Statistical Tests for Detecting Positive Selection by Utilizing High Frequency SNPs

Kai Zeng[*,†]

Suhua Shi[†]

Yunxin Fu[‡,§]

Chung-I Wu[*]

[*] Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

[†] State Key Laboratory of Biocontrol and Key Laboratory of Gene Engineering of the Ministry of Education, Sun Yat-sen University, Guangzhou 510275, China

[‡] Human Genetics Center, School of Public Health, University of Texas at Houston, Houston, Texas 77030

[§] Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming, China

Running head: positive selection

Key words: positive selection, background selection, neutrality test, demography, frequency spectrum

Corresponding authors:

Chung-I Wu, Ph.D.

1101 E 57th Street, Chicago, IL 60637, U.S.A

Tel: 1-773-702-2565

Fax: 1-773-702-9740

Email: ciwu@uchicago.edu

OR

YunXin Fu, Ph.D.

Human Genetics Center

School of Public Health

University of Texas at Houston

1200 Herman Pressler

Houston, Texas 77030

Email: yunxin.fu@uth.tmc.edu

ABSTRACT

By comparing the low-, intermediate- and high-frequency parts of the frequency spectrum, we gain information on the evolutionary forces that influence the pattern of polymorphism in population samples. We emphasize the high-frequency variants on which positive selection and negative (background) selection exhibit the primary different effects. We propose a new estimator of q (the product of effective population size and neutral mutation rate), q_L, which is sensitive to the changes in high frequency variants. The new q_L allow us to revise Fay and Wu's H test by normalization. To complement the existing statistical tests, we further propose a new test, E, which is shown to be powerful for detecting the presence of selection when the sample is taken from the recovery phase after the loss of genetic diversity due to the fixation of an advantageous allele. Extensive simulations were carried out to compare the powers of these tests for detecting positive selection. Their sensitivity to background selection and demographic changes are also compared.

Detecting the footprint of positive selection is an important task in evolutionary genetic studies. Many statistical tests have been proposed for this purpose. Some use only divergence data between species (see NEI and KUMAR 2000; and YANG 2003 for reviews) while others use both divergence and polymorphism data (for example, HUDSON *et al.* 1987; MCDONALD and KREITMAN 1991; BUSTAMANTE *et al.* 2002). Those that rely only on polymorphism data can be classified into either "haplotype" tests (for example, HUDSON *et al.* 1994; SABETI *et al.* 2002; VOIGHT *et al.* 2006) or the site-by-site "frequency spectrum" tests (for example, TAJIMA 1989; FU and LI 1993).

Frequency spectrum is the distribution of the proportion of sites where the mutant is at frequency *x* and is the focus of this study. We may divide the spectrum broadly into three parts: the low-, intermediate- and high- frequency variant classes. Comparing the three parts of the spectrum, we shall have a fuller view of the configuration of polymorphisms. Tajima's D (1989) and Fu and Li's D (1993) ask whether there are too few or many more rare variants than common ones. Fay and Wu's H (2000) takes into consideration the abundance of very high-frequency variants relative to the intermediate-frequency ones. Thus far, there is not a test which addresses the relative abundance of the very high and very low frequency classes although this comparison may be most informative about new mutations. After all, new mutations are most likely to be found in the very low class and least abundant in the very high class.

In this study, we first introduce a new estimator of θ ($\theta = 4N\mu$ where N is the

effective population size and μ is the neutral mutation rate of the gene). On the basis of this new estimator, we revise Fay and Wu's H test which is not normalized. We also introduce a new test (E) which contrasts the high- and low-frequency variants. Extensive simulations were conducted to compare the power to detect positive selection as well as sensitivity to background selection and demographic factors.

GENERAL BACKGROUND

Let ξ_i be the number of segregating sites where the mutation occurs i times in the sample. Following FU (1995), we shall refer to the class of mutations with i-occurrences as mutations of size i. Given the population mutation rate θ, FU (1995) showed that

$$E(x_i) = q / i, \quad i = 1, ..., n-1 \tag{1}$$

for a sample of size n. Since each class of mutations contains information on θ, there can be many linear functions of ξ_i which are unbiased estimators of θ, depending on how each frequency class is weighted. A general form is

$$S = \sum_{i=1}^{n-1} c_i x_i \tag{2}$$

with E(S) = θ, where c_i's are weight constants. A few well-known examples are Watterson's θ_W (1975), Tajima's θ_π (1983), Fu and Li's ξ_e (FU and LI 1993), and Fay and Wu's θ_H (FAY and WU 2000):

$$
\begin{aligned}
q_W &= \frac{1}{a_n}\sum_{i=1}^{n-1} x_i \\
q_p &= \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i(n-i)x_i \\
x_e &= x_1 \\
q_H &= \binom{n}{2}^{-1}\sum_{i=1}^{n-1} i^2 x_i
\end{aligned}
\tag{3}
$$

where $a_n = \sum_{i=1}^{n-1}(1/i)$. Theoretical variances of these estimators are given in Figure 1. Among them, θ_W has the smallest variance.

Different estimators have varying sensitivities to changes in different parts of the frequency spectrum. For example, ξ_e and θ_W are sensitive to changes in low frequency variants, θ_π to changes in intermediate frequency variants, and θ_H to high frequency ones. When the level of variation is influenced by different population genetic forces, different parts of the spectrum are affected to different extents (SLATKIN and HUDSON 1991; FU 1996; FU 1997; FAY and WU 2000; GRIFFITHS 2003). The difference between two θ estimators can thus be informative about such forces. For example, rapid population growth tends to affect low-frequency variants more than it affects high-frequency ones. As a result, θ_W tends to be larger than θ_π. The first test to take advantage of the differences between estimators is Tajima's D test (TAJIMA 1989), as shown below

$$
D = \frac{q_p - q_W}{\sqrt{Var(q_p - q_W)}} \tag{4}
$$

Many others have since been proposed (FU and LI 1993; FU 1996; FU 1997; FAY and WU 2000).

Among the evolutionary forces that may cause the frequency spectrum to deviate

from the neutral equilibrium, hitchhiking (MAYNARD SMITH and HAIGH 1974) has attracted much attention. A salient feature of positive selection is the excess of high frequency variants (FAY and WU 2000) which is the focus of our analysis.

ANALYTICAL RESULTS

Measurements of variation based on very high frequency variants – θ_H and θ_L

To capture the dynamics of high frequency variants, we need to put the most weights on high frequency variants in the estimation of θ. FAY and WU (2000) used θ_H (3). Its variance term, however, is not easy to obtain. Here we propose a new estimator, θ_L. The variance of θ_L can be easily obtained and then used to calculate the variance of θ_H.

Consider a genealogy of n genes from a non-recombining region. The mean number of mutations accumulated in each gene since the most recent common ancestor (MRCA) of the sample can be calculated as

$$q_L^{'} = \frac{1}{n}\sum_{i=1}^{n-1} i x_i \qquad (5)$$

Since $E(\xi_i) = \theta / i$, it follows that

$$E(q_L^{'}) = \frac{n-1}{n} q \qquad (6).$$

Although θ'_L is an asymptotically unbiased estimator of θ, it is more convenient to work with its canonical form

$$q_L = \frac{1}{n-1}\sum_{i=1}^{n-1} i x_i \qquad (7)$$

which has $E(\theta_L) = \theta$. It can be shown that

$$Var(q_L) = \frac{n}{2(n-1)}q + [2(\frac{n}{n-1})^2(b_{n+1}-1)-1]q^2 \quad (8)$$

where $b_n = \sum_{i=1}^{n-1}(1/i^2)$. Since $i^2 = n \cdot i - i(n-i)$, it is easy to see that $q_H = 2q_L - q_p$. From this property, we obtain

$$Var(q_H) = q + 2 \cdot \frac{36n^2(2n+1)b_{n+1} - 116n^3 + 9n^2 + 2n - 3}{9n(n-1)^2} \cdot q^2 \quad (9)$$

(see Appendix).

The normalized Fay and Wu's H statistic – contrasting high- and intermediate-frequency SNPs

Recall that Fay and Wu's H (2000) is defined as H = $\theta_\pi - \theta_H = 2(\theta_\pi - \theta_L)$. Now we can write the normalized H statistic as

$$H = \frac{q_p - q_L}{\sqrt{Var(q_p - q_L)}} \quad (10)$$

where

$$Var(q_p - q_L) = \frac{n-2}{6(n-1)}q + \frac{18n^2(3n+2)b_{n+1} - (88n^3 + 9n^2 - 13n + 6)}{9n(n-1)^2}q^2 \quad (11)$$

(see Appendix).

A new E test – contrasting high- and low- frequency SNPs

As mentioned before, there is no test contrasting the low- and high-frequency part of the spectrum. Both D and H use intermediate frequency variants as a benchmark for comparison. Nevertheless, there are occasions when it is informative to contrast high- and low-frequency variants. For example, when most variants are lost (*e.g.* due to selective sweep), the recovery of the neutral equilibrium is most rapid for the low-frequency ones and slowest for the high-frequency ones (see Figure 2B below).

Taking advantage of the newly derived θ_L, we propose a new test statistic

$$E = \frac{q_L - q_W}{\sqrt{Var(q_L - q_W)}} \tag{12}$$

where

$$\begin{aligned} &Var(q_L - q_W) \\ &= \left(\frac{n}{2(n-1)} - \frac{1}{a_n} \right) q + \left(\frac{b_n}{a_n^2} + 2\left(\frac{n}{n-1} \right)^2 b_n - \frac{2(nb_n - n + 1)}{(n-1)a_n} - \frac{3n+1}{n-1} \right) q^2 \end{aligned} \tag{13}$$

(see Appendix). Thus, the 3 tests, D, H and E contrast the 3 parts of the frequency spectrum, low-, intermediate- and high-frequency variants, in a pairwise manner. Using different combinations of these tests we may be able to capture most information contained in the frequency spectrum about the underlying evolutionary process.

APPLICATIONS

In this section, we apply the tests discussed in the previous section to data simulated under a variety of conditions, in particular, positive selection for the purpose of evaluating their statistical powers. Such simulations have been done extensively in the past for studying existing tests (BRAVERMAN *et al.* 1995; SIMONSEN *et al.* 1995; FU 1997; PRZEWORSKI 2002). We shall focus on the tests modified in this study. Since the standardized H test is always more powerful than the original one (varying from about 1% to 5%), thanks to the normalization, we will report only the results of the new H test. Details of the simulation algorithms and other related issues are described in the Appendix.

The Power to detect positive selection before and after fixation

We define $R(i) = E(x_i')/E(x_i)$, where $E(x_i')$ and $E(x_i)$ are the expected number of segregating sites of size *i* under selection and neutrality, respectively (FU 1997). Figure 2 shows the dynamics of frequency spectrum changes before and after the fixation of an advantageous mutation. The reduction in intermediate frequency variants and the accumulation of high frequency variants start at the early stage of the sweep (Figure 2A, $f = 0.3$). When the advantageous mutation reaches 60% in the population, the spectrum is already highly deviant from neutrality, losing many of the intermediate frequency variants (Figure 2A, $f = 0.6$). By the time of fixation, the population has lost virtually all the intermediate frequency variants, and mutations appear at either very high or very low frequency (Figure 2B, $\tau = 0$). In neutral equilibrium, there should be 22.4 mutations in the sample on average for $\theta = 5$ and $n = 50$. However, only 11.4 mutations are seen in the sample at $\tau = 0$ (Figure 2B), among which 57% has a frequency lower than 10%, and 31% has a frequency above 80%. Strikingly, the mean number of mutations of size 49 is 7.2 times that of the neutral expectation. When $\tau > 0$, these high frequency variants quickly drift to fixation and no longer contribute to polymorphism. For a long period after fixation, fewer than expected numbers of intermediate to high frequency mutations are observed (Figure 2B).

Figure 3 shows how the tests behave before and after the fixation of the advantageous mutation. Before fixation (Figures 3A and 3B, panels on the left), the power of D and H increases rapidly as the frequency of the advantageous mutation increases, while the E test has no power. This is expected as the reduction of moderate

frequency variants and the accumulation of high frequency variants characterize this stage (Figure 2A). Sometimes, H and D reach their peak of power before fixation. This happens when too much variation is removed by the selective sweep, as is the case in Figure 3A. After fixation (Figures 3A and 3B, panels on the right) the E test quickly becomes the most powerful test ($\tau \approx 0.1$, in Figure 3A; $\tau \approx 0.15$, in Figure 3B). From Figure 2B, we can see that after fixation the low frequency part of the spectrum recovers first, and the high frequency part returns to its equilibrium level last. The result of the E test fits this observation. Furthermore, since the recovery phase is much longer than the selective phase when selection is strong, the E test can be useful for detecting sweep.

In summary, different tests are effective in different phases of the hitchhiking process. No single test can capture all the changes in different parts of the frequency spectrum. Since H and E are mutually exclusive (no single overlapping case in all simulations), these two tests in combination appear to cover the process of selective sweep and the subsequent recovery rather nicely.

Sensitivity of the tests to other driving forces

A good test of a particular population genetic force, say, positive selection, should be sensitive to that force, and that force only. Although a test that is sensitive to many different factors can be a useful general tool, it is ultimately uninformative about the true underlying force. Hence, power can in fact be a blessing in disguise. We examine the sensitivity of these tests to forces other than positive selection below.

Background selection: Selection against linked deleterious mutations, often

referred to as background selection, can have similar effects on the level of genetic diversity as does the selective sweep (CHARLESWORTH *et al.* 1993). The distinction between selective sweep and background selection is therefore crucial. Fortunately, the two modes of selection often have very different effects on the frequency spectrum (FU 1997). For example, background selection is not likely to have any effect on high-frequency variants (FU 1997) and its effect on the low-frequency variants depends on U and N, where U is the mutation rate per diploid genome and N is the effective population size.

Table 1 summarizes the power of the tests to detect background selection. The power increases as U increases but this increase also depends on N. For a given U, the power of the tests decreases as N increases. Overall, D and E perform similarly, however, E is slightly more powerful (or sensitive). Significantly, H is not affected in all cases, and hence is discriminatory between the two modes of selection. See (FU 1997) for a more thorough discussion on background selection.

Population growth: When population increases in size, it tends to have an excess of low frequency SNPs (SLATKIN and HUDSON 1991; FU 1997; GRIFFITHS 2003). Both D and E are sensitive to this type of deviation (Figure 4). E is the most sensitive test because high frequency variants are the last to reach the new equilibrium after expansion. For the same reason, H is nearly unaffected.

Population shrinkage: When the population decreases in size, the number of low frequency SNPs tends to be smaller than that of the intermediate and high frequency SNPs (FU 1996). Thus, H can be sensitive to population shrinkage, whereas

D and E are largely unaffected (Figure 5).

Population subdivision: When the population is structured, one commonly uses the fixation index F_{ST} as a measure of genetic differences among subpopulations. In the symmetric two-deme model, $F_{ST} = 1 / (1 + 8Nm)$, where m is the fraction of new migrants in the population (NORDBORG 1997). When there is a strong population subdivision ($F_{ST} \geq 0.33$ or $4Nm \leq 1$, for example) and samples do not come evenly from all populations, the frequency spectrum is deviant from the neutral equilibrium (results not shown). Figure 6A gives an example where all samples come from one population. In that case, H is most sensitive to population subdivision. The D test is also sensitive, but only when the subdivision is very strong ($4Nm < 1$, or $F_{ST} > 0.33$). Note that, for relatively mild population subdivisions with $F_{ST} < 0.2$ (or $4Nm > 2$, as in humans or *Drosophila melanogaster*), neither is notably affected. Interestingly, E is completely insensitive to population subdivision. Figure 6B shows the effect of sampling on the power of the tests. We use $4Nm = 0.2$, as Figure 6A shows that D and H are most sensitive at this level of subdivision. It is clear that, when samples are more evenly distributed among populations, the tests become progressively less sensitive to population structure.

CONCLUSION

One important conclusion from this study is that no single statistical test is capable of detecting positive selection in all different phases, or different types of natural selection. Table 2 summaries qualitatively the power of the tests against

different driving forces. By using the three tests, D, H and E in combination, we gain a more comprehensive view of the changes in all parts of the frequency spectrum. Thus, for example, before selection is completed, D and H are effective in detecting the loss of the intermediate frequency variants but D and E, especially E, are effective in detecting the slow recovery to equilibrium after selection is completed. D is indeed a "general purpose" test as it is sensitive to many perturbations. E and H are negatively correlated and, hence, overlap with D under different conditions.

ACKNOWLEDGEMENT

We thank Drs. J. Braverman and R. R. Hudson for their helpful discussions about the simulation algorithms. Kai Zeng is supported by Sun Yat-sen University. Suhua Shi is supported by grants from the National Natural Science Foundation of China (30230030, 30470119, 30300033, 30500049). Yun-Xin Fu is supported by National Institutes of Health grants (GM 60777 and GM50428) and fund from Yunnan University, China. Chung-I Wu is supported by National Institutes of Health grants and an OOCS grant from the Chinese Academy of Sciences.

LITERATURE CITED

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783-796.

BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in arabidopsis. Nature **416:** 531-534.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619-1632.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

FU, Y. X., 1995 Statistical properties of segregating sites. Theor. Popul. Biol. **48:** 172-197.

FU, Y. X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557-570.

FU, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics **147:** 915-925.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

GRIFFITHS, R. C., 2003 The frequency spectrum of a mutation, and its age, in a general

diffusion model. Theor. Popul. Biol. **64:** 241-251.

HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp. 23-36 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARK. Sinauer, Sunderland, MA.

HUDSON, R. R., 2002 Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics **18:** 337-338.

HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (*sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329-1340.

HUDSON, R. R., and N. L. KAPLAN, 1994 Gene trees with background selection, pp. 140-153 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. GOLDING. Chapman & Hall, London.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887-899.

MARKOVTSOVA, L., P. MARJORAM and S. TAVARÉ 2001 On a test of Depaulis and Veuille. Mol. Biol. Evol. **18:** 1132-1133.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23-35.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *adh* locus in drosophila. Nature **351:** 652-654.

NEI, M., and S. KUMAR, 2000 *Molecular Evolution and Phylogenetics.* Oxford University Press, New York.

NORDBORG, M., 1997 Structured coalescent processes on different time scales. Genetics **146:** 1501-1514.

PRZEWORSKI, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179-1189.

SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419:** 832-837.

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.

STEPHAN, W., T. H. E. WIEHE and M. W. LENZ, 1992 The effect of strongly selected substitutions on neutral polymorphism - analytical results based on diffusion-theory. Theor. Popul. Biol. **41:** 237-254.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437-460.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent

positive selection in the human genome. PLoS Biol. **4:** e72.

WALL, J. D., and R. R. HUDSON, 2001 Coalescent simulations and statistical tests of neutrality. Mol. Biol. Evol. **18:** 1134-1135.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276.

YANG, Z., 2003 Adaptive molecular evolution, pp. 229-254 in *Handbook of Statistical Genetics,* edited by D. BALDING, M. BISHOP and C. CANNINGS. Wiley, New York.

APPENDIX

Analytical results

We list some useful properties of the θ estimators. See Y. FU (in preparation) for mathematical details.

$$\begin{aligned} Cov(q_p, q_L) &= \frac{n+1}{3(n-1)}q + \frac{7n^2+3n-2-4n(n+1)b_{n+1}}{2(n-1)^2}q^2 \\ Cov(q_L, q_W) &= \frac{q}{a_n} + \frac{(nb_n - n + 1)}{a_n(n-1)}q^2 \end{aligned} \quad \text{(A1)}$$

Since $q_H = 2q_L - q_p$, using (A1), we have

$$Var(q_H) = q + 2 \cdot \frac{36n^2(2n+1)b_{n+1} - 116n^3 + 9n^2 + 2n - 3}{9n(n-1)^2} \cdot q^2 \quad \text{(A2)}$$

$Var(q_q - q_L)$ and $Var(q_L - q_W)$ are calculated in a similar manner.

Simulation algorithms

Positive selection: We used the coalescent process with a selective phase (KAPLAN *et al.* 1989; STEPHAN *et al.* 1992; BRAVERMAN *et al.* 1995) to simulate random samples and estimate the power of various tests. The model assumes selection at the favored locus is additive and the selective pressure is strong ($2Ns >> 1$). In this case, the trajectory of the frequency of the advantageous allele in the interval with endpoints ε and $1 - \varepsilon$ is deterministic (STEPHAN *et al.* 1992). The choice of ε has little effect on the results (BRAVERMAN *et al.* 1995). We used $\varepsilon = 1 / (2N)$ in this study. Our implementation of the algorithm follows the description in (BRAVERMAN *et al.* 1995), except for the following features: 1), we use a given value of the population mutation rate ($\theta = 4Nm$), rather than a fixed number of segregating sites; 2) we allow intragenic recombination within the neutral locus to

happen through the whole process (*i.e.*, during both neutral and selective phase); 3), the genetic distance between the selective locus and the neutral locus (c_{bet}), selection coefficient (s) and time (τ, measured in units of $4N$ generations) at which an advantageous mutation arises are given values rather than random variables.

We extended the coalescent simulation algorithm described above to incorporate the case of incomplete hitchhiking. Suppose that the frequency of the favored allele is f. By solving the deterministic equation given by STEPHAN *et al.* (1992), one may obtain the time since the birth of this advantageous mutation and the corresponding frequency trajectory. In each replicate of simulation, we assign the number of genes linked to the advantageous allele by a binomial sampling with parameters n and f. Then the process, going backward in time, follows the same procedure described above.

Background selection and demographic scenarios: We simulated background selection using coalescent method (HUDSON and KAPLAN 1994; CHARLESWORTH *et al.* 1995). The software package kindly provided by R. HUDSON (2002) was used to simulate all demographic scenarios.

Critical values of the tests

The critical values of a test were determined from its empirical distribution obtained by 100,000 rounds of coalescent simulations with no recombination and $q = \hat{\theta}_W$ (FU and LI 1993; FU 1997). Another method to determine critical values was proposed by HUDSON (1993). The differences between these two methods have been discussed extensively (MARKOVTSOVA *et al.* 2001; WALL and HUDSON 2001).

However, we found that tests are consistently more powerful if critical values are determined by the former method. Therefore, we only report those generated by the first method. All tests were done at a 5% lower-tail level in this study.

Table 1. Powers of D, E and H in detecting background selection.

U	N	q	D	E	H
0.01	2500	1	0.06	0.07	0.06
		10	0.06	0.06	0.05
	50000	1	0.05	0.06	0.05
		10	0.05	0.05	0.05
0.1	2500	1	0.09	0.10	0.04
		10	0.34	0.34	0.03
	5000	1	0.06	0.06	0.05
		10	0.23	0.25	0.03
	10000	1	0.03	0.03	0.05
		10	0.16	0.18	0.04
	20000	1	0.02	0.02	0.07
		10	0.10	0.12	0.04
	50000	1	0.02	0.02	0.07
		10	0.07	0.09	0.05
0.2	2500	1	0.10	0.10	0.03
		10	0.60	0.61	0.02
	5000	1	0.07	0.07	0.03
		10	0.48	0.50	0.03
	10000	1	0.04	0.04	0.03

		10	0.35	0.35	0.03
	20000	1	0.02	0.02	0.03
		10	0.22	0.22	0.03
	50000	1	0.01	0.01	0.04
		10	0.10	0.10	0.04

10, 000 samples were simulated for each parameter set.

U is the mutation rate per diploid genome.

N is the effective population size.

Table 2. A qualitative summary of the powers of the 3 tests to detect various population genetic forces.

Driving force	D	H	E
Positive selection (before fixation)	+	+	-
Positive selection (after fixation)	+	-	+
Background selection	+	-	+
Population growth	+	-	+
Population shrinkage	-	+	-
Subdivision	+	+	-

- denotes the lack of power under the specified condition.

+ denotes moderate to strong power.

FIGURE LEGENDS

Figure 1. Variance of the five estimators of θ. Sample size (n) is 50.

Figure 2. (A) Changes in $R(i) = E(x_i')/E(x_i)$ as the advantageous mutation increases in frequency (f). (B) Changes in $R(i)$ at different times after fixation of the advantageous mutation. Time (τ) is measured in units of $4N$ generations. In all simulations, the parameters are defined as follows: $\theta = 4N\mu$ where μ is the mutation rate for the linked neutral locus; s is the selective coefficient of the advantageous mutation and c is the recombination rate (between the neutral variation under investigation and the advantageous mutation nearby), which is usually scaled by the selective coefficient. In Figure 2, the parameter values are $\theta = 5$, $s = 0.001$, $c/s = 0.02$, and sample size (n) is 50. In the simulation for hitchhiking, we also incorporated intragenic recombination among the neutral variants under investigation. The intragenic recombination rate of the neutral locus, multiplied by $4N$, is 25 in Figure 2 and 3. Intragenic recombination in other cases has negligible effect on the results and was not incorporated.

Figure 3. Power of the tests before and after hitchhiking is completed. The x-axis on the left panel represents the increase in the frequency of the advantageous mutation; on the right panel is the time after fixation. (All parameter values, unless specified, are the same as those of Figure 2.) (A) $c/s = 2\times10^{-5}$; (B) $c/s = 0.02$.

Figure 4. Sensitivity (or power) of the tests to population expansion. We assume that the effective population size increases 10 times instantaneously at time 0 to $\theta = 5$. Sample size (n) is 50.

Figure 5. Sensitivity (or power) of the tests to population shrinkage. We assume that the effective population size decreases 10 times instantaneously at time 0 to $\theta = 2$. Sample size (n) is 50.

Figure 6. Sensitivity (or power) of the tests to population subdivision. A two-deme model with $\theta = 2$ per deme was simulated. Populations are assumed to be in drift-migration equilibrium with symmetric migration at a rate of m, which is the fraction of new migrants each generation. Sample size (n) is 50. (A) Sensitivity as a function of the degree of population subdivision, expressed as $4Nm$ on the x-axis. All genes were sampled from one subpopulation. (B) Sensitivity as a function of the sampling skewness; for example, 5/45 means 5 genes are sampled from one subpopulation and 45 from the other. In this case, $4Nm = 0.2$, a value at which the tests are most sensitive to population subdivision.

Figure 1.

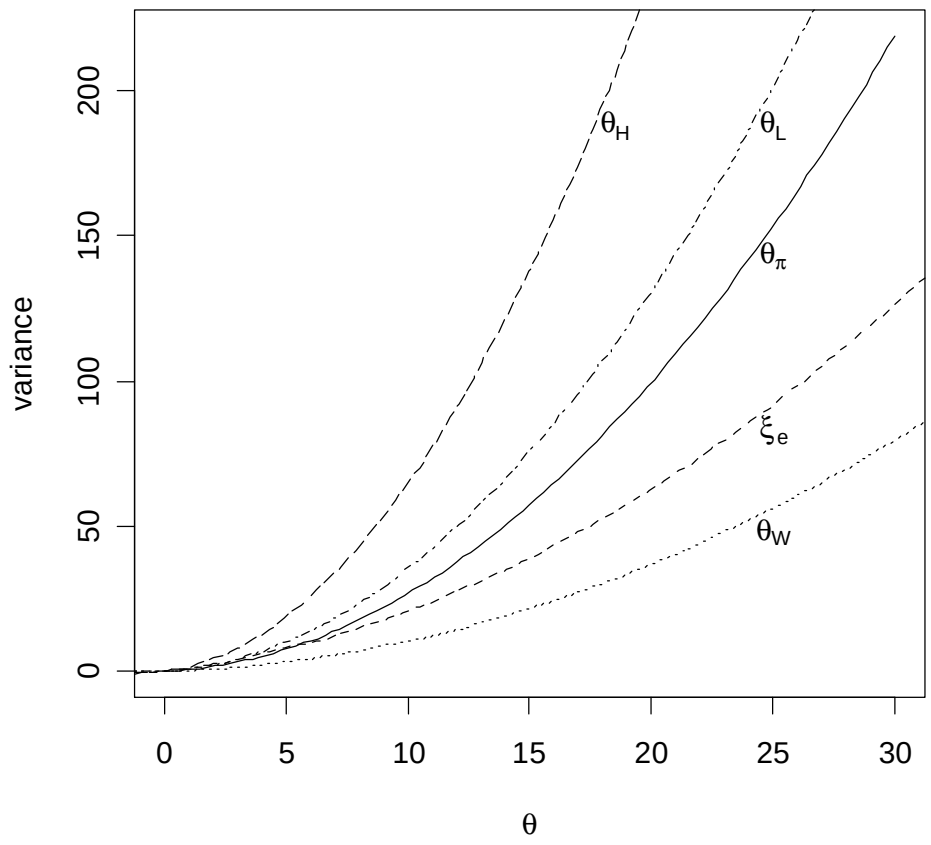

Figure 2A

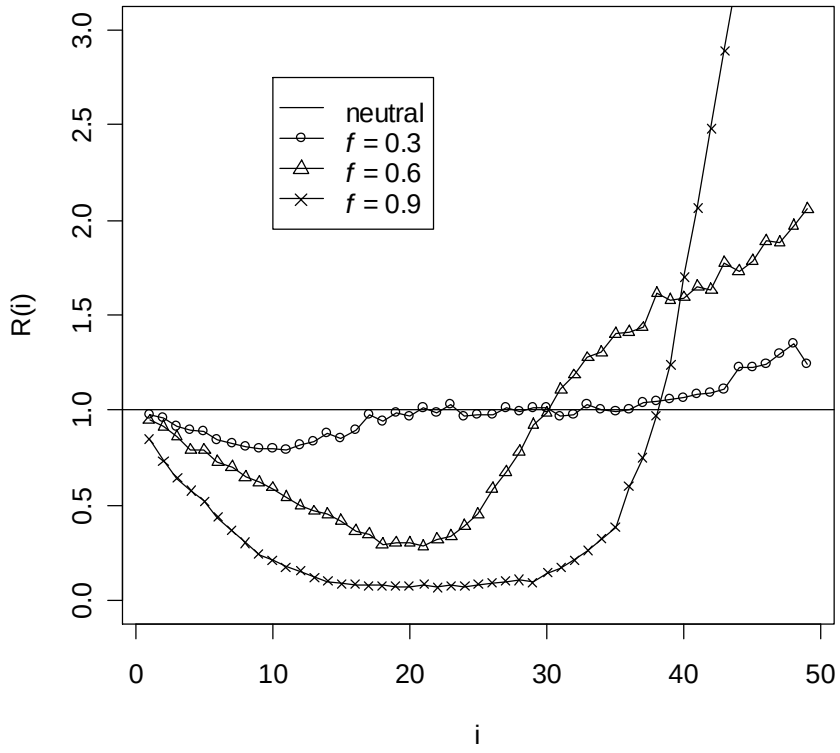

Figure 2B

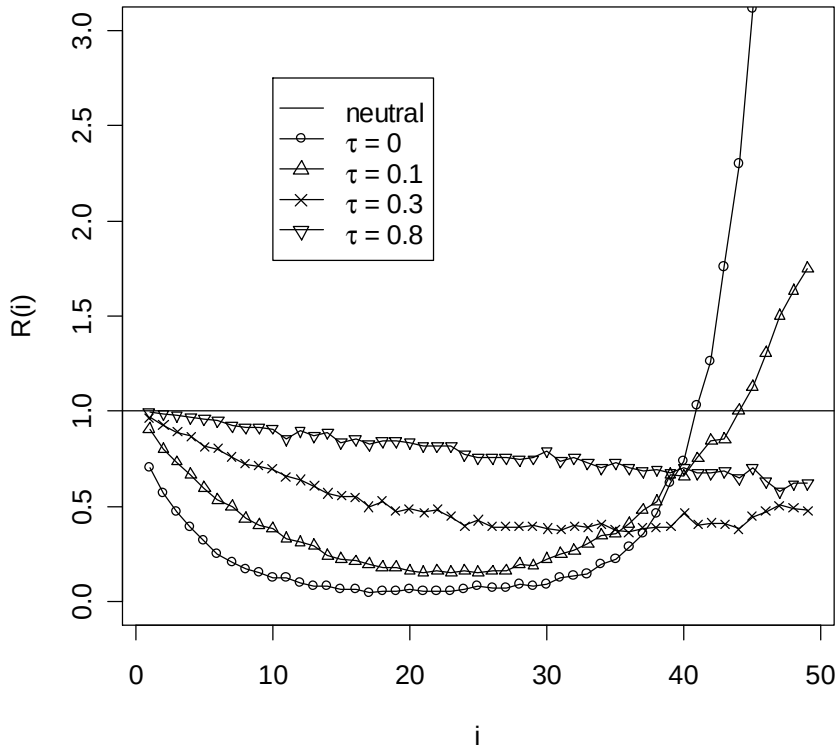

Figure 3A

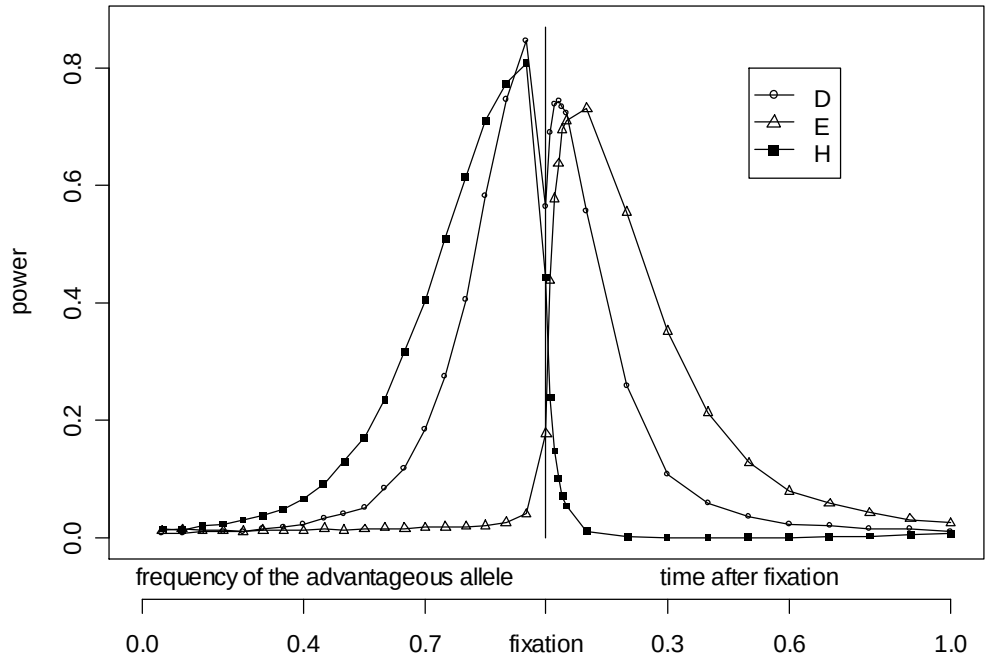

Figure 3B

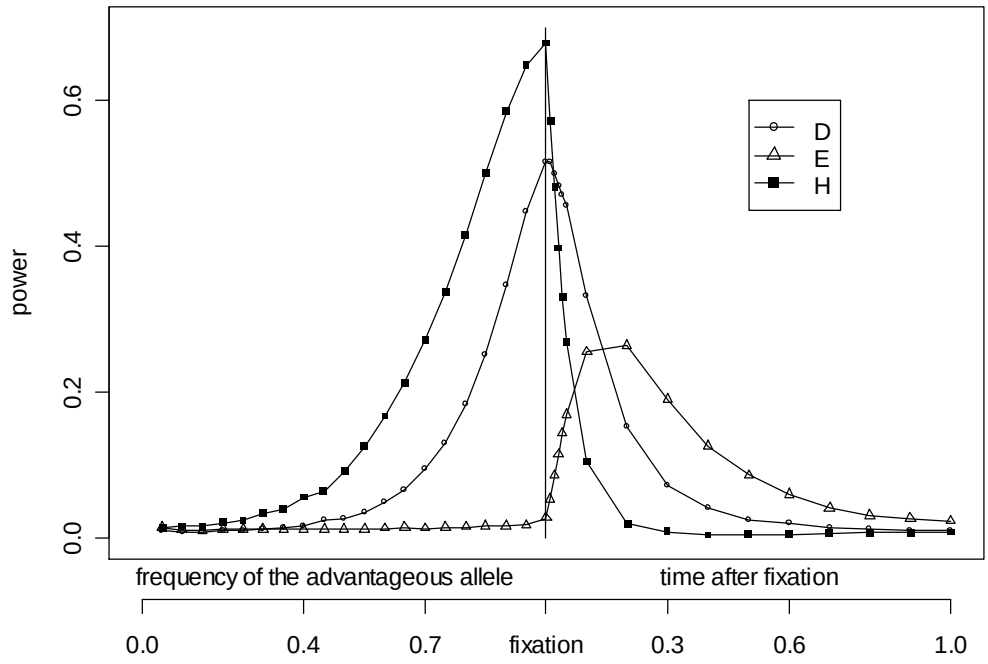

Figure 4

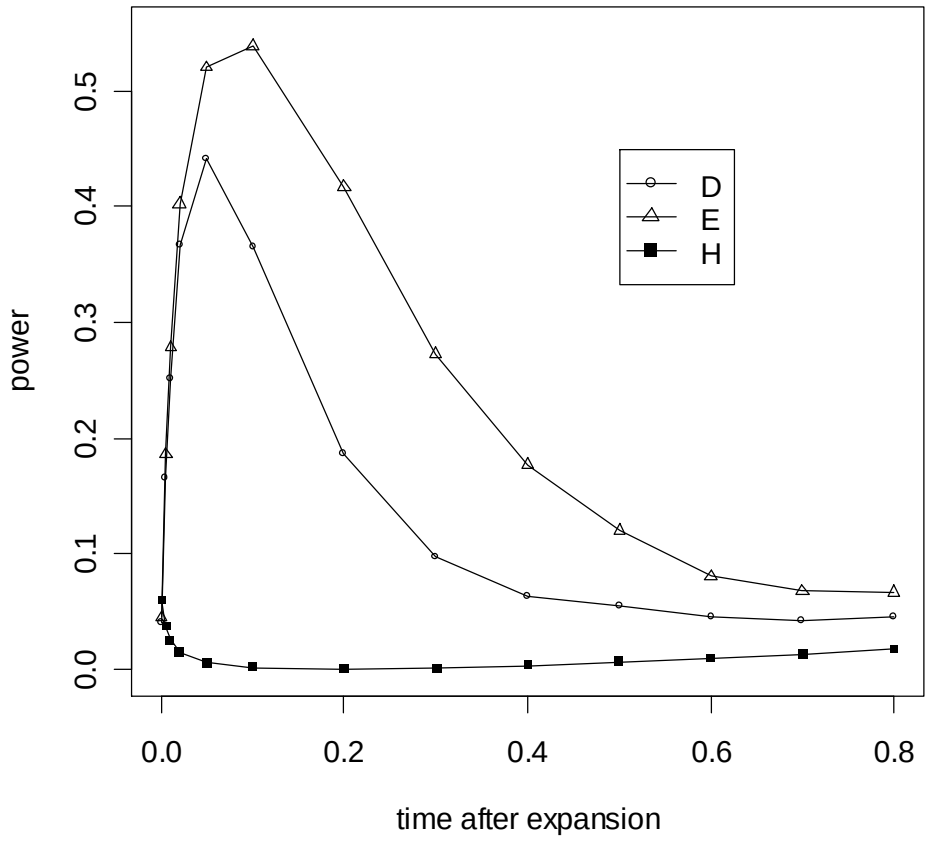

Figure 5

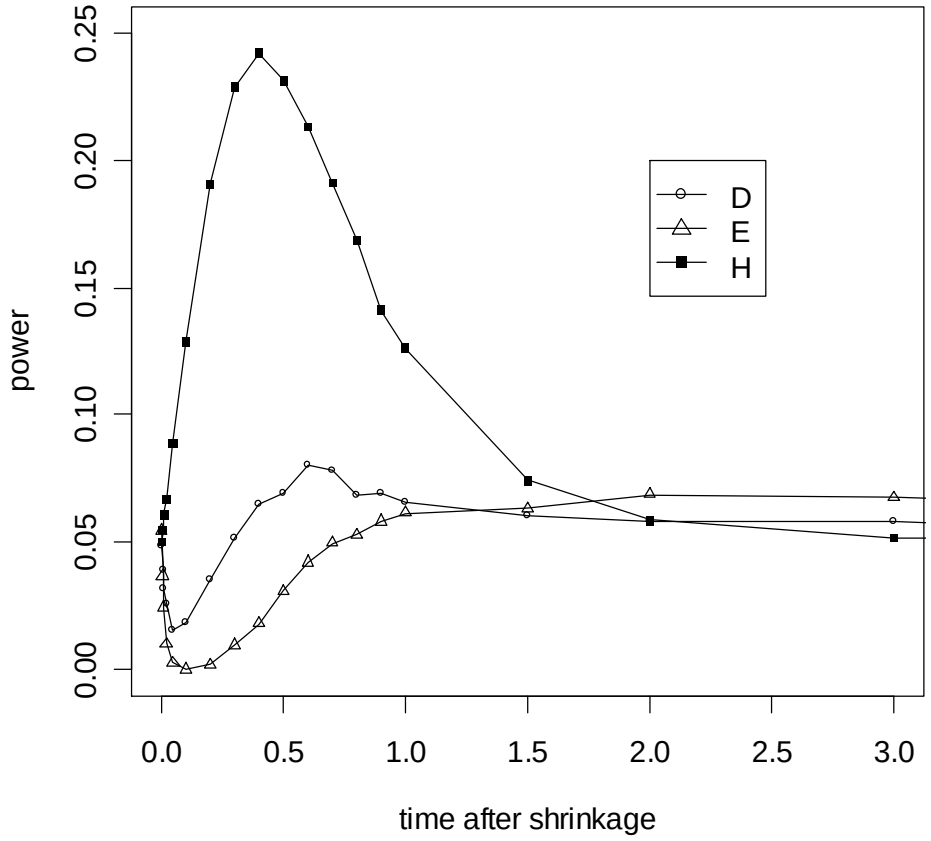
power
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
time after shrinkage
D
E
H

Figure 6A

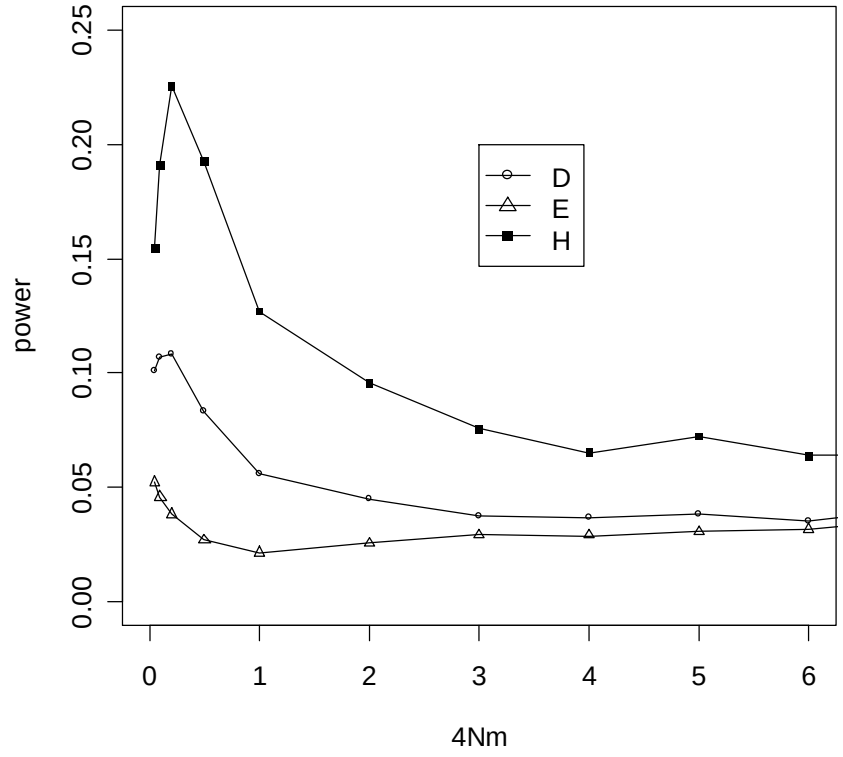

Figure 6B