mean \( 5 \times 10^{-4} \); details are given below.) From the resulting 30 allele size distributions we computed the test statistics, \( \log \hat{\beta}_1, \log \hat{\beta}_2, g, \text{ and } k \).

We simulated 1,000 replicates for each null hypothesis. Critical values for the tests based on \( \hat{\beta} \) and \( \hat{g} \) were determined by the 0.05 quantile of the resulting empirical distributions of each test statistic. The cutoff for the kurtosis test was based on the binomial distribution with \( n = 30 \) and \( p = 0.515 \). We chose to reject the null hypothesis if 11 or fewer loci yielded positive...
$k$ values, resulting in a significance level of approximately 0.074, slightly higher than the type 1 error rate of the other three tests.

For each alternative hypothesis, we simulated 1,000 replicates for 1,000 realizations of each test statistic. The power of each test was estimated by the proportion of these 1,000 realizations that were less than or equal to the critical value.

Power to Detect Stepwise Expansion

Let $N(t)$ denote the size of the population $t$ generations in the past (so that $N(0)$ gives the population size at the time of sampling), and consider the null hypothesis $H_0: N(t) = N$ for all $t \geq 0$. Using the procedure just
described, we estimated the power of the four statistical tests to detect alternatives of the form

\[ N(t) = \begin{cases} 
2,500 & \text{if } t \geq t_e \\
N & \text{if } t < t_e,
\end{cases} \]

for a range of expansion times \( t_e \). Mutation-drift equilibrium at time \( t_e \) and mutation rate \( \nu = 5 \times 10^{-4} \) are assumed. The power estimates for three stepwise expansion scenarios, corresponding to 2-fold, 10-fold, and 100-fold growth, are shown in figure 4.

Power to Detect Exponential Expansion

As an example of more gradual expansion, we considered a population which grew exponentially from mutation-drift equilibrium. With the same null hypothesis in mind, we tested alternatives of the form

\[ N(t) = \begin{cases} 
2,500 & \text{if } t \geq t_e \\
N_{\text{est}} & \text{if } t < t_e,
\end{cases} \]

for a range of expansion times \( t_e \) and for the same final population sizes as in the previous section, i.e., \( N = 5,000, 25,000, \) and \( 250,000 \). Since we require \( N(t_e) = 2,500 \), each expansion time \( t_e \) specifies a different exponential rate parameter, namely, \( \alpha = (1/t_e) \ln(2,500/N) \). The power estimates are plotted in figure 5.

Expansion with Variable Mutation Rate

We also estimated the power of the four tests to detect stepwise expansion in the presence of mutation rate variability and examined the effect of this variability on the significance level of the test procedure. Specifically, we assumed a lognormal distribution for \( \nu \) (with mean \( \mu = 5 \times 10^{-4} \) and standard deviation \( \sigma \)) and tested the null hypothesis

\[ H_0: \quad \nu \sim \text{lognormal}(\mu, \sigma), \quad N(t) = 25,000 \quad \text{for all } t \geq 0 \]

against alternatives of the form

\[ \nu \sim \text{lognormal}(\mu, \sigma), \quad N(t) = \begin{cases} 
2,500 & \text{if } t \geq t_e \\
25,000 & \text{if } t < t_e,
\end{cases} \]

for a range of expansion times \( t_e \). Mutation-drift equilibrium at the time of expansion is assumed.

Power estimates for these scenarios are shown in figure 6. In panel a, \( \nu \) is sampled from a lognormal distribution with mean \( \mu = 5 \times 10^{-4} \) and standard deviation \( \sigma = 1 \times 10^{-4} \). The interquartile range for this distribution is \((4.3 \times 10^{-4}, 5.6 \times 10^{-4})\). Panel b corresponds to \( \mu = 5 \times 10^{-4}, \sigma = 3 \times 10^{-4} \), and interquartile range \((2.9 \times 10^{-4}, 6.2 \times 10^{-4})\), while panel c represents \( \mu = 5 \times 10^{-4}, \sigma = 6 \times 10^{-4} \), and interquartile range \((1.7 \times 10^{-4}, 6.1 \times 10^{-4})\).

Table 1 shows the effect of mutation rate variability on the 0.05 quantiles of \( \ln \hat{\beta}_1 \), \( \ln \hat{\beta}_2 \), and \( \ln g \) under the null hypothesis of constant population size over a range of values for \( \theta \) and \( \sigma/\mu \). Since the 0.05 quantiles of all three statistics increase with mutation rate variability, we conclude that tests based on these statistics are conservative (i.e., have significance levels lower than the nominal 0.05 level). Reich, Feldman, and Goldstein (1999) have shown that tests based on \( k \) are insensitive to mutation rate variability as long as \( \theta > 1 \).
Table 2
Estimated 0.05 Quantiles as a Function of Multistep Mutation Probability

<table>
<thead>
<tr>
<th>Probability of 10-Step Mutation</th>
<th>0.00</th>
<th>0.00125</th>
<th>0.0025</th>
<th>0.005</th>
<th>0.01</th>
<th>0.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta = 1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ln \hat{\beta}_1$</td>
<td>$-0.32$</td>
<td>$-0.34$</td>
<td>$-0.31$</td>
<td>$-0.27$</td>
<td>$-0.22$</td>
<td>$-0.06$</td>
</tr>
<tr>
<td>$\ln \hat{\beta}_2$</td>
<td>$-0.65$</td>
<td>$-0.65$</td>
<td>$-0.64$</td>
<td>$-0.62$</td>
<td>$-0.58$</td>
<td>$-0.53$</td>
</tr>
<tr>
<td>$\ln g$</td>
<td>$-1.08$</td>
<td>$-1.07$</td>
<td>$-1.04$</td>
<td>$-0.95$</td>
<td>$-0.85$</td>
<td>$-0.33$</td>
</tr>
<tr>
<td>$\theta = 5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ln \hat{\beta}_1$</td>
<td>$-0.17$</td>
<td>$-0.15$</td>
<td>$-0.11$</td>
<td>$-0.10$</td>
<td>$0.00$</td>
<td>$0.33$</td>
</tr>
<tr>
<td>$\ln \hat{\beta}_2$</td>
<td>$-0.57$</td>
<td>$-0.58$</td>
<td>$-0.55$</td>
<td>$-0.54$</td>
<td>$-0.52$</td>
<td>$-0.42$</td>
</tr>
<tr>
<td>$\ln g$</td>
<td>$-0.73$</td>
<td>$-0.73$</td>
<td>$-0.62$</td>
<td>$-0.52$</td>
<td>$-0.28$</td>
<td>$0.34$</td>
</tr>
<tr>
<td>$\theta = 10$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\ln \hat{\beta}_1$</td>
<td>$-0.02$</td>
<td>$0.04$</td>
<td>$0.02$</td>
<td>$0.11$</td>
<td>$0.25$</td>
<td>$0.68$</td>
</tr>
<tr>
<td>$\ln \hat{\beta}_2$</td>
<td>$-0.53$</td>
<td>$-0.51$</td>
<td>$-0.52$</td>
<td>$-0.48$</td>
<td>$-0.44$</td>
<td>$-0.33$</td>
</tr>
<tr>
<td>$\ln g$</td>
<td>$-0.40$</td>
<td>$-0.31$</td>
<td>$-0.27$</td>
<td>$-0.09$</td>
<td>$0.11$</td>
<td>$0.62$</td>
</tr>
</tbody>
</table>

Note.—Quantiles are based on 1,000 simulations, each consisting of a sample of $n = 40$ individuals sampled from a population of constant size and typed at $L = 30$ unlinked loci. Mutation rate $u$ is fixed at $5 \times 10^{-5}$. Conditional on a mutation event occurring, the allele size change $U$ is $10$ with probability $p$, and $1$ with probability $1 - p$.

Impact of Multistep Mutations

We have thus far assumed that all mutations involve only single-step changes in allele size. Since some microsatellites exhibit multistep mutations, we examined the effect of such events on the null distributions of the four test statistics. Allele size distributions were simulated under the assumption that mutations are multistep with probability $p$ and single-step with probability $1 - p$.

Figure 7 plots the estimated quantiles of $\ln \hat{\beta}_1$, $\ln \hat{\beta}_2$, and $\ln g$ as a function of the probability $p$ of a two-step mutation. As $p$ ranges from 0 to 0.2, the quantiles of $\ln g$ are roughly constant, while the quantiles of the statistics based on $\hat{\beta}$ increase with $p$. This effect is more pronounced for larger mutation step sizes, as shown for 10-step mutations in Figure 8. In this case, even the $g$ statistic is affected, and again the effect is conservative for all three statistics. Table 2 presents estimated 0.05 quantiles for $\ln \hat{\beta}_1$, $\ln \hat{\beta}_2$, and $\ln g$ as a function of $\theta$ and the probability $p$ of a 10-step mutation. Reich, Feldman, and Goldstein (1999) have shown that the $k$ test is nonconservative in the presence of multistep mutations.

The conservative effect of multistep mutations on the $\beta$ and $g$ tests is accompanied by a reduction in power. Nevertheless, unless a substantial fraction of mutations are multistep (say >5% for 2-step mutations or >0.5% for 10-step mutations), this reduction is very slight.

Discussion

The sensitivity of all four tests depends on the configuration of the genealogy associated with each locus in the sample. Sensitivity is greatest when the most ancient coalescence times (say, $T_2$ and $T_3$) reflect the initial population size while the more recent ones reflect the expanded population size, as shown, for example, in figure 2b. On the other hand, if the expansion is either so recent that the sampled genealogies reflect only the pre-expansion population size or so ancient that they reflect only the post-expansion demography, any test will be noninformative. The interval of sensitivity defined by these extremes depends on the test, as does the power to detect expansion.

When the mutation rate $u$ is fixed, statistics based on the imbalance index $\beta$ are more sensitive to population increase than the $g$ and $k$ statistics, and they maintain their sensitivity over a longer time interval. The power estimates for $\ln \hat{\beta}_1$ and $\ln \hat{\beta}_2$ nearly coincide. For 100-fold stepwise growth (fig. 4c), all three tests achieve high power over a range of times which are relevant to the evolution of early human populations. In all three stepwise scenarios, the $\beta$ statistics are more responsive than $g$ and $k$ to expansions of recent origin. As the expansion time becomes more remote, tests based on $\beta$ maintain power comparable or superior to that of the $g$ and $k$ tests.

The exponential expansion scenario (fig. 5) yields results similar to those of the stepwise case, with the following exception. The expansion signal appears later and persists longer in all four tests. Both facts are related to the gradual nature of exponential versus stepwise growth. In the stepwise scenario, coalescence times begin to reflect the final population size $N$ immediately after the expansion event. If the growth is exponential, however, the coalescence process is governed by values of $N(t) \ll N$ for some time after the expansion, and the signal arises more slowly. This fact also accounts for the persistence of the signal under exponential expansion. Stepwise growth goes undetected if $\sum_{k=2}^{\infty} T_k^e < t_e$. Under the analogous exponential expansion (with the same values for $N_0$, $N$, and $t_e$), the most remote coalescence times will reflect values of $N(t)$ only slightly greater than $N_0$, corresponding to the early stages of growth. Sampled genealogies will be qualitatively similar to the one in figure 2b, and the expansion will be detectable with positive probability.
Microsatellite-Based Tests for Population Growth

It should be stressed that in figure 5, increasing values of \( t_0 \) correspond to smaller exponential growth rates. This fact explains the eventual loss of power in all tests. If we fix the growth rate \( \alpha \) (so that the final population size increases with \( t_0 \)), we would expect the expansion signal to persist for all time. It can be shown, for example, that under exponential growth, the imbalance index \( \beta(t) \rightarrow 1/# \) as \( t \rightarrow \infty \) (unpublished data).

When the mutation rate \( \nu \) varies across loci, the power to detect expansion also depends on how each statistic incorporates information from multiple loci. If the variation is relatively small, the results are similar to those for the case of constant mutation rate (compare figs. 6a and 4b). As \( \text{Var}(\nu) \) increases, tests based on \( \ln \beta_1 \) and \( g \) lose sensitivity, while the power of tests based on \( \ln \beta_2 \) and \( k \) is maintained. In the terminology of Reich and Goldstein (1998), the latter two quantities are “within-locus” statistics: when \( L \) loci are sampled, they combine \( L \) single-locus estimates, each of which is informative provided \( \nu \) is not too small. The tests which suffer, however, depend on the “interlocus” statistics \( \ln \beta_1 \) and \( g \), both of which combine multilocus information in the numerator and separately in the denominator.

It should be noted that the power analyses performed in the present work were under the assumptions of the SSMM. Kimmel and Chakraborty (1996) showed that when individual mutations may result in multistep repeat size changes, at mutation-drift equilibrium (see eq. 4), the within-population allele size variance has the form \( V(\infty) = 4Nn\psi'(1) \), where \( \psi'(1) \) is the second moment of the distribution of repeat size changes produced by individual mutations. An analogous closed-form expression for the homozygosity, \( P_0(\infty) \), does not exist under a generalized stepwise mutation model (Kimmel and Chakraborty 1996). Therefore, although any contraction-expansion bias of mutations does not affect \( V(\infty) \) or \( P_0(\infty) \), the sampling properties of all of the statistics considered here are only approximate if mutation events do not obey the SSMM. Empirical data on mutations, although sparse at present, show a reasonable degree of agreement with the SSMM assumption. For example, nearly 84% of the mutation events encountered in parental testing laboratories (which use tri- and tetranucleotide repeat loci) involve only single-repeat size changes (American Association of Blood Banks 1998). Likewise, Brinkmann et al. (1998) reported that 22 of 23 mutation events observed during parentage testing experiences in Germany are single-step changes, with the remaining event being a double-step change. Even for loci with considerably higher mutation rates, the SSMM approximation may be reasonable. For example, Deka et al. (1999b) showed that for the gene-associated CAG repeat locus ERDA1, where the mutation rate is over 6%, only 11 of 46 observed mutation events were multistep. Our analysis of the impact of multistep allele size changes indicates that a substantial fraction of such mutations is required to significantly reduce the power of the \( \beta \) and \( g \) tests.

In summary, the \( \ln \beta_2 \) test is the most sensitive to population expansion over all the scenarios considered here. This sensitivity may not always be a virtue. When only ancient, massive expansions are of interest, the \( g \) or \( k \) statistic may be preferred.

LITERATURE CITED


WOLFGANG STEPHAN, reviewing editor

Accepted August 7, 2000