Perturbation-Reperturbation Test of Selection *vs.* Hitchhiking of the Two Major Alleles of *Esterase-5* in *Drosophila pseudoobscura*

Einar Árnason

Institute of Biology, University of Iceland, Grensásvegur 12, 108 Reykjavík, Iceland

Manuscript received May 10, 1990
Accepted for publication May 23, 1991

ABSTRACT

A perturbation-reperturbation tests selective neutrality of *100/100/100/100/100* and *106/100/100/100/100*, the two most common alleles at the highly polymorphic *X*-linked locus *Esterase-5* in *Drosophila pseudoobscura*. A total of 22 replicate populations are set up in cages, 11 start at a high frequency of 76% (U) and 11 at a low frequency of 21% (N) of the *106* allele. Allele frequencies change directionally and decrease in both U and N populations as groups and reach equilibria of 60 and 14%, respectively, after 200–300 days. These changes suggest natural selection. A hypothesis of balancing selection accounts for the pattern and predicts a dynamic equilibrium. A rival neutral hypothesis accounts for the pattern equally well by postulating hitchhiking and breakup of linkage leaving the *Est-5* variants to drift at neutral equilibria. A reperturbation of allele frequencies in each population, creating 22 additional reperturbed populations EN and EU, with the original populations as controls, directly addresses the question of balancing selection or hitchhiking and breakup of linkage effects. Allele frequencies do not change directionally among the reperturbed populations as a group. The hypothesis of balancing selection is rejected in favor of the hypothesis of initial hitchhiking and dissipated linkage effects. The power of the experimental design to detect selection is studied by simulation. Within the limits of power set by the design, it is concluded that the *100* and *106* are iso-fitness alleles of *Est-5* under the environmental conditions of the laboratory populations. The requirements of a method of perturbation and reperturbation are discussed.

THE hypotheses of selection or neutrality (KIMURA 1983; LEWONTIN 1974) of variation at allozyme loci demand a test: is this variation neutral in its effect on fitness or is it maintained by some form of balancing selection (KIMURA and OHTA 1971; KIMURA 1983; LEWONTIN 1974). Despite numerous studies, however, the issue has not been resolved. One major problem is how to disassociate the effects of a single locus from the background effects of linked loci. Associative overdominance at a neutral allozyme locus may arise from linkage to either truly overdominant background loci (FRYDENBERG 1963; OHTA and KIMURA 1971) or to detrimental mutations at background loci (OHTA and KIMURA 1971). And even though linkage disequilibrium between allozyme loci is not prevalent in natural animal populations (CHARLESWORTH and CHARLESWORTH 1973; MUKAI 1985; LANGLEY, TOBARI and KOJIMA 1974) it still can be a problem in experimental studies because in setting up an experiment one necessarily reduces population size. Any such bottleneck will generate variance in random linkage disequilibria (OHTA and KIMURA 1971; OHTA 1971).

A perturbation experiment using laboratory population cages is a standard method to test for selection. The rationale is to perturb allele frequencies away from equilibrium frequencies in nature and monitor subsequent changes. If balancing selection is operating, allele frequencies will change directionally back to an equilibrium. If variation at the locus is neutral, allele frequencies will change due to drift, but the direction of change will be indeterminate. Directionality thus is the criterion that can be used to distinguish between selection and drift. Allele frequencies at a strictly neutral locus undergoing drift in a single population will change in one direction or the other. Due to autocorrelation of allele frequencies in a single population, a large change taken once and preceded/succeeded by smaller changes or a succession of small changes in the same direction will both result in directionality of change in that population over the timespan of a few generations; therefore, changes due to drift can appear directional if viewed in a single population on a short time scale. Experimentally, therefore, directionality or lack of it can only be judged in replicate populations as a group. Furthermore, balancing selection predicts an internal equilibrium that should be approached both from above and below; therefore, a perturbation experiment is often started with a set of populations at a high frequency and another set at a low frequency (KOJIMA and TOBARI 1970; PROUT 1971a,b; BARKER and EAST 1980). To avoid initial linkage disequilibria randomization of variation at the locus in question relative to back-

ground variation is done by combining a number of independently derived lines carrying that allele from a natural population (YAMAZAKI 1971; LEWONTIN 1974; JONES and YAMAZAKI 1974; ÁRNASON 1982; MUKAI 1985). The locus interacts with background variation and marginal effects at the marker locus are looked for. From the rate of change of frequencies one may estimate fitness by fitting various models of selection (WILLIAMS, ANDERSON and ARNOLD 1990), for example, a model of selection by frequency-dependent fitnesses (WRIGHT and DOBZHANSKY 1946) or a model of selection by constant fitness (DUMOUCHEL and ANDERSON 1968; WILSON *et al.* 1982; WATTERSON 1982; WILSON and OAKESHOTT 1985).

A follow-up reperturbation experiment is a test of whether balancing selection or linkage disequilibria are responsible for allele frequency changes at an allozyme marker locus in a perturbation experiment. If balancing selection is responsible for the equilibrium, allele frequencies will return to the equilibrium upon reperturbation. If linkage effects are responsible for the approach of an allozyme locus to equilibrium, the equilibrium has resulted from a dissipation of that association; the equilibrium is then neutral and allele frequencies at the allozyme locus are not expected to change directionally upon reperturbation. A reperturbation involves a reduction in population size, however, and therefore linkage disequilibria may be generated anew in the process of reperturbation. Such secondarily generated linkage disequilibria will have a random sign, by expectation half are positive and half negative and if a reperturbation is done on a number of independent populations the effects of secondarily generated linkage disequilibria would mimic the effects of drift and result in no overall change in the ensemble. Directionality of allele frequency changes in the reperturbed populations would then be evidence for selection at the marker locus or for a tight association that had not dissipated yet or, perhaps, for a block of tightly linked and interacting genes.

To judge directionality of change among the reperturbed populations one cannot, however, rely entirely on an analysis of mean behavior or the behavior of pooled data, for although newly generated linkage disequilibria are random in sign they are not necessarily equal in magnitude in different populations; therefore, with a skewed distribution of effects the mean behavior could show directionality that was due to a single replicate. An analysis of the individuality or heterogeneity of replicate populations and the variability among the populations along with an analysis of the ensemble behavior is therefore a strategy for judging directionality. This can be done by estimating various statistics of selection and neutrality in the individual populations as well as in the ensemble (WRIGHT and DOBZHANSKY 1946; WILSON and OAKESHOTT 1985) along with statistics of overdispersion (BRIER 1980; MCCULLAGH and NELDER 1989; WILLIAMS, ANDERSON and ARNOLD 1990) and heterogeneity of theory and data (see appendix).

The *X*-linked genetic locus *Esterase-5* in *Drosophila pseudoobscura* is one of the most polymorphic allozyme loci in Drosophila (LEWONTIN and HUBBY 1966; PRAKASH, LEWONTIN and HUBBY 1969; COYNE, FELTON and LEWONTIN 1978; KEITH 1983). KEITH (1983) thus finds a total of 41 alleles in two natural populations approximately 400 km apart, two major alleles *100/100/100/100/100* at 35% and *106/100/100/100/100* at 21%, and a third, the *112/100/102/102/100*, at 5% polymorphic frequencies; the frequencies of the other alleles range from 4 to 0.4%. So far most of the experimental work addressing the question of the adaptive significance of *Est-5* variation has dealt with the *100* class of allelic variants versus the *112* class (YAMAZAKI 1971; JONES and YAMAZAKI 1974; ÁRNASON 1982; KEITH 1983). The high frequency and wide distribution of the *100* and *106* classes of alleles in nature (KEITH 1983), however, *a priori* make the *100* and the *106* alleles the most likely candidates as alleles of a balanced polymorphism at the *Est-5* locus.

This paper reports the results of an experiment of perturbation-reperturbation using laboratory population cages to test the rival hypothesis of selection *vs.* hitchhiking of the two major alleles of *Esterase-5*.

MATERIALS AND METHODS

Genetic strains: Nine isogenic lines were obtained from TIM P. KEITH as material for the present study. The lines originated from flies collected at the James Reserve in the San Jacinto Mountains in Southern California (KEITH 1983). Five lines (JR38, JR52, JR113, JR178, JR464) were isogenic for the *100/100/100/100/100* allele (hereafter referred to as the *100* allele) and four lines (JR106, JR129, JR146, JR395) were isogenic for the *106/100/100/100/100* allele (hereafter the *106* allele). These isogenic lines were constructed from holding stocks by KEITH in July to October 1980 (KEITH 1983). In October to December 1984 seven additional isogenic lines were constructed from other holding stocks supplied by KEITH. The holding stocks were either segregating for one of the active and a *null Est-5* allele or fixed for one or the other allele. Five males and females were pair-mated from each holding line, allowed to lay eggs, and subsequently both parents were electrophoresed. If the phenotype of both parents was an active enzyme, ten male offspring of that mating were electrophoresed. If all ten had an active enzyme phenotype, the line was carried forward as an isogenic line. The probability that the female of a line thus carried forward was heterozygous for an active and the *null* allele is $(1/2)^{10} = 0.00098$. Thus three lines (JR34, JR144, JR249) were made isogenic for the *100* allele and four lines (JR96, JR104, JR181, JR478) isogenic for the *106* allele of KEITH (1983).

The genetic material for the experiment was based on these 16 independently derived isogenic lines, eight for each allele, thus randomizing the *X* chromosomal genetic back-

ground of the alleles (LEWONTIN 1974; JONES and YAMAZAKI 1974). Chromosomes carrying the *Sex Ratio* inversions (DOBZHANSKY and EPLING 1944) were excluded from the material (TIM P. KEITH, personal communication). The region around the *Est-5* locus in this genetic material is expected to have been representative of wild-caught flies from the James Reserve. Recombination during the extraction and holding stages may, however, have introduced genetic material from other populations, as there are no balancers available for extraction of the *D. pseudoobscura X* chromosome. In the method of constructing the isogenic lines (KEITH 1983), the wild chromosomes from James Reserve were held against a chromosome with an *Est-5 null* allele (line FC51) originating from Furnace Creek, California (COYNE, FELTON and LEWONTIN 1978), A chromosome from Strawberry Canyon, California, (the 116 allozyme standard, COYNE, FELTON and LEWONTIN 1978) also came into contact with the wild chromosomes during the extraction (TIM P. KEITH and RICHARD C. LEWONTIN, personal communication). Thus parts of the genetic background in this experiment may have originated from Furnace Creek and Strawberry Canyon, although the latter is less likely as the genetic material from Strawberry Canyon only came into brief contact.

Perturbation experiment: To minimize the generation of initial linkage disequilibria flies from isogenic lines were mass-mated to produce large numbers of flies of various genotypes used to start population cages. A set of five females and five males from each isogenic line representing each allele were mated in a number of bottles. The offspring from all bottles within an allelic class yielded the *100/100* and the *106/106* homozygotes. Heterozygous *100/106* flies were obtained by crossing in a number of bottles ten and in other bottles five virgin females of each line isogenic for the *100* allele and ten males from each line isogenic for the *106* allele and the reverse cross of virgin females from each line isogenic for the *106* allele and males from each line isogenic for the *100* allele. The offspring from all bottles of each type of cross were combined and thus a large number of *100/100* and *106/106* homozygotes and *100/106* heterozygotes were obtained and used to set up egg-laying cages. In February 1985 two population cages were initiated for collecting eggs used to found the isolated replicate populations. In order to start with equal frequencies in both sexes 20, 160 and 320 virgin females of respectively the *100/100, 100/106,* and *106/106* genotypes and 59 and 239 males of respectively the *100/Y* and the *106/Y* genotypes were introduced into the egg-laying cage used to found the populations starting at a high frequency of the *106* allele. The mnemonic is U populations (Table 1). The frequency of the *106* allele among both sexes combined in the egg-laying cage was thus 0.800 ± 0.0111 ($q \pm$ SD). For the N populations (Table 1) 241, 85 and eight virgin females of respectively the *100/100, 100/106* and *106/106* genotypes and 204 and 37 males of respectively the *100/Y* and *106/Y* genotypes were introduced into the egg-laying cage used to found the populations starting at a low frequency of the *106* allele. The frequency of the *106* allele among both sexes combined in the egg-laying cage was thus 0.152 ± 0.0119.

On 19 February 1985 ten cups with medium were introduced into each egg-laying cage. Three days later each set of ten cups was used to found replicates starting with a high density of eggs and larvae; these are referred to as the NH and UH replicates (Table 1). They were maintained throughout the experiment to see if different methods of founding made a difference in the results. By not using the first cups to found replicates 1–10 the egg-laying cages were given an opportunity, for example by allowing some time for mating and remating, to overcome any possible initial effects of handling the flies. On 22 February ten fresh cups were introduced to the egg-laying cages and eggs collected for four days. In each frequency category ten replicate populations N1, . . ., N10 and U1, . . ., U10, were founded with one cup. Fresh cups were reintroduced to the egg-laying cages and the above process repeated until 12 March by which time each replicate had four cups of different ages giving each replicate population an age structure from the start.

TABLE 1

Names, mnemonics and descriptions of replicates

Replicates	Mnemonic	Description
N1–N10, NH	N	Perturbation experiment, starts at a low frequency; H replicate founded with an initial high density
EN1–EN10, ENH	EN	Reperturbation of the N replicates to frequencies below N controls
U1–U10, UH	U	Perturbation experiment, starts at a high frequency; H replicate founded with an initial high density
EU1–EU10, EUH	EU	Reperturbation of the U replicates to frequencies below U controls
Pooled	ALL	The pooled data of all replicates within a group

The date of founding of the egg-laying cage, 22 February 1985, was counted as day zero of the experiment. Samples of eggs were taken from the founding cups and a first sample of eggs and larvae laid by the adults that eclosed from the founding cups was taken from all replicates including the H populations on day 43. This was also the date of the first available sample from the H populations which therefore have two degrees of freedom less in the analysis presented. Subsequently samples were taken approximately every generation until 22 June 1987, day 850, when the experiment was terminated.

When a large number of adults had eclosed from the first two cups in all population cages, three fresh cups were introduced into each cage and a schedule of cup replacement started. Due to differential mortality or fertility in the egg-laying cages the mean allele frequency at day zero was somewhat lower for the U populations and higher for the N populations than the allele frequency among the respective adult egg-laying populations. The total mortality in the egg-laying cages from 22 February to 12 March was about 29%.

The populations were sampled regularly by scraping eggs and first instar larvae from the medium and letting them develop in vials outside the cages. The cups with medium were kept in the cage for two to four days before a sample was scraped off. The eggs and larvae were put onto fresh medium in vials where they developed under less crowded conditions than they would have in cups in the cage. Several vials were used for a sample from each cage. Adults were collected and combined from all vials of a sample and frozen for later electrophoresis. No attempt was made to keep track of vials and thus any vial effect in a sample was lost.

The population cages made according to the design of ÁRNASON (1986) were kept in a room with controlled light, temperature, and humidity. The light cycle was ten hours

fluorescent lights and 14 hours darkness. Temperature was about 23° with a 1° fluctuation. A hygrostat controlling a humidifier was set at 55% relative humidity. The schedule of cup replacement was three cups once a week, with a four week rotation and 12 cups per cage. The medium was an agar-high yeast-brown sugar-cornmeal medium in the following proportions (per liter): water, 450 ml of medium; agar, 12 ml; dried flaked yeast, 24 ml; brown sugar, 80 g; cornmeal, 60 g; and propionic acid, 6 ml, was added to control mold. About 90 ml of medium were poured into disposable 180-ml plastic cups produced for commercial yogurt. The cups measured 78 mm high with a 54-mm bottom diameter and a 70-mm top diameter. The surface area of the medium in a cup was approximately 28 cm^2 and thus approximately 84 cm^2 became available for egg-laying weekly.

Reperturbation experiment: On day 468 of the experiment new replicate populations were founded by perturbing the frequency of a sample from each replicate; the original N and U populations were continued as controls. Between 130 and 150 mated females were sampled from each cage, allowed to lay eggs singly in a vial, and electrophoresed. Vials were chosen on the basis of female genotype, and a biased sample used to found a new population, the EN and EU groups of populations (Table 1). No attempt was made to equalize the number of flies from each vial. A sample of males was not electrophoresed so the male allele frequency at this time was unknown. Also unknown was the allele frequency among sperm carried by these females. The allele frequencies among females prior to taking a biased sample of vials is the best available estimate of the frequency among sperm carried by the biased sample of founding females. The method of reperturbation thus creates a disparity in the frequency among the sexes.

On average 41 vials were used to found the reperturbed EU populations and 81 to found the EN populations. Each vial chosen carried three independently derived chromosomes from a population if one assumes that each original female carried sperm from only a single male. Based on this assumption the harmonic mean number of chromosomes used to found the EU population was 117 (the 95% confidence limits are 100 and 140 calculated as the reciprocals respectively of the upper and lower limits of the mean of the reciprocals) and 232 for the EN populations (the 95% confidence limits are 184 and 314). If females carried sperm from more than one male the numbers would be higher. The harmonic means were reasonably large and thus the bottleneck of the genetic background was small in the reperturbation to judge from the estimated effective number of chromosomes used for founding.

Electrophoresis: Electrophoresis followed the general procedures of KEITH (1983) and ÁRNASON and CHAMBERS (1987) with modifications. During electrophoresis EST-5 exists in a dynamic equilibrium between monomers and dimers whose position shifts towards the monomer as the running temperature of the gel is increased (ÁRNASON and CHAMBERS 1987; CANN 1987). To take advantage of this behavior all gels were run at 22° which yielded monomers. Electrophoresis takes a shorter time at 22° than at the conventional 4°. Separation of bands is greater and scoring of gels is easier and more reliable, especially scoring of heterozygotes.

Normally 68 females were electrophoresed from each population for a sampling date. Not all samples yielded 68 females, however, so the number was less in some instances but larger in other instances. Altogether 40,118 females were scored in the experiment. Males have only been run for some sampling dates and the results are not reported here.

Statistical model fitting: The fitting of statistical selection models followed methods developed by WILSON (1980) and WILSON *et al.* (1982) who extended methods considered by SCHAFFER, YARDLEY and ANDERSON (1977) which were originally developed by FISHER and FORD (1947). For each population the data comprised a sequence of allele frequencies p_t at different times t, $p_t = y_t/n_t$ where y_t is the observed number of the allele in question in the sample and n_t is the total number of alleles in the tth sample. The allele frequency was subjected to the following angular (arcsine) transformation (FREEMAN and TUKEY 1950):

$$Y_t = \sin^{-1}\sqrt{\frac{y_t}{n_t + 1}} + \sin^{-1}\sqrt{\frac{y_t + 1}{n_t + 1}}.$$

The transformation stabilizes the variance of Y_t to a value that is less dependent on the true frequency than are the untransformed allele frequencies. In general the variance of a frequency thus transformed is $1/(n + \frac{1}{2})$.

For each population the frequencies are represented by a vector $\mathbf{Y}$ whose c elements, Y_t, are the transformed values at the various sampling dates d_t. The variance-covariance matrix for the transformed frequencies is the $c \times c$ symmetric matrix $\mathbf{V}$; c is the number of sampling dates. The variance of a sample arises from the finite sample size as well as from the effects of drift due to a finite population size. The diagonal elements of $\mathbf{V}$ are given by

$$V_{t,t} = \frac{1}{n_t + \frac{1}{2}} + \left(1 - \frac{1}{n_t + \frac{1}{2}}\right)\left[1 - \left(1 - \frac{1}{N + \frac{1}{2}}\right)^{(d_t - g)/g}\right].$$

The covariance of the sample allele frequencies of any two dates arises entirely from random drift, both samples being descended from the true frequency in the generation before the earlier sampling date. If a true frequency is due to selection, a deviation from that true frequency is due to drift; in a population which drifts above/below the true frequency at a particular generation the population frequency is likely to be on the same side of the true frequency in the next generation and that creates a correlation between the allele frequencies of the samples from these two generations. Selection, on the other hand, will tend to bring a population back to the true frequency. Thus the covariance of sample frequencies at two dates is affected only by drift and not by selection. Therefore the off-diagonal elements of $\mathbf{V}$ are given by

$$V_{t,t+r} = V_{t+r,t} = 1 - \left(1 - \frac{1}{N + \frac{1}{2}}\right)^{(d_t - g)/g}$$

(SCHAFFER, YARDLEY and ANDERSON 1977; WILSON 1980; WILSON *et al.* 1982; NEI and TAJIMA 1981; POLLAK 1983). In this formulation d_t is the sampling day in days from the start of the experiment $\{t = 1 \ldots c\}$, g is the generation length (taken to be 28 days which was the schedule of cup-replacement in the cages), N is the effective number of alleles in the population ($N/2$ is the effective population size N_e), and n_t the number of alleles scored at time d_t as already mentioned. The models of drift and selection considered here assume discrete generations but the populations are continuously breeding. To reconcile these a linear approximation of time is used (*cf.* DUMOUCHEL and ANDERSON 1968).

A genetic selection model m is fitted to the data and gives rise to an expected transformed allele frequency e_t for sampling day d_t. The e_t's are the elements of the vector of

expectations ${}^{m}\mathbf{E}$ for model m. For each population a difference vector, ${}^{m}\mathbf{D}$, is formed for a given genetic model m

$$ {}^{m}\mathbf{D} = \mathbf{Y} - {}^{m}\mathbf{E}. $$

From this is found a likelihood-ratio statistic referred to as the deviance KENDALL and STUART 1961; NELDER and WEDDERBURN 1972; BRIER 1980; WILSON 1980; WILSON *et al.* 1982; MCCULLAGH and NELDER 1989; WILLIAMS, ANDERSON and ARNOLD 1990)

$$ {}^{m}X^2 = ({}^{m}\mathbf{E} - \mathbf{Y})'\mathbf{V}^{-1}({}^{m}\mathbf{E} - \mathbf{Y}) $$
$$ = {}^{m}\mathbf{D}'\mathbf{V}^{-1}{}^{m}\mathbf{D} $$

which is distributed approximately as a χ^2 variate with degrees of freedom given by ${}^{m}df = c - {}^{m}\nu - f(V_{i,j})$ $\{i \neq j\}$. Here c is the number of sampling points, ${}^{m}\nu$ the number of separate parameters fitted under genetic model m, and $f(V_{i,j})$ $\{i \neq j\}$, a function of the covariance matrix, is a number by which the df are reduced due to the autocorrelation of allele frequencies at the various sampling dates resulting from drift. Inspecting $\mathbf{V}$ it is fairly safe to assume that the reduction in the df due to autocorrelation from drift is small. Furthermore, the results are in most instances such that a small reduction in df would not make a qualitative difference to the interpretation of statistical significance. Therefore the reduction of df is ignored henceforth.

To test the hypothesis that a population drifts around the estimated initial allele frequency (q_0), a vector is formed ${}^{q_0}\mathbf{Q} = q_0\mathbf{1}$, where $\mathbf{1}$ is a vector of ones, and transformed using the above angular transformation to yield ${}^{q_0}_{T}\mathbf{Q}$. The difference vector is ${}^{q_0}\mathbf{D} = \mathbf{Y} - {}^{q_0}_{T}\mathbf{Q}$ and ${}^{q_0}X^2 = {}^{q_0}\mathbf{D}'\mathbf{V}^{-1}{}^{q_0}\mathbf{D}$ with ${}^{q_0}\nu = 1$ and thus $df = c - 1$. Similarly, to test the hypothesis that a population drifts around the estimated mean frequency $\bar{q} = \mathbf{1}'\mathbf{V}^{-1}\mathbf{Y}/\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}$, the difference vector ${}^{\bar{q}}\mathbf{D} = \mathbf{Y} - \bar{q}\mathbf{1}$ yields the test statistic ${}^{\bar{q}}X^2 = {}^{\bar{q}}\mathbf{D}'\mathbf{V}^{-1}{}^{\bar{q}}\mathbf{D}$ with ${}^{\bar{q}}\nu = 1$ and $df = c - 1$ (WILSON 1980; WILSON *et al.* 1982). The selection models are tested in a similar fashion with ${}^{s}X^2 = {}^{s}\mathbf{D}'\mathbf{V}^{-1}{}^{s}\mathbf{D}$ and ${}^{s}\nu = 3$ and therefore $df = c - 3$.

If a hierarchy of models is tested such that model j is a subset of model m, one may test whether the more specific model (j) gives a better fit to the data than the more general model (m). The test statistic is

$$ {}^{m-j}X^2 = {}^{m}X^2 - {}^{j}X^2 $$

which is distributed approximately as a χ^2 with $df = {}^{j}\nu - {}^{m}\nu$ under the null hypothesis of no improvement in fit with the more highly parameterized model (BILLINGSLEY 1961; WILSON 1980; WILSON *et al.* 1982). Specifically, the above tests of neutrality are more general than the tests of selection; one can thus test the goodness-of-fit of a model of neutrality, a model of selection, and finally the model of selection under consideration *vs* the model of neutrality. Strictly speaking the models considered are not nested, but one may argue that fitting the selection model is a good approximation to fitting a selection model and drift because $4N_e s \gg 1$ (KIMURA 1983) as it turned out.

Heterogeneity: Directionality of allele frequency changes among the ensemble, a property to be looked for as evidence for selection over neutrality, was studied in the mean of the replicates and in the pooled data represented by the fictitious population ALL (Table 1). In addition the heterogeneity, indicative of the individuality of the behavior of the populations within each category, was also informative about the processes going on in these populations; therefore attention was also given to a study of individual populations and the details of their behavior. The overdispersion of the data was measured by the heterogeneity of the replicates and the behavior of the ensemble is studied relative to this heterogeneity (BRIER 1980; MCCULLAGH and NELDER 1989; WILLIAMS, ANDERSON and ARNOLD 1990).

TABLE 2

Deviance statistics of heterogeneity of theory and data with associated degrees of freedom

Source	Deviance	*df*
Total	${}_{T}X^2 = \sum_i^k(\mathbf{Y}_i - {}_{T}\mathbf{E})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - {}_{T}\mathbf{E})$	$kc - {}^{s}\nu$
Within	${}_{W}X^2 = \sum_i^k(\mathbf{Y}_i - \mathbf{E}_i)'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{E}_i)$	$k(c - {}^{s}\nu)$
Between	${}_{B}X^2 = \sum_i^k(\mathbf{E}_i - {}_{T}\mathbf{E})'\mathbf{V}_i^{-1}(\mathbf{E}_i - {}_{T}\mathbf{E})$	$(k-1){}^{s}\nu$
Heterogeneity	${}_{H}X^2 = \sum_i^k(\mathbf{Y}_i - {}_{T}\mathbf{Y})'\mathbf{V}_i^{-1}(\mathbf{Y}_i - {}_{T}\mathbf{Y})$	$(k-1)c$
Marginal	${}_{M}X^2 = ({}_{T}\mathbf{Y} - {}_{T}\mathbf{E})'{}_{T}\mathbf{V}^{-1}({}_{T}\mathbf{Y} - {}_{T}\mathbf{E})$	$c - {}^{s}\nu$

See text for explanation of symbols.

The behavior of the ensemble and the question of heterogeneity or individuality within a group of replicates is addressed with the summary deviance statistics defined in Table 2. The definitions are analogous to LEWONTIN's partitioning of *G*-statistics in the appendix. The between replicate deviance ${}_{B}X^2$ (Table 2) measures the heterogeneity of the local theoretical expectations relative to the global theoretical expectations, whereas ${}_{H}X^2$ measures the usual heterogeneity of data or the deviations of the local observations from the pooled observations. Similarly ${}_{W}X^2$ measures the variability of local theoretical fit within a group. The amount of overdispersion or variance inflation is estimated from these statistics (WILLIAMS, ANDERSON and ARNOLD 1990). The dispersion parameters are ${}_{W}\hat{C} = {}_{W}X^2/{}_{W}Wdf$ (MCCULLAGH and NELDER 1989; WILLIAMS, ANDERSON and ARNOLD 1990) and ${}_{H}\hat{C} = {}_{H}X^2/{}_{H}df$ (BRIER 1980). The dispersion parameters $\hat{C}$ are used to correct the goodness-of-fit deviances of theoretical expectations, as well as reported standard errors and deviance residuals. The corrected $X^2_{\nu}/\hat{C}$ goodness-of-fit statistics are distributed as a χ^2_{ν} (BRIER 1980; MCCULLAGH and NELDER 1989; WILLIAMS, ANDERSON and ARNOLD 1990).

Selection models: Two models of selection are considered in this paper, a model of frequency dependent fitnesses, and a model of constant fitnesses. The analysis employs the standard autosomal selection dynamics (WRIGHT and DOBZHANSKY 1946) and does not develop the *X*-linked model as males were not electrophoresed for all samples. As shown already by WRIGHT and DOBZHANSKY (1946) predictions of models assuming sex-dependent fitnesses are indistinguishable from the standard dynamics unless populations are far from equilibria. The use of the approximations of the standard dynamics therefore seemed justified. There is also an assumption of no random drift in male allele frequencies creating a disparity in the frequencies of the sexes which would influence allele frequencies among females and the variance-covariance matrix.

In the model fitting the female genotypes *100/100, 100/106, 106/106* whose frequencies are $(1-q)^2$: $2(1-q)q$:q^2 have fitnesses $W_{100/100}$:$W_{100/106}$:$W_{106/106}$. Under the frequency-dependent model fitnesses are a linear function of the allele frequency: $1 - a + bq$:1:$1 + a - bq$ (WRIGHT and DOBZHANSKY 1946). This is the simplest frequency-dependent model with heterozygotes exactly intermediate in fitness thus assuming additivity of gene action in fitness. The mnemonic for this model is *F*. More complicated models, such as letting fitness depend on the frequency of the genotype, are not considered in this paper. Under this model a change in allele frequency per generation due to selection is $\Delta q = q(1-q)(a - bq)/(1 - (a - bq)(1 - 2q))$ and the predicted equilibrium $\hat{q} = a/b$ (WRIGHT and DOBZHANSKY 1946).

Under the constant-fitness model (mnemonic C) the fitnesses of the three genotypes are $1 - t$:1:$1 - s$ respectively, the change in allele frequency per generation due to selection is $\Delta q = -q(1 - q)(sq - t(1 - q))/(1 - sq^2 - t(1 - q))$ and the predicted equilibrium $\hat{q} = s/(s + t)$ (WRIGHT and DOBZHANSKY 1946).

Given a selection model and given an initial starting allele frequency the task of the estimation is to find by some estimation procedure the coefficients, a and b in one instance and s and t in the other, that give the best fit to the data. One problem that immediately arises is to decide what initial frequency to use; the goodness-of-fit of the data to the selection curve critically depends on where that curve starts. The usual practice has been to use the known or assumed-known initial allele frequency among the egg-layers or founders of the experiment (*e.g.*, DUMOUCHEL and ANDERSON 1968; WILSON *et al.* 1982). I call it assumed-known to underscore my view that the allele frequency among the eggs that one gets out of the egg-layers may be quite different from the input allele frequencies among egg-layers. Such an initial change in allele frequency from egg-layers to eggs may be called natural selection but it is uninteresting for it may be caused by the experimenter in handling the flies or be peculiar to the conditions of egg-laying in the egg-laying cages. Being thus a unique founding event it is not of general interest in a causal analysis of selection at the locus. Therefore, instead of assuming a known initial frequency, q_0, it was estimated along with the coefficients of both selection models and thus three parameters were estimated under both models.

For each population the parameters were estimated by maximizing the function

$$L = -(\mathbf{Y} - {}^{m}\mathbf{E})'\mathbf{V}^{-1}(\mathbf{Y} - {}^{m}\mathbf{E})$$
$$= -{}^{m}\mathbf{D}'\mathbf{V}^{-1}{}^{m}\mathbf{D}$$

which is equivalent to minimizing $-L$. The algorithm AMOEBA of PRESS *et al.* (1986) which implements the downhill simplex method of minimization was used.

The error distribution of the estimated fitness and initial frequency was obtained by Monte Carlo simulations as described by (PRESS *et al.* 1986, p. 529 *ff.*). For each population 200 random synthetic data sets were generated using the estimated fitnesses and throwing in both random drift for each generation prior to a sample as well as errors of sampling. Changes in allele frequency due to drift were simulated by binomial sampling in a population of size N_e = 1200. To simulate errors of sampling, genotypic numbers of each sampling date d_i were synthesized as binomial samples of frequencies $(1 - q)^2$:$2(1 - q)q$:q^2 of n_i individuals. The algorithm BNLDEV was used. The random number generator was RAN3 in all instances (PRESS *et al.* 1986).

Residuals: The adequacy of the selection model was checked by examining deviance residuals (DRAPER and SMITH 1981; MCCULLAGH and NELDER 1989; WILLIAMS, ANDERSON and ARNOLD 1990). A deviance residual is defined as $r_D^* = \text{sign}(Y_i - E_i)\sqrt{d_i}$ where $\text{sign}(x) = x/|x|$ and ${}^{s}X^2 = \sum d_i$ (MCCULLAGH and NELDER 1989; WILLIAMS, ANDERSON and ARNOLD 1990). Standardized deviance residuals adjusted by the overdispersion factor $r_D = r_D^*/\sqrt{\hat{C}}$ were plotted in time series against the arcsine transformed model expectations.

Robustness to variation in N_e: The effect of effective population numbers on the estimates was studied, first, by applying the amoeba minimization to estimate N_e under the model of neutrality around the mean frequency. The estimates were very low, as expected, for this only translates the large directional allele frequency changes due to selection into drift expressed in the currency of an effective population number. Second, the robustness of the estimators to variation in N_e was studied by doing the minimization for each replicate using six different population sizes N_e = {300, 600, 1200, 2400, 4800, 9600}. The statistics and estimates of selection and neutrality reported in the paper are the ones obtained with an N_e = 1200.

Convergence: The convergence of the minimizations was tested by starting the AMOEBA with four different sets of starting values for each replicate and each of the six effective population numbers used for the variance-covariance matrix. After reaching a minimum each run was restarted as recommended by PRESS *et al.* (1986) with one vertice of the restart simplex at the minimum and the other vertices from the original starting values of that run. The restarts always converged back to the minimum within the 10^{-6} tolerance limits of the algorithm. In all but two of the N and U populations the four different sets of starting values gave very similar estimates, again within tolerance limits of the algorithms. In the EN and EU population, however, the four sets of starting values sometimes gave very different results. The results reported here are the ones based on the lowest observed local minima from the $6 \times 4 \times 2 = 48$ minimization of each replicate. Most of the other observed local minima gave estimates, such as negative fitnesses, which were not biologically meaningful.

Power: The power of the experimental design to detect selection was studied by Monte Carlo simulations (*cf.* DYKHUIZEN and HARTL 1980). Given the frequency-dependent model of selection, an initial frequency of $q_0 = 0.75$ and an equilibrium frequency of $\hat{q} = 0.50$, a total of 1000 random data sets were generated assuming random drift (N_e = 1200) and variance of sampling (n = 68) with the sampling dates of the U populations and binomial sampling. Samples were generated using true fitness differences between heterozygous and homozygous genotypes ranging from 0.5 to 10.0% at the initial frequency. Fitness and deviance statistics were estimated in the random samples and used to make statements about power of the design.

All calculations were done on an AST Premium/386 equipped with an 80387 math coprocessor using Turbo Pascal 5.0 *extended* and *longint* variables where applicable for precision in the algorithms. The algorithms and computer programs were tested with known data. A test of these is also implied in a comparison of the original estimates with the mean estimates of randomly generated data sets (see Tables 7 and 8) and in the estimates of power (Table 3). These two are the same or very similar except where the distribution of the estimates of the randomly generated data are skewed.

RESULTS AND ANALYSIS

The perturbation experiment: In both the N and the U populations as a group, the frequency of the *106* allele decreased initially and subsequently reached equilibria (Figure 1).

The frequency of the *106* allele among the N populations as a group decreased from a mean starting frequency of 0.214 ± 0.0104 (n = 11) at the first sampling date to an equilibrium frequency of 0.140 ± 0.0103 (n = 66; the mean ± SE of the last six sampling dates after day 468 (Figure 1) a date arbitrarily chosen by inspection as representing an equilibrium frequency. This represents a mean allele fre-

TABLE 3

The power of the experimental design to detect selection

Fitness difference			Selection: frequency dependence			Neutrality					
						Initial frequency			Mean frequency		
True	Estim.										
ΔW_{q_0}	$\widehat{\Delta W}_{q_0}$	$s\widehat{\Delta W}$	$\overline{{}^{s}X^2_{16}}$	${}^{s}P$	% > 0.05	$\overline{{}^{q_0}X^2_{18}}$	${}^{q_0}P$	% > 0.05	$\overline{X^2_{18}}$	${}^{\hat{q}}P$	% > 0.05
0.005	0.024	0.090	9.7	0.883	99.9	11.8	0.858	98.2	10.5	0.914	100.0
0.010	0.037	0.107	9.8	0.879	99.9	13.2	0.781	97.4	11.2	0.884	100.0
0.020	0.030	0.056	10.1	0.863	99.7	15.3	0.644	97.1	13.5	0.760	99.6
0.030	0.034	0.028	10.1	0.862	99.8	18.3	0.437	95.1	16.0	0.592	99.2
0.040	0.046	0.032	10.1	0.863	100.0	22.5	0.210	82.1	18.7	0.411	96.9
0.050	0.054	0.023	10.3	0.852	99.7	26.4	0.091	61.6	21.3	0.264	90.4
0.060	0.064	0.027	10.3	0.851	99.8	30.3	0.035	39.0	23.5	0.172	80.8
0.070	0.076	0.030	10.3	0.853	99.9	34.6	0.011	22.0	25.5	0.112	67.8
0.080	0.085	0.030	10.2	0.854	99.8	37.8	0.004	11.5	27.0	0.079	58.6
0.090	0.096	0.033	10.3	0.851	100.0	41.6	0.001	5.0	28.5	0.055	46.0
0.100	0.106	0.034	10.1	0.859	99.9	44.9	0.000	2.9	29.5	0.043	39.5

Given a true frequency-dependent fitness differential of $\Delta W = a - b_{q_0}$ between homozygous and heterozygous genotypes at an initial frequency of $q_0 = 0.75$ and given an equilibrium frequency $\hat{q} = 0.50$, a total of 1000 random samples were generated assuming $N_e = 1200$ and the sampling scheme of the U populations. The table shows the estimated mean fitness differentials, $\widehat{\Delta W}$ and their standard errors, the mean deviance statistics, the probability associated with the mean deviance, and the percentage of the 1000 deviances exceeding a nominal 0.05 level of significance for both the selection and neutral models.

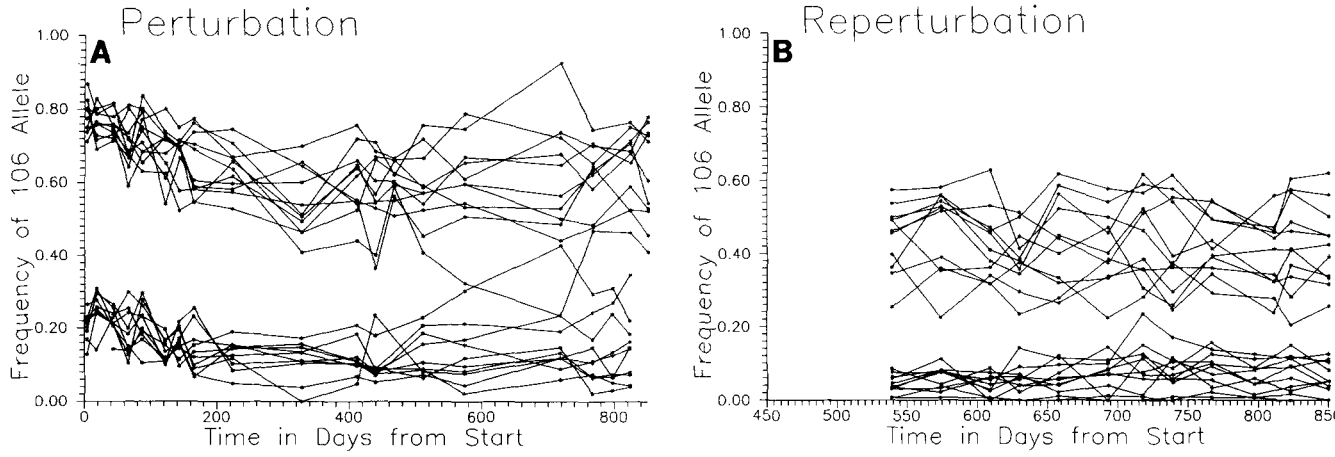

FIGURE 1.—Frequency of the *106* allele on all sampling dates among females in the replicate N and U populations (A) and the replicate EN and EU populations (B). Frequencies in individual populations are traced from one sampling date to another.

quency change of about 7% in roughly 17 generations. The individual populations mostly behaved in this fashion with notable exceptions analyzed further below. The pattern of change of allele frequencies was similar among males and among sexes combined both among the populations as a group as well as individually as was evident from the limited amount of male data available (data not shown). The frequencies among sexes combined showed the least variation between the populations as expected from the highest number of alleles sampled (data not shown).

The frequency of the *106* allele among females in the U populations decreased from a starting frequency of 0.756 ± 0.0102 (mean ± SE; $n = 11$) at the first sampling date to an equilibrium frequency of 0.602 ± 0.0115 ($n = 99$; the mean ± SE of the last nine sampling dates after day 368, a date again arbitrarily chosen as equilibrium (Figure 1); this is a mean allele frequency change of about 15% in roughly 13 generations. The individual populations mostly behaved in this fashion but again as with the N populations there were notable exceptions that are analyzed below. The pattern of change of allele frequency was similar among males and among sexes combined both among the populations as a whole as well as individual populations (data not shown).

The reperturbation experiment: The reperturbation started with the offspring of a biased sample of vials on 18 July 1986, day 511 of the experiment. The mean female frequency of the biased sample of vials was 0.0070 ± 0.000041 (mean ± SE). The allele frequency of females among the unbiased sample of vials

TABLE 4

The evolution of frequency-dependent fitness differences between genotypes

	Generation										
	0	2	4	6	8	10	12	14	16	18	20
ΔW_q	0.050	0.046	0.043	0.039	0.036	0.033	0.030	0.027	0.025	0.022	0.020
q	0.750	0.731	0.713	0.695	0.679	0.663	0.649	0.635	0.623	0.611	0.601

The fitness differences (ΔW_q) between homozygous and heterozygous genotypes decreases as the population evolves from an initial frequency of $q_0 = 0.75$ towards an equilibrium frequency of $\hat{q} = 0.50$. The median number of generations of the U populations in the actual experiment is 16.

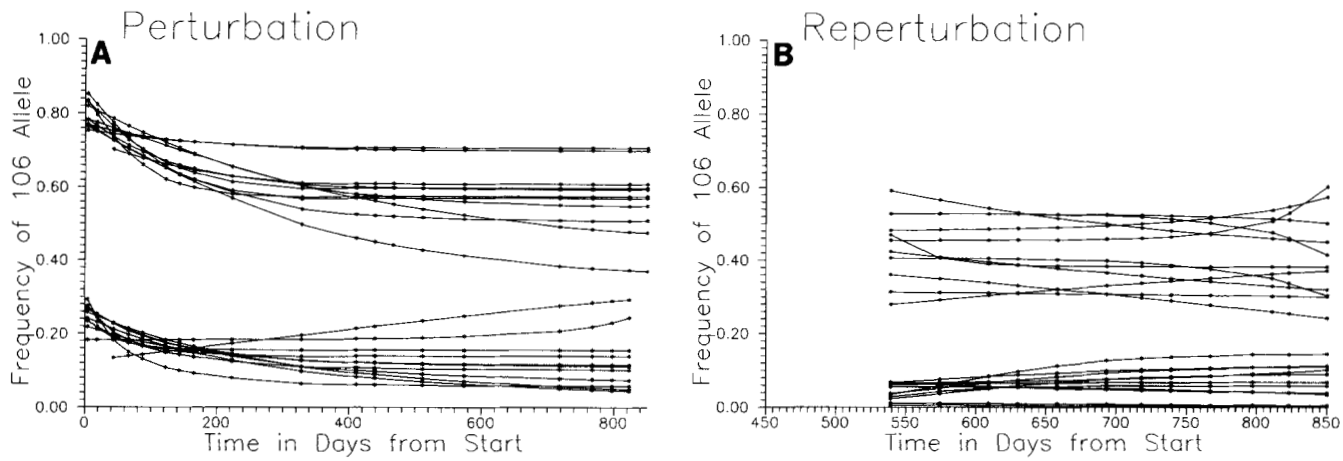

FIGURE 2.—Fitted selection curves of the *106* allele among females in all replicate N and U populations (A) and in all replicate reperturbed EN and EU populations (B). Fitted curves are based on a model of selection by frequency-dependence. See text for further details.

of the N populations was 0.129 ± 0.0174 and this is the best estimate of the allele frequency among the males that had mated with these females and thus of the allele frequency among the sperm carried by females in the biased sample of vials. The method used for the reperturbation thus created a disparity in allele frequency between the sexes. Due to random mating at the *X*-linked *Est-5* locus the disparity is halved every generation (barring selection and random drift). The expected frequency based on random mating is the mean of $\frac{1}{3}$ frequency among males and $\frac{2}{3}$ frequency among females, which in the EN populations yields an expectation of 0.080 ± 0.0116.

The reperturbation successfully altered allele frequencies in the reperturbed EN populations relative to the N population controls (Figure 1). The frequencies in populations EN1 and EN3 almost went to fixation during the reperturbation (Figure 1). Including these two populations in the analysis did not qualitatively alter the results. The mean frequency over all sampling dates of the reperturbed EN populations (Figure 1) was 0.084 ± 0.0048 among females (0.072 ± 0.0063 among males and 0.081 ± 0.0043 among sexes combined, data not shown). The overall mean frequency was thus very similar to the predicted mean equilibrium frequency of 0.080 due to random mating in the absence of selection. The mean frequency of the *106* allele in the reperturbed low populations (Figure 1) leveled off at a mean frequency about 6% below the mean frequency of the N population controls.

In the reperturbed EU populations the mean frequency of the *106* allele among females in the biased sample of vials was 0.354 ± 0.0133. The predicted frequency among males or sperm based on the original unbiased sample of females was 0.58 ± 0.031. The predicted mean frequency at equilibrium due to random mating was 0.502 ± 0.0285.

The reperturbation successfully altered allele frequencies in the reperturbed EU populations relative to the U population controls (Figure 1). The overall mean frequency of the *106* allele over all sampling dates among females (Figure 1) was 0.425 ± 0.0092. The reperturbation thus resulted in a greater downward shift than was predicted based on the biased sample of vials. Overall the frequencies of the *106* allele did not change among the EU populations and by inspection of the raw data (Figure 1) there was no hint of a return to an equilibrium given by the mean of the U populations.

Statistical analysis of selection and neutrality

Power of experiment to detect selection: Under the frequency-dependent model of selection, fitness differences between genotypes are largest at the initial

TABLE 5

X^2 goodness-of-fit test-statistics and probabilities for models of neutrality and selection by frequency-dependent fitnesses in the N and EN populations

	Selection: frequency dependence		Neutrality					Selection *vs.* neutrality	
			Initial frequency		Mean frequency				
Population	${}^{1}X^2_{15}/\hat{C}$	${}^{1}P$	${}^{q_0}X^2_{17}/\hat{C}$	${}^{q_0}P$	${}^{\hat{q}}X^2_{17}/\hat{C}$	${}^{\hat{q}}P$	$\hat{N}_e$	${}^{\hat{q}-1}X^2_2/\hat{C}$	${}^{\hat{q}-1}P$
N1	8.8	0.887	31.0	0.020	19.5	0.298	3	10.7	0.005
N2	13.7	0.549	34.8	0.007	26.1	0.073	6	12.4	0.002
N3	5.8	0.983	76.1	0.000	26.4	0.067	4	20.7	0.000
N4	20.2	0.165	45.9	0.000	25.7	0.080	5	5.5	0.064
N5	16.0	0.383	36.3	0.004	23.2	0.142	3	7.2	0.027
N6	16.8	0.329	28.5	0.040	19.2	0.316	5	2.4	0.302
N7	14.7	0.470	37.0	0.003	28.3	0.042	3	13.5	0.001
N8	10.7	0.772	19.0	0.326	14.5	0.632	3	3.8	0.152
N9	29.4	0.014	45.7	0.000	35.2	0.006	5	5.9	0.054
N10	17.6	0.287	18.5	0.356	18.5	0.356	5	1.0	0.614
NH	9.3	0.747	16.8	0.329	14.6	0.478	7	5.3	0.071
NALL	20.4	0.158	73.9	0.000	38.9	0.002	3	18.5	0.000
	${}^{1}X^2_{6}/\hat{C}$	${}^{1}P$	${}^{q_0}X^2_{11}/\hat{C}$	${}^{q_0}P$	${}^{\hat{q}}X^2_{11}/\hat{C}$	${}^{\hat{q}}P$	$\hat{N}_e$	${}^{\hat{q}-1}X^2_2/\hat{C}$	${}^{\hat{q}-1}P$
EN1	5.7	0.773	5.9	0.879	6.0	0.876	12	0.3	0.864
EN2	5.8	0.763	6.3	0.850	6.3	0.851	12	0.6	0.759
EN3	6.2	0.715	6.6	0.833	6.2	0.862	11	0.1	0.964
EN4	14.5	0.104	14.6	0.202	14.2	0.221	14	0.3	0.849
EN5	11.3	0.255	13.3	0.274	12.8	0.307	12	1.5	0.475
EN6	8.6	0.479	24.5	0.011	20.9	0.034	9	12.4	0.002
EN7	5.7	0.772	8.8	0.641	8.1	0.704	15	2.4	0.296
EN8	11.0	0.278	12.5	0.330	12.2	0.351	18	1.2	0.548
EN9	9.9	0.359	17.2	0.101	15.3	0.167	21	5.4	0.066
EN10	9.3	0.413	17.0	0.107	14.5	0.208	10	5.2	0.074
ENH	11.1	0.268	12.2	0.352	12.4	0.336	15	1.3	0.530
ENALL	12.0	0.212	14.4	0.214	14.3	0.219	6	2.2	0.327

The X^2 deviance statistics have been corrected by the estimated overdispersion factor $\hat{C}$. The P values are the probabilities of the corrected $X^2/\hat{C}$ statistics. N_e was estimated by applying the AMOEBA minimization to the model of neutrality around mean frequency.

frequency and evolve towards zero at equilibrium. For example (Table 4) given an initial frequency $q_0 = 0.75$ and an equilibrium frequency of $\hat{q} = 0.50$ an initial fitness difference of 5% has evolved towards 2.5% at generation 16 which was the median generation of the U populations which served as a baseline for studies on power (approximately day 450 slightly before the reperturbation). Thus a stated power is contingent on the differences in initial and equilibrium allele frequency as well as on the reference generation. Given a true fitness difference ΔW_{q_0} at q_0 ranging from 0.5% to 10% the simulations showed that the selection model always gave a very good fit (Table 3). For fitness differences of 0.5–1% the neutral model of drift around the mean frequency provided an equally good fit to the selection model as measured by the probability of the mean deviance statistic (Table 3). The probability for the neutral model dropped off as the true fitness difference increased and the model is rejected at the 5% level given a true fitness difference of 9–10%. Furthermore, selection was a better model at 3% true fitness difference (${}^{\hat{q}-s}X^2_2 = 16 - 10.1 = 5.9$, $P \approx 0.05$; Table 3), or from a different vantage point, selection is a preferred model at a fitness difference at the median generation of about 1.5%. The model of neutrality around the estimated initial frequency fares much worse and its probability drops off at a more rapid rate (Table 3). If another criterion for power is used, such as the percentage of statistics exceeding a nominal 5% level of significance, the same qualitative difference between the models is found although this criterion does not drop off as rapidly with increasing fitness difference as does the probability of the mean deviance goodness-of-fit criterion. Overall, therefore, the design is expected to be able to reveal selective differences of a few percent.

Models of selection and neutrality and goodness-of-fit: In the perturbation experiment there was, overall, evidence of strong and significant selection against the *106* allele with attainment of equilibria (Figure 2; Tables 5 and 6). Fitnesses at the initial frequency were very similar and significantly different from neutrality for the N and U populations pooled data ALL or 0.82 and 0.83, respectively (Tables 5–8). The larger overall

TABLE 6

X^2 goodness-of-fit test-statistics and probabilities for models of neutrality and selection by frequency-dependent fitnesses in the U and EU populations

	Selection: frequency dependence		Neutrality					Selection *vs.* neutrality	
			Initial frequency		Mean frequency				
Population	$^{s}X^2_{16}/\hat{C}$	^{s}P	$^{90}X^2_{18}/\hat{C}$	^{90}P	$^{\bar{q}}X^2_{18}/\hat{C}$	$^{\bar{q}}P$	$\hat{N}_e$	$^{\bar{q}-s}X^2_2/\hat{C}$	$^{\bar{q}-s}P$
U1	13.9	0.606	15.4	0.631	14.5	0.693	5	0.6	0.723
U2	25.3	0.065	27.8	0.065	26.3	0.094	6	1.0	0.616
U3	18.5	0.295	41.6	0.001	33.8	0.013	6	15.3	0.000
U4	7.5	0.962	35.7	0.008	22.7	0.204	5	15.1	0.001
U5	15.9	0.458	81.5	0.000	39.0	0.003	4	23.1	0.000
U6	23.4	0.105	90.8	0.000	39.9	0.002	4	16.6	0.000
U7	19.7	0.234	62.5	0.000	45.2	0.000	7	25.5	0.000
U8	14.9	0.530	35.2	0.009	22.2	0.221	5	7.3	0.026
U9	4.5	0.998	34.4	0.011	18.9	0.401	3	14.4	0.001
U10	11.1	0.806	56.1	0.000	22.8	0.197	5	11.8	0.003
UH	19.3	0.154	27.9	0.033	21.6	0.156	7	2.3	0.316
UALL	15.8	0.465	98.4	0.000	50.8	0.000	4	34.9	0.000
	$^{s}X^2_9/\hat{C}$	^{s}P	$^{90}X^2_{11}/\hat{C}$	^{90}P	$^{\bar{q}}X^2_{11}/\hat{C}$	$^{\bar{q}}P$	$\hat{N}_e$	$^{\bar{q}-s}X^2_2/\hat{C}$	$^{\bar{q}-s}P$
EU1	7.5	0.589	13.4	0.269	11.6	0.391	21	4.2	0.124
EU2	6.9	0.647	7.1	0.795	7.0	0.797	14	0.1	0.943
EU3	6.8	0.654	9.3	0.597	8.7	0.648	10	1.9	0.391
EU4	8.0	0.533	10.5	0.482	10.5	0.488	10	2.5	0.291
EU5	7.5	0.589	28.4	0.003	9.6	0.564	12	2.2	0.340
EU6	6.9	0.651	10.8	0.456	10.5	0.484	15	3.7	0.161
EU7	7.5	0.586	11.4	0.413	10.9	0.455	12	3.4	0.186
EU8	12.6	0.184	12.6	0.320	12.7	0.315	15	0.1	0.941
EU9	10.0	0.354	12.9	0.303	12.2	0.348	9	2.3	0.324
EU10	10.0	0.348	12.5	0.324	12.5	0.325	15	2.5	0.287
EUH	15.4	0.081	17.2	0.101	17.2	0.102	11	1.8	0.406
EUALL	13.7	0.132	15.3	0.171	14.9	0.189	7	1.1	0.568

See Table 5 for explanation of symbols.

shift in allele frequency among the U populations (Figure 1) was thus not due to greater fitness differences but rather due to the higher genetic variance at those frequency ranges. In the reperturbation on the other hand, there was no overall significant selection (Figure 2; Tables 5–8) nor was there evidence for a return to the equilibria of control populations.

The detailed analysis of the individual replicates bears out the overall picture as revealed by the selection curves, standard errors and goodness-of-fit statistics for the individual populations (Figure 3; Tables 5–8). In the perturbation experiment selection by the frequency-dependent model gave in most replicates a better fit than neutrality around the estimated initial frequency or around the mean frequency. The former neutral model gave in most instances a very poor fit. The neutral model around mean frequency gave in two instances (N10 and U1) a better fit than the selection model, but considered as a whole selection was a better model in the perturbation.

In the reperturbation this picture is reversed. Both neutral models were viable alternatives to the model of selection and gave a superior fit in many instances. (Figure 2; Tables 5–8). Some of the EN populations seemed to return to the equilibria that they were perturbed from (*e.g.*, EN9), but in spite of some increase in allele frequency early on in ENALL the overall frequency levelled off at an equilibrium frequency of 7%, or 5% below the equilibrium of the NALL control (Figure 3; Table 7). A greater variability in the direction of estimated changes was evident in the EU reperturbation, which showed no indication of a return to equilibria of control populations.

With notable exceptions the selection curves of the constant fitness model were in most instances confluent with the curves of the model of selection by frequency-dependence (Figure 3 dashed lines) and the goodness of fit statistics were almost identical (data not shown). The exceptions (replicates N10, U2 and U6) in which the two selection models do not agree were best described as showing a decrease and a subsequent increase of allele frequency through the same range of frequencies by simple inspection of the data points. Under the selection models, however, the fitted curves can only be monotone and this explains the lack of fit (Tables 5, 6; see for example replicate

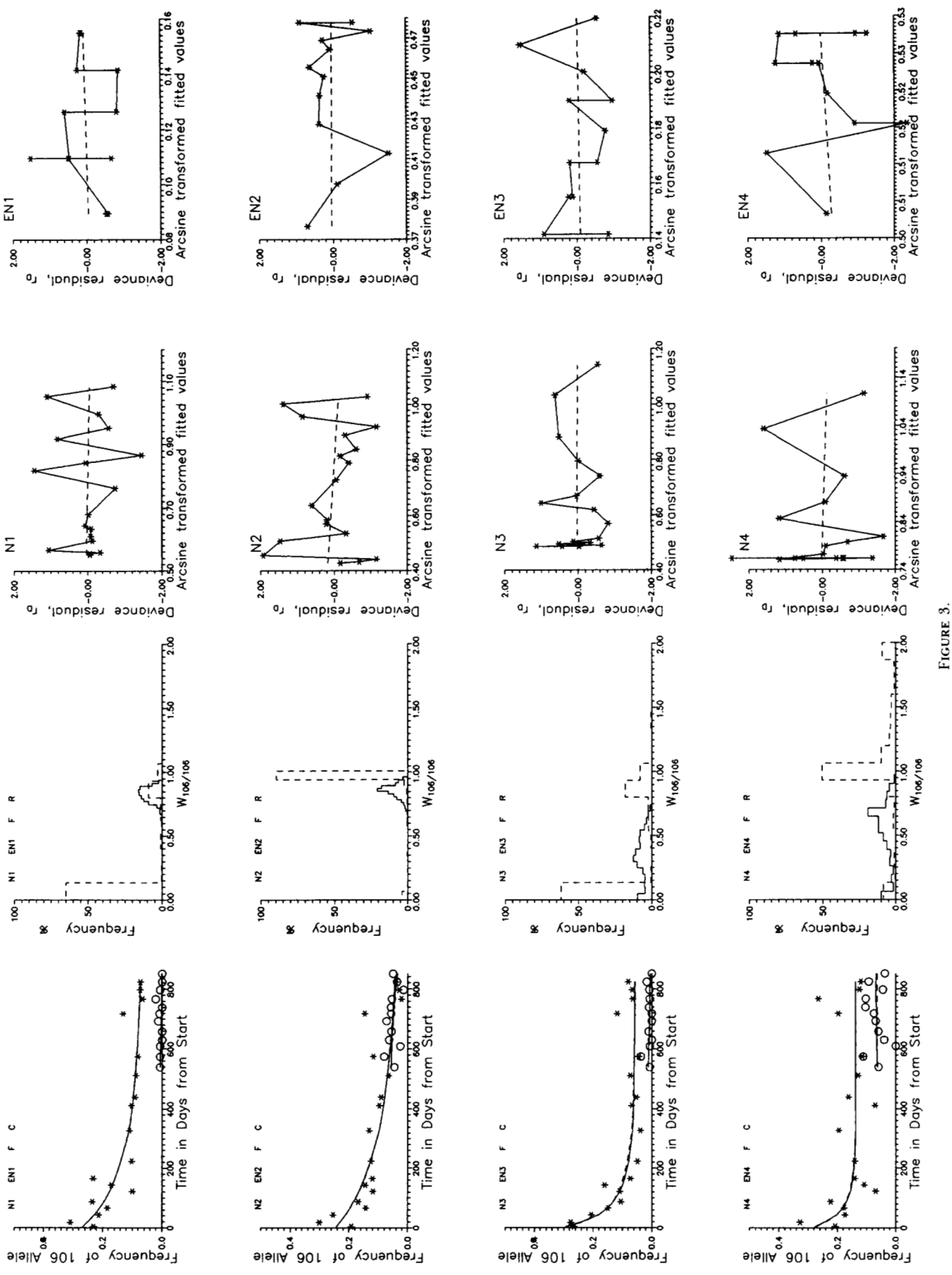
N1 EN1 F C
Frequency of 106 Allele
Time in Days from Start
N1 EN1 F R
Frequency %
W106/106
N1
Deviance residual, r_D
Arcsine transformed fitted values
EN1
N2 EN2 F C
N2 EN2 F R
N2
EN2
N3 EN3 F C
N3 EN3 F R
N3
EN3
N4 EN4 F C
N4 EN4 F R
N4
EN4

FIGURE 3.

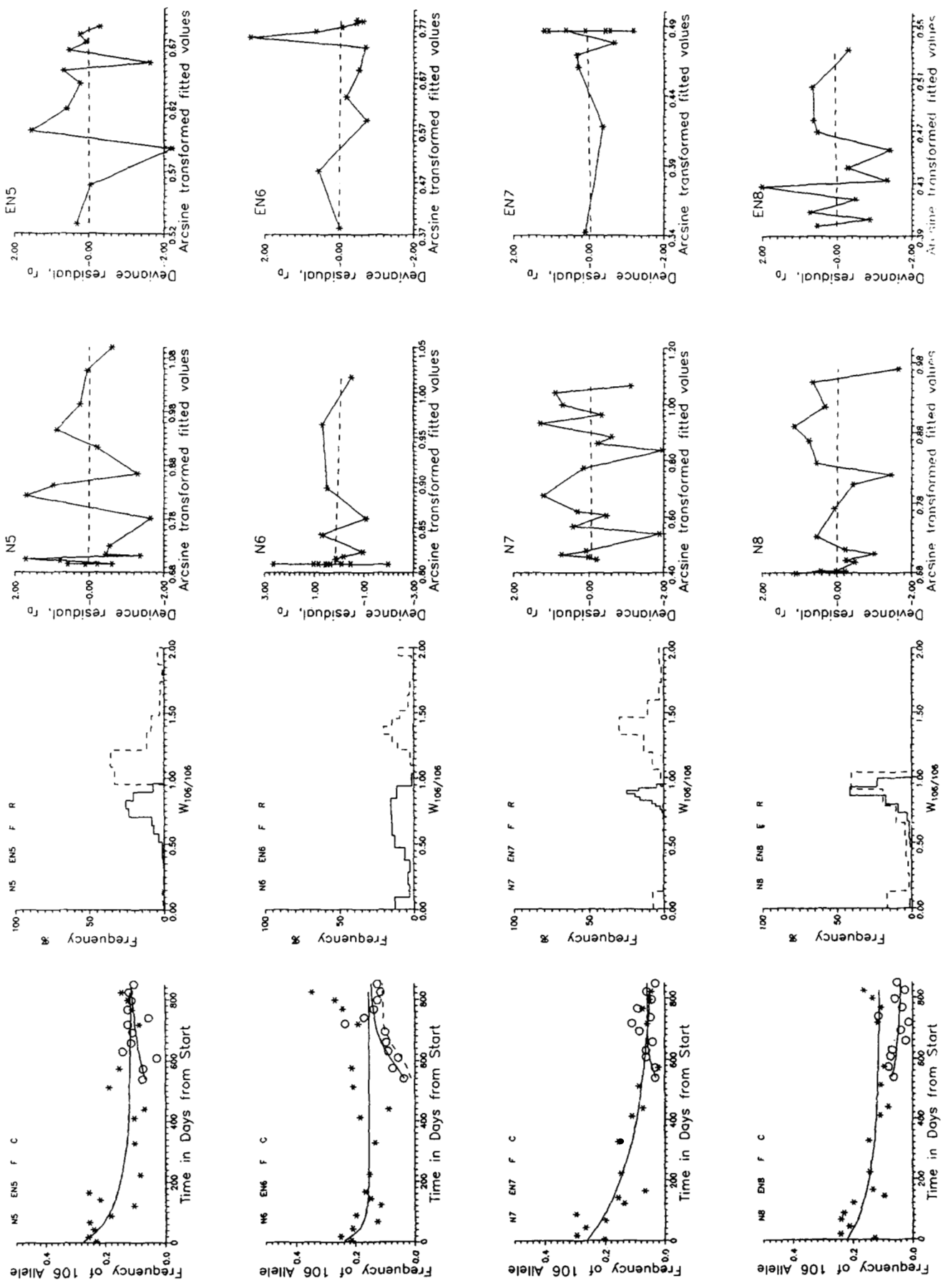

FIGURE 3.—Continued

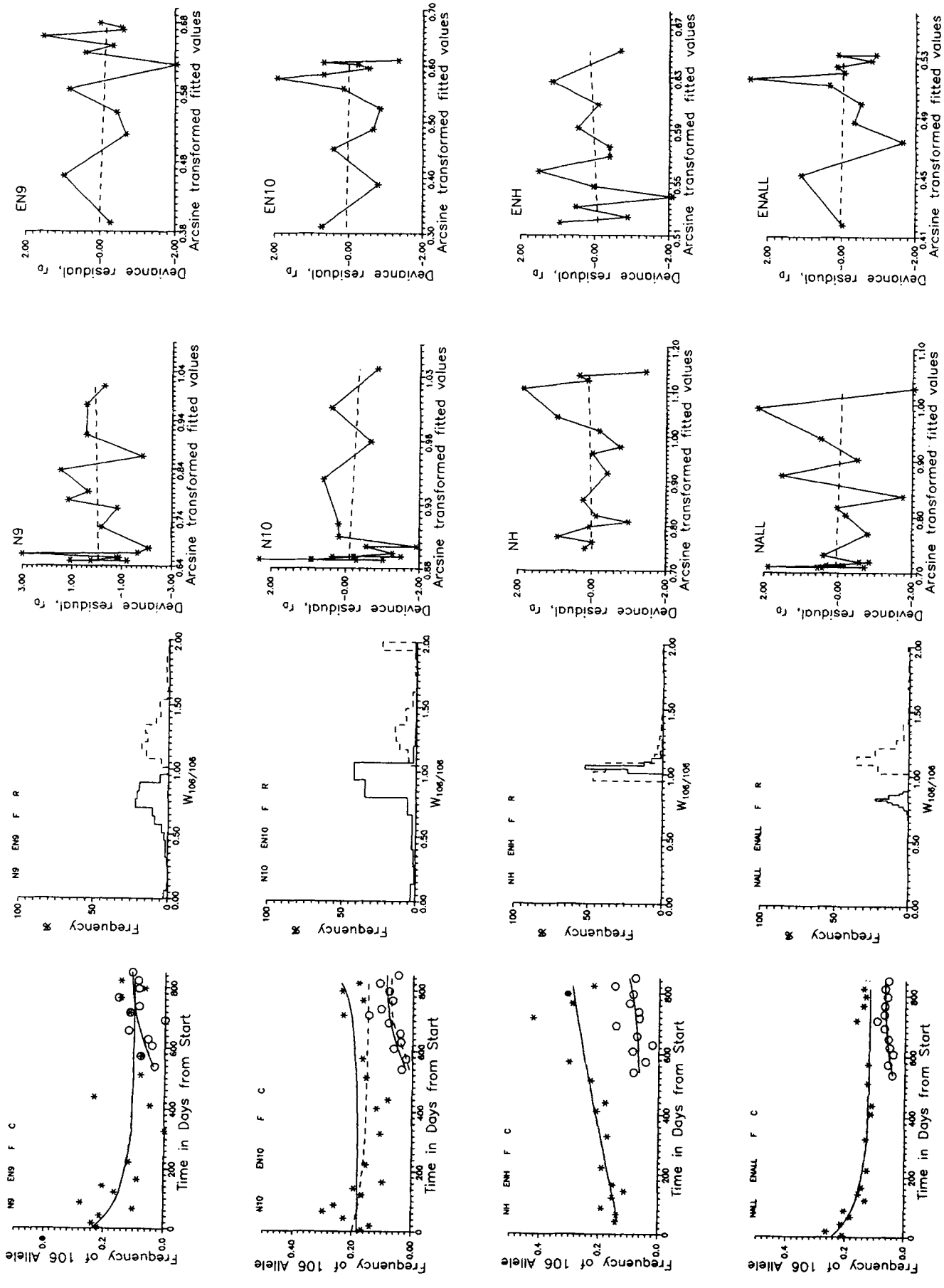

FIGURE 3.—Continued

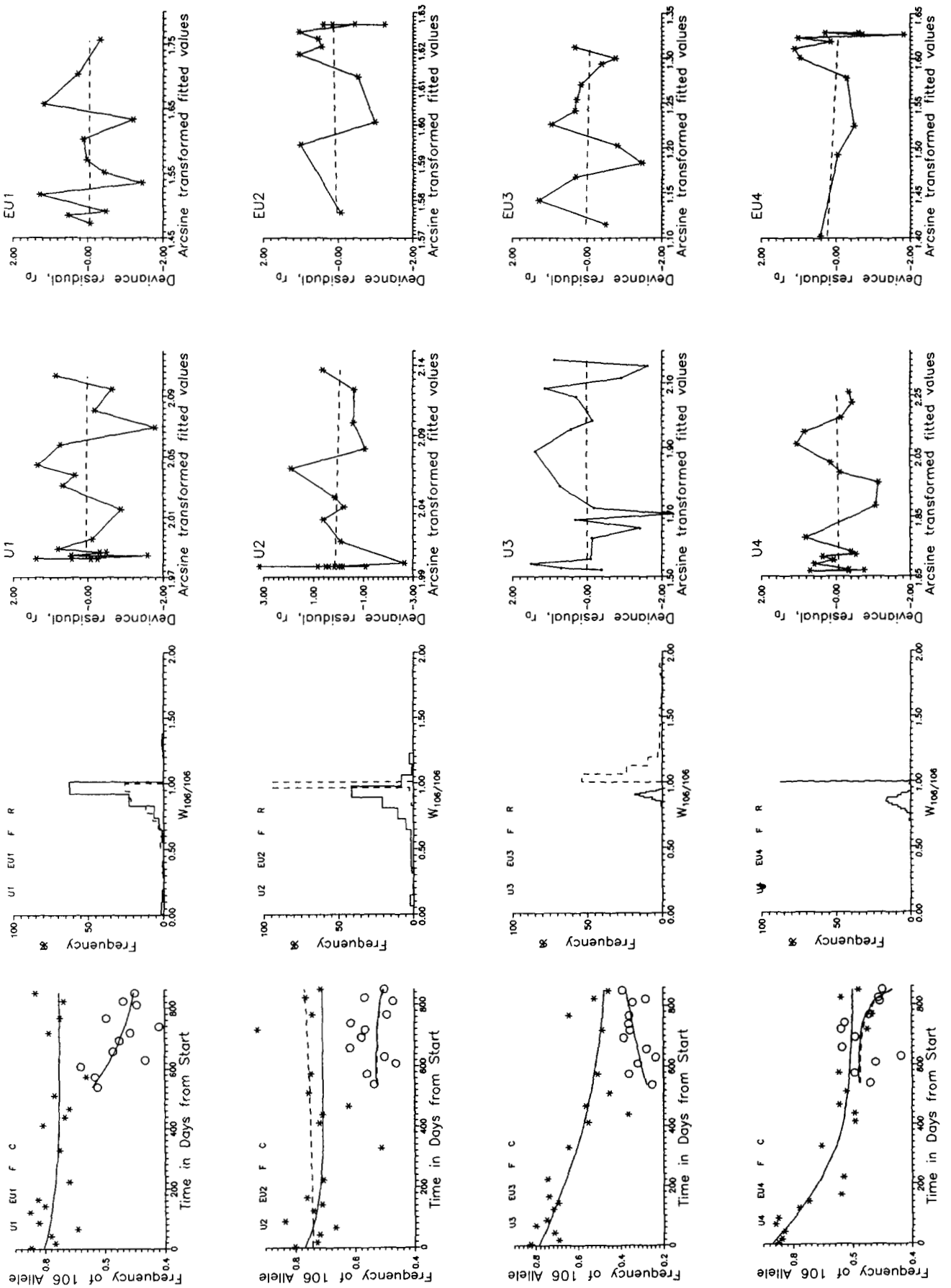

FIGURE 3.—Continued

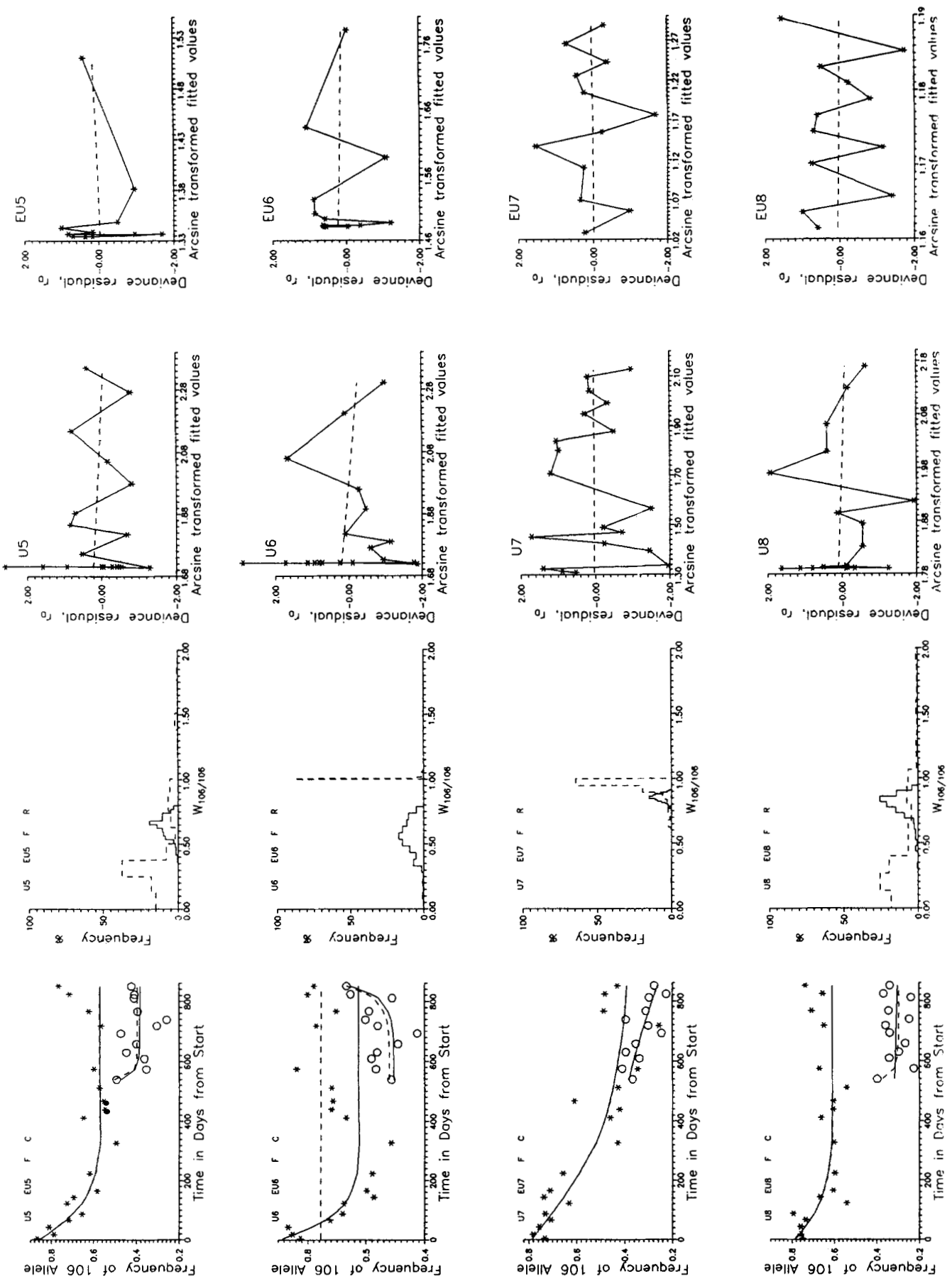

FIGURE 3.—Continued

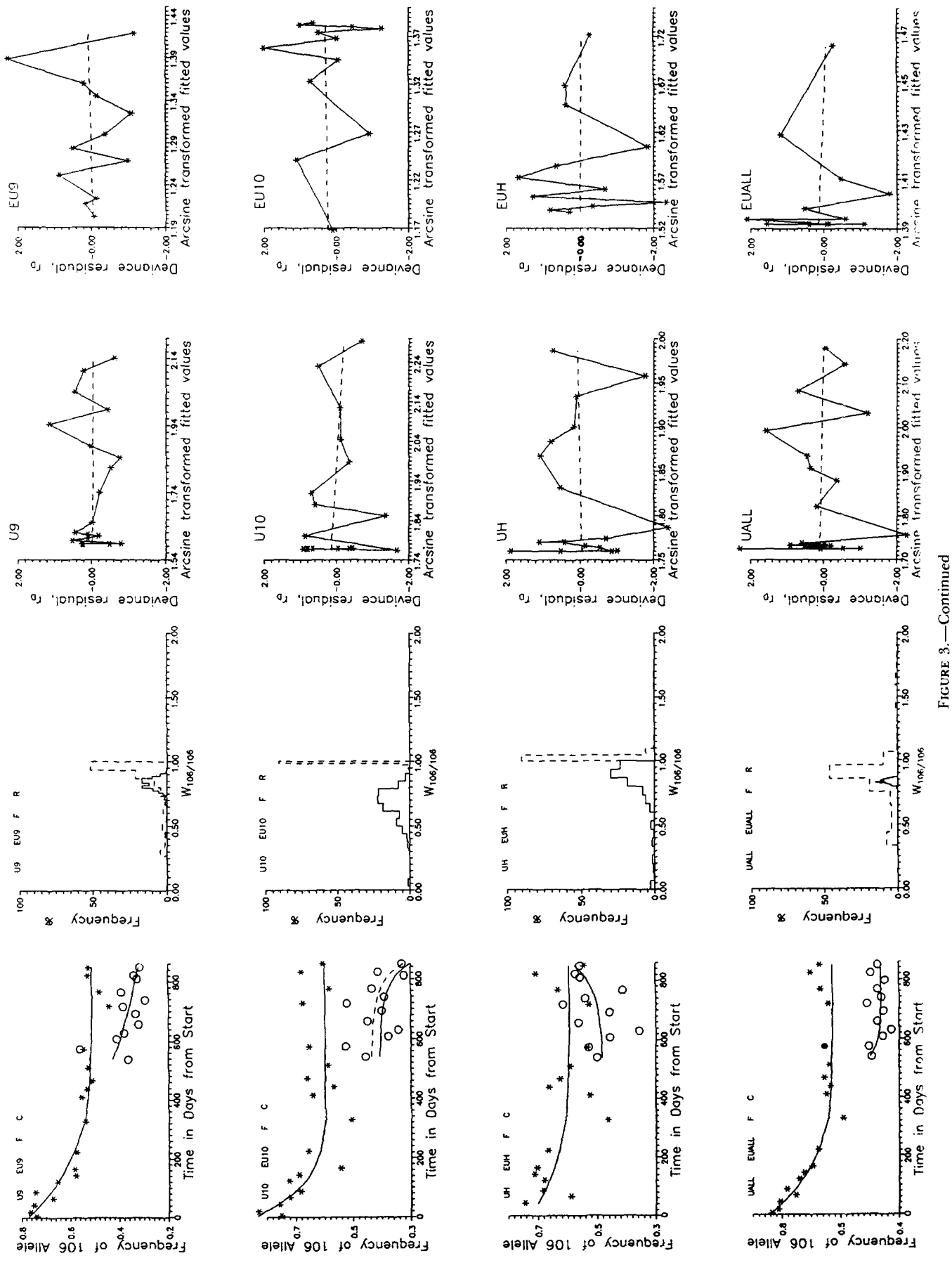

FIGURE 3.—Continued

N9 in which no model gave a satisfactory fit). As an example, inspection of the actual data points of N6 (Figure 3) revealed the pattern of initial decrease of allele frequency with a subsequent increase and a quadratic ($\hat{Y}_t = 18.35 - 3.18d_t + 4.00d_t^2$) gave a best-fit polynomial regression of arcsine transformed frequencies on time (SOKAL and ROHLF 1981) explaining 62% of the variation; and in U6 a cubic ($\hat{Y}_t = 30.36 - 611d_t + 5.63d_t^2 - 5.05d_t^3$) gave a best fit polynomial with 75% of variation explained. This suggests that genetic models allowing nonmonotone allele frequency changes (multiallele or multilocus models with quadratic or higher order terms) would improve fit. This was not formally tested here.

Overdispersion and heterogeneity of theory and data: The heterogeneity of data and theory over and above the overdispersion (Table 9) provides a formal analysis of the individuality of the populations. The two different estimates of overdispersion parameters, ${}_H\hat{C}$ and ${}_W\hat{C}$ were numerically very similar within all groups of replicates but differed slightly between groups. The pooled data of the perturbation showed significant marginal effects whereas the reperturbation did not (population ALL in Tables 5, 6, 9). In addition, there was significant heterogeneity in the local theoretical fit between both the N and the U perturbations (${}_B X^2/\hat{C}$ in Table 9). Thus the selective effects in the N and U populations were to some extent unique to each population even though all replicates within a group were founded from the same egg-laying cage. The heterogeneity was not due to the H populations (which were founded differently) on the one hand and the rest of the populations on the other hand, for although NH was different from the majority of the N replicates, the UH was typical of the majority of U replicates. The heterogeneity of selective effects were not significant in the reperturbation and thus the unique effects of the perturbation seemed to have dissipated and new ones were not generated in the bottleneck of the reperturbation.

Residuals: The standardized or normalized deviance residuals, r_D, plotted against the arcsine transformed fitted values revealed an adequate fit to the model overall (Figure 3) and thus the adequacy of the variance function. Although there were rare exceptions such as EU8 which by inspection of the figure showed an apparent negative temporal correlation of residuals, the details of the individual populations did not qualitatively change the overall view of no systematic departure from the model. The overall mean was close to zero and the linear regressions not significant; most values lie in the range from −2 to +2 with occasional values between −3 and +3, especially in ill-behaved populations such as U6 (Figure 3). There were other exceptions to this overall pattern in the group of populations showing a decrease/increase cycle (*e.g.*, N10 and U6) in which the pattern of residuals implied that a quadratic term would improve fit; this reiterates the point made earlier.

Robustness of estimators to variation in effective number: Applying the AMOEBA minimization to the neutral model of drift around the mean frequency gave estimates of $\hat{N}_e$ less than ten in the perturbation and less than 20 in the reperturbation (Tables 5, 6), which simply indicates that the large observed allele frequency differences due to selection have been converted to drift effects and expressed in the currency of N_e. Rather than estimating effective population sizes, a more useful approach is to ask what effect different realistic values of N_e would have on the estimates; therefore the estimator was run using six different values of $N_e = \{300, 600, 1200, 2400, 4800, 9600\}$ for each replicate. Different values of N_e had no significant effects on the estimates of fitness (Figure 4). If an N_e less than sample size was used, however, the estimator yielded negative fitnesses which are not biologically feasible (results not shown). Having established the robustness of estimates to differences in realistic values of N_e an $N_e = 1200$ was used for all statistics reported here to make them comparable between replicates and categories. This value was chosen by noting that the population cages supported actual populations of 3000–5000 and assuming an effective size of $\frac{1}{3}$ actual size as a reasonable figure. The results were robust in any case.

Summary: In summary there was evidence for strong selection against the *106* allele in the original perturbation experiment (N and U) and no evidence for these same selective effects in the reperturbation (EN and EU). Individuality in the evolution of the populations is a striking feature of the experiment and is shown clearly by the significant heterogeneity among the populations in the perturbation. Tests of selection show adequacy of models and variance func-

FIGURE 3.—Actual data points, fitted selection curves, fitness distributions and deviance residuals of all individual replicates. In the figure on the left side (first column) of each row are the actual observed frequencies given by an asterisk (*) for the perturbation and an open circle (O) for the reperturbation. Also given in this figure are the fitted selection curves under the model of selection by frequency-dependent fitnesses (solid lines) and under the model of selection by constant fitnesses (dashed lines); the curves are confluent in most instances. In the second figure from the left (second column) are the distributions of fitness from the 200 randomly generated data sets; solid lines: perturbation; dashed lines: reperturbation. In the third and fourth figure in each row (third and fourth column) are the standardized deviance residuals plotted against the arcsine transformed fitted frequencies in the corresponding perturbation (third column) and corresponding reperturbation (fourth column); solid lines connect data points in time series; dashed lines are least squares linear regressions. See text for further details.

TABLE 7

Estimated parameters in the N and EN populations

	Fitness at q_0			Initial frequency			Equilibrium frequency		
Population	W_{106}	${}_R\bar{W}_{106}$	$\sqrt{\hat{C}}\mathrm{SD}_{RW106}$	q_0	${}_R\bar{q}_0$	$\sqrt{\hat{C}}\mathrm{SD}_{Rq0}$	$\hat{q}$	${}_R\bar{\hat{q}}$	$\sqrt{\hat{C}}\mathrm{SD}_{R\hat{q}}$
N1	0.83	0.82	0.081	0.27	0.27	0.030	0.06	0.06	0.096
N2	0.87	0.85	0.075	0.25	0.25	0.029	0.00	0.06	0.277
N3	0.37	0.33	0.246	0.31	0.31	0.046	0.06	0.05	0.011
N4	0.60	0.53	0.329	0.29	0.29	0.052	0.14	0.14	0.021
N5	0.79	0.76	0.171	0.28	0.28	0.044	0.11	0.11	0.029
N6	0.69	0.56	0.383	0.24	0.25	0.042	0.15	0.15	0.014
N7	0.88	0.87	0.062	0.26	0.26	0.029	0.00	0.03	0.186
N8	0.89	0.85	0.183	0.22	0.22	0.036	0.11	0.11	0.070
N9	0.79	0.72	0.235	0.24	0.25	0.035	0.10	0.10	0.025
N10	0.95	0.89	0.370	0.20	0.20	0.042	0.15	0.15	0.093
NH	1.05	1.06	0.058	0.13	0.12	0.025	0.41	0.33	0.323
Mean	0.79	0.75	0.199	0.24	0.25	0.037	0.11	0.12	0.104
SE	0.06	0.06	0.037	0.01	0.02	0.003	0.04	0.02	0.033
NALL	0.82	0.82	0.047	0.25	0.25	0.011	0.12	0.12	0.006
EN1	0.91	0.65	1.013	0.01	0.04	0.134	0.02	0.13	0.300
EN2	0.99	0.99	0.233	0.06	0.06	0.027	0.06	0.05	0.023
EN3	0.94	0.90	1.006	0.01	0.01	0.057	0.03	0.06	0.232
EN4	1.02	1.05	0.570	0.06	0.06	0.050	0.07	0.07	0.068
EN5	1.11	1.20	0.313	0.06	0.05	0.033	0.12	0.12	0.138
EN6	1.39	1.42	0.195	0.02	0.02	0.015	0.15	0.15	0.041
EN7	1.42	1.36	0.558	0.02	0.05	0.158	0.06	0.06	0.013
EN8	0.89	0.71	0.456	0.08	0.09	0.055	0.03	0.06	0.110
EN9	1.26	1.30	0.223	0.03	0.03	0.015	0.11	0.13	0.185
EN10	1.33	1.37	0.213	0.02	0.01	0.014	0.09	0.09	0.050
ENH	1.02	1.05	0.145	0.06	0.06	0.017	0.05	0.07	0.074
Mean	1.12	1.09	0.448	0.04	0.04	0.052	0.07	0.09	0.112
SE	0.06	0.08	0.095	0.01	0.01	0.015	0.01	0.01	0.028
ENALL	1.14	1.18	0.193	0.04	0.04	0.010	0.07	0.07	0.024

Estimates of frequency-dependent fitness (W) of the *106/106* homozygote at the initial frequency, q_0, and estimates of the initial (q_0) and equilibrium ($\hat{q}$) frequencies of the *106* allele in the original N and EN replicates. Also given are the means of these estimates in Monte Carlo data sets (R) generated using the estimated fitnesses assuming random genetic drift in a population of size $N_e = 1200$ and binomial errors of sampling with the actual sample sizes of the respective population. Also given are the standard deviation (SD) of the distribution of estimates of the randomly generated data sets scaled by $\sqrt{\hat{C}}$, a correction factor obtained from the estimated overdispersion parameter. The mean and standard error (SE) over replicates 1–10 and H are given along with a fictitious population ALL which is an unweighted sum of all samples.

tion and the results are robust with respect to variation in effective population number.

DISCUSSION

Power: An experiment designed to detect selection over hitchhiking and neutrality, like the one presented here, must specify its statistical power to detect a selective difference. The power of a statistical test is a function of the true fitness difference as well as the variance. The simulations assuming binomial variance reveal a power to detect selection of a few percentage points depending on which criterion and vantage point of allele frequency and generation one uses. This is less power, albeit not much, than the simulated power of the *Escherichia coli* bacterial chemostat (DYKHUIZEN and HARTL 1980). The realized power of a statistical test is, however, a function of the actual variance of the experiment and not the nominal variance of an assumed distribution. Because overdispersion is the rule rather than an exception (MCCULLAGH and NELDER 1989) the power of the method to detect selection should also be assessed by taking overdispersion into account. Thus in this experiment the estimated standard error of fitness at q_0 adjusted by overdispersion is 4.8% in the NALL population and as low as 2.6% in the UALL population. These low standard errors in the U populations, which result from having a large number of replicates, a large number of sampling dates, and a large number of individuals sampled, bring the realized power to the level of the simulated power.

Requirements of a method: The results of this study have implications about the requirements of a method. First, due to autocorrelation, directionality of frequency changes cannot be judged in a single or a few populations. Directionality must be judged in an ensemble population. Second, due to linkage, a population that reaches an equilibrium must be reper-

TABLE 8

Estimated parameters in the U and EU populations

Population	Fitness at q_0			Initial frequency			Equilibrium frequency		
	W_{106}	${}_R\bar{W}_{106}$	$\sqrt{\hat{C}SD}_{RW106}$	q_0	${}_R\bar{q}_0$	$\sqrt{\hat{C}SD}_{Rq_0}$	$\hat{q}$	${}_R\bar{\hat{q}}$	$\sqrt{\hat{C}SD}_{R\hat{q}}$
U1	0.96	0.91	0.224	0.76	0.76	0.042	0.70	0.70	0.099
U2	0.93	0.85	0.262	0.77	0.78	0.040	0.71	0.71	0.036
U3	0.90	0.90	0.030	0.79	0.79	0.020	0.43	0.42	0.095
U4	0.85	0.84	0.052	0.82	0.83	0.025	0.55	0.54	0.032
U5	0.64	0.64	0.107	0.86	0.86	0.028	0.57	0.57	0.021
U6	0.58	0.55	0.108	0.84	0.85	0.042	0.58	0.58	0.022
U7	0.86	0.85	0.037	0.77	0.77	0.024	0.34	0.34	0.068
U8	0.81	0.79	0.127	0.79	0.79	0.031	0.61	0.61	0.025
U9	0.83	0.83	0.064	0.77	0.78	0.028	0.51	0.50	0.029
U10	0.70	0.66	0.186	0.84	0.85	0.045	0.60	0.60	0.023
UH	0.90	0.80	0.289	0.73	0.77	0.096	0.59	0.60	0.078
Mean	0.82	0.78	0.142	0.79	0.80	0.038	0.56	0.56	0.048
SE	0.04	0.04	0.028	0.01	0.01	0.006	0.03	0.03	0.009
UALL	0.83	0.83	0.024	0.79	0.79	0.009	0.58	0.58	0.009
EU1	0.90	0.85	0.188	0.62	0.63	0.082	0.40	0.43	0.256
EU2	1.00	0.99	0.068	0.53	0.53	0.037	0.53	0.53	0.039
EU3	1.05	1.12	0.240	0.27	0.25	0.074	0.49	0.35	0.209
EU4	1.00	1.00	0.005	0.52	0.53	0.023	0.52	0.53	0.028
EU5	0.29	0.38	0.403	0.68	0.67	0.197	0.38	0.38	0.038
EU6	1.00	1.01	0.018	0.47	0.47	0.026	0.47	0.47	0.111
EU7	0.96	0.94	0.121	0.37	0.38	0.060	0.50	0.34	0.256
EU8	0.29	0.44	0.511	0.55	0.55	0.260	0.29	0.29	0.028
EU9	0.95	0.86	0.258	0.44	0.46	0.103	0.12	0.38	0.236
EU10	1.00	0.99	0.027	0.43	0.44	0.031	0.43	0.43	0.079
EUH	1.00	1.03	0.128	0.48	0.47	0.051	0.48	0.49	0.120
Mean	0.86	0.87	0.179	0.49	0.49	0.086	0.42	0.42	0.127
SE	0.09	0.07	0.049	0.03	0.03	0.023	0.04	0.02	0.029
EUALL	0.88	0.84	0.289	0.48	0.49	0.086	0.41	0.41	0.020

See Table 7 for explanation of symbols.

TABLE 9

X^2 deviance statistics, degrees of freedom and variance inflation factors, $\hat{C}$, from estimates of selection by constant fitnesses in all groups of populations

Populations	Total		Between		Within			Marginal		Heterogeneity		
	${}_TX^2$	${}_Tdf$	${}_BX^2$	${}_Bdf$	${}_WX^2$	${}_Wdf$	${}_W\hat{C}$	${}_MX^2$	${}_Mdf$	${}_HX^2$	${}_Hdf$	${}_H\hat{C}$
N	367.6	193	83.9	30	290.9	163	1.78	36.4	15	335.2	178	1.88
EN	205.7	129	54.8	30	151.5	99	1.53	18.4	9	186.8	120	1.56
U	417.0	204	83.1	30	311.4	174	1.79	28.3	16	392.5	188	2.09
EU	238.8	129	71.5	30	167.5	99	1.69	23.2	9	215.0	120	1.79

See Table 2 for definitions of the deviance statistics.

turbed away from that equilibrium to test whether the first approach to the equilibrium is due to linkage disequilibria that have dissipated or whether the equilibrium is actively maintained by selection. It is a test of the dynamic sufficiency (LEWONTIN 1974) of a single locus perturbation experiment. A reperturbation is accompanied by a reduction in population size which means that linkage disequilibria are generated anew. Any new linkage equilibria will be random in sign, however, and will mimic drift in its effects on the locus under study. Therefore directionality in an ensemble reperturbed population cannot be due to such secondarily generated linkage disequilibria. The method therefore demands a reperturbation of a number of independent populations. But since newly generated linkage disequilibria are not necessarily equal in magnitude, the judgement of overall directionality has to take into account the individuality of the populations. The requirements of a method of perturbation-reperturbation have not always been appreciated. Some of the re-extractions of populations for studying inversion polymorphisms are done with

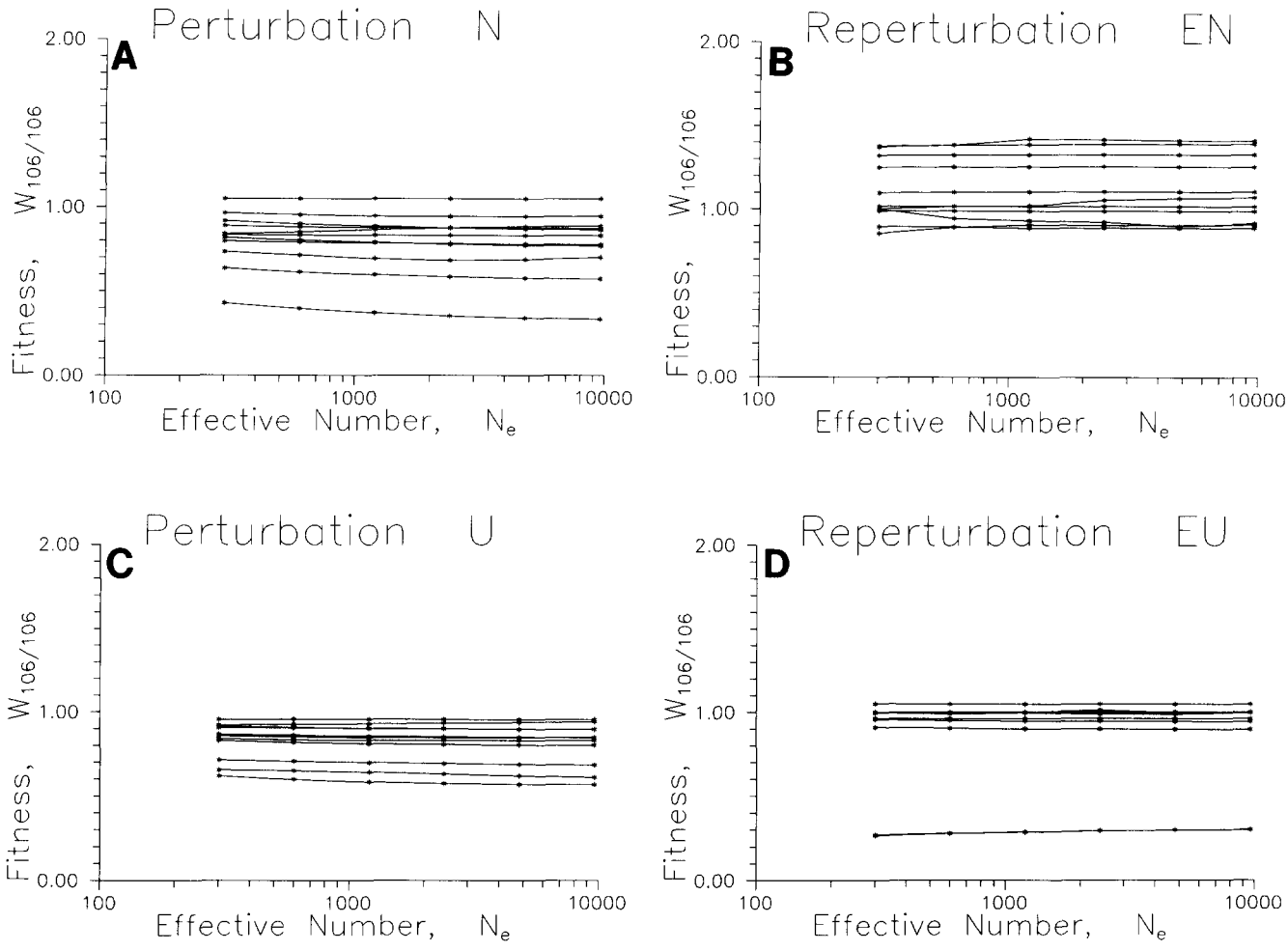

FIGURE 4.—Estimated fitness as a function of assumed effective numbers in all perturbation and reperturbation replicates.

few individuals (SPIESS 1966) or few populations (LEWONTIN 1958). In this study, however, these requirements are met.

The test of selection *vs.* hitchhiking: The frequencies of the *106* allele decrease overall in both the U and N populations and reach equilibria of approximately 60% and 14% respectively 200–300 days into the experiment. The date of attainment of equilibrium for the populations as a whole is arbitrarily defined for it varies among the populations in each category. The decrease is due to natural selection. Nevertheless, the results of the perturbation do not allow a choice to be made between the rival selectionist and neutralist explanations because both provide reasonable explanations for the directional change of allele frequencies and the attainment of equilibria. A selectionist explanation is that balancing selection is operating. One possibility is that there are two equilibria at 60 and 14% and that the initial starting frequencies of the populations were on the upper side of the domains of attraction of both equilibria. Another possibility is a single equilibrium around 14% and that the U populations have slowed down due to a shallow fitness potential around a frequency of 60%. This would imply a frequency-dependent fitness potential. A neutralist explanation is that in spite of the multiple and independently derived founding lines the *106* allele is initially preferentially associated with some negative selective force, an association that has broken up leaving the two *Est-5* alleles to drift at neutral equilibria.

In the reperturbation, which directly addresses this question of balancing selection or dissipated linkage disequilibria and subsequent neutrality, the frequencies in the EN and EU populations do not change overall in a directional way and no selection is detectable among the pooled data. Thus balancing selection is rejected as an explanation for the initial decrease of allele frequency among the perturbation N and U populations and for the maintenance of equilibria. Therefore, the initial selection is due to the generation of linkage disequilibria in the initial egglaying population–linkage effects that have mostly dissipated after 200–300 days leaving the variants to drift at neutral equilibria. The nature and origin of the dissipated selective effect can only be speculated on. It may have

come from the James Reserve population and be an integral part of the chromosomes used. Or it may have come from Furnace Creek or Strawberry Canyon having entered the genetic material in the process of making isogenic lines. Therefore, within the limits of power set by the experimental design to detect selection the *100* and the *106* alleles of *Est-5* are neutral with respect to each other under the environmental conditions of the laboratory populations.

The *Est-5* polymorphism: Of all the alleles known at *Est-5* (COYNE, FELTON and LEWONTIN 1978; KEITH 1983) the *100* and the *106* alleles are *a priori* the best candidates as alleles of a balanced polymorphism because they are widespread and common in natural populations. The implication of this study is that they are neutral or iso-fitness alleles. Naive pan-selectionists may argue that these two alleles would be selected under natural environmental conditions or under a different set of laboratory conditions from the ones used here. To the proponents of such hypotheses I can only recommend the method of perturbation-reperturbation used in this work.

Although neutrality of these two alleles is a likely explanation a saving statement can be made for selection by assuming that the monitored *Est-5* variation is participating in epistatic interactions with other loci and not simply being carried along. The evident dynamic insufficiency (LEWONTIN 1974) of single locus theory to explain the behavior of some populations can be due to two reasons. First, there may have been hidden heterogeneity among the original *100* and *106* lines used not revealed by the sequential electrophoresis of KEITH (1983) or variation at the DNA level (KREITMAN 1983), so that there were multiple alleles in the material; this could account for the heterogeneity. Second, there may be epistatic interactions of the *100* and *106* variants with control regions or the *Est-5* gene or larger parts of the region around the *Est-5* locus. Of course there may not be a difference between these explanations of multiple alleles and epistasis. That is a question of the unit of selection which this experiment does not resolve.

The problem remains to explain the functional significance, if any, of the other alleles at the *Est-5*, and the fitness relationship of these two iso-fitness alleles (ÁRNASON 1982), *100* and *106*, to all the other alleles found in natural populations (KEITH 1983; COYNE, FELTON and LEWONTIN 1978). Are all or most of the other alleles heterotic (LEWONTIN, GINZBURG and TULJAPURKAR 1978), are they all neutral or slightly deleterious (OHTA 1973; KING and OHTA 1975; KIMURA 1983), is there only purifying selection (KEITH 1983), or are the alleles clustered into fewer iso-fitness alleles that are maintained with some form of balancing selection—is there neutrality within balanced alleles (ÁRNASON 1982).

At the present time one cannot intelligently advance testable predictions about the fitness relationships of the numerous alleles at the *Est-5* locus. What is needed is information, preferably DNA sequence data, that can be used to reveal the geneology of all the alleles. Only with such information can one cluster the alleles based on their mutational history (ÁRNASON 1982) and advance testable predictions by letting major branches of an allelic geneological tree stand for iso-fitness alleles. Information of this kind is available for the polymorphism of *Adh* gene in *Drosophila melanogaster* (KREITMAN 1983; AQUADRO *et al.* 1986; KREITMAN and HUDSON 1991), the Hemoglobin *S* and *E* and thalassemia polymorphisms in humans (ANTONORAKIS *et al.* 1982; PAGNIER *et al.* 1984; ANTONORAKIS, KAZAZIAN and ORKIN 1985; HILL 1986; CHAN *et al.* 1987), and the *t* chromosomes of mice (GOLUBIĆ *et al.* 1987). Further tests of selection among *Est-5* alleles will have to wait for such information.

Analysis of selection: The estimates of selection are robust over a wide range of effective population sizes and the analysis of residuals shows that the variance function is adequate (MCCULLAGH and NELDER 1989) in most replicates. Using the method, however, it is very difficult to distinguish between models of selection based on constant fitness and those of frequency-dependence, a point already made by WRIGHT and DOBZHANSKY (1946). It is likely that the results of an *X*-linked model fitted to these data would be qualitatively similar to the results obtained using the standard dynamics although this remains to be done, for it is only at frequencies far from equilibria that one can hope to distinguish between various models with data from a single sample per generation (WRIGHT and DOBZHANSKY 1946; PROUT 1971a,b). If an interesting selective effect had been found a selection component analysis would have been appropriate (PROUT 1971a,b; CHRISTIANSEN and FRYDENBERG 1973; CLARK and BUNDGAARD 1984; ANDERSON *et al.* 1986; WILLIAMS, ANDERSON and ARNOLD 1990). The results show, however, that the changes at the *Est-5* are caused by general and heterogeneous background effects which dissipate in most instances (*cf.* CLARK and BUNDGAARD 1984) and thus a component analysis would not be informative for assigning causality for observed selective effects to specific genes.

Heterogeneity of theory: Why is there heterogeneity of theory between replicates in the perturbation—why do the replicates not behave as statistical replicates? The N and U replicates were founded by taking samples of eggs in cups from a single egglaying cage in both groups of populations and after founding, the replicates received the same average environment and treatment. Yet there is great individuality among the populations and a group of populations behave quite differently from the others. The difference does

not lie between on one hand the NH and UH populations, which were founded differently, and the rest of the populations on the other hand, because NH belongs to the former group but UH to the latter. The heterogeneity among the populations therefore must be explained as heterogeneity of genetic material in the various founding cups taken from the egglaying cages. Such heterogeneity is hardly due to differential segregation of different alleles at the *Est-5* locus. Instead it is likely due to the generation of a variety of specific combinations by recombination—a sudden release among the egglayers of variability stored in the original chromosomes (*cf.* SPASSKY *et al.* 1958). Some of these associations are short-lived and disappear in less than 200 days while others survive the reperturbation and last till the termination of the experiment at 850 days, thus pointing to a variety of specific associations. There was no significant heterogeneity of theory in the reperturbation so such associations were not generated in the bottlenecks which occurred in the founding of the reperturbation. Rather, they occur in the beginning of the experiment and thus seem to result from the mixing of isochromosomal lines. Thus most of the selection takes place in the early part of the experiments.

I thank RICHARD C. LEWONTIN in whose laboratory I did parts of the work. I thank him and TIM P. KEITH for material and advice. I also thank ADALGEIR ARASON, ARA OSHAGAN, BJARNHEIDUR GUDMUNDSDOTTIR, EIRÍKUR STEFÁNSSON, EVA THÓRDARDÓTTIR, HELGI B. MAGNÚSSON, HLYNUR ÓSKARSSON, INGIBJÖRG KOLBEINS, KRISTINN P. MAGNÚSSON, RALPH TIEDEMANN, SNÆBJÖRN PÁLSSON, THORSTEINN NARFASON, STEFÁN GISSURASON and ÚLFAR BERGTHÓRSSON who all assisted with technical work. I thank JUDITH and JAMES VOKAC for correcting American spelling and usage and R. C. LEWONTIN and two anonymous reviewers for very helpful comments on the manuscript and analysis. And I thank BERJOUG for her encouragement and understanding.

This work was supported by National Institute of Health grant to R. C. LEWONTIN, and by a U.S. Public Health Service Fogarty International Fellowship number 1 F05 TW03027-01, by a Fulbright Senior Scholar Award, by a Senior Scholar Fulbright Research Award, by grants from the Icelandic Science Foundation (Vísindasjódur Íslands), by grants from the University of Iceland Research Fund (Rannsóknarsjódur Háskóla Íslands), and by funds from the Institute of Biology, University of Iceland to the author.

LITERATURE CITED

ANDERSON, W. W., J. ARNOLD, S. A. SAMMONS and D. G. YARDLEY, 1986 Frequency-dependent viabilities of *Drosophila pseudoobscura* karyotypes. Heredity **56:** 7–17.

ANTONORAKIS, S. E., H. H. KAZAZIAN, JR. and S. H. ORKIN, 1985 DNA polymorphism and the molecular pathology of the human globin gene clusters. Human Genetics **69:** 1–14.

ANTONORAKIS, S. E., S. H. ORKIN, H. H. KAZAZIAN, JR., S. C. GOFF, C. D. BOEHM, P. G. WABER, J. P. SEXTON, H. OSTRER, V. F. FAIRBANKS and A. CHAKRAVARTI, 1982 Evidence for the multiple origins of the β^E-globin gene in Southeast Asia. Proc. Natl. Acad. Sci. USA **79:** 6608–6611.

AQUADRO, C. F., S. F. DESSE, M. M. BLAND, C. H. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of the *Alcohol dehydrogenase* gene region of *Drosophila melanogaster*. Genetics **114:** 1165–1190.

ÁRNASON, E., 1982 An experimental study of neutrality at the *Malic Dehydrogenase* and *Esterase-5* loci in *Drosophila pseudoobscura*. Hereditas **96:** 13–27.

ÁRNASON, E., 1986 Yet another population cage. Drosophila Inform. Serv. **63:** 137.

ÁRNASON, E., and G. K. CHAMBERS, 1987 Macromolecular interactions and the electrophoretic mobility of *Esterase-5* from *Drosophila pseudoobscura*. Biochem. Genet. **25:** 287–307.

BARKER, J. S. F., and P. D. EAST, 1980 Evidence of selection following perturbation of allozyme frequencies in a natural population of *Drosophila*. Nature **284:** 166–168.

BILLINGSLEY, P., 1961 *Statistical Inference for Markov Processes*. The University of Chicago Press, Chicago.

BRIER, S. S., 1980 Analysis of contingency tables under cluster sampling. Biometrika **67:** 591–596.

CANN, J. R., 1987 Theory of electrophoresis of hybridising enzymes with kinetic control: implications for population genetics of electrophoretic markers. J. Theor. Biol. **127:** 461–477.

CHAN, V., T. K. CHAN, F. F. CHEBAB and D. TODD, 1987 Distribution of β-thalassemia mutations in South China and their association with haplotypes. Am. J. Hum. Genetics **41:** 678–685.

CHARLESWORTH, B., and D. CHARLESWORTH, 1973 A study of linkage disequilibrium in populations of *Drosophila melanogaster*. Genetics **73:** 351–359.

CHRISTIANSEN, F. B., and O. FRYDENBERG, 1973 Selection component analysis of natural polymorphisms using population samples including mother-offspring combinations. Theor. Popul. Biol. **4:** 425–445.

CLARK, A. G. and J. BUNDGAARD, 1984 Selection components in background replacement lines in *Drosophila*. Genetics **108:** 181–200.

COYNE, J. A., A. A. FELTON and R. C. LEWONTIN, 1978 Extent of genetic variation at a highly polymorphic esterase locus in *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. USA, **75:** 5090–5093.

DOBZHANSKY, TH., and C. EPLING, 1944 Contributions to the genetics, taxonomy, and ecology of *Drosophila pseudoobscura* and its relatives. Carnegie Inst. Wash. Publ. **554:** 1–183.

DRAPER, N. R., and H. SMITH, 1981 *Applied Regression Analysis*. John Wiley & Sons, New York.

DUMOUCHEL, W. H., and W. W. ANDERSON, 1968 The analysis of selection in experimental populations. Genetics **58:** 435–449.

DYKHUIZEN, D., and D. L. HARTL, 1980 Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. Genetics **96:** 801–817.

FISHER, R. A., and E. B. FORD, 1947 The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* L. Heredity **1:** 143–174.

FREEMAN, M. F., and J. W. TUKEY, 1950 Transformations related to the angular and the square root. Ann. Math. Statistics **21:** 607–611.

FRYDENBERG, O., 1963 Population studies of lethal mutations in *D. melanogaster*. I. Behavior in populations with discrete generations. Heredity **50:** 89–116.

GOLUBIĆ, M., O. BUDIMIR, R. SCHOEPFER, M. KASAKAVA, W. E. MAYER, F. FIGUEROA and J. KLEIN, 1987 Nucleotide sequence analysis of class II genes borne by mouse *t* chromosomes. Genet. Res. **50:** 137–146.

HILL, A. V. S., 1986 The population genetics of α-thalassemia and the malaria hypothesis. Cold Spring Harbor Symp. Quant. Biol. **51:** 489–498.

JONES, J. S., and T. YAMAZAKI, 1974 Genetic background and fitness of allozymes. Genetics **78:** 1185–1189.

KEITH, T., 1983 Frequency distributions of *Esterase-5* alleles in two populations of *Drosophila pseudoobscura*. Genetics **105:** 135–155.

KENDALL, M. G., and A. STUART, 1961 *The Advanced Theory of Statistics.* Charles Griffin & Co. Ltd., London.

KIMURA, M., 1983 *The Neutral Theory.* Cambridge University Press, Cambridge.

KIMURA, M., and T. OHTA, 1971 *Theoretical Aspects of Population Genetics.* Princeton University Press, Princeton.

KING, J. L., and T. OHTA, 1975 Polyallelic mutational equilibria. Genetics **79:** 681–691.

KOJIMA, K., and Y. N. TOBARI, 1970 The pattern of viability changes associated with genotype frequency at the *Alcohol dehydrogenase* locus in a population of *D. melanogaster.* Genetics **61:** 201–209.

KREITMAN, M., 1983 Nucleotide polymorphism at the *Alcohol dehydrogenase* locus of *Drosophila melanogaster.* Nature **304:** 412–417.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

LANGLEY, C. H., Y. N. TOBARI and K. I. KOJIMA, 1974 Linkage disequilibrium in natural populations of *Drosophila melanogaster.* Genetics **78:** 921–936.

LEWONTIN, R. C., 1958 Studies of heterozygosity and homeostasis. II. Loss of heterosis in a constant environment. Evolution **12:** 494–503.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

LEWONTIN, R. C., L. R. GINZBURG and S. D. TULJAPURKAR, 1978 Heterosis as an explanation for large amounts of genic polymorphism. Genetics **88:** 149–170.

LEWONTIN, R. C., and J. L. HUBBY, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura.* Genetics **54:** 595–609.

MCCULLAGH, P., and J. A. NELDER, 1989 *Generalized Linear Models.* Chapman & Hall, London.

MUKAI, T., 1985 Experimental verification of the neutral theory, pp. 125–145 in *Population Genetics and Molecular Evolution,* edited by T. OHTA and K. AOKI. Japan Scientific Societies Press, Tokyo.

NEI, M., and F. TAJIMA, 1981 Genetic drift and estimation of effective population size. Genetics **98:** 625–640.

NELDER, J. A., and R. W. M. WEDDERBURN, 1972 Generalized linear models. J. R. Stat. Soc. A. **135:** 370–384.

OHTA, T., 1971 Associative overdominance caused by linked detrimental mutations. Genet. Res. **18:** 277–286.

OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96–98.

OHTA, T., and M. KIMURA, 1971 Behavior of neutral mutants influenced by associated overdominant loci in finite populations. Genetics **69:** 247–260.

PAGNIER, J., J. G. MEARS, O. DUNDA-BELKHODJA, K. E. SCHAEFER-REGO, C. BELDJORD, R. L. NAGEL and D. LABIE, 1984 Evidence for the multicentric origin of the sickle cell hemoglobin gene in Africa. Proc. Natl. Acad. Sci. USA **81:** 1771–1773.

POLLAK, E., 1983 A new method for estimating effective population size from allele frequency changes. Genetics **104:** 531–548.

PRAKASH, S., R. C. LEWONTIN and J. L. HUBBY, 1969 A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura.* Genetics **61:** 841–858.

PRESS, W. H., B. P. FLANNERY, S. A. TEUKOLSKY and W. T. VETTERLING, 1986 *Numerical Recipes.* Cambridge University Press, Cambridge.

PROUT, T., 1971a The relation between fitness components and population prediction in Drosophila. I. The estimation of fitness components. Genetics **68:** 127–149.

PROUT, T., 1971b The relation between fitness components and population prediction in Drosophila. II. Population prediction. Genetics **68:** 151–167.

SCHAFFER, H. E., D. YARDLEY and W. W. ANDERSON, 1977 Drift or selection: a statistical test of gene frequency variation over generations. Genetics **87:** 371–379.

SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry.* W. H. Freeman, San Francisco.

SPASSKY, B., N. SPASSKY, H. LEVENE and TH. DOBZHANSKY, 1958 Release of genetic variability through recombination. I. *Drosophila pseudoobscura.* Genetics **43:** 844–867.

SPIESS, E. B., 1966 Chromosomal fitness changes in experimental populations of *Drosophila persimilis* from Timberline. II. Reextracted chromosomes. Evolution **20:** 82–91.

WATTERSON, G. A., 1982 Testing selection at a single locus. Biometrics **38:** 323–331.

WILLIAMS, C. J., W. W. ANDERSON and J. ARNOLD, 1990 Generalized linear modeling methods for selection component experiments. Theor. Popul. Biol. **37:** 389–423.

WILSON, S. R., 1980 Analyzing gene-frequency data when the effective population size is finite. Genetics **95:** 489–502.

WILSON, S. R., and J. G. OAKESHOTT, 1985 Statistical analysis of data from gene frequency perturbation experiments. Biometrical J. **27:** 623–633.

WILSON, S. R., J. G. OAKESHOTT, J. B. GIBSON and P. R. ANDERSON, 1982 Measuring selection coefficients affecting the *Alcohol dehydrogenase* polymorphism in *Drosophila melanogaster.* Genetics **100:** 113–126.

WRIGHT, S., and TH. DOBZHANSKY, 1946 The genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of *Drosophila pseudoobscura.* Genetics **31:** 125–156.

YAMAZAKI, T., 1971 Measurement of fitness at the *Esterase-5* locus in *Drosophila pseudoobscura.* Genetics **67:** 579–603.

Communicating editor: J. R. POWELL

APPENDIX

Einar's Goodness-of-fit Problem

Richard C. Lewontin

Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts

Heterogeneity statistics: We have k separate experiments and each one is fitted to a model whose parameters are estimated from the data. Then there will be a different set of expected values for each experiment. We can also fit the model to the entire set of data pooled over all the experiments. Problem: If we calculate a χ^2 goodness-of-fit test for each experiment separately and for the pooled data, what is the relation between these tests? In what follows, we use G rather than X^2 as a statistic for the χ^2 distribution. To simplify notation and subscripts, we consider only the contribution to G of one particular class in all the samples.

The result will apply to the contributions to G of all classes and thus to the total G over classes.

Within sample i whose total size is N_i the G for goodness-of-fit is

$$G_i = n_i \ln p_i - n_i \ln \pi_i \tag{1}$$

where n_i is the observed number in the ith sample, in the chosen class, $p_i = n_i/N_i$ is the observed proportion of that sample in the chosen class and π_i is the theoretical proportion in the chosen class in the ith sample, where the theoretical proportion is calculated from the model and parameters for the ith experiment.

Now, the total G over all experiments

$$G_W = \sum G_i = \sum_{i=1}^{k} n_i \ln p_i - \sum_{i=1}^{k} n_i \ln \pi_i. \tag{2}$$

On the other hand, the G calculated from the pooled data

$$\bar{G} = n_T \ln \bar{p} - n_T \ln \pi_T = n_T \ln \bar{p} - \sum_{i=1}^{k} n_i \ln \pi_T \tag{3}$$

where $\bar{p} = n_T/N_T$ is the proportion in the chosen class for the pooled sample. The difference between G_W and $\bar{G}$, the difference between the total G and the G of the total is from (2)–(3) and reordering the terms:

$$G_W - \bar{G} = \left[\sum_{i=1}^{k} n_i \ln p_i - n_T \ln \bar{p}\right] - \left[\sum_{i=1}^{k} n_i \ln \pi_i - n_i \ln \pi_T\right] \tag{4}$$

$$G_W - \bar{G} = \left[\sum_{i=1}^{k} n_i (\ln p_i - \ln \bar{p})\right] - \left[\sum_{i=1}^{k} n_i (\ln \pi_i - \ln \pi_T)\right]. \tag{5}$$

But the first quantity in brackets on the right hand side of (5) is simply the usual heterogeneity G of the data, putting aside any theoretical model. The second quantity measures, by analogy, the heterogeneity of the theoretical model values among samples! So, we may say that the *total goodness-of-fit* G minus the *goodness-of-fit for the total* is equal to the *heterogeneity of the data* minus the *heterogeneity of the model expectations.*

We can carry the process on in another way, analogous to the analysis of variance.

Consider the goodness-of-fit of each sample data point to the theoretical values derived from the pooled data.

$$G_T = \sum_{i=1}^{k} n_i \ln p_i - n_T \ln \pi_T = \sum_{i=1}^{k} n_i (\ln p_i - \ln \pi_T). \tag{6}$$

This can be analyzed into two components:

a. The deviation of each sample point from its own theoretical values, summed over samples (the total G on the previous page).

$$G_W = \sum G_i = \sum n_i (\ln p_i - \ln \pi_i) \tag{7}$$

b. The deviation of each sample theoretical value from the theoretical values of the merged sample, which we identified above as the "heterogeneity of theoretical values."

$$G_B = \sum n_i (\ln \pi_i - \ln \pi_T). \tag{8}$$

Then, obviously, $G_T = G_W + G_B$.

Finally, we remember $\bar{G}$ from above

$$\bar{G} = \sum_{i=1}^{k} n_i (\ln \bar{p} - \ln \pi_T), \tag{9}$$

the goodness-of-fit of the pooled data. But, from (5)

$$G_W - \bar{G} = \text{heterogeneity } G - G_B. \tag{10}$$

Therefore,

$$G_W + G_B = G_T = \text{heterogeneity } G + \bar{G}. \tag{11}$$

That is, the total G of all the individual data points from the theoretical value estimated from the pooled data consists of two components: the heterogeneity of the data plus the goodness-of-fit of the pooled data to the pooled theory. It is the sum of a goodness-of-fit component and a heterogeneity of data component.

Degrees of freedom: With k experiments, c classes in each experiment and P parameters estimated from the data, we have for various components the following degrees of freedom:

Component	d.f.	
1. G_W	$k(c - P - 1)$	
2. $\bar{G}$	$(c - P - 1)$	
3. $G_W - \bar{G}$	$(k - 1)(c - P - 1)$	(1) – (2)
4. Het. of data	$(k - 1)(c - 1)$	
5. Het. of parameters	$(k - 1)P$	(4) – (3)