The effect of population history on the distribution of the Tajima's D statistic

Deena Schmidt and John Pool

May 17, 2002

Abstract

The Tajima's D test measures the allele frequency distribution of nucleotide sequence data. This statistic can be influenced by both population history and natural selection. The purpose of this study is to distinguish between these two evolutionary forces. We observe the simulated distribution of Tajima's D under an equilibrium model, an exponential growth model, and a population bottleneck model. We note the probability of rejection under Tajima's published critical values, and subsequently generate a corrected set of critical values for each demographic scenario. Utilizing these adjusted values should increase the power and reduce the chance that demography will cause the rejection of the null hypothesis of the Tajima's D test.

1 Introduction

The Tajima's D test is a widely used test of neutrality in population genetics. This statistic illustrates the allele frequency distribution of nucleotide sequence data and is based on the difference between two estimators of θ (the population mutation rate $4N_e\mu$): (1) Tajima's estimator, which is based on the average number of pairwise differences between sequences, and (2) Watterson's estimator, which is based on the number of segregating sites in the sample. Note that Tajimas estimator takes into account allele frequency when comparing pairwise differences between sequences whereas Wattersons estimator does not. A segregating site counts as any point where there are differing nucleotides between sequences in the data set, independent of the actual number of differences and hence independent of allele frequency. The difference between these two estimators of θ is scaled by the standard deviation of their difference. A positive value indicates an excess of intermediate frequency (polymorphic) alleles, while a negative value indicates an excess of rare alleles. The null hypothesis of the Tajima's D test is neutral evolution in an equilibrium population. This implies that no selection is acting at the locus and that the population has not experienced any recent growth or contraction (Tajima 1989).

Other tests of neutrality for nucleotide sequence data exist, such as Fu and Li's D^*. However, Tajima's D test seems to have more power to detect selective sweeps than

other tests of its kind. Still, Tajima's original critical values may be overly conservative (Simonsen, Churchill, and Aquadro 1995). A more accurate set of critical values would further increase the power of this test.

Although the Tajima's D statistic is commonly employed as a test for natural selection, this is confounded by the influence of population history. For instance, balancing selection may yield a positive value for Tajima's D. However, demographic processes such as population reduction (Maruyama and Fuerst 1985), population subdivision, a recent bottleneck (Maruyama and Fuerst 1985), or migration can also produce this result. Similarly, linkage to a selective sweep or purifying selection at the studied locus can generate a negative value for Tajima's D. Yet, this pattern can also be due to population growth (Maruyama and Fuerst 1984), a less recent bottleneck (Maruyama and Fuerst 1985), or migration. We see both positive and negative D values with varying levels of migration. To cite an example, Wall and Przeworksi (2000) examined human nuclear sequence data to evaluate hypotheses involving both selection and demography. If they were able to test these factors separately, more confidence could be placed in their explanations.

The intent of this study is to discriminate between the influences of population history and natural selection. To do this, we will first simulate DNA sequence data under various demographic scenarios and study the distribution of Tajima's D values. We will also note the frequency of rejection under Tajimas published critical values. Lastly, we will produce adjusted critical values specific to the demographic parameters defined in the model.

2 Methods

We utilize a computer program that simulates DNA sequence data under an infinite sites model. The program generates genealogies based on coalescent theory. Coalescence times are randomly chosen according to an exponential distribution, and mutations are then added in randomly across the genealogy. This process is repeated for 100,000 cycles for each case. The program then calculates Tajima's D for each replicate and compiles a distribution of these values. We note the mean and variance of this distribution, and observe the histogram of Tajima's D values produced by the program. The program also returns the probability of rejection under Tajima's published critical values, along with modified rejection regions to fit the simulated data.

With this program, we perform simulations under an equilibrium model, an exponential growth model, and a population bottleneck model. Each of these models is investigated with per locus q values of 1, 10, and 50. Most runs use a sample size of n = 10, and a few cases are explored with n = 50.

The exponential growth model describes an expanding population. In addition to q and n, this model incorporates the parameter a. Slatkin and Hudson (1991) define a as the product of the initial population size N_0 and the population growth rate r. Five positive values of a are considered: $0.1, 0.5, 1.0, 2.0$, and 5.0. These values are large enough to have a measurable effect on the data, but still lie within a reasonable biological range (Innan and Stephan, 2000). The scenario of population reduction

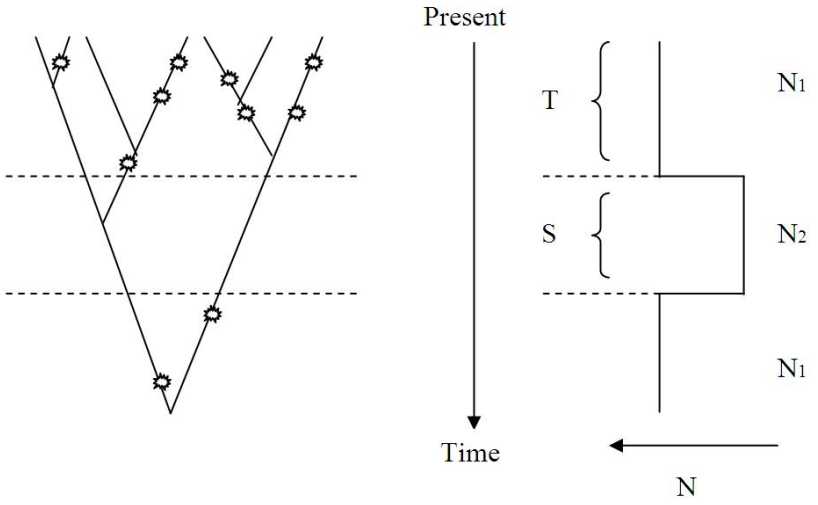

Figure 1: A graphical depiction of the computer simulated bottleneck model. At a time T before the present, the population size is reduced from N1 to N2 for a period of time S, during which no mutations are permitted. The genealogy on the left illustrates that no mutations occur during time S.

cannot be examined under the constraints of the exponential growth model, and is not investigated during this study.

The population bottleneck model involves a sudden reduction in population size, followed by an instantaneous recovery. This represents an oversimplification of a natural bottleneck, but is expected to behave similarly for our purposes. This model incorporates two new parameters, T and S. T denotes the time since the occurrence of the bottleneck, scaled in units of $2N$ generations. S measures the strength of the bottleneck, which includes both the degree of population size reduction and the duration of the bottleneck. The program models a bottleneck by turning off mutations, but still allowing coalescences, for a period of time equal to S in units of $2N$ generations. This bottleneck model is described by Galtier, Depaulis, and Barton (2000), and the process is depicted in Figure 1. We investigate a weak bottleneck with $S = 0.5$ and a strong bottleneck with $S = 1.0$. For each of these scenarios, we examine a sequence of times after a bottleneck where $T = 0, 0.01, 0.05, 0.10, 0.25$, and 0.50. These values are expected to give a variety of results for Tajima's D, with the most recent bottlenecks producing positive values for D and less recent bottlenecks generating negative values for D. An increase in the variance of Tajima's D is also expected (Maruyama and Fuerst 1985).

The program gives us the proportion of D values that fall above Tajima's published upper 2.5% critical value and below the lower 2.5% critical value (Tajima 1989). A similar idea is presented by Innan and Stephan (2000). We combine these two proportions into a total rejection percentage. The program also returns the upper and lower 2.5% cutoff values for the simulated Tajima's D distribution. This represents a set of

corrected critical values for the given demographic model.

3 Results

A numerical summary of our results is presented in Table 1. As expected, the equilibrium simulations give mean values of Tajima's D close to zero. Notably, we observe a slightly negative skew in the mean value of D. This effect is more pronounced for higher values of q, in agreement with the results of Tajima (1989). Higher values of q are also associated with a modestly reduced variance in D. The distribution of Tajima's D values for each equilibrium trial is presented in Figure 2. The overall rejection region is less than 5% in each case, which supports the claim that Tajima's published critical values are generally too conservative (Simonsen, Churchill, and Aquadro 1995). In most cases, we generated a corrected rejection region smaller than the published rejection region. However, for the case of q = 10, we find the probability of rejection in the negative direction is 0.0256. Since this value should not exceed 0.025, our result is a slight non-conservative deviation. This demonstrates that Tajima's critical values may not be completely reliable, even under equilibrium conditions.

As predicted, population growth models produce increasingly negative mean values for Tajima's D. However, the reduced variance at higher a values overwhelms the negatively shifting mean. Figure 3 illustrates this pattern for q = 50, which produces the most notable difference from equilibrium. As shown in Table 1, the cases of q = 1 and q = 10 produce a similar trend, but to a lesser degree. The negatively skewed but narrowing distribution described above has some interesting consequences with respect to rejection regions. Tajima's published critical values are conservative in every case examined, with overall rejection rates varying from 0.2% to 2.4%. For the corrected critical values, the lower 2.5% value stayed fairly constant between trials. This can be seen in figure 3, where the negatively skewed mean for a = 5 is offset by the reduced variance. In contrast, the upper 2.5% critical value is significantly reduced for higher a. The probabilities of rejection in the upper direction are correspondingly reduced, to the point of being undetectable in our simulations for a = 5 and q = 50.

The expected pattern of average Tajima's D values through time after a bottleneck includes an initial period of positive D values, declining toward a period of negative D values, and finally turning back toward equilibrium (Maruyama and Fuerst, 1985). This is consistent with our observations for both weak and strong bottlenecks (Figure 4). Weak bottlenecks start out with higher D values, while strong bottlenecks produce more sharply negative results.

The most extreme results for weak and strong bottlenecks are seen for q = 50, where we observe both the most positive and most negative mean Tajima's D values, along with the highest rate of rejection. The patterns we see for q = 1 and q = 10 follow a similar trend, but to a lesser extreme. As shown in Table 1, q = 1 shows a less dramatic shift in mean D value, and also a substantially less inflated, relatively constant variance over the range of T values we consider. Notably, Tajima's critical values are conservative for every case of q = 1. q = 10 gives somewhat more negative mean D value, a higher variance, and elevated rates of rejection (in both directions).

Other than the magnitude of these changes, the main difference in q values appears to be the rate at which these simulated populations go through and recover from the bottleneck cycle. As might be expected, a higher population mutation rate first leads to a quicker reduction in mean D value, and subsequently a faster recovery to equilibrium. Since q = 50 gave the most illustrative trends, we focus much of our examination on this case.

For a weak bottleneck with q = 50, Figures 5 and 7 depict the distribution of Tajima's D values for a series of T values. The mean D value starts out rather high (0.652) at $T = 0$. The variance also starts out relatively high, and the total rejection rate starts out at 27.6%. All three of these measures decline as T progresses through the remaining trials. The rate of rejection in the positive direction is high at $T = 0$ and decreases thereafter. In contrast, the negative rejections start out moderately high, then peak around $T = 0.05$ and $T = 0.10$ before declining at higher values of T. By $T = 0.50$, both positive and negative rejections have fallen off to a total rejection rate of 3.0%. However, even at this stage (which represents N generations after a weak bottleneck since T is scaled in units of $2N$ generations), Tajima's lower critical value is slightly non-conservative (0.0255).

The results for a strong bottleneck with q = 50 are in some ways even more extreme, as illustrated in Figures 6 and 8. However, we should first note that at $T = 0$, the mean D value (0.544) is less elevated than for a weak bottleneck (0.652). This would make sense if during the strong bottleneck, more of our replicate cycles coalesced down to one lineage than in the weak bottleneck. Our data support this claim. Immediately after the strong bottleneck, 36.1% of Tajima's D values were centered at zero. This is consistent with having only one allele or lineage survive the bottleneck. For the weak bottleneck, this proportion was only 9.4%. Variance starts off very high at $T = 0$, and actually expands to 2.44 at $T = 0.01$ before contracting throughout the remaining trials. Due to the sharply elevated variances, the overall rejection rates are very high (above 25

A few representative cases are studied with n = 50 (all with q = 10), and Table 1 displays the results. These trials test for the influence of a larger sample size on the distribution of Tajima's D. For the equilibrium case, the variance is nearly identical to the n = 10 case. The mean is slightly more negative (-0.0995), but this does not lead to non-conservative rejection rates. For the case of a = 1, the mean is also more negatively skewed, but the distribution is not greatly affected. For the bottleneck trials, n = 50 serves to exaggerate the trends already observed. Positive mean D values are increased, up to 0.962 for $T = 0.01$ with a weak bottleneck. Negative mean D values are further reduced, down to -0.864 for $T = 0.25$ with a strong bottleneck. These represent the most extreme Tajima's D values recorded in this study, and also produce increased rejection rates over the corresponding n = 10 simulations.

4 Discussion

In this study we use computer simulations to examine the distribution of the Tajima's D test statistic under an equilibrium model, an exponential growth model, and a pop-

ulation bottleneck model. For the equilibrium model, results are much as expected, and match the findings of Tajima (1989). We note that Tajima's published critical values are quite conservative. Using corrected critical values such as those presented in Table 1 should add more power to reject the null hypothesis of this test. For example, in the equilibrium trial with n = 10 and q = 50, Tajima's rejection region is (-1.73, 1.98). The corrected range is (-1.71, 1.60). The exception where Tajima's lower critical value was not conservative is worth noting. In this case, our corrected value slightly extends the negative 2.5 percentile (to -1.74). This example provides another reason to use corrected critical values from computer simulations, even when a population is thought to be at equilibrium.

The effect of increasing values of q is primarily to enhance the magnitude of trends observed at lower q values. For the equilibrium case, higher q leads to a more negative mean and slightly contracted variance. Increasing the sample size to n = 50 also reduced the mean, but had no apparent effect on the variance. We focus in particular on cases where q = 50, which often provide the most dramatic results. Also, such per locus values of q are likely to be of increasing importance in population genetics as sequencing larger regions becomes more feasible. This suggests that the extreme results generated by q = 50 trials will be relevant to many future studies.

For the exponential growth model, we observe that increasing a values lead to a negatively shifted but significantly contracted distribution of Tajima's D values. The observed reduction in mean D value is expected due to the input of new mutations in the expanding population. The lower variance would make sense if the increased rate of lineage survival served to decrease the stochastic role of genetic drift in the population. Because of the reduced variance, both Tajima's upper and lower critical values are always conservative for these situations. Therefore, if a studied population is experiencing exponential growth, it may be very difficult to reject the null hypothesis, particularly in the positive direction. As an extreme example, when q = 50 and a = 5.0, there is no perceptible probability of exceeding Tajima's upper critical value under the null hypothesis, and less than a 0.6

Our study supports the bottleneck model of Maruyama and Fuerst (1985). The period immediately after the bottleneck is characterized by an excess of intermediate frequency alleles as rare alleles are lost from the population due to genetic drift. However, as new mutations enter the re-expanded population, the distribution shifts toward an excess of rare alleles. Finally, the population returns to equilibrium. Hence the pattern shown in Figure 4, initially positive Tajima's D values becoming negative, then turning back toward zero.

In general, Tajima's published critical values are not conservative with regard to either of the bottleneck scenarios investigated in this study. In the cases of q = 10 and q = 50, the null hypothesis is rejected at surprisingly high rates (up to 34.2%). The weak bottleneck ($S = 0.5$) rejects in the positive direction more often than the strong bottleneck ($S = 1.0$), which in turn rejects more frequently in the negative direction. This is probably related to the higher proportion of genealogies that coalesce down to one lineage during the strong bottleneck, which lasts for $2N$ generations. The differences among genealogies between those that coalesce completely during the bottleneck and those in which multiple lineages survive may help account for the dramatically

elevated variance, which is noted particularly for the strong bottleneck. The case of n = 50 enhances these deviations and increases rejection rates. Like large q values, large sample sizes may become increasingly relevant in future sequencing studies.

The extreme rejection rates observed for population bottleneck scenarios indicate that it is very unwise to use Tajima's D as a test of neutrality with published critical values if a bottleneck is suspected in the population's history. Recent bottlenecks produce substantial deviations from the equilibrium distribution of this test statistic. Even for much less recent bottlenecks, Tajima's published critical values may be non-conservative. Due to these inconsistencies, an adjusted set of critical values like those presented in Table 1 should be utilized. Since the variance is so high, these critical values are widely separated. For instance, when q = 50 and $T = 0$ after a strong bottleneck, the corrected range is (-2.14, 2.78). Although the power to reject these critical values may be much less than under an equilibrium model, their use should dramatically reduce the probability of rejecting the null hypothesis due to demography.

The goal of this study is to separate the effects of demographic history and natural selection in the examination of DNA sequence data. By simulating the distribution of the Tajima's D test statistic, we were able to observe the influence of exponential growth and population bottlenecks, as compared to an equilibrium model. We find that Tajima's D is conservative for exponential growth, but that population bottlenecks severely violate the expected distribution if an equilibrium model is assumed. From the simulated distributions obtained here, we record the 2.5 percentile on each tail of the distribution, yielding a corrected set of critical values for each set of demographic parameters. In order to apply this method, accurate estimates of these parameters are needed. Connecting this study to valid methods of parameter estimation presents a possible avenue for future study. Whenever such estimates are available, the use of these new critical values should increase the power and avoid a rejection due to population history. Only with an accurate set of critical values that account for the demographic history of a given population can the Tajima's D test be legitimately used as a test of selective neutrality.

5 References

Galtier, N., F. Depaulis, and N. H. Barton, 2000. Detecting Bottlenecks and Selective Sweeps from DNA Sequence Polymorphism. Genetics 155: 981-987.

Innan, H., and W. Stephan, 2000. The Coalescent in an Exponentially Growing Metapopulation and Its Application to Arabidopsis thaliana. Genetics 155: 2015-2019.

Maruyama, T, and P. A. Fuerst, 1984. Population Bottlenecks and Non-Equilibrium Models in Population Genetics. I. Allele Numbers when Populations Evolve from Zero Variability. Genetics 108: 745-763.

Maruyama, T, and P. A. Fuerst, 1985. Population Bottlenecks and Non-Equilibrium Models in Population Genetics. II. Number of Alleles in a Small Population that was

Formed by a Recent Bottleneck. Genetics 111: 675-689.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro, 1995. Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. Genetics 141: 413-429.

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 Methods for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123: 585-595.

Wall, J. D., and M. Przeworksi, 2000. When Did the Human Population Size Start Increasing? Genetics 155: 1865-1874.