

# Distinguishing Between Selection and Population Expansion in an Experimental Lineage of Bacteriophage T7

Matthew W. Hahn,<sup>1</sup> Mark D. Rausher and Clifford W. Cunningham

*Evolution, Ecology, and Organismal Biology Group, Department of Biology, Duke University, Durham, North Carolina 27708*

Manuscript received September 21, 2001

Accepted for publication February 5, 2002

## ABSTRACT

Experimental evolution of short-lived organisms offers the opportunity to study the dynamics of polymorphism over time in a controlled environment. Here, we characterize DNA polymorphism data over time for four genes in bacteriophage T7. Our experiment ran for 2500 generations and populations were sampled after 500, 2000, and 2500 generations. We detect positive selection, purifying (“negative”) selection, and population expansion in our experiment. We also present a statistical test that is able to distinguish demographic from selective events, processes that are hard to identify individually because both often produce an excess of rare mutations. Our “heterogeneity test” modifies common statistics measuring the frequency spectrum of polymorphism (*e.g.*, Fu and Li’s *D*) by looking for processes producing different patterns on nonsynonymous and synonymous mutations. Test results agree with the known conditions of the experiment, and we are therefore confident that this test offers a tool to evaluate natural populations. Our results suggest that instances of segregating deleterious mutations may be common, but as yet undetected, in nature.

UNCOVERING the processes that generate the patterns of variation we observe in nature is one of the major objectives for evolutionary biology. Selective, demographic, and random processes can all play important parts in shaping patterns of DNA sequence polymorphism. A number of statistical tests have been developed to detect the effects of some of these processes from a sample of DNA sequences (*e.g.*, HUDSON *et al.* 1987; TAJIMA 1989; McDONALD and KREITMAN 1991; FU and LI 1993). As more genes are sequenced and analyzed in this population genetic framework, it is necessary to distinguish exactly which evolutionary forces our statistical tools detect. The interplay among these processes in natural populations means that no one process will be responsible for all the genetic variation we observe, but it will be useful to evolutionary studies to determine those with the most significant effect.

Many different processes can produce similar patterns of DNA sequence polymorphism. Many currently used statistical tools can detect deviations from neutrality or from a Wright-Fisher population model, but cannot distinguish between alternative mechanisms that may cause these deviations. For example, the ability to distinguish between the reduced genetic variability at loci linked to selectively favored alleles (“hitchhiking”; MAYNARD SMITH and HAIGH 1974) and loci linked to deleterious alleles maintained by mutation (“background selection”;

CHARLESWORTH *et al.* 1993) in regions of low recombination has been a point of contention in population genetics for close to 10 years (*e.g.*, BEGUN and AQUADRO 1992; LANGLEY *et al.* 1993; CHARLESWORTH *et al.* 1993, 1995; HUDSON and KAPLAN 1995; STEPHAN *et al.* 1998; FAY and WU 2000).

Some popular tests of the neutrality of mutations (TAJIMA 1989; FU and LI 1993; FU 1996) are unable to distinguish between true selective departures from neutrality and demographic processes, such as population expansion, that may produce similar effects. Purifying selection, population expansion, and selective sweeps can all produce an excess of rare alleles, *i.e.*, alleles at low frequencies. Statistical tests that look at the frequency spectrum of alleles to detect departures from neutrality, *e.g.*, TAJIMA’s *D* (1989), FU and LI’s *D*, *D\**, *F*, and *F\** (1993), and FU’s *F<sub>s</sub>* (1996), can be significant under purifying selection, population expansion, or selective sweeps, although each statistic may be best at detecting one of these forces (BRAVERMAN *et al.* 1995; SIMONSEN *et al.* 1995; FU 1997; FU and LI 1999).

One major class of effects that leads to an excess of rare alleles may be termed homogeneous effects because different types of mutations (nonsynonymous and synonymous) at a locus are affected equally. Both population expansion after a bottleneck and the sweep to fixation of a favored allele have homogeneous effects across a locus; the new mutations that arise in the population are both nonsynonymous and synonymous. While these two homogeneous processes have similar effects on mutations at a single locus, they have different effects across multiple loci (*e.g.*, GALTIER *et al.* 2000). Demographic processes such as population expansion will cause an

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AF419412–AF419511.

<sup>1</sup>Corresponding author: Department of Biology, Box 90338, Duke University, Durham, NC 27708. E-mail: mwh3@duke.edu

increase of rare alleles at all loci in a genome. By contrast, selective sweeps will normally affect at most only a few loci, and with recombination these effects may be limited to one locus.

The other major class of effects may be termed heterogeneous because nonsynonymous and synonymous mutations are affected unequally. Purifying selection acts to eliminate deleterious nonsynonymous mutations, but is presumed to have little or no effect on neutrally evolving synonymous mutations (but see, for example, AKASHI 1995). VERRELLI and EANES (2000) take advantage of heterogeneous effects on *Drosophila simulans* DNA to detect differential effects on nonsynonymous and synonymous changes. Looking at the frequency spectrum of polymorphism for the entire Pgm gene they do not detect a significant excess of low-frequency mutations. Examining solely nonsynonymous mutations, however, they find a significant excess of low-frequency polymorphism consistent with segregating deleterious mutations and purifying selection acting on nonsynonymous mutations.

In this article we show that processes that act on a population with similar effects can be distinguished from one another and that we can also detect multiple processes acting at the same time. We use a simple method of modifying statistical tests that look at the frequency spectrum of polymorphism to distinguish among different effects selective and demographic forces have on DNA sequence polymorphism. We use an experimental system of bacteriophage T7 virus in which initial conditions, demographic patterns, and selective environment are all known and/or standardized. While computer modeling of DNA sequence evolution also offers the ability to test the limits of statistical tools, untested assumptions and unconsidered processes are not included. Unlike two earlier sets of experimental lineages of T7 (e.g., HILLIS *et al.* 1992; CUNNINGHAM *et al.* 1997), which periodically bottlenecked populations to a single individual, our lineage was propagated with a minimum amount of bottlenecking. This should reduce the amount of genetic drift and increase the efficacy of selection. Other advantages of this experimental system are high rates of molecular evolution in the virus, a known ancestral genotype, and high levels of recombination.

## MATERIALS AND METHODS

**Bacteriophage T7:** Starting from a single plaque (designated WT), a population of T7 was propagated for 500 lytic cycles,  $\sim 2500$  generations, by C. W. Cunningham and J. J. Bull at the University of Texas, Austin (Figure 1). At each lytic transfer  $\sim 2 \mu\text{l}$  of the 2-ml lytic culture of viruses ( $\sim 10^5$  individuals) was passaged to the next tube, and at no point was the lineage bottlenecked to a single individual. The lineage was sampled at three time points named for the age of the lineage in number of lytic cycles (one lytic cycle is  $\sim 5$  generations; J. J. BULL, personal communication): populations CW100, CW400, and CW500. The genes sequenced (identified in DUNN and STUDIER

1983) from the single ancestral plaque and the descendant populations were the first 285 bp of 0.3, which inactivates host restriction (the rest of the gene was truncated by a deletion event, as in CUNNINGHAM *et al.* 1997); 17.0 (1662 bp), a tail fiber protein; 17.5 (204 bp), which is associated with lysis; and 18.0 (270 bp), a DNA maturation protein. Sequences are GenBank nos. AF419412–AF419511. T7 was grown in 2-ml cultures of *Escherichia coli* strain W3110 in the presence of the mutagen *N*-methyl-*N'*-nitro-*N*-nitrosoguanidine (20  $\mu\text{g}/\mu\text{l}$ ). See HILLIS *et al.* (1992) and BULL *et al.* (1993) for further details on the growth and maintenance of the T7 phage.

**Modification of current tests:** As an exemplar, we illustrate our method using Fu and Li's *D* test (1993), but it is important to point out that the same idea applies to Fu and Li's *F*, *D*<sup>\*</sup>, *F*<sup>\*</sup> (1993), and TAJIMA's *D* tests (1989). The main idea behind our method is that under selective neutrality of polymorphism the distribution of nonsynonymous and synonymous mutations should be proportional across a genealogy. In terms of Fu and Li's *D* test, this means that the ratio of nonsynonymous mutations on internal branches of a genealogy to those on external branches should equal the ratio of synonymous mutations on internal to external branches. As a result of homogeneous processes, such as population expansion or a selective sweep, there is an excess of mutations on external branches (i.e., rare alleles) but this affects both nonsynonymous and synonymous mutations equally. Purifying selection and any resulting segregating deleterious mutations have heterogeneous effects across a locus; nonsynonymous and synonymous mutations are not affected equally, so the distribution of mutations is disproportional. Under purifying selection there will be an excess of mutations on external branches, but nonsynonymous mutations will be disproportionately represented because they are being actively selected against and thus kept at low frequencies. To test for heterogeneous effects, therefore, we calculated *D* for two sets of data: nonsynonymous and synonymous mutations.

Our procedure (heterogeneity test) for testing for differences in Fu and Li's *D* between synonymous and nonsynonymous mutations was relatively simple. First we calculated, for each gene, *D* and  $\theta_w$  (the population mutation parameter,  $2N_e\mu$ , based on the number of segregating sites; WATTERSON 1975) separately for the nonsynonymous and synonymous data sets, and then we calculated  $\Delta D$  (synonymous *D* – nonsynonymous *D*). Using a PERL version of the make\_tree program of HUDSON (1990; available upon request or on the Web at <http://www.duke.edu/~mwh3>), we conducted Monte Carlo coalescent simulations of 10,000 gene genealogies with no recombination; the assumption of no recombination makes our test conservative. Each of the 10,000 genealogies was simulated with the values of both synonymous and nonsynonymous  $\theta_w$ . For each tree the value of Fu and Li's *D* was then calculated for both synonymous and nonsynonymous mutations and the difference,  $\Delta D$ , was recorded. This distribution of the values of  $\Delta D$  was then used to calculate the probability, *P*, of observing a difference in *D* values between synonymous and nonsynonymous mutations as great or greater than the observed difference. A one-tailed test is used because we have an *a priori* expectation that *D* for nonsynonymous mutations will be more negative due to segregating deleterious mutations. This program can also be used on Fu and Li's *F*, *D*<sup>\*</sup>, *F*<sup>\*</sup>, and Tajima's *D* statistics.

**Data analysis:** Sequences used in this study were visually aligned; there were no gaps in any of the aligned sequences we used. Calculations of Fu and Li's *D*,  $\pi$  (the average number of pairwise nucleotide differences per site; TAJIMA 1983),  $\pi_n/\pi_s$  (the ratio of pairwise nonsynonymous and synonymous differences per site), and  $\theta_w$  were done using DNAsp 3.5 (ROZAS and ROZAS 1999). The outgroup used for the calculation of *D* was the known ancestral sequence (WT).

The population recombination parameter,  $\gamma$  ( $2N_c c$ ), was analyzed using SITES (HEY and WAKELEY 1997). This is used because HUDSON's  $C$  (1987) is unreliable for small sample sizes (HEY and WAKELEY 1997; HUDSON 1987). For some populations SITES (HEY and WAKELEY 1997) cannot calculate  $\gamma$ , and in these cases  $C$  is used; these are marked by a superscript  $a$  in Table 2. SITES cannot generate an estimate of  $\gamma$  for some data sets either because they have too few informative sites that are shared in subsets of four lines or because of the spacing of those sites with regard to whether or not they show evidence of recombination (HEY and WAKELEY 1997; J. HEY, personal communication). Estimates of  $C$  are almost always  $>\gamma$  because of error involved in calculating  $C$  from small sample sizes.

MCDONALD and KREITMAN (1991) suggested a comparison of the ratio of polymorphism to fixed differences of both synonymous and nonsynonymous mutations as a statistical test for evaluating the role of natural selection in causing substitutions in protein-coding genes. This test suggests the action of positive selection when there is a relative excess of nonsynonymous fixed differences (MCDONALD and KREITMAN 1991). We performed the McDonald and Kreitman (M-K) test using Fisher's exact test to evaluate significance; fixed differences were calculated between the WT ancestral sequence and the evolved populations.

## RESULTS

**Molecular evolution of bacteriophage T7:** A total of four complete genes were sequenced in each of three populations and the single ancestor. Almost 2.5 kb was obtained from each of 25 individuals in our experiment. In addition to the WT ancestor, we sequenced 9 individuals from the CW100 population, 6 from CW400, and 9 from CW500 (Figure 1). All four genes show considerable amounts of nucleotide variation with differences present between genes. There was a general pattern of change in the amount of variation over time, with both  $\pi$  and  $\theta_w$  increasing as more polymorphisms appeared in the majority of sequences (Table 1). As expected, the number of fixed differences also increased over time. It should be noted that we have observed a few instances where a mutation that was counted as fixed in one population was found to be polymorphic hundreds of generations later in another population. We suggest that this is because our sampling of alleles was not extensive enough to uncover all polymorphisms in the population. We still counted differences as fixed, however, even if they were revealed to be polymorphic in a later population. We do this for consistency and also because we could not rule out the possibility of a back mutation. Tests performed that use the number of fixed differences in their calculations should not be biased by this method, as synonymous and nonsynonymous differences are affected equally by the sampling strategy.

**Recombination:** The population recombination parameter is an important underlying factor that may affect patterns of polymorphism. This quantity is given per site per generation for each locus separately in each of the three populations. The results of our analysis are presented in Table 2. There is a pattern of increase in

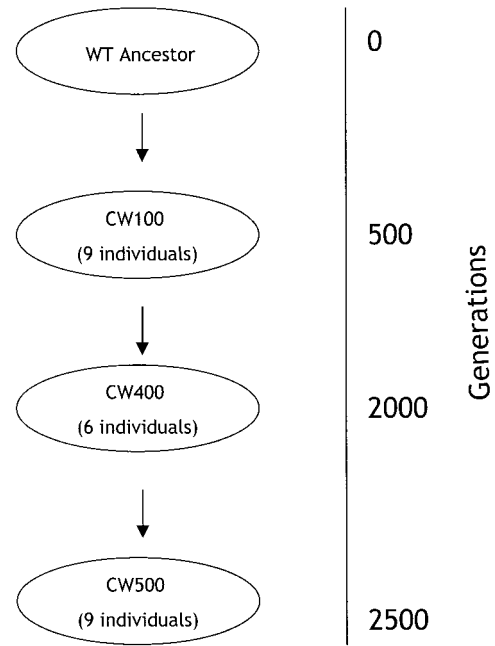


FIGURE 1.—Experimental setup and sampling design. The single ancestral plaque of bacteriophage T7, designated “WT,” was evolved for 2500 generations. The large population of phage was sampled at three points for sequencing; these sample populations are designated CW100, CW400, and CW500. The number of individuals sampled in each population is given in parentheses.

recombination rate over time among the four genes. We believe that this is an artifact of the estimation methods and that the recombination rate is actually constant over time. The pattern of increasing recombination rate may be due to two distinct possibilities. First, the small number of polymorphisms available just after a population expansion may make it harder to detect recombination events and will lead to smaller estimates of  $\gamma$  and  $C$ . Second, the effect of changing population size or the nonneutrality of mutations will violate the assumptions made in the estimation of recombination parameters and the results may be sensitive to these violations (HEY and WAKELEY 1997). The true value of the population recombination rate, for these reasons, is likely closest to that given by population CW500 ( $\gamma = 0.127\text{--}0.440$ ). The double-stranded linear genome of bacteriophage T7 is only 39,937 bp (DUNN and STUDIER 1983) and is inserted into the *Escherichia coli* host and replicated by the host's own cellular machinery. In our experimental evolution there are a large number of phage relative to *E. coli* and, therefore, a high number of phage infecting a single host (“multiplicity of infection”). Because of this high multiplicity of infection and the nature of the T7 genome, there is a large amount of recombination in T7 between multiple naked genomes floating free in the cytosol. The population recombination rate calculated for the four genes in the CW500 population

**TABLE 1**  
Nucleotide variation

Population	Gene 0.3	Gene 17.0	Gene 17.5	Gene 18.0
CW100	$\pi = 0.020$ $\theta_w = 0.026$	0.015 0.020	0.022 0.023	0.007 0.011
CW400	$\pi = 0.053$ $\theta_w = 0.055$	0.036 0.037	0.018 0.017	0.019 0.018
CW500	$\pi = 0.057$ $\theta_w = 0.058$	0.035 0.040	0.019 0.020	0.026 0.027

Standard measures of nucleotide diversity are calculated by DNAsp 3.5 (ROZAS and ROZAS 1999), where  $\pi$  is taken from the average number of pairwise nucleotide differences per site (TAJIMA 1983) and  $\theta_w$  is based on the number of segregating mutations per site (WATTERSON 1975).

varies from two to six times the highest rate found for a *Drosophila* gene (Table 2; SCHAEFFER and MILLER 1993; HEY and WAKELEY 1997).

**Population expansion in population CW100 and constraint in all populations:** In our experimental system, we know that there was a population expansion after the single WT ancestor was allowed to multiply. Therefore, we can test the sensitivity of Fu and Li's  $D$  statistic to detect this event. Three genes out of four sequenced (0.3, 17.0, 18.0) show significantly negative  $D$  values in the CW100 population (Table 3), indicating the presence of an excess of rare alleles inconsistent with neutral processes in a stable population, but consistent with either demographic or selective processes. Because the CW100 population was sampled not long after the single WT ancestral plaque was propagated in cultures of *E. coli* (Figure 1), this pattern is most likely produced by the dramatic population expansion. The likelihood of the pattern having been generated by selective processes is also low because the effect was seen at multiple loci and because polymorphism is recovered over time from the monomorphic ancestor (Table 1).

To further test the hypothesis that the  $D$  values were produced by a homogeneous process such as population expansion, which would affect all mutations equally, we partitioned the data for each gene in population CW100 into nonsynonymous and synonymous data sets.  $D$  was negative in value for both classes of data for all genes (Table 4). For genes 0.3, 17.0, and 18.0 the synonymous data sets are all significantly negative (Table 4). For genes 17.0 and 18.0 the nonsynonymous data is also significant. It is not certain why gene 17.5 shows little pattern of population expansion, but some possibilities are advanced in the DISCUSSION. These negative values of  $D$  are generally consistent with a homogeneous process such as population expansion even though the values were significant for only three genes in the synonymous data set and for two genes in the nonsynonymous

**TABLE 2**  
Recombination

Population	Gene 0.3	Gene 17.0	Gene 17.5	Gene 18.0
CW100	0.045	0.031	0.312 <sup>a</sup>	0.048 <sup>a</sup>
CW400	0.162	0.099	0.188 <sup>a</sup>	0.036
CW500	0.214	0.127	0.440	0.280

The population recombination parameter is calculated as  $\gamma$  using the program SITES (HEY and WAKELEY 1997). For some populations SITES cannot calculate  $\gamma$ , and in these cases Hudson's  $C$  (HUDSON 1987) is used. Estimates of  $C$  are almost always  $> \gamma$  would be because of error involved in calculating it from small sample sizes (HUDSON 1987; HEY and WAKELEY 1997).

<sup>a</sup> Denotes value of  $C$ , not  $\gamma$ .

data set. Our statistical test for heterogeneity fails to find differences in the magnitude of  $D$  between nonsynonymous and synonymous mutations for any of the four genes in population CW100 (Table 4), consistent with the action of a homogeneous process.

Although population expansion, most likely, has created an excess of low-frequency mutations for all mutations in population CW100, constraint on nonsynonymous mutations is still evident. Table 5 shows that the values of  $\pi_a/\pi_s$  (the ratio of pairwise nonsynonymous to synonymous differences per site) are  $\ll 1$  for every gene. This pattern reveals the action of purifying selection constraining the available neutral or advantageous mutations at nonsynonymous sites (HUGHES and NEI 1988). This effect of purifying selection can be seen for every gene in every population (Table 5).

**Positive selection and weakly deleterious mutants in populations CW400 and CW500:** Positive selection is apparent in two genes in our experiment. A comparison of the number of nonsynonymous and synonymous polymorphisms and fixed differences allows a test of the neutral mutation hypothesis (MCDONALD and KREITMAN 1991). In the results shown in Table 6, every population's polymorphism is compared to the number of fixed differences from the WT ancestor. Genes 17.0 and 17.5 reject the null model of the M-K test in both the CW400 and CW500 populations. They both appear to have too many nonsynonymous fixed differences relative to synonymous fixed differences, suggesting positive selection. Neither gene rejects the null hypothesis of the M-K test in population CW100, however. We suggest that this is because there has not been time for a large enough number of advantageous mutations to fix. The 5 nonsynonymous mutations that are eventually fixed in gene 17.5 are all present as polymorphisms in the CW100 population; 7 of the 10 nonsynonymous mutations that are fixed in gene 17.0 are also present as polymorphisms or already fixed. Although a short time to fixation is predicted for advantageous mutations (KIMURA 1983),

**TABLE 3**  
**Fu and Li's  $D$  for all mutations**

Population	Gene 0.3	Gene 17.0	Gene 17.5	Gene 18.0
CW100	-1.93 ( $P = 0.014$ )	-2.22 ( $P = 0.005$ )	-0.17	-2.17 ( $P = 0.036$ )
CW400	-0.28	-0.05	0.08	-0.51
CW500	-0.21	-0.06	0.18	-0.14

The values for Fu and Li's (1993)  $D$  statistic are shown for all mutations combined.  $D$  is calculated using the WT ancestor as the outgroup by the program DNAsp 3.5 (ROZAS and ROZAS 1999). Negative values indicate an excess of low-frequency mutations.

there are only 500 generations between the start of this experiment and population CW100.

Although the homogeneous effects of population expansion are no longer detectable by Fu and Li's  $D$  in any gene for populations CW400 and CW500, our statistical test of heterogeneity suggests the action of a heterogeneous process—most likely purifying selection on weakly deleterious mutants (Table 4). Values for Fu and Li's  $D$  statistic computed over whole genes are not significant in any gene for populations CW400 and CW500 (Table 3). However, when the four genes are divided into nonsynonymous and synonymous data sets there are skewed (excess of low-frequency mutants) values for nonsynonymous mutations only, as expected from purifying selection. Table 4 shows values for  $D$  for both nonsynonymous and synonymous mutations, as well as the probability of observing both values from the same locus (heterogeneity test). While none of the values are significantly different from one another at a single locus, they are in the right direction (combined probability of  $P = 0.09$  in population CW500; SOKAL

and ROHLF 1995). For genes 0.3, 17.0, and 18.0, in population CW500, the values for  $D$  at nonsynonymous mutations are always more negative than those at synonymous mutations. Gene 0.3 has values of  $-0.95$  and  $-0.06$  for nonsynonymous and synonymous mutations, respectively, 17.0 has values of  $-1.05$  and  $-0.71$ , and 18.0 has values of  $-1.31$  and  $-0.27$ . Gene 17.5 does not show significant heterogeneity between nonsynonymous and synonymous mutations in any population. For gene 17.5 there is no nonsynonymous polymorphism segregating in either population CW400 or CW500. The same is true for gene 18.0 in population CW400. A major reason heterogeneity may not be detected in these instances is that Fu and Li's  $D$  statistic, as well as other tests that utilize frequency distributions, can detect only relatively weak purifying selection (BRAVERMAN *et al.* 1995; CHARLESWORTH *et al.* 1995). Strong purifying selection allows little or no polymorphism to segregate in a population and consequently will not be detected by these types of tests. In addition, gene 17.5 appears to be under positive selection. Reasons why departures

**TABLE 4**  
**Fu and Li's  $D$  for nonsynonymous and synonymous mutations**

Population	Gene 0.3	Gene 17.0	Gene 17.5	Gene 18.0
CW100				
Nonsynonymous	-1.31	-1.89*	-0.12	-1.75**
Synonymous	-2.06*	-2.39**	-0.03	-1.85*
Heterogeneity test	$P = 0.79$	$P = 0.81$	$P = 0.48$	$P = 0.46$
CW400				
Nonsynonymous	-0.28	-0.41	No polymorphism	No polymorphism
Synonymous	-0.83	-0.06	0.25	1.08
Heterogeneity test	$P = 0.78$	$P = 0.19$	NA	NA
CW500				
Nonsynonymous	-0.95	-1.05	No polymorphism	-1.31
Synonymous	-0.06	-0.71	0.02	-0.27
Heterogeneity test	$P = 0.11$	$P = 0.19$	NA	$P = 0.20$

Values for Fu and Li's (1993)  $D$  statistic are calculated separately for nonsynonymous and synonymous data sets.  $D$  is calculated using the WT ancestor as the outgroup by the program DNAsp 3.5 (ROZAS and ROZAS 1999). The heterogeneity test addresses differences between values of  $D$  for nonsynonymous and synonymous mutations (discussed fully in MATERIALS AND METHODS). Significant values of the heterogeneity test indicate that different processes appear to be acting on nonsynonymous and synonymous mutations. When there is no polymorphism for one class of mutations, and hence no calculable value for  $D$ , no comparison can be made. NA, not applicable. \* $P < 0.05$ , \*\* $P < 0.01$ .

**TABLE 5**  
Ratio of  $\pi_a$  to  $\pi_s$

Population	Gene 0.3	Gene 17.0	Gene 17.5	Gene 18.0
CW100	0.245	0.216	0.164	0.077
CW400	0.200	0.155	0	0
CW500	0.105	0.142	0	0.008

The ratio of pairwise nonsynonymous and synonymous differences per site,  $\pi_a/\pi_s$  (HUGHES and NEI 1988), calculated by DNAsp 3.5 (ROZAS and ROZAS 1999), is shown. Values  $<1$  indicate constraint on the available neutral mutations at nonsynonymous sites. Values of 0 occur when there was no nonsynonymous polymorphism in a population.

from neutrality may not be detected by Fu and Li's  $D$  in this case are suggested in the DISCUSSION.

## DISCUSSION

Demographic, selective, and random processes can all determine the pattern of polymorphism in a genome. In this experimental system we are able to observe these processes and their effects on variation in a population over time. Our goal in carrying out this experiment was to attempt to tease apart the effects of each process on sequence polymorphism using current or modified statistical tests commonly used by population geneticists. Because this is an experimental system, the conditions under which this experiment was carried out made this task easier, as ancestral genotype and demographic history were known, and selective environment was similar throughout. Positive selection and purifying selection,

on strongly and weakly deleterious mutants (evident as segregating nonsynonymous mutations), are both found in this experiment; population expansion was also detected. We are also able to distinguish between weak purifying selection and population expansion using a method that modifies current tests of the frequency spectrum of mutations. Finally, we are able to show that the relative importance of these processes in shaping the pattern of polymorphism shifts over time.

**Positive selection:** Polymorphism and fixation data for two of the genes in our study, 0.3 (the gene that inactivates host restriction) and 18.0 (a gene involved in DNA maturation), do not allow us to reject the null hypothesis of neutrality with the M-K test (MCDONALD and KREITMAN 1991) for any population examined. For genes 17.0 and 17.5 (a tail fiber gene and a gene associated with lysis, respectively), however, the M-K test rejects neutrality in the CW400 and CW500 populations (Table 6). Because the CW lineage was propagated with a minimum of bottlenecking, the ability of positive selection to act is increased. Although the fixation of mildly deleterious alleles in small populations is a possibility, theory shows that advantageous mutations are actually more likely to be fixed in a growing population (OTTO and WHITLOCK 1997).

Using our experimental system we were able to test for positive selection not just at one point in time, but at multiple points in time under a similar selective regime. Data for the CW100 population do not reject the M-K test for any gene. We suggest that this is because advantageous mutations have either not arisen or have not had enough time to go to fixation; the latter hypothesis is strongly supported by our data. Of the five mutations

**TABLE 6**  
Nonsynonymous and synonymous polymorphism and divergence

Population	Gene 0.3		Gene 17.0		Gene 17.5		Gene 18.0	
	N.S.	Syn.	N.S.	Syn.	N.S.	Syn.	N.S.	Syn.
CW100								
Polymorphic	9	11	37	54	5	8	2	6
Fixed	0	0	2	0	0	0	0	0
CW400								
Polymorphic	13	21	45	90	0	9	0	11
Fixed	2	2	10	6	5	1	0	2
			$P = 0.024$		$P = 0.002$			
CW500								
Polymorphic	14	30	58	120	0	11	1	19
Fixed	2	1	9	5	5	1	0	1
			$P = 0.020$		$P = 0.001$			

A comparison of nonsynonymous polymorphism and fixed differences to synonymous polymorphism and fixed differences as suggested by MCDONALD and KREITMAN (1991) is shown. Departures from the neutral expectation are evaluated using Fisher's exact test. Fixed differences are counted as differences from the WT ancestor. N.S., nonsynonymous mutations; Syn., synonymous mutations;  $P$  = significance of  $2 \times 2$  test of independence using Fisher's exact test.

that are fixed in gene 17.5 in population CW400, all are present as polymorphisms in population CW100. One of these segregating polymorphisms in population CW100 is a mutation encoding methionine at the first position of the protein encoded by gene 17.5; our WT ancestor has a nonoptimal valine as the start codon. This mutation is expected to have a large, positive effect on transcription rates and, therefore, is expected to be advantageous. Also no nonsynonymous fixed differences appear between populations CW400 and CW500 for either gene; all fixations occurred between CW100 and CW400. This supports the hypothesis that advantageous mutations had the opportunity to arise but had not gone to fixation early in the experiment.

The only significant result using Fu and Li's  $D$  test from either gene 17.0 or 17.5 is from 17.0 in population CW100 (Tables 5 and 6). This result appears to be due to the population expansion from the WT ancestor. It is noteworthy, then, that positive selection at these loci fails to result in detectable hitchhiking (selective sweep) events, although it is possible that such events are unseen in gene 17.0 against the backdrop of population expansion. It is likely that the high level of recombination in our experiment reduces the region swept to fixation to a very small piece of DNA, therefore reducing the loss of polymorphism. Because little polymorphism is lost, the expected excess of low-frequency mutants with hitchhiking (FU and LI 1993; BRAVERMAN *et al.* 1995; FU 1997) is absent or undetectable.

As expected (HUGHES 1999), the M-K test is more powerful in detecting the action of positive selection than a comparison of the ratio of nonsynonymous to synonymous fixed differences ( $K_a/K_s$ ). For both populations CW400 and CW500 the  $K_a/K_s$  ratio has a value  $>1$  for gene 17.5 ( $K_a/K_s = 1.45$  for both). But for gene 17.0  $K_a/K_s$  is  $<1$  for both populations ( $K_a/K_s = 0.50$  and  $0.54$ ). The sensitivity of  $K_a/K_s$  ratios to detect positive selection is low and the criterion of  $K_a/K_s > 1$  as evidence for positive selection is very stringent, especially over long genes (HUGHES 1999). We suggest that the length of gene 17.0 (1662 bp) compared to the length of gene 17.5 (204 bp) has caused the inequality in the detection of positive selection using a  $K_a/K_s$  ratio. Because the M-K test takes advantage of more information (if available) it has the power to detect positive selection in both genes (AKASHI 1999).

**Distinguishing population expansion from purifying selection:** Some currently used tests for the neutrality of mutations (TAJIMA 1989; FU and LI 1993) attempt to detect an excess of low- or high-frequency mutants in the distribution of polymorphism. An excess of low-frequency mutations is generally interpreted as being due to purifying selection and the resulting weakly deleterious mutations segregating in a population; these tests assume a population at equilibrium. However, population expansion after a genetic bottleneck can give results similar to those caused by purifying selection

(TAJIMA 1989). Under purifying selection an excess of low-frequency mutants is expected because selection is acting to remove new mutations that have deleterious effects. During a population expansion there is an excess of low-frequency mutants because all mutations are new and therefore rare. Because Fu and Li's and Tajima's tests combine nonsynonymous and synonymous mutations it is impossible to separate effects of heterogeneous (*e.g.*, purifying selection) and homogeneous (*e.g.*, population expansion) processes.

To distinguish between homogeneous and heterogeneous processes, our heterogeneity test takes advantage of the fact that selective forces acting on DNA polymorphism often have one set of effects on nonsynonymous changes and another on synonymous changes. The reason for this is the magnitude of selective consequence of these two types of mutations. Nonsynonymous changes often have a much larger effect on the fitness of an organism and so are much more strongly influenced by natural selection (whether positive or negative). Synonymous changes more often than not have small or no fitness effects and so are much more strongly influenced by random genetic drift. These heterogeneous processes can be distinguished, then, from homogeneous processes, such as population expansion, that have equal effects on all types of mutations. An important aspect of this test is that different levels of selection on closely linked sites show independent effects on the pattern of polymorphism (AKASHI 1999; PRZEWORSKI *et al.* 1999).

We separated individual genes into nonsynonymous and synonymous mutations and then applied Fu and Li's  $D$  statistic to our data. We found two other instances where this simple partitioning has been done (RAND and KANN 1996; VERRELLI and EANES 2000). For example, RAND and KANN (1996) found no significant result using Tajima's  $D$  for the ND5 gene in *D. melanogaster* when their data included all types of mutations. Separating nonsynonymous from synonymous mutations, however, they found significant results for only nonsynonymous mutations and interpreted this as the heterogeneous process of purifying selection.

We have taken this approach a step further by providing a statistical framework in which to compare the magnitude of differences of various statistics between nonsynonymous and synonymous mutations. In particular, we compared values of Fu and Li's  $D$  statistic in a framework that allows an inference to be made about whether the values were likely to have been pulled from the same neutral distribution. Without a method for determining whether  $D$  is significantly different between nonsynonymous and synonymous mutations, simply partitioning the data offers ambiguous results: a heterogeneous process cannot be distinguished from a homogeneous process. For example, if a partitioned data set reveals a significant value for nonsynonymous mutations and a nonsignificant value for synonymous mutations, it may be that the value for synonymous mutations is just

slightly less negative than that needed for significance; therefore, they should not be considered heterogeneous simply because one is significant and the other is not. In addition, because our statistic conservatively assumes no intralocus recombination, it can be applied to comparisons of other classes of mutations within a locus that may be expected to have different selective constraints (*e.g.*, binding sites *vs.* nonbinding sites in a promoter). See HEY (1997) for the case where comparisons are made between loci.

Examining the data in RAND and KANN (1996) and VERRELLI and EANES (2000) provides good examples of the utility of our statistical test. Rand and Kann examined polymorphism at the ND5 gene in *D. melanogaster* and found only significant results in nonsynonymous mutations for Tajima's *D*. Applying our test of heterogeneity to their data (sample size = 59, nonsynonymous segregating sites = 8, synonymous segregating sites = 13,  $\Delta D = 0.93$ ) finds that there is no significant difference between the values of *D* for synonymous and nonsynonymous data sets ( $P = 0.15$ ). These results, therefore, can be interpreted as due to a homogeneous process. Verrelli and Eanes examined polymorphism at the Pgm locus in *D. simulans* and found no significant skew toward low-frequency mutants for the gene as a whole. Dividing the gene into nonsynonymous and synonymous mutations, however, they found a significant value for Fu and Li's *D* for the nonsynonymous data set. Applying our test to their data (sample size = 13, nonsynonymous segregating sites = 5, synonymous segregating sites = 64,  $\Delta D = 2.90$ ) we get a significant value of  $P = 0.0006$ . This supports the hypothesis of a heterogeneous process such as purifying selection.

In our data, in population CW100 three genes have significant Fu and Li *D* values for the combined-mutation data. For these three genes, 0.3, 17.0, and 18.0, our modified Fu and Li statistic gives results consistent with a homogeneous process acting in bacteriophage T7: none of the sets of values are significantly heterogeneous (Table 4). Moreover, the value of *D* is more negative for synonymous mutations in these genes, indicating the absence of the effects of significant purifying selection. It should be pointed out that neither Fu and Li's test nor even the more powerful  $F_s$  statistic (FU 1997) is able to detect this dramatic population expansion 2000 generations after propagation from a single ancestral individual ( $F_s = -1.0$ ,  $P = 0.18$ , population CW400; Table 3). Other polymorphism data also support the interpretation of this homogeneous process as population expansion. In all genes the value of *D* becomes less negative and nonsignificant in population CW400 (Table 3). This trend coincides with an increase in  $\pi$  relative to  $\theta_w$  as populations recover polymorphism from the monomorphic ancestral clone (Table 1). In addition, these patterns appear at multiple loci in the same organism. Some homogeneous processes work at only a single locus (*e.g.*, selective sweeps), but demographic

forces affect all loci in a population. It is possible, however, that purifying selection acting at a single gene in a larger linkage group may result in a homogeneous pattern at all loci ("background selection"; CHARLESWORTH *et al.* 1993). High recombination rates in our system suggest that this is unlikely to be a factor. We cite the rates found in population CW500 as the most likely to be the actual population recombination rates (see RESULTS), but even the values calculated for population CW100 are approximately equal to the highest *Drosophila* rate. Polymorphism data at multiple loci allow us to distinguish among the various possibilities of selective and demographic processes, and our data conform to the hypothesis of population expansion.

Not all of the loci studied in population CW100 give the same story; gene 17.5 does not show the effects of population expansion. Because mutation and genetic drift are random phenomena, there will always be some loci that do not exhibit the expected pattern. For gene 17.5 drift has allowed a number of synonymous mutations to climb to intermediate frequencies, giving only a slightly negative value for synonymous mutations in Fu and Li's *D* statistic (Table 4). Positive selection on nonsynonymous mutations, on the other hand, has had only a minor effect on the distribution of mutations: there are no nonsynonymous mutations at high frequencies. The three nonsynonymous mutations present in population CW100 that are fixed in populations CW400 and CW500 are all present at frequencies of 0.22 (two of nine individuals) while the other mutations are present in only single individuals. This combination of processes on nonsynonymous mutations, namely genetic drift, population expansion, and positive selection, results in a slightly negative value of *D* (Table 4). Fu and Li's *D* statistic on all mutations, therefore, is close to zero and nonsignificant.

Moving from population CW100 through CW400 to CW500 for all genes, a pattern appears. The homogeneous effects of the population expansion begin to resolve themselves into the heterogeneous effects of purifying selection. By population CW500 there are suggestive differences in genes 0.3, 17.0, and 18.0 in the distribution of frequencies of nonsynonymous and synonymous mutations (Table 4; gene 17.5 could not be tested). A combined probability test (SOKAL and ROHLF 1995) in population CW500 results in a probability of heterogeneity of  $P = 0.09$ . The transition from low-frequency mutations in all classes due to homogeneous processes to low-frequency nonsynonymous mutations due to only heterogeneous processes results in a time when neither population expansion nor purifying selection can be detected. This transition period in population CW400 shows only a slight skew at synonymous mutations as the result of demographic processes and only a slight difference between mutational classes due to selective processes. This type of complex result may



be the most challenging in attempting to resolve demographic, random, and selective effects.

We think that the VERRELLI and EANES (2000) data and our own data reveal a common occurrence: namely, hidden departures from neutrality for nonsynonymous mutations in many genes. Because Fu and Li's and Tajima's tests are most often applied to data that combine synonymous and nonsynonymous mutations, significant departures from neutrality for only one set of mutations (which is more likely to occur for nonsynonymous mutations) may be swamped out by the combined distribution of mutations. We predict that future analyses that partition the data into synonymous and nonsynonymous mutations will reveal heretofore unseen instances of segregating deleterious mutants, consistent with a nearly neutral theory of molecular evolution (OHTA 1992). This result is already hinted at in data from large human single nucleotide polymorphism data sets (CARGILL *et al.* 1999; HALUSHKA *et al.* 1999). If this prediction is true, it will have a large impact not only on the neutralist-selectionist debate (KIMURA 1983; GILLESPIE 1991; OHTA 1992), but also on tests of neutrality that use polymorphism as a putatively neutral standard (*e.g.*, McDONALD and KREITMAN 1991). If one believes that some fraction of the amino acid mutations segregating in a population will not go to fixation because they are deleterious and therefore are not a phase in neutral evolution (KIMURA 1983), then they can be eliminated from comparisons of polymorphism to divergence. The removal of these polymorphisms will then provide for more power to detect changes in the rate of fixation of nonsynonymous mutants (FAY *et al.* 2001). This may also reveal positive selection as a much more common phenomenon (FAY *et al.* 2001).

The authors thank J.J. Bull for propagation of the CW lineage and for making it available to us. M.W.H. especially thanks J. Stajich for help with the programming. We also thank J. Balhoff and T. Oakley for comments and suggestions on the manuscript and the Duke University Population Biology group for helpful criticism in the early stages of analysis. J. Wares, L. Moyle, and the members of MEDGAR all also made welcome suggestions, and J. Hey and two anonymous reviewers helped to improve the statistical test and the manuscript. C.W.C. was supported by National Science Foundation grant DEB-9615461, M.D.R. by National Science Foundation grant MCB-0110596, and M.W.H. by a National Institutes of Health training grant administered by the Duke University Program in Genetics.

#### LITERATURE CITED

- AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139**: 1067–1076.
- AKASHI, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**: 221–238.
- BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila melanogaster*. *Nature* **356**: 519–520.
- 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.
- BULL, J. J., C. W. CUNNINGHAM, I. J. MOLINEUX, M. R. BADGETT and D. M. HILLIS, 1993 Experimental molecular evolution of Bacteriophage T7. *Evolution* **47**: 993–1007.
- CARGILL, M., D. ALTSHULER, J. IRELAND, P. SKLAR, K. ARDLIE *et al.*, 1999 Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**: 231–238.
- 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.
- CUNNINGHAM, C. W., K. JENG, J. HUSTI, M. BADGETT, I. J. MOLINEUX *et al.*, 1997 Parallel molecular evolution of deletions and nonsense mutations in Bacteriophage T7. *Mol. Biol. Evol.* **14**: 113–116.
- DUNN, J. J., and F. W. STUDIER, 1983 Complete nucleotide sequence of Bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol.* **166**: 477–535.
- FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.
- FAY, J. C., G. J. WYCKOFF and C. I. WU, 2001 Positive and negative selection on the human genome. *Genetics* **158**: 1227–1234.
- FU, Y., and W. LI, 1993 Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- 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, 1999 Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Popul. Biol.* **56**: 1–10.
- GALTIER, N., F. DEPAULIS and N. H. BARTON, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155**: 981–987.
- GILLESPIE, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, New York.
- HALUSHKA, M. K., J. B. FAN, K. BENTLEY, L. HSIE, N. P. SHEN *et al.*, 1999 Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**: 239–247.
- HEY, J., 1997 Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* **14**: 166–172.
- HEY, J., and J. WAKELEY, 1997 A coalescent estimator of the population recombination rate. *Genetics* **145**: 833–846.
- HILLIS, D. M., J. J. BULL, M. E. WHITE, M. R. BADGETT and I. J. MOLINEUX, 1992 Experimental phylogenetics: generation of a known phylogeny. *Science* **255**: 589–592.
- HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**: 245–250.
- HUDSON, R. R., 1990 Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**: 1–44.
- HUDSON, R. R., and N. L. KAPLAN, 1995 Deleterious background selection with recombination. *Genetics* **141**: 1605–1617.
- HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.
- HUGHES, A. L., 1999 *Adaptive Evolution of Genes and Genomes*. Oxford University Press, New York.
- HUGHES, A. L., and M. NEI, 1988 Pattern of nucleotide substitution at major histocompatibility complex class-I loci reveals overdominant selection. *Nature* **335**: 167–170.
- KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.
- LANGLEY, C. H., J. MACDONALD, N. MIYASHITA and M. AGUADE, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* **90**: 1800–1803.
- MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favorable 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.
- OHTA, T., 1992 The nearly neutral theory of molecular evolution. *Annu. Rev. Ecol. Syst.* **23**: 263–286.
- OTTO, S. P., and M. C. WHITLOCK, 1997 The probability of fixation in populations of changing size. *Genetics* **146**: 723–733.
- PRZEWORSKI, M., B. CHARLESWORTH and J. D. WALL, 1999 Genealogies and weak purifying selection. *Mol. Biol. Evol.* **16**: 246–252.
- RAND, D. M., and L. M. KANN, 1996 Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* **13**: 735–748.
- ROZAS, J., and R. ROZAS, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174–175.
- SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. *Genetics* **135**: 541–552.
- 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.
- SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry*. W. H. Freeman, New York.
- STEPHAN, W., L. XING, D. A. KIRBY and J. M. BRAVERMAN, 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. *Proc. Natl. Acad. Sci. USA* **95**: 5649–5654.
- 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.
- VERRELLI, B. C., and W. F. EANES, 2000 Extensive amino acid polymorphism at the *Pgm* locus is consistent with adaptive protein evolution in *Drosophila melanogaster*. *Genetics* **156**: 1737–1752.
- WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**: 256–275.

Communicating editor: J. HEY