BRIEF COMMUNICATIONS

Evolution, 53(2), 1999, pp. 600–604

APPLICABILITY OF THE HYPERGEOMETRIC PHENOTYPIC MODEL TO HAPLOID AND DIPLOID POPULATIONS

MAX SHPAK[1] AND ALEXEY S. KONDRASHOV[2,3]
[1]*Department of Biology, Yale University, New Haven, Connecticut 06520*
[2]*Section of Ecology and Systematics, Cornell University, Ithaca, New York 14853*
[3]*E-mail: ask3@cornell.edu*

Abstract.—We show that the phenotypic hypergeometric model of a quantitative trait can exactly describe both haploid and diploid populations. The condition necessary for this is equiprobability of genotypes within each phenotype. This requires equal allele frequencies across the loci, which may be the case when the population is under disruptive selection.

Key words.—Allele frequencies, linkage disequilibrium, phenotypic model, selection.

Received February 6, 1998. Accepted October 19, 1998.

Analyzing the dynamics of quantitative traits is essential for many problems in evolutionary genetics. Unfortunately, models with genotype frequencies as dynamic variables are intractable. Even if n, the number of loci contributing to variability in a trait, is moderate, the number of genotypes 2^n (assuming haploidy and two alleles, A0 and A1, at each locus) is too large.

Thus, it is desirable to find a less numerous set of more inclusive variables that will nevertheless provide a dynamically sufficient (Lewontin 1974) description of a quantitative trait. If alleles at different loci are distributed independently (no linkage disequilibrium), then allele frequencies at n loci can be used, instead of any $2^n - 1$ genotype frequencies. Deviations from this independence can be often taken into account, with good precision, only by $n(n - 1)/2$ pairwise linkage disequilibria, because the distribution of phenotypes (breeding values) of a trait may remain close to Gaussian even under selection (Turelli and Barton 1994).

However, assortative mating and/or very strong selection can create substantial higher-order linkage disequilibria, thus making the phenotype distribution significantly non-Gaussian. In such cases, for example during sympatric speciation, the whole distribution of frequencies of the $n + 1$ phenotypes (assuming additive loci with equal effects, so that a phenotype is determined by the number of A1 alleles in the genotype) must be specified.

In fact, even this distribution alone is generally not dynamically sufficient. To infer the overall phenotype distribution in the next generation from the current one, we need to know the distributions of phenotypes in the offspring resulting from the matings of any two parental phenotypes. Generally, these distributions depend on the genotype frequencies within each parental phenotype.

Still, there are at least two important special cases when the phenotypic description is dynamically sufficient, regardless of higher-order disequilibria: (1) populations in which alleles of the same type (say, A1) are frequent at all the loci (see Kimura and Maruyama 1966); and (2) populations at the final stage of sympatric speciation when only the extreme phenotypes (and genotypes) of the incipient species are frequent and at least one partner in every mating has such a phenotype (see Kondrashov 1983).

Recently, a phenotypic model based on the hypergeometric distribution was proposed by Barton (1992) for diploid populations reproducing through a gamete pool. Doebeli (1996a,b,c) independently proposed such a model for haploid populations and used it to study various problems. However, the conditions required for applicability of this model were not investigated. We do this below, as well as present the exact hypergeometric model for diploid populations where reproduction occurs through formation of breeding pairs.

EQUIPROBABLE GENOTYPES WITHIN PHENOTYPES

Let us assume haploidy and suppose that all possible genotypes within each particular phenotype have the same frequency. This, of course, implies that frequencies of alleles A1 are the same at all the loci within each phenotype (and, thus, within the population as the whole) and all alleles are distributed independently within each phenotype (intraphenotype linkage equilibrium). If so, symmetry considerations prove that equiprobability of genotypes within each phenotype will persist in the next generation if segregation is Mendelian and recombination is free, whereas selection and/or assortative mating depend on phenotypes only.

Consider a mating between two parents drawn randomly from the population in which all constituent genotypes are equiprobable within every phenotype. The parents have some arbitrary phenotypes i and j and each parent of a given phenotype is equally likely to carry A1 at any particular locus. Thus, because the progeny inherit an allele at any locus independently of the alleles at its other loci, an offspring of phenotype k is also equally likely to carry its k alleles A1 at any loci, which preserves intraphenotype equiprobability of genotypes within the progeny. Because this holds for any pair of parental phenotypes, this equiprobability is also preserved for the total pool of offspring, under any assortative mating based on phenotypes.

Selection acting on phenotypes also does not disturb intraphenotype equiprobability of genotypes because the proportions within each phenotype remain the same in spite of

changes in phenotype frequencies. Thus, although change in phenotype frequencies creates linkage disequilibria at the population level, intraphenotype linkage equilibrium and equal allele frequencies will continue to hold.

Hypergeometric Model for Haploids

Consider a mating between phenotypes i and j. If the number of loci at which both parents have A1 (genetic overlap) is $L \leq \min(i, j)$, they both have A0 at the other $n - i - j + L$ loci, and different alleles at the remaining $i + j - 2L$ loci. Assuming Mendelian segregation and free recombination, an offspring inherits maternal and paternal alleles independently with probabilities 0.5 at each of these $i + j - 2L$ loci, and the probability that the offspring will have phenotype k is given by a binomial distribution:

$$R_{i,j,L}(k) = \binom{i + j - 2L}{k - L}(1/2)^{i+j-2L}. \quad (1)$$

Here and below

$$\binom{a}{b} = 0 \quad \text{if} \quad b < 0 \quad \text{or} \quad a < b.$$

The respective formulae (10), (8), and (8) in Doebeli (1996a,b,c) are mistyped. With equiprobability of genotypes within phenotypes, the distribution of genetic overlap L is hypergeometric (see eqs. [12] and [10] in Doebeli 1996a,c):

$$P_{i,j}(L) = \frac{\binom{i}{L}\binom{n - i}{j - L}}{\binom{n}{j}}. \quad (2)$$

Thus, $P_{i,j}(L) > 0$ only if $\max(0, i + j - n) \leq L \leq \min(i, j)$. Equation (2) is true because an individual of the phenotype i carries exactly L A1 at those loci where an individual of phenotype j also has A1 with the same probability that there are L black balls in a sample of size j drawn without replacement from an urn containing i black balls and $n - i$ white balls (Feller 1950, p. 33; Ross 1994, p. 172).

A mating between parents of the phenotypes i and j with unknown genetic overlap L will produce an offspring with phenotype k with probability

$$R_{i,j}(k) = \sum_{L=\max(0,i+j-n)}^{\min(i,j)} P_{i,j}(L)R_{i,j,L}(k). \quad (3)$$

Of course, $R_{j,i}(k) = R_{i,j}(k)$. The phenotype distribution after reproduction in generation $t + 1$, $p'_{t+1}(i)$, is connected to the phenotype distribution after selection in the parental generation t, $p_t(i)$, by:

$$p'_{t+1}(k) = \sum_{i=0}^{n}\sum_{j=0}^{n} A(i, j)R_{i,j}(k), \quad (4)$$

where $A(i, j)$ is the frequency of matings between ordered phenotypes i and j in generation t (with random mating $A(i, j) = p_t(i)p_t(j)$, but any phenotypic nonrandom mating can also be incorporated).

Phenotypic selection $w(i)$ after reproduction completes the model:

$$p_{t+1}(i) = p'_{t+1}(i)w(i) \Big/ \sum_{i=0}^{n} p'_{t+1}(i)w(i). \quad (5)$$

Equations (1)–(5) constitute the exact haploid hypergeometric model of Doebeli (1996a,b,c). Kondrashov (1984) derived an approximate hypergeometric model, which he erroneously believed to be exact, and used it assuming global linkage equilibrium, which he erroneously believed to be necessary for its applicability.

Although the original approximate model was diploid, let us first consider its haploid form. Again, consider two parents of phenotypes i and j. Their minimal possible genetic overlap is $L_{\min} = \max(0, i + j - n)$. When $i + j \geq n$, an offspring will surely inherit A1 at $L_{\min}$ loci, and there are equal numbers of A0 and A1 at the remaining $h = n - L_{\min}$ loci of the zygote. With intraphenotype equiprobability of genotypes, Mendelian segregation, and free recombination, an offspring inherits A1 at these h loci approximately with the same probability with which there will be k black balls in a sample of h balls drawn without replacement from an urn containing h black and h white balls. This probability is given by the hypergeometric distribution:

$$H(h, k) = \frac{\binom{h}{k}\binom{h}{h - k}}{\binom{2h}{h}}. \quad (6)$$

The case of $i + j < n$ is similar, because an offspring surely inherits A0 at $n - i - j$ loci, and there are equal numbers of A0 and A1 at the remaining $h = i + j$ loci of the zygote. Thus,

$$^{app}R_{i+j}(k) = H(h, \max[k, k + i + j - n]). \quad (7)$$

It takes $\sim n^3$ operations with equation (7), rather than $\sim n^4$ operations with equation (3), to compute the offspring phenotype distribution, although this difference can be eliminated if R is calculated just once and stored. Whereas equation (3) is exact, (7) ignores the distribution of A0 and A1 at loci where homozygosity is not guaranteed. Thus, equation (7) depends only on the sum of parental phenotypes and treats the choice of A1 as equiprobable and independent at each potentially heterozygous locus, which is invalid because an offspring cannot get both alleles at a locus. In contrast, equation (3) depends on each phenotype separately and takes the distribution of A0 and A1 into account by explicitly treating different genetic overlaps. However, with large n the nonindependence in distributions of A0 and A1 within the zygote, which is imposed by phenotypes of the parents, become increasingly less important, and equation (7) provides a close approximation to (3) for $n > 10$.

Hypergeometric Model for Diploids

The description of the phenotype of a diploid individual by its total number of A1 is not dynamically sufficient even if this number determines mate choice and fitness, because

diploid genotypes are not always equiprobable within a phenotype. For example, if individuals with n A1 were produced exclusively in the mating $0 \times 2n$, they all are heterozygous at all n loci. Obviously, the distribution of the number of A1 per genotype in offspring from mating between such individuals is different from that in offspring from mating between individuals carrying n A1, which can be homozygous at some loci.

Still, an exact hypergeometric model can be constructed if we describe a diploid individual by a pair (i, j) ("phenotype"), the numbers of A1 in its maternal and paternal genomes. If all possible maternal and paternal genotypes are equiprobable within individuals with a particular (i, j) (gametic equiprobability), such individuals will again produce all possible gametes with a particular number of A1 with the same probability. Thus, gametic equiprobability is preserved.

Function $R_{i,j}(k)$ now describes the gamete production (Barton 1992, eq. A2). Thus, in a mating between phenotypes (i, j) and (k, l) an offspring (x, y) is produced with probability

$$_{2n}R_{i,j,k,l}(x, y) = R_{i,j}(x)R_{k,l}(y). \quad (8)$$

Mating and selection are described by the equations analogous to (4) and (5), respectively.

In contrast, the approximate diploid hypergeometric model of Kondrashov (1984) retains the description of the phenotype of a diploid by the total number of A1. In this case, the mating between individuals with phenotypes i and j ($i, j = 0, \ldots, 2n$) produces an offspring with phenotype k with probability

$$^{\text{app}}_{2n}R_{i,j}(k) = \sum_{l=0}^{i} {}^{\text{app}}R_i(l)\,{}^{\text{app}}R_j(k - l), \quad (9)$$

whereas the rest is the same, as in the haploid case.

Another diploid hypergeometric model, recently proposed by Doebeli (1997), also uses unidimensional description of phenotypes. This model depends on the assumption that all diploid genotypes within a phenotype are equiprobable. As we have seen, this assumption is not exact. However, this model appears to provide an even better approximation than equation (9), because it is based on explicit consideration of all possible numbers of homozygous loci.

Approximate diploid hypergeometric models require, if $^{\text{app}}_{2n}R$ is precalculated, $\sim n^3$ operations per generation and thus have a big computational advantage over the exact model that requires $\sim n^6$ operations, if $_{2n}R$ is precalculated.

If diploid individuals do not form pairs for reproduction, but, instead, the gametes produced by all individuals merge into a single pool (see Korzukhin and Kaganova 1982), a diploid population can be described by frequencies of "phenotypes" (i.e., of classes of genotypes with various numbers of A1) among gametes. Thus, even an exact hypergeometric model of a diploid population can be based on a unidimensional distribution (Barton 1992). However, reproduction through breeding pairs can make a significant difference, if mating is nonrandom and selection is strong.

Applicability of Hypergeometric Model

With haploidy, equiprobability of genotypes within a phenotype makes the description based on any n phenotype frequencies and hypergeometric phenotype inheritance dynamically sufficient. Such equiprobability requires equal allele frequencies at all loci within a phenotype. Equal allele frequencies in the whole population do not guarantee their equality within the phenotype. For example, consider a population where all individuals of phenotype i have A0 at locus 3 and A1 at locus 4, with a reciprocal situation in phenotype j. In the population consisting solely of these individuals, in equal proportions, the allele frequencies at loci 3 and 4 are equal in the population in spite of different intraphenotype frequencies.

We have shown that, if initially present, equiprobability of genotypes within a phenotype will persist, as long as recombination is free and selection and assortative mating depend only on phenotypes. Apparently, this equiprobability can be stable only if selection is disruptive, because only disruptive selection maintains equality of allele frequencies at all loci in the whole population. This is the case because disruptive selection favors deviating phenotypes and the equality of allele frequencies leads to the highest phenotypic variance. However, under invariant disruptive selection a population rapidly loses genetic variability, due to fixation of alleles of the same type (A0 or A1) at all loci. Thus, a polymorphic equilibrium in a population under disruptive selection is possible only if selection is also negatively frequency-dependent, that is, if rare phenotypes are favored (Udovic 1980). In this case, frequencies of A1 at different loci, both global and within phenotypes, rapidly approach equality; intraphenotypic linkage disequilibria, when initially present, disappear, even if overall linkage disequilibrium persists.

If so, overall equality of A1 frequencies across the loci guarantees that, perhaps after some generations, the hypergeometric model becomes dynamically sufficient. To support this assertion, we compared the results of deterministic iterations of the hypergeometric model and the results of stochastic experiments with an individual-based model with free recombination (see Kondrashov et al. 1998). Both models are implemented as Macintosh THINK C programs and are available on request.

Figure 1 shows the equilibrium distributions of phenotypes under the strength of disruptive selection close to that required for sympatric speciation (10% of individuals with the lowest and the highest phenotype values had fitness 1, whereas the remainder have fitness $d = 0.16$; weak frequency dependence was added to keep the population polymorphic; see Kondrashov et al. 1998). Speciation occurred (i.e., only the two reproductively isolated extreme phenotypes persisted at equilibrium) with $d \leq 0.15$ in the case of the exact hypergeometric model and individual-based model, and $d \leq 0.14$ in the case of the approximate hypergeometric model; simplified analytical theory predicts speciation if $d \leq 0.20$ (Kondrashov et al. 1998). The average frequency of A1 across the loci was very close to 0.5, and the standard deviation of the distribution of these frequencies at the individual loci was ~ 0.02. The outcome of individual-based simulations was the same regardless of the initial allele frequencies, as long as the initial population was not too close to speciation, because frequency-dependent disruptive phenotypic selection rapidly (usually, in less than 100 generations) led to almost uniform allele frequencies (data not reported). Strictly speaking, the

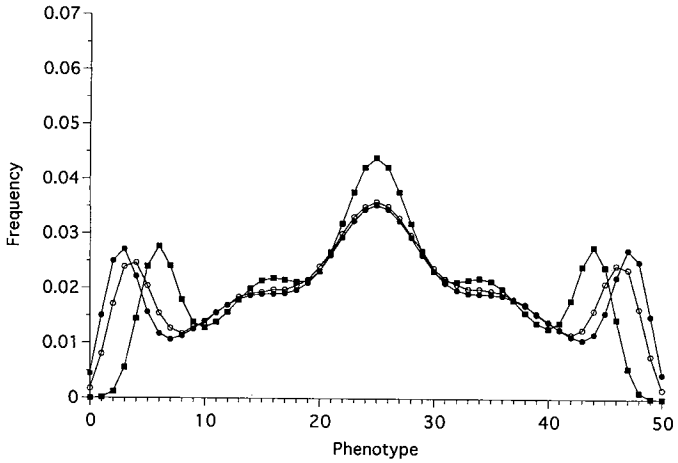

FIG. 1. Equilibrium phenotype distributions in a haploid population with $n = 50$ and reproductive isolation between the opposite extreme phenotypes with no restrictions on any other matings described by the exact hypergeometric model (solid circles), approximate hypergeometric model (solid squares), and individual-based model (open circles, the average phenotype frequencies in generations 100–150 are reported for the population of 10,000 individuals).

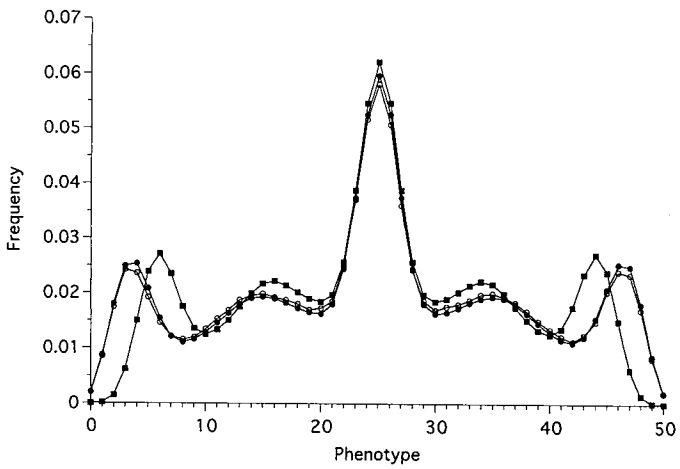

FIG. 2. The same as Figure 1, but the population is diploid with $n = 25$.

distribution in an individual-based model presented in Figure 1 must be regarded as only a quasi-equilibrium, because eventually some allele (A0 or A1) will be fixed at every locus. However, under even moderately strong frequency-dependent disruptive selection such fixations occur extremely rarely, and even very slow mutation is enough to counterbalance them.

Figure 2 shows the analogous results for the diploid population under the same phenotypic selection. In the case of the exact hypergeometric model, the phenotype of an individual is the total number of A1 in its maternal and paternal genotypes. Speciation occurred with $d \leq 0.15$ in all three models. In both the haploid and diploid cases, equilibrium variance under the approximate model was slightly less than under the exact model, because the approximation implied by equation (7) overestimates the degree of genetic exchange among organisms. However, the difference is rather small.

Unfortunately, the hypergeometric model turns out to be sensitive to violations of the equal allele frequency prerequisite. Although providing a good approximation when the frequencies of A1 vary within the range of 0.4–0.6 or even 0.3–0.7, it cannot be used when the frequencies are more variable (Fig. 3).

PHENOTYPIC DESCRIPTION REQUIRES EQUAL ALLELE FREQUENCIES

Instead of equiprobability of genotypes within each phenotype, a less restrictive assumption of intraphenotype linkage equilibrium might be enough to make the phenotypic description dynamically sufficient. Here we will show that, unfortunately, with assortative mating and/or strong selection, such linkage equilibrium generally does not persist when allele frequencies at different loci are unequal, either in the whole population or even only within individual phenotypes. We present two examples in which intraphenotypic linkage equilibrium is destroyed in one generation. In both cases, we assume free recombination and $n = 3$ because with $n = 2$ there could be no linkage disequilibrium within a phenotype.

First, assume that A1 frequencies at the three loci have some generic values $q1$, $q2$, and $q3$, while the population is in global linkage equilibrium (which also implies linkage equilibrium within every phenotype). A single generation of complete assortative mating (where matings occur only within the same phenotype) will lead to linkage disequilibria within phenotypes 1 and 2. Computing the genotype frequencies in phenotype 1 offspring (which could result from either a 1 × 1 or a 2 × 2 mating) is enough to see this.

Second, assume that the population after selection contains only phenotypes 1 and 2 with equal frequencies (0.5). Suppose that all phenotype 1 individuals have genotype 100, whereas all phenotype 2 individuals have genotype 011. Thus, overall allele frequencies are 0.5 at all three loci, while

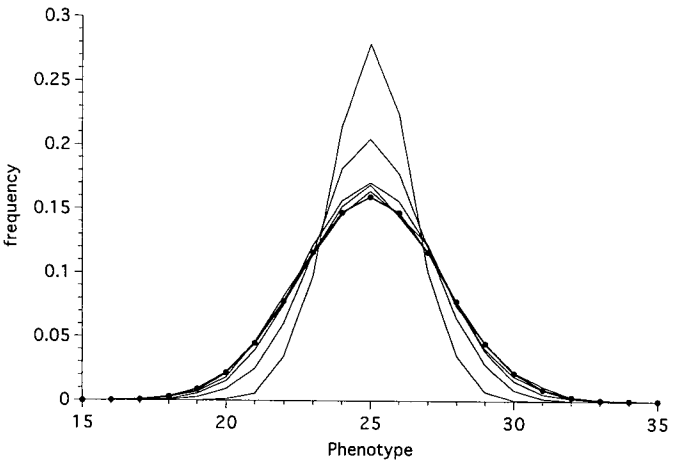

FIG. 3. Phenotype distributions in the offspring of parents that were all of the phenotype 25 (haploidy, $n = 50$) and at linkage equilibrium in the exact hypergeometric model (curve with solid circles) and in individual-based simulations (10,000 individuals) where the frequency of A1 among the parents was p at 25 loci and $1 - p$ at the other 25. The peaks of the curves that correspond to $p = 0.5$, 0.6, 0.7, 0.8, and 0.9 are of the increasing heights. With $p = 1.0$, the whole distribution is concentrated at 25 (not shown).

the population is in (trivial) intraphenotype linkage equilibrium because frequency of each genotype is the product of the frequencies of the constituent alleles within its phenotype. If mating is random, crosses 011 × 011 occur with probability 0.25 and produce offspring 011 with probability 1.00, whereas crosses 100 × 011 occur with probability 0.5 and produce offspring 110 or 011 with probability 0.125 each. Thus, in the next generation, there will be no linkage equilibrium within phenotype 2 (as well as within phenotype 1). In this case, however, intraphenotype allele frequencies will eventually become equal, and intraphenotype linkage equilibrium will be restored (see above).

In the general case, the genotype frequencies within each phenotype are the dynamical property of the previous generation's genotype frequencies. Unequal allele frequencies within phenotypes give rise to intraphenotype linkage disequilibrium, which make it impossible to describe the genotype frequencies as a predictable proportion of each phenotype. As a result, a phenotype-based description becomes dynamically insufficient. Thus, with overall linkage disequilibrium the phenotypic description is dynamically sufficient only if intraphenotype linkage equilibrium is supplemented by equal allele frequencies within each phenotype. Together, these assumptions imply intraphenotype equiprobability of genotypes. Thus, the hypergeometric model appears to be the only purely phenotypic model possible under a generic distribution of phenotypes.

Feasibility of Equal Allele Frequencies

Because stabilizing selection leads to fixations of different alleles at different loci (see Barton 1989), it is reasonable to assume that equality of allele frequencies may occur if selection in the equilibrium population is disruptive. Stability of such polymorphic equilibrium requires, of course, that selection is also frequency-dependent (Udovic 1980). Perhaps, if selection does not eventually lead to monomorphism even in the absence of mutation, it will always cause equality of equilibrium allele frequencies at all loci, as well as coupling linkage disequilibria among the loci.

This hypothesis, which is supported by the results of our individual-based simulations, implies that the hypergeometric model can describe the third phase of Wright's shifting balance (Barton 1992) and sympatric speciation (Doebeli 1996c). This is important, because a simpler linear phenotypic model (Kondrashov 1983) describes only the final stage of the process. In the case of Doebeli (1996b) and Doebeli (1996c) the hypergeometric model probably works if the availability of resources is a bimodal or uniform function of the phenotype (Doebeli 1996b) because this leads to selection against the more frequent intermediate phenotypes. Even when this function is unimodal (Doebeli 1996a) the hypergeometric model may still be applicable if strong competition among similar phenotypes leads to disruptive selection at equilibrium (Doebeli 1996a; Fig. 2). In contrast, under stabilizing selection phenotypic variability is due to various rare alleles (see Barton 1989), which often have rather different frequencies at different loci (e.g., Kondrashov and Yampolsky 1996), thus making the hypergeometric model useless.

Thus, it is desirable to create a "hybrid" model that will take into account both phenotype and allele frequencies. Apparently, such a model cannot be exact, unlike the hypergeometric model. Still, the knowledge of overall allele frequencies may be enough to infer approximately (unequal) genotype frequencies within each phenotype and, thus, to calculate the distributions of phenotypes in the offspring of different matings. A tractable dynamical model of a quantitative trait applicable under more or less general conditions is badly needed, so that the efforts in this direction will be worthwhile, although the problem seems to be difficult.

Acknowledgments

We are grateful to N. Barton and M. Doebeli for helpful comments and to S. Kuznetsov for pointing out that equation (7) is only an approximation. The work was supported by the NSF grant DEB 9417753.

Literature Cited

Barton, N. H. 1989. The divergence of a polygenic system subject to stabilizing selection, mutation and drift. Genet. Res. 54:59–77.

———. 1992. On the spread of new gene combinations in the third phase of Wright's shifting-balance. Evolution 46:551–557.

Doebeli, M. 1996a. An explicit genetic model for ecological character displacement. Ecology 77:510–520.

———. 1996b. Quantitative genetics and population dynamics. Evolution 50:532–546.

———. 1996c. A quantitative genetic competition model for sympatric speciation. J. Evol. Biol. 9:893–909.

———. 1997. Genetic variation and the persistence of predator-prey interactions in the Nicholson-Bailey model. J. Theor. Biol. 188:109–120.

Feller, W. 1950. An introduction to the probability theory and its applications. Vol. 1. John Wiley, New York.

Kimura, M., and T. Maruyama. 1966. The mutation load with epistatic gene interactions in fitness. Genetics 54:1337–1351.

Kondrashov, A. S. 1983. Multilocus model of sympatric speciation. I. One character. Theor. Pop. Biol. 24:121–135.

———. 1984. On the intensity of selection for reproductive isolation at the beginnings of sympatric speciation. Genetika 20: 408–415.

Kondrashov, A. S., and L. Y. Yampolsky. 1996. High genetic variability under the balance between symmetric mutation and fluctuating stabilizing selection. Genet. Res. 68:157–164.

Kondrashov, A. S., L. Y. Yampolsky, and S. A. Shabalina. 1998. On the sympatric origin of species by means of natural selection. Pp. 90–98 *in* D. Howarth, ed. Endless forms: species and speciation. Oxford Univ. Press, Oxford.

Korzukhin, M. D., and O. Z. Kaganova. 1982. Genotype dynamics in populations with breeding pairs. J. Theor. Biol. 96:201–209.

Lewontin, R. C. 1974. The genetic basis of evolutionary change. Columbia Univ. Press, New York.

Ross, S. 1994. A first course in probability. 4th ed. Macmillan, New York.

Turelli, M., and N. H. Barton. 1994. Genetic and statistical analyses of strong selection on polygenic traits: what, me normal? Genetics 138:913–941.

Udovic, D. 1980. Frequency-dependent selection, disruptive selection and evolution of reproductive isolation. Am. Nat. 116:621–641.

Corresponding Editor: W. T. Starmer