Conservation Genetics (2005) 6:551–562
DOI 10.1007/s10592-005-9009-5

Comparison of methods for detecting bottlenecks from microsatellite loci

Ellen G. Williamson-Natesan[1,2,3]
[1]*Department of Integrative Biology, University of California, Berkeley, CA, 94720, USA;* [2]*American Association for the Advancement of Science assigned to the National Center for Environmental Assesment, Office of Research and Development, US Environmental Protection Agency, Washington, DC, 20460, USA (e-mail: natesan.ellen@epa.gov)*

Received 13 July 2004; accepted 29 January 2005

Key words: allele frequency distribution, population size reduction, effective population size, extinction risk

Abstract

This paper describes simulation tests to compare methods for detecting recent bottlenecks using microsatellite data. This study considers both type I error (detecting a bottleneck when there wasn't one) and type II error (failing to detect a bottleneck when there was one) under a variety of scenarios. The two most promising methods were the range in allele size conditioned on the number of alleles, M_k, and heterozygosity given the number of alleles, H_k, under a two-phase mutation model; in most of the simulations one of these two methods had the lowest type I and type II error relative to other methods. M_k was the method most likely to correctly identify a bottleneck when a bottleneck lasted several generations, the population had made a demographic recovery, and mutation rates were high or pre-bottleneck population sizes were large. On the other hand H_k was most likely to correctly identify a bottleneck when a bottleneck was more recent and less severe and when mutation rates were low or pre-bottleneck population sizes were small. Both methods were prone to type I errors when assumptions of the model were violated, but it may be easier to design a conservative heterozygosity test than a conservative ratio test.

Introduction

Many natural populations have recently experienced effective size reductions (population bottlenecks) as a result of over-exploitation, habitat destruction or modification, and population fragmentation. While all small populations may be at risk because fluctuations in population size can lead to extinction, recently reduced populations may be at increased risk due to genetic factors. A sudden reduction in population size can lead to loss of genetic variation and increased inbreeding, and in typically outbreeding species these factors may contribute to population extinction (Lande 1994; Frankham 1995; Newman and Pilson 1997; Frankham 1999). It is therefore important to identify populations that have recently experienced a bottleneck in order to effectively prioritize the allocation of conservation resources and develop management strategies aimed at mitigating the potentially negative consequences of population bottlenecks. Unfortunately, it is often difficult to directly determine if a population has recently experienced a bottleneck because the historical population size is generally unknown. When temporally spaced population samples are available, genetic effective population size between sampling events can be inferred from genetic drift (Waples 1989; Williamson and Slatkin 1999; Anderson et al. 2000). For populations that have not experienced drastic changes in population size and have remained in approximate mutation-drift equilibrium, a coalescent-based approach for detecting historical changes in population size from a single population sample has been developed by Beaumont (1999). Methods for determining population size in recently bottlenecked non-equilibrium populations have also been developed (Schwartz et al. 1998). However, these methods, which are based on gametic disequilibrium and heterozygote excess, require large sample sizes and are typically associated with large confidence intervals (Schwartz et al. 1998). Since estimating the size of a bottleneck from single samples has proven difficult, a number of methods have also been developed to simply detect recent population bottlenecks in non-equilibrium populations using molecular marker data obtained from a single population sample (Cornuet and Luikart 1996; Luikart et al. 1998; Garza and Williamson 2001). In this paper I compare methods from this last category by applying them to simulated data under a variety of mutation and demographic scenarios.

The methods examined in this paper for identifying recent population bottlenecks using microsatellite data make several assumptions about the mutation process of microsatellite loci (Cornuet and Luikart 1996; Luikart et al. 1998; Garza and Williamson 2001). This paper examines type I and type II errors that might occur as a result of not knowing the mutation rate, population size, or details of the mutation process.

For most statistical tests, a certain degree of error must be tolerated. For example, selecting a critical value at the 95th percentile of a distribution implies that type I error of 5% is permissible. There are often tradeoffs between type I and type II errors and the amount and type of error that is acceptable depends on the question being asked. For example, the goal may be to identify bottlenecked populations with a high degree of certainty. In this situation the researcher would probably prefer a conservative test (low type I error) with low power (high type II error) over a test that has high type I error but low type II error. On the other hand, if the goal is to detect all populations that might be bottlenecked, type I error might be considered an acceptable tradeoff if the result is a test with more power to detect bottlenecks. The objective of this paper is to provide a preliminary comparison of bottleneck detection methods in terms of type I and type II errors through a simulation study. The hope is to help inform empirical researchers that apply these methods to real data and must choose between these types of tradeoffs.

Methods

Simulations

My simulation approach is similar to that used in Garza and Williamson (2001). I used the two-phase model (TPM) of microsatellite evolution introduced in (Di Rienzo et al. 1994). In this model, a fixed proportion of mutations are single steps (where alleles either increase or decrease by one repeat unit) and the mutation sizes for remaining mutations are drawn from a geometric distribution. In this model the majority of mutations are single steps or small, but larger mutations do occur. This model appears to be provide a reasonable fit to empirical evidence about the mutation process (Di Rienzo et al. 1994; Garza and Williamson 2001).

The TPM has three parameters: $\theta = 4N_e\mu$, where N_e is the effective population size, and μ is the mutation rate; p_g, the percent of mutations that are larger than single steps; and δ_g, the mean size of the mutations larger than single steps. The distribution of larger mutations is given by a two-sided geometric distribution (Di Rienzo et al. 1994; Garza and Williamson 2001). The probability of a mutation of size x is

$$P(x) = \frac{1}{2(\delta_g - 1)}\left(1 - \frac{1}{\delta_g - 1}\right)^{|x|-2}, |x| \geq 2. \tag{1}$$

The mean of x is zero, but the mean of $|\ x|$ is δ_g. The variance of this distribution is

$$\sigma_g^2 = 2\delta_g^2 - 3\delta_g + 2 \tag{2}$$

and thus the variance in mutation size in the TPM is

$$\sigma_m^2 = (1 - p_g) + p_g\sigma_g^2 \tag{3}$$

where $(1-p_g)$ is the proportion of mutations that are single steps.

From direct observations of the mutation process in the literature Garza and Williamson

(2001) estimated that $\delta_g = 2.8$ and $p_g = 0.12$. Because these parameter values are derived from a small amount of data, I use them only as a starting point in this paper. Empirically observed microsatellite distributions from numerous species are also consistent with these values, but also with other combinations of δ_g and p_g (Garza and Williamson 2001). In this paper I use a variety of parameter values that broadly encompass the realistic range of possible parameter values that would be consistent with this empirical data, namely, δ_g between 1.0 and 3.5 and p_g between 0 and 0.2. Note that setting either $\delta_g = 1.0$ or $p_g = 0$ reduces the TPM to the SSM as a special case.

For these simulations I assume a Wright–Fisher model of evolution with mutation. During a bottleneck, I used an exact simulation of the Wright–Fisher process. Each generation, alleles were selected with multinomial probability from the allele frequencies of the previous generation (Ewens 1979).

Before and after the bottleneck, a full Wright–Fisher process is an inefficient simulation approach because this method becomes exponentially slower with larger population sizes. Coalescent models offer a way to quickly approximate the Wright–Fisher process. Coalescent models are an accurate approximation to the Wright–Fisher process when population size is large relative to sample size (Tavare 1984).

Coalescent models track the ancestors of a sample into the past. The waiting time between events (in units of $1/2N_e$) between events is distributed exponentially with mean $1/(2j(j+\theta-1))$ where j is the number of lineages and $\theta = 4N_e\mu$. Events can be either two lineages coalescing (sharing a common ancestor) with probability $(j-1)/(j+\theta-1)$ or a single lineage mutating with probability $\theta/(j+\theta-1)$. The distribution of mutation sizes were given by the TPM described above. A sample drawn from this simulation process would have allele sizes specified relative to the allele size of the most recent common ancestor for the sample. Since I only needed information on heterozygosity, allele number and range in allele size, relative allele sizes were sufficient. Further details of the coalescent model and the validity of this approximation to the Wright–Fisher model are given in Tavaré (1984). A description of how I handled transitions between the full Wright–Fisher bottleneck process and the pre- and post-bottleneck coalescent process is given in the Appendix.

Bottleneck detection methods examined

L from Luikart et al. (1998)

The first method examined was a method developed by Luikart et al. (1998), to test for distortion of allele frequency distributions. The basis of the test is the observation that rare alleles are more likely to be lost during a bottleneck than common alleles. The test can be applied to data from not only microsatellite loci, but also to loci with very different mutation processes such as allozyme loci. To use the Luikart et al. (1998) test all the loci in a sample are pooled and then alleles are binned by frequency into 10 allele frequency classes. If fewer alleles are found in the rare frequency category than any other category (the distribution is not L-shaped), then this test "detects" a bottleneck.

M from Garza and Williamson (2001)

$$M = \frac{k}{r} \tag{4}$$

where k is the number of alleles and r is the range in allele size. While k decreases quickly during a bottleneck due to increased genetic drift, only the loss of the smallest or largest allele will lead to a reduction in r. To test for a bottleneck, an expected distribution for M under equilibrium conditions is generated using simulations. The critical value is at the lower 95th percentile of this distribution. A bottleneck is "detected" whenever a sample or test value of M is lower than this critical value.

H_k from Cornuet and Luikart (1996)

Cornuet and Luikart (1996) developed a test for heterozygosity excess. Since more rare alleles than common alleles are expected to be lost during a bottleneck, bottlenecked populations are expected to have increased levels of heterozygosity, H, given the number of alleles, k, relative to equilibrium populations. Given allele frequency data, the heterozygosity of a sample is

$$H = 1 - \sum x_i^2 \tag{5}$$

where x_i is the allele frequency of the ith allele. Cornuet and Luikart (1996) assumed that mutation at the sampled locus could be described by either the stepwise mutation model (SMM) or the infinite alleles model. θ is the only parameter that needs to be specified to describe evolution under either of these models. For loci evolving according to the SMM, they assumed a uniform distribution for $\log(\theta)$ and used computer simulations to numerically generate expected distributions for H given k. To do this test, observed heterozygosity is compared to the simulation-generated expected distribution of H given the observed number of alleles in the sample.

H_k extended to the TPM

Because the TPM is probably closer to the true mode of mutation at microsatellite loci than the SMM, I extended the method of Cornuet and Luikart (1996) to the TPM. This is similar to the extension to the TPM used by Piry (1999) but I allowed θ and p_g to vary and also used a geometric distribution (Equation (1), Di Rienzo et al. 1994) instead of a uniform distribution for selecting the size of larger mutations. I used a uniform distribution between −3 and 2 for $\log(\theta)$ and a uniform distribution between 0 and 0.2 for p_g. These distributions were chosen to represent a wide range of values for these parameters that might be found in natural populations (Cornuet and Luikart 1996; Garza and Williamson 2001).

Although it would be preferable to allow θ, p_g and δ_g to all be variable, this would make the method too computationally intensive for this type of simulation analysis. Allowing p_g alone to vary offered a significant improvement over assuming strict SMM because p_g and δ_g are similar parameters; increasing either p_g or δ_g causes an increase in the variance in mutation size (Equation (3)), and subsequently an increase in the number of alleles given θ. To simplify the computational intensity of the method, δ_g was held constant at 2.8, the value that seems most consistent with empirical data (Garza and Williamson 2001).

I simulated samples from an equilibrium population with θ and p_g chosen randomly from these distributions. I repeated these simulations until I had 10,000 simulation replicates for each value of k. From these results, I derived an expected distribution of H for each value of k under equilibrium conditions.

M_k and the TPM

Given k, the statistic M is essentially equivalent to the range in allele size, r. However for notational consistency I will use M_k rather than r_k to denote M given k. Initial simulation tests suggested that the distribution of M in an equilibrium population is strongly determined by the parameterization of the mutation model (Garza and Williamson 2001). Since k is also influenced by the parameterization of the mutation model, conditioning on k may remove some of this undesirable sensitivity. I used the same general approach and parameterization as that for H_k described above.

Combining loci

To combine loci with different number of alleles, I followed Cornuet and Luikart (1996) and assumed that the expected distributions for M_k and H_k are approximately normal. The test uses the following approximation

$$T_S = \frac{1}{\sqrt{L}} \sum_{i=1}^{L} \frac{(\hat{S} - S_e)}{\sigma_i}$$

(Cornuet and Luikart 1996; Test 2)

(6)

where L is the number of loci sampled, $\hat{S}$ is the observed statistic (M_k or H_k), S_e is the mean of the expected distribution of the statistic, and σ_i is the standard deviation of the equilibrium distribution of the statistic. At equilibrium, T should approximate a normal $N(0,1)$ distribution. I used a critical value at the 95th percentile of the normal distribution. The H_k test detected a bottleneck when $T_{Hk} > 1.645$, and the M_k test detected a bottleneck when $T_{Mk} < -1.645$. If the normal approximation is valid, these tests should have 5% type I error.

Note

A Bayesian approach for detecting population expansion or decline was also developed by Beaumont (1999). This method uses coalescent theory to assign posterior distributions to demographic parameters given the observed allele frequency distributions at the sampled loci. The posterior distributions were used to distinguish population expansion from decline. This approach is computationally quite intensive and not well suited to extensive computer simulation tests.

Furthermore, this method relies on coalescent theory, which is a good model of equilibrium populations but is not a good model of bottlenecked populations. For example, a bottleneck generally only lasts a few generations during which time polymorphism declines very quickly and many alleles coalesce (are derived from a common parental allele) simultaneously. However, in coalescent theory, at most one pair of alleles can coalesce in a single generation. Thus, the Beaumont (1999) method is best suited to detect long term and gradual changes in population size, whereas the other statistics described above are specifically designed to detect recent and dramatic departures from equilibrium. For these reasons, I did not consider the Beaumont (1999) method in this paper.

Testing the models

I tested how differences between the mutation model used to generate expected distributions for each test statistic and the mutation model used to generate the sample data can result in type I or type II errors for these methods. First, I tested these methods on data sampled from populations simulated with fixed parameter values θ, p_g and δ_g (Tables 1 and 2). These fixed parameter values included values of θ and p_g that were smaller and larger than the values of θ and p_g used to derive expected distributions for the test statistics. Note that although p_g and δ_g take fixed values, they are parameters describing a probability distribution of mutation sizes, so mutation sizes were not fixed (except if $p_g=0$, which reduces the TPM to the SSM). I tested the methods on data taken from equilibrium and bottlenecked populations to determine type I and type II error.

Next, I simulated several bottleneck scenarios and allowed mutation rate to vary among loci (Figure 1a–h). To generate these data, I drew θ from a gamma distribution with the mean equal to the variance. I modeled two different gamma distributions of θ, one with mean $\theta=2$ and one with mean $\theta=10$. The bottlenecks were size $2N_e=20$ and lasted from 1 to 7 generations. I computed the power to detect a bottleneck as a function of the duration of the bottleneck. I also tested the methods on populations that had had some time to recover after a bottleneck. Because pre- and post-bottleneck populations were modeled with coalescent simulations, this recovery time $T=0.002$ is in units of generations/$2N_e$ (if post-bottleneck $N_e=5000$ this is 20 generations).

Table 1. Type I error. Percent of the time a test gave a positive result (detected a bottleneck) when test data is drawn from an equilibrium population

θ	p_g	L	M	M_{kS}	M_{kT}	H_{kS}	H_{kT}
7 Loci							
2	0	0.06	0.00	0.09	0.00	0.04	0.08
2	0.1	0.03	0.01	0.54	0.08	0.01	0.03
2	0.2	0.01	0.09	0.79	0.26	0.00	0.01
10	0	0.01	0.00	0.06	0.00	0.03	0.11
10	0.1	0.00	0.06	0.39	0.06	0.00	0.02
10	0.2	0.00	0.20	0.62	0.18	0.00	0.01
20	0	0.00	0.00	0.06	0.00	0.03	0.11
20	0.1	0.00	0.11	0.29	0.06	0.01	0.03
20	0.2	0.00	0.33	0.49	0.14	0.00	0.01
14 Loci							
2	0	0.02	0.00	0.08	0.00	0.05	0.14
2	0.1	0.00	0.00	0.73	0.08	0.01	0.04
2	0.2	0.00	0.09	0.95	0.37	0.00	0.01
10	0	0.00	0.00	0.06	0.00	0.04	0.23
10	0.1	0.00	0.05	0.58	0.06	0.00	0.03
10	0.2	0.00	0.31	0.86	0.27	0.00	0.01
20	0	0.00	0.00	0.05	0.00	0.04	0.23
20	0.1	0.00	0.15	0.46	0.06	0.00	0.04
20	0.2	0.00	0.53	0.74	0.20	0.00	0.01

The sample size was 70 (35 diploid individuals). The test data was generated using the fixed parameter values given in the table. In all simulations $\delta_g=2.8$. For example, in the first row, the 7 loci in the sample were drawn from an equilibrium population in which $\theta=2$, $p_g=0$, and $\delta_g=2.8$. Each cell represents the result from 10,000 simulation replicates. The method tested is listed at the top of each column. L is the L-shape test. M is the ratio test, where the expected distribution of the ratio M was derived from an equilibrium model with fixed parameters $\theta=10$, $p_g=0.1$, and $\delta_g=2.8$. M_{kS} and M_{kT} are the ratio M conditional on the number of alleles, k. H_{kS} and H_{kT} are the heterozygosity test conditional on k. The expected distributions of M_{kS} and H_{kS} were derived using a Single Step Model (SSM) and a uniform distribution for log(θ) from −3 to 2. The expected distributions of M_{kT} and H_{kT} were derived using a Two-phase model (TPM) and a uniform distribution of p_g from 0 to 0.2, $\delta_g=2.8$, and a uniform distribution for log(θ) from −3 to 2.

Results

For the parameter values I examined, the test for distortion of allele frequency distribution developed by Luikart et al. (1998) consistently had low type I error (detecting a bottleneck in an equilibrium population) (Table 1). However, this test had high type II error (Table 2). The power, to detect a

Table 2. Type II error. Percent of the time a test failed to detect a bottleneck given test data drawn from a bottlenecked population

θ	p_g	L	M	M_{kS}	M_{kT}	H_{kS}	H_{kT}
7 Loci							
2	0	0.52	1.00	0.52	0.94	0.57	0.48
2	0.1	0.58	0.92	0.17	0.66	0.62	0.52
2	0.2	0.63	0.75	0.06	0.39	0.67	0.57
10	0	0.73	0.83	0.11	0.61	0.56	0.35
10	0.1	0.81	0.34	0.01	0.17	0.61	0.41
10	0.2	0.86	0.10	0.00	0.03	0.65	0.43
20	0	0.92	0.41	0.02	0.25	0.57	0.33
20	0.1	0.96	0.05	0.00	0.02	0.60	0.36
20	0.2	0.98	0.01	0.00	0.00	0.61	0.37
14 Loci							
2	0	0.58	1.00	0.34	0.95	0.29	0.18
2	0.1	0.65	0.94	0.04	0.51	0.37	0.23
2	0.2	0.70	0.68	0.00	0.17	0.45	0.28
10	0	0.79	0.76	0.01	0.42	0.27	0.08
10	0.1	0.89	0.13	0.00	0.03	0.36	0.11
10	0.2	0.92	0.01	0.00	0.00	0.39	0.13
20	0	0.97	0.18	0.00	0.06	0.28	0.07
20	0.1	0.99	0.00	0.00	0.00	0.32	0.09
20	0.2	1.00	0.00	0.00	0.00	0.34	0.09

The bottleneck was size $N_e=10$ ($2N_e=20$) and lasted two generations. The sample size is 70 (35 diploid individuals). The test data is generated using the fixed parameter values given in the table. In all cases $\delta_g=2.8$. Each cell represents the result from 10,000 simulation replicates. Each column represents a different type of test which is described in Table 1.

two-generation bottleneck of $N_e=10$ was lowest when mutation rates were high ($\theta=20$, Table 2). This test was also less likely to detect the bottleneck when mutations were larger ($p_g=0.2$, Table 2). Adding more loci did not reduce type II error (Table 2). The L-test appears to be a conservative test with limited power to detect bottlenecks.

The ratio test is based on the idea that drift removes alleles randomly with respect to allele size, and so the number of alleles at a locus, k, decreases more quickly than the total range in allele size, r. Therefore, bottlenecked populations are likely to have a larger range for a given number of alleles, or a smaller ratio $M=k/r$, than equilibrium populations. In other words, genetic drift creates gaps in the distribution of alleles arranged by size. However, a higher mutation rate or a larger probability of multi-step mutations (larger p_g) also lead to a smaller ratio by making gaps in the total allele size distribution more likely. As a result, I expected to find increased type I error when the expected distribution of M was generated assuming a lower θ or p_g than the θ and p_g that were used to generate the data.

As expected, the ratio test was prone to type I error when the test data had values of θ or p_g that were higher than the values of θ and p_g that were used to determine critical values (Table 1). For example, when the expected distribution and critical values for M were generated assuming an equilibrium population with parameters $\theta=10$ and $p_g=0.1$ (the M column) type I error was highest for test data taken from populations with the largest values of θ and p_g ($\theta=20$ and $p_g=0.2$, Table 1). The type I error was even higher for the ratio test that assumed the SSM ($p_g=0$) (M_{kS} column, Table 1). Although assuming the TPM reduced type I error (M_{kT} column, Table 1), the ratio test was still prone to high (greater than 5%) type I error if the parameter p_g was underestimated. For example, the M_{kT} test assumed p_g was uniformly distributed between 0 and 0.2. For the rows in which test data was simulated with $p_g=0.2$, the M_{kT} ratio test had high type I error. Underestimating pre-bottleneck δ_g produced similar results (not shown). Not too surprisingly, more data does not reduce errors when the errors stem from violations of model assumptions. Type I error was even larger when more loci were sampled.

Symmetrically, type II error was highest for the ratio test when test data had values of θ or p_g that were lower than the values of θ and p_g that were used to determine critical values (columns M, M_{kS}, and M_{kT}, Table 2). Reduced type I error appears to have the tradeoff of increased type II error. Parameter combinations that increase the probability of correctly detecting a bottleneck (Table 2) also increase the probability of falsely "detecting" a bottleneck in an equilibrium population (Table 1).

In contrast to the ratio test, type I error was highest for the heterozygosity test when test data had values of θ or p_g that were lower than the values of θ and p_g that were used to determine critical values (columns H_{kS}, and H_{kT}, Table 1). For example, the highest type I error is found when the TPM is assumed (H_{kT}) but the test data is strictly SSM ($p_g=0$, Table 1). The heterozygosity test is based on the idea that bottlenecks eliminate rare alleles. Low mutation rates and low values of p_g can also make rare alleles less likely at

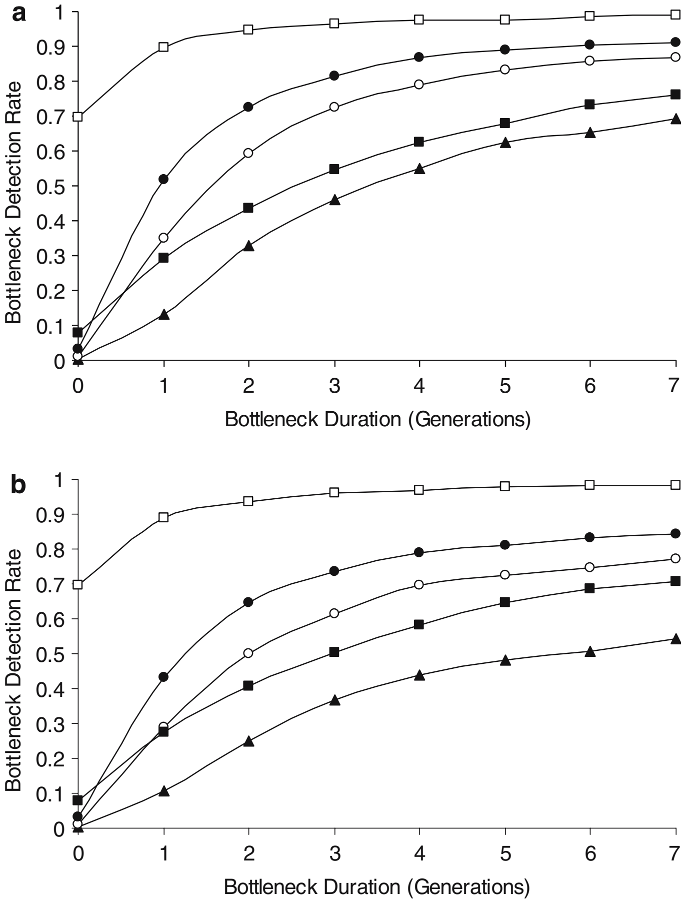

Figure 1. The proportion of the time a bottleneck is detected, as a function of the number of generations the bottleneck lasted. The leftmost point on each plot, with no bottleneck, is the type I error. The remaining points are the power to detect a bottleneck (1 – type II error) after successive generations of bottleneck. Each point is an average over 10,000 simulated data sets. For computing the test (equilibrium/null) distribution I assumed the following: a uniform prior distribution between −3 and 2 for log(θ), a uniform prior distribution between 0 and 0.2 for p_g, and $\delta_g = 2.8$. Simulated test data were drawn from a population with mutation rates variable among loci (θ was drawn randomly from a gamma distribution with mean = $\overline{\theta}$ = variance) and a fixed probability distribution for mutation sizes (fixed p_g and δ_g). During the bottleneck $2N_e = 20$. Sample size was 70 (35 diploid individuals) and 14 loci. Solid circles are H_k assuming TPM, open circles are H_k assuming SMM, solid squares are M_k assuming TPM, open squares are M_k assuming SSM, and triangles are the L method. In Figure 1(e)–(h) the larger solid circles are H_k assuming TPM tested on data with $\delta_g = 3.5$, the smaller solid circles are H_k assuming TPM tested on data with $\delta_g = 1.5$, the larger solid squares are M_k assuming TPM tested on data with $\delta_g = 3.5$, the smaller solid squares are M_k assuming TPM tested on data with $\delta_g = 1.5$. The parameters describing the test data for each figure are given below. T is the post-bottleneck recovery time in units of generations/$2N_e$. For example, if post-bottleneck $N_e = 5000$. Then $T = 0.002$ is equivalent to 20 generations. (a–h) $P_g = 0.12$, (a–d) $\delta_g = 2.8$, (e–f) $\delta_g = 1.5$ or 3.5, (a) $\overline{\theta} = 2$, $T = 0.0$, (b) $\overline{\theta} = 10$, $T = 0.0$, (c) $\overline{\theta} = 2$, $T = 0.002$, (d) $\overline{\theta} = 10$. $T = 0.002$. (e) $\overline{\theta} = 2$, $T = 0.0$, (f) $\overline{\theta} = 10$, $T = 0.0$, (g) $\overline{\theta} = 2$, $T = 0.002$, (h) $\overline{\theta} = 10$, $T = 0.002$.

equilibrium because mutations are more likely to generate allele sizes that already exist. As in the case of the ratio tests, increasing the number of loci only exacerbates the type I error that is due to violations of model assumptions (Table 1). Again, there is a tradeoff between type I and type II error. Increasing the power of correctly identifying a bottleneck by changing the assumed underlying mutation model also increases the chances of falsely detecting a bottleneck when there isn't one (Tables 1 and 2).

I next explored the performance of these bottleneck tests on data taken from populations that had experienced different durations of bottleneck and also allowed the mutation rate to vary among loci. The results for these simulation tests are

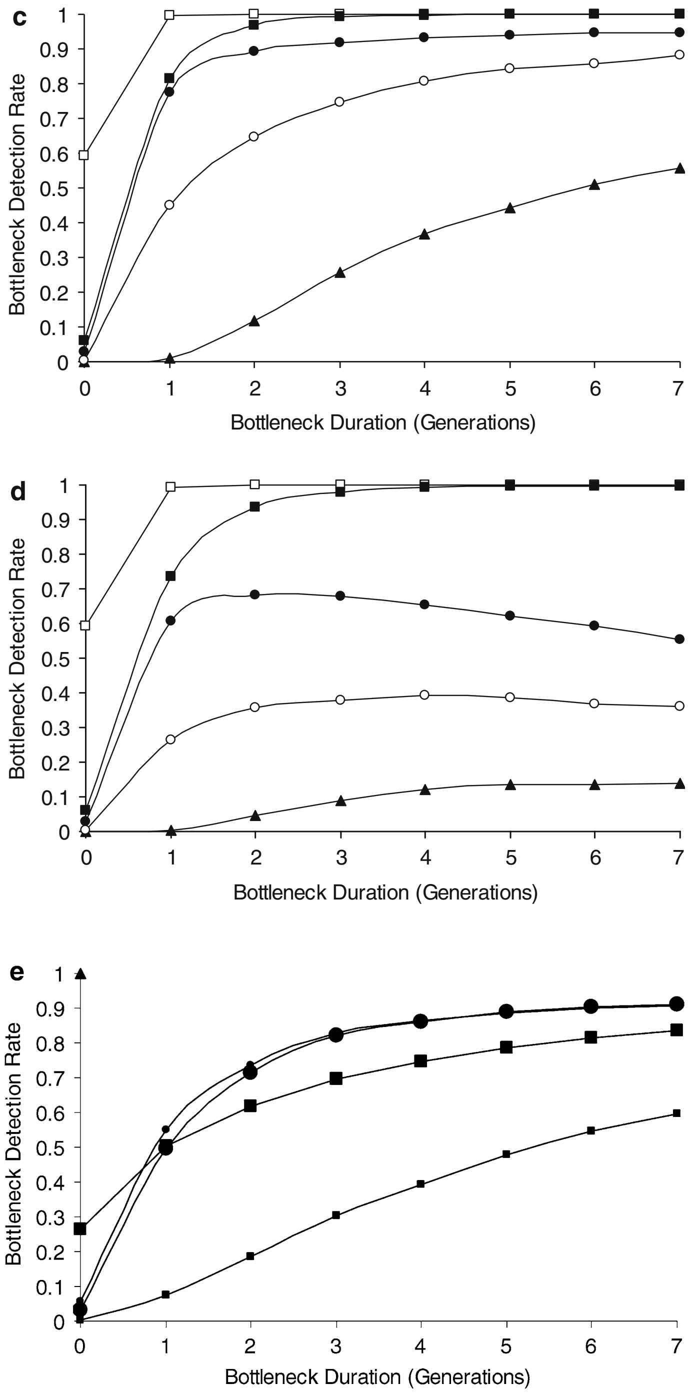

Figure 1. (Continued)

shown in Figure 1(a)–(h). As I had found in earlier simulations (Tables 1 and 2), the M_k method was prone to very high type I error when the SSM was assumed (Figure 1a–d, open squares, bottleneck duration = 0). Together these simulations suggest that the M_k method assuming the SSM is probably not a useful method because type I error is predictably high. On the other hand, also similar to

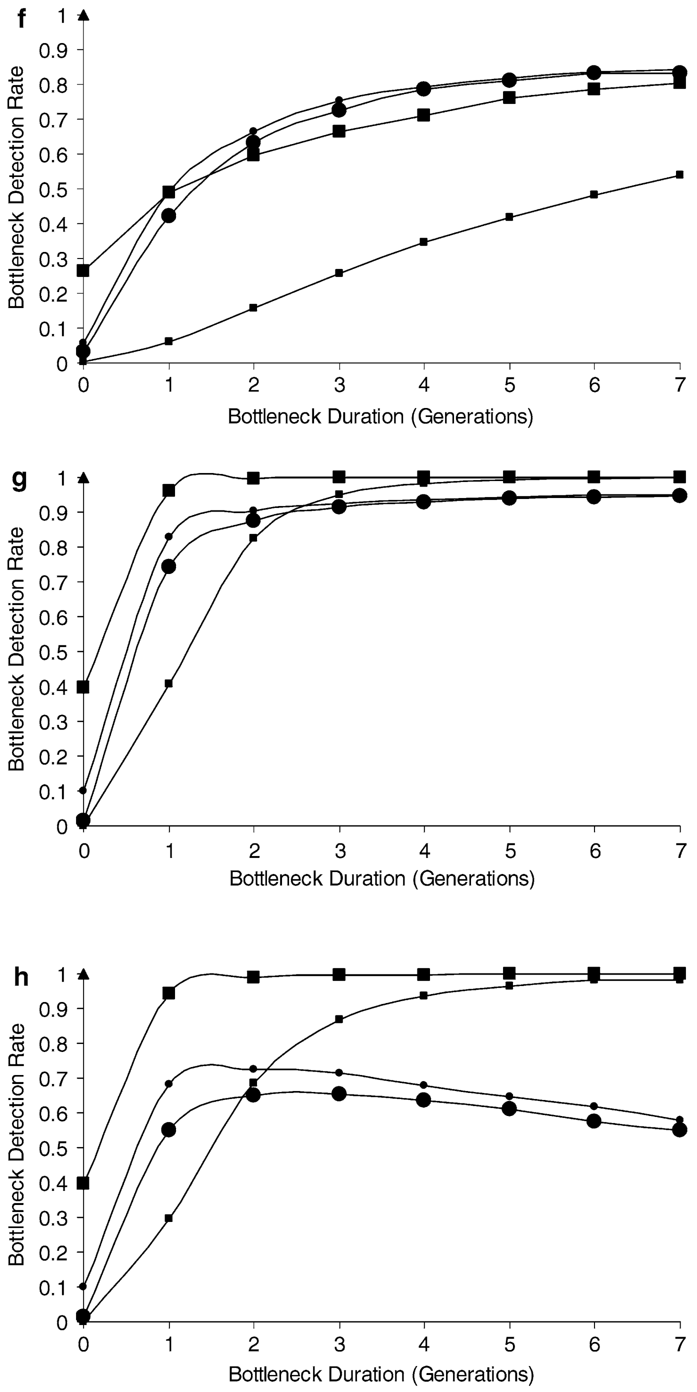

Figure 1. (Continued)

earlier results (Tables 1 and 2), the *L*-test consistently had low type I error but also a consistently low probability of correctly identifying a bottleneck (Figure 1a–d, triangles).

Type I error for the ratio method assuming the TPM was also larger than 5%, especially when θ was small (solid squares, bottleneck duration = 0, Figure 1a–d). This is because the expected equilibrium distribution of M_k is not normally distributed for small k, and small θ increases the chances of small k. Preliminary results suggested that type I error can be reduced by discarding loci with $k=2$ and 3. However, this seemed to be a strategy that would be of limited use for most real data sets, so I did not pursue this approach further.

When $\overline{\theta} = 2$, results for simulations with population recovery (Figure 1b) were very similar to results with bottleneck alone (Figure 1). For low mutation rates, $T=0.002$ is apparently not sufficient time to accumulate enough mutations to influence the bottleneck tests. On the other hand, when $\overline{\theta} = 10$, both the H_k and L methods were less able to detect a bottleneck after a brief population recovery (Figure 1a–d). This may be because these tests detect a shortage of rare alleles, and rare alleles may accumulate quickly during population recovery.

Generally, these simulation tests suggest that H_k assuming the TPM is more likely to correctly identify a bottleneck than the other methods if the bottleneck is very recent, or less severe, or if the pre-bottleneck value of θ was small (solid circles, Figure 1a, b). On the other hand, if the pre-bottleneck θ was large, the bottleneck lasted several generations, and the population has had some time to recover, then M_k assuming the TPM seems to be the method the most likely to correctly identify a bottleneck (solid squares, Figure 1c, d).

Finally, I tested the methods using simulated data with $\delta_g=1.5$ and $\delta_g=3.5$ but otherwise the same parameterization as before (Figure 1e–h). I only show the M_k and H_k methods (assuming TPM) because they typically had lower error than the other methods (see Figure 1a–d).

As before, H_k assuming the TPM was more likely to correctly identify a bottleneck when the bottleneck was very recent, or less severe, or if the pre-bottleneck value of θ was small (Figure 1e, f, circles). M_k assuming the TPM was more likely to correctly identify a bottleneck when the pre-bottleneck θ was large, the bottleneck lasted several generations, and the population had some time to recover (Figure 1g, h, squares).

The M_k method had high type I error when δ_g was underestimated (Figure 1e–h, large squares, bottleneck duration = 0). This is similar to the type I errors in the ratio method when p_g was underestimated (Table 1). Allowing p_g to have a variable distribution reduces this type I error, however 0–0.2 may be too small of a range for p_g to compensate for $\delta_g=3.5$. For example, the largest variance in allele size with $\delta_g=2.8$ and p_g between 0 and 0.2, is $\sigma_m{}^2=2.656$ when $p_g=0.2$, whereas if $p_g=0.12$ and $\delta_g=3.5$ then $\sigma_m{}^2=2.8$ (Eq. 3). These results suggest that to the ratio test is reliably conservative only if you use a large value for δ_g or a broad distribution for p_g when generating the expected distribution for M_k. Since good empirical estimates of p_g and δ_g are generally not available, and there can be a tradeoff of increased type II error if p_g and δ_g are underestimated (Table 2 and Figure 1e–h, small squares), it is difficult to know where to set these values to best balance these two types of error.

Similarly, the H_k method had type I error above 5% when δ_g was overestimated (Figure 1e–h, small circles, bottleneck duration = 0), just like when p_g was overestimated (Table 1). However, in these simulations this type I error was typically much smaller than the type I errors seen in the ratio method. It it is possible to force the H_k method to be conservative by simply assuming the SSM, however there may be tradeoffs in terms of increased type II error (open circles, Figure 1a–d).

Discussion

If the assumptions of the models are met, these simulation results suggest that H_k is the method most likely to correctly detect a bottleneck if the bottleneck was very recent, less severe, and the pre-bottleneck value of θ was small. If, on the other hand, the pre-bottleneck θ was large, the bottleneck lasted several generations, or the population made a demographic recovery, then the bottleneck is more likely to be correctly detected from the value of M_k.

Conditioning the ratio M on the number of alleles, k, (M_k) reduces the type I error that arises from not knowing pre-bottleneck θ. However, the M_k method is still prone to type I error if the true pre-bottleneck p_g or δ_g is higher than the of p_g or δ_g used to generate the expected equilibrium distribution of M_k. Due to this sensitivity to p_g and δ_g, the utility of the M_k statistic for analyzing natural populations depends on how much variation there is among microsatellite loci in these parameters. If p_g and δ_g can be expected to be below some upper bound, then the ratio M could be a very informative statistic. The distribution of p_g and δ_g among loci in natural populations is an empirical question that will hopefully be resolved as more data become available.

Heterozygosity conditioned on the number of alleles, originally proposed by Cornuet and Luikart (1996), had improved power to detect bottlenecks when the equilibrium distribution for H_k was generated using the TPM rather than the SMM. However, assuming the TPM can increase type I error if of p_g or δ_g are underestimated. Assuming the SSM is a more conservative approach, but there may be less power to detect bottlenecks that have actually occurred than if the TPM was used.

This paper only investigates the error in bottleneck detection methods that might be introduced by incorrect parameterization of the mutation model. Other violations of model assumptions, such as population subdivision and admixture or differential selection of alleles could also be major sources of error. To assess appropriateness of any method a researcher must evaluate how well a specific population fits the assumptions of the method and also decide the degree and type of error that is acceptable. Hopefully this analysis will help inform this process.

Acknowledgements

This study was supported by NIH grant GM40282 to M. Slatkin. I would like to thank J.C. Garza, C. Muirhead, R.D. Schnabel, M. Slatkin, and T.J. Ward for many helpful comments and suggestions.

Appendix

The transition between an equilibrium population to a bottlenecked population required the transition from a coalescent simulation to a full Wright–Fisher simulation. To do this, I generated a sample of size $2N_b$ using the coalescent simulation. This sample represented the entire population during the first generation of the bottleneck.

Simulating the recovery from a bottleneck to a large population size required the transition from the full Wright–Fisher simulation to the coalescent simulation. This transition was more complicated because coalescent simulations keep track of only the sample and ancestors of the sample backward in time while full Wright–Fisher simulations require information on the entire population and work forward in time. First, using the size of the sample and the time since the bottleneck as input parameters, I used a coalescent simulation to determine the number of ancestors, n, of the sample, that were present immediately after recovery from the bottleneck. Thus, the coalescent simulation of the recovered population was stopped after a specific amount of time, T, unlike the pre-bottleneck simulation, which was stopped when all lineages coalesced to a single common ancestor. Allele sizes of these n ancestors were then determined by sampling (with replacement) n individuals from the bottlenecked population (the last generation of the full Wright–Fisher imulation). Finally, allele sizes for the sample from the recovered population were determined by the mutations and coalescent events dictated by the coalescent simulation that determined n.

References

Anderson EC, Williamson EG, Thompson EA (2000) Monte carlo evaluation of the likelihood for N_e from temporally spaced samples. *Genetics*, **156**, 2109–2118.

Beaumont MA (1999) Detecting population expansion and decline using microsatellites *Genetics*, **153**, 2013–2029.

Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB (1994) Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA*, **91**, 3166–3170.

Ewens WJ (1979) *Mathematical Population Genetics*, Springer-Verlag, New York.

Frankham R (1995) Conservation genetics *Annu. Rev. Genet.*, **29**, 305–327.

Frankham R (1999) Quantitative genetics in conservation biology *Genet. Res.*, **74**, 237–244.

Garza JC, Williamson EG (2001) Detection of reduction in population size using data from microsatellite loci. *Mol. Ecol.*, **10**, 305–318.

Lande R (1994) Risk of population extinction from fixation of new deleterious mutations *Evolution*, **48**, 1460–1469.

Luikart GL, Allendof FW, Cornuet JM, Sherwin WB (1998) Distortion of allele frequency distributions provides a test for recent population bottlenecks. *J. Hered.*, **89**, 238–247.

Newman D, Pilson D (1997) Increased probability of extinction due to decreased genetic effective population size: Experimental population of Clarkia pulchella. *Evolution*, **51**, 345–362.

Piry S, Luikart G, Cornuet JM (1999) BOTTLENECK: A computer program for detecting recent reductions in the effective population size using allele frequency data. *J. Hered.*, **90**, 502–503.

Schwartz MK, Tallmon DA, Luikart G (1998) Review of DNA-based census and effective population size estimators. *Animal Conserv.*, **1**, 293–299.

Tavare S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models *Theor. Popul. Biol.*, **26**, 119–164.

Waples RS (1989) A generalized approach for estimating effective population size from temporal changes in allele frequency *Genetics*, **121**, 379–391.

Williamson EG, Slatkin M (1999) Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics*, **152**, 755–761.