J. Genet., Vol. 64, No. 1, July 1985, pp. 21–29. © Printed in India.

A sampling theory for local selection

STANLEY SAWYER* and DANIEL HARTL

Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA.

* Also at Department of Mathematics, Washington University, St. Louis, MO 63130, USA.

MS received 9 March 1985

Abstract. We have examined a model of selection (local selection) in which successive favorable alleles enter into a population by displacing a random fraction of each of the pre-existing alleles. When the distribution of fitness among newly arising favorable mutations is given by a power law, then the distribution of allele frequencies in the population converge to a Poisson–Dirichlet limit, and the sampling distribution of alleles is a Ewens distribution. This property leads to a convenient algorithm for simulating random equilibrium frequencies of alleles within samples. The model can also be interpreted in terms of species abundances when each invading species displaces a random fraction of each pre-existing species, or in terms of age structures in populations subjected to random catastrophes.

Keywords. Population structure; allele displacement; species abundance.

1. Introduction

We examine a model of selection appropriate to a widespread haploid organism such as *Escherichia coli* in which new favorable mutations displace only parts of local populations. As new favorable mutations occur and supplant random fractions of the pre-existing alleles, the number of alleles in the population grows, and the population structure becomes increasingly more complicated. The limiting population structure depends on the distribution of the fitnesses of the newly arising favorable mutations. If this distribution is a power law (see below for more detail), the joint distribution of the allele frequencies at equilibrium is the Poisson–Dirichlet, and the distribution of allelic configuration in random samples is the Ewens distribution. This model is also appropriate for the distributions of species abundances when each new invading species displaces a random fraction of each of the resident species in a habitat. It also applies to age structures in populations subject to occasional catastrophes, assuming that mortality is independent of age and juvenile individuals replace those that have died.

2. Details of the model and results

The geographical structure we envision corresponds to a haploid asexual organism dispersed in two dimensions. (Haploidy and asexuality are assumed mostly for definiteness.) The population is sufficiently large so that, in the absence of mutation, the relative frequency of each genotype will remain constant from generation to generation. From time to time, new favorable mutations occur and replace a fraction of each of the pre-existing genotypes by local selection. By *local selection* we mean that local ecological or geographic factors allow the invasion of a favorable mutant in parts of a habitat, but prevent it from being fixed over the entire range. The unreplaced genotypés have safe sites or refuges which cannot be invaded by the new mutant. These sites are not permanent safe harbors, however. At some future time a different favorable mutation may take advantage of forcuitous ecological or climatic conditions and succeed in invading or further invading a refuge. These safe sites need not be geographical. They might also be resources that one genotype could utilize more efficiently than others, or as techniques for neutralizing particular environmental hazards. In this sense, the model is more like an infinite-dimensional model.

When a favorable mutation occurs, it replaces a fraction of each of the types already existing in the population, where the fraction depends on the degree to which the new mutant is globally favored. To be concrete, assume that initially the population is fixed for a single type. A new favorable mutation arises and displaces a random fraction Z_1 of the pre-existing type, yielding frequencies Z_1 and $1-Z_1$ of the two types. A second favorable mutation then arises and displaces a random fraction Z_2 of each of the resident alleles, yielding frequencies Z_2, $(1-Z_2)Z_1$, and $(1-Z_2)(1-Z_1)$ for the three types. This process continues with the occurrence of a series of mutants that, in turn, displace fractions $\{Z_k\}$ of each of the pre-existing types. Each new mutant type is assumed to be entirely new to the population, and the time scale of displacement is sufficiently short relative to mutation so that, for purposes of the model, displacement occurs instantaneously. This selection process is illustrated in figure 1. In (a), the habitat is shown as being occupied by four distinct genotypes, indicated by the open and filled shapes. In (b), a new favorable mutation has occurred and displaced a fraction (in this case one-fourth) of each of the resident alleles (shaded area). It should be emphasized that the geographical areas occupied by each allele in figure 1 are shown as contiguous

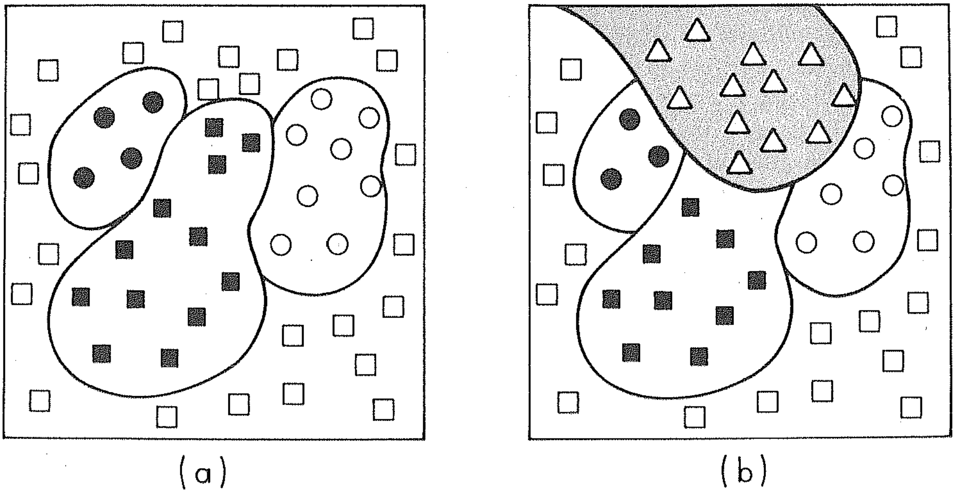

Figure 1. Local selection model. (**a**) Habitat is occupied by four distinct types of haploid organisms, indicated by the open and filled shapes. (**b**) A new favorable mutation occurs (triangles) and displaces a random fraction (in this case one-fourth) of each of the pre-existing types (shaded area).

only for the sake of convenience. In the model, the region occupied by each allele would actually be dispersed. Indeed, as noted above, the two-dimensional interpretation of the model is only one of many possibilities. The key assumption in the model is that each new allele displaces exactly the same proportion of each of the pre-existing alleles and that these proportions are the same as its overall frequency in the entire habitat.

The numbers $\{Z_k\}$ are the analogues of fitnesses in classical population genetics models. These will be assumed to be independent random variables with a common distribution subject to the restriction $Z_k > 0$. After the arrival of the nth new mutation, the frequencies of the n most recently arrived alleles will be

$$Z_n,\ Y_n Z_{n-1},\ Y_n Y_{n-1} Z_{n-2},\ \ldots,\ Y_n Y_{n-1} \ldots Y_2 Z_1 \tag{1}$$

where $Y_k = 1 - Z_k$.

We now consider the characteristics of samples from a population undergoing the evolutionary process described above. Suppose that a sample of size r has individuals distributed among k different types with, specifically, n_1 individuals of type 1, n_2 individuals of type 2, ..., n_k individuals of type k. We describe sample configurations by functions $\beta = \beta(x) = \text{card}\ \{i\colon n_i = x\}$, where the designation "card" means "number of". Thus $\beta(x)$ is the number of distinct types that are represented by exactly x individuals in the sample. Let $P(n, r, \beta)$ be the probability that a sample of size r taken from the model after the introduction of the nth allele has the configuration β. Since the fitnesses $\{Z_k\}$ are independent and identically distributed, the frequencies in (1) have the same joint distribution as

$$Z_1,\ Y_1 Z_2,\ Y_1 Y_2 Z_3,\ \ldots,\ Y_1 Y_2 Y_3 \ldots Y_{n-1} Z_n \tag{2}$$

where, for example, the second term in (2) always refers to the second most recently arrived species. The effect of increasing n is just to add more terms to the end of (2). Thus

$$P(n, r, \beta) \to P(r, \beta, F) \quad \text{as} \quad n \to \infty$$

where $P(r, \beta, F)$ is the configuration probability distribution defined by multinomial sampling from the infinite sequence of frequencies

$$Z_1,\ Y_1 Z_2,\ Y_1 Y_2 Z_3,\ \ldots,\ Y_1 Y_2 Y_3 \ldots Y_{n-1} Z_n,\ \ldots \tag{3}$$

and

$$F(\mathrm{d}x) = \text{Prob}\,[Z_k \in \mathrm{d}x]$$

is the probability distribution function of Z.

In a sample of size r, the probability that exactly x are of the same allelic type as the most recently arrived type is $\binom{r}{x} Z^x (1-Z)^{r-x}$, where Z is the frequency of this type. This consideration leads to the recurrence relation

$$P(n+1, r, \beta) = \sum_{x=0}^{r} \binom{r}{x} E[Z^x (1-Z)^{r-x}]\, P(n, r-x, \beta - x^1) \tag{4}$$

where $\beta - x^1$ refers to the configuration β with one fewer $x$plet i.e., less a type with exactly x representatives in the sample). If β has no $x$plets, the corresponding term in (4) does not occur. Since $P(n, r, \beta) \to P(r, \beta, F)$ as $n \to \infty$, the sampling formula $P(r, \beta, F)$ also satisfies (4), and consequently

$$P(r, \beta, F) = \sum_{x=1}^{r} \binom{r}{x} \frac{E[Z^x (1-Z)^{r-x}]}{1 - E[(1-Z)^r]}\, P(r-x, \beta - x^1, F) \tag{5}$$

where the 0th term has been subtracted from both sides of (4) and the equation divided through by $1-E(1-Z)^r$. The sampling formula $P(r, \beta, F)$ is thus uniquely determined by (5) and the boundary conditions

$$P(0, \phi, F) = P(1, 1^1, F) = 1$$

where ϕ denotes the empty configuration and 1^1 represents a sample consisting of a single copy of a single allele, which is the only possible configuration of a sample of size 1. The coefficients in (5) depend of course on $F(\mathrm{d}x)$. We were not able to find a closed-form expression for $P(r, \beta, F)$ for an arbitrary fitness distribution $F(\mathrm{d}x)$. However

Theorem 1. *The sampling distribution $P(r, \beta, F)$ is*

$$P(r, \beta, F) = \frac{\theta^k r!}{L_r(\theta) n_1 \dots n_k \beta(1)! \dots \beta(r)!} \tag{6}$$

where $L_r(\theta) = \theta(\theta+1)\dots(\theta+r-1)$ if and only if

$$F(\mathrm{d}x) = \mathrm{Prob}[Z_k \in \mathrm{d}x] = \theta(1-x)^{\theta-1}\,\mathrm{d}x \tag{7}$$

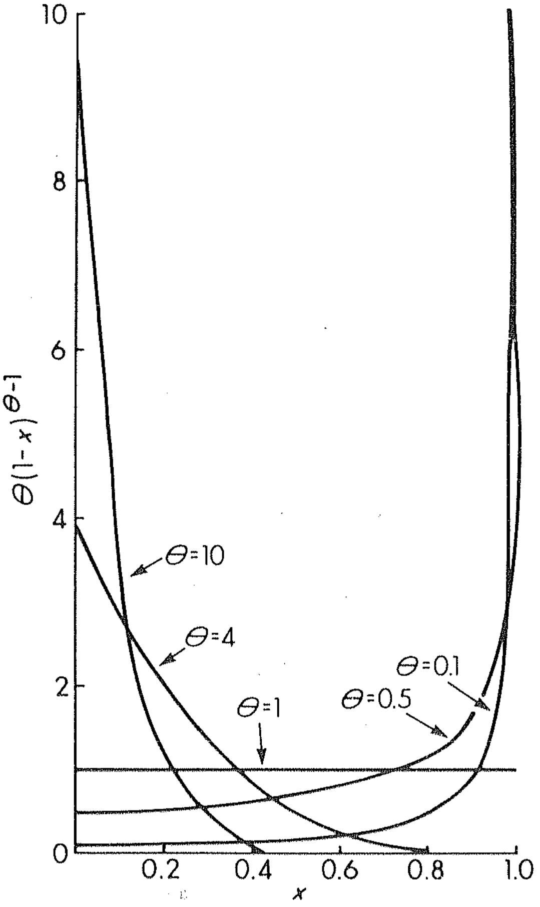

Figure 2. Examples of the power law distribution defined in (7) for various values of θ.

The sampling distribution (6) has come to be known in population genetics as the Ewens distribution (Ewens 1972, 1979a,b), but it is usually associated with selective neutrality. The family of fitness distributions (7) allows a wide scope in overall shape. Some examples are shown in figure 2 and range from a uniform distribution ($\theta = 1$) to distributions skewed to the left or to the right.

One direction of theorem 1 is easy to verify. Given (7),

$$\begin{aligned} E[Z^x(1-Z)^{r-x}] &= \theta \int_0^1 y^x(1-y)^{r-x+\theta-1}\,\mathrm{d}y \\ &= \theta\frac{\Gamma(x+1)\Gamma(r-x+\theta)}{\Gamma(r+1+\theta)} = \frac{\theta}{r+\theta}\,\frac{x!\,L_{r-x}(\theta)}{L_r(\theta)} \end{aligned} \tag{8}$$

In particular, $E[(1-Z)^r] = \theta/(r+\theta)$. Rearranging (5) and substituting from (8),

$$\begin{aligned} &\sum_{x=1}^{r}\binom{r}{x}\frac{E[Z^x(1-Z)^{r-x}]}{1-E[(1-Z)^r]}\,\frac{P(r-x,\beta-x^1,\theta)}{P(r,\beta,\theta)} \\ &\quad=\frac{\theta}{r}\sum_{x=1}^{r}\frac{P(r-x,\beta-x^1,\theta)L_{r-x}(\theta)/(r-x)!}{P(r,\beta,\theta)L_r(\theta)/r!} \end{aligned} \tag{9}$$

However, (6) implies

$$\frac{P(r-x,\beta-x^1,\theta)L_{r-x}(\theta)/(r-x)!}{P(r,\beta,\theta)L_r(\theta)/r!} = \frac{x\beta(x)}{\theta} \tag{10}$$

which, when substituted into the right-hand side of (9), yields 1 because $\sum x\beta(x) = r$. The boundary conditions are straightforward to verify.

Thus, the unique sampling distribution corresponding to the fitness distribution (7) is (6). Conversely, we show in the appendix that the sampling formula (6) implies that the fitness distribution must have been the power law (7).

2.1 *Remarks*

(i) Kingman (1975, 1978) has shown that (6) uniquely characterizes random sampling from a population whose type frequencies have the Poisson–Dirichlet joint distribution. Thus the frequencies in (3) must be the Poisson–Dirichlet type frequencies, although not necessarily in decreasing order. Indeed, Watterson (1976) obtained the joint distribution of the Poisson–Dirichlet frequencies when arranged in decreasing order, which turns out to be much more complicated than (3), (7).

(ii) The representation (3) suggests an efficient algorithm for simulating the Ewens distribution for a given value of θ. Stewart (1977) gives an algorithm for simulating (6) conditional on a fixed value of k. If one simulates population frequencies $\{p_n\}$ according to (3) and then uses Stewart's technique for multinomial sampling from $\{p_n\}$, the resulting configurations will be distributed according to the Ewens sampling formula.

(iii) Let $\{p_n\}$ be the frequencies in (3) for an arbitrary fitness distribution $F(\mathrm{d}x)$. Then by (3)

$$\lim_{n\to\infty}\frac{\log p_n}{n} = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\log Y_k = E[\log(1-Z)] \tag{11}$$

almost surely, by the law of large numbers. Thus, for an arbitrary fitness distribution F (dx), the frequencies $\{p_n\}$ decay approximately exponentially fast, and have the same approximate asymptotic behavior as the Poisson–Dirichlet frequencies with $\theta = 1/E[-\log(1-Z)]$. In contrast, the population frequencies defined by Zipf's Law or by the stable subordinators of Kingman (1975) decay only algebraically fast. To this extent, the sampling formulas obtained for an arbitrary fitness distribution F(dx) may not be very different from the Ewens distribution corresponding to the same value of $E(Z)$.

(iv) It might seem that one could generate sampling formulas for a more general family of Z by equating the terms in (10) to $x\beta(x)/\theta$ and solving for $P(r, \beta, F)$. However, the sampling distributions generated in this way would satisfy

$$P(r, \beta, F)\frac{x\beta(x)}{r} = c(r, x)P(r-x, \beta - x^1, F) \tag{12}$$

for functions $c(r, x)$. Note that (12) says that the distribution of $\beta - x^1$, where x^1 is the allelic class of a gene chosen at random from the sample, is $P(r-x, \beta - x^1)$. Unfortunately, this is a characterization of the Ewens sampling formula due to Kingman (1978), and leads to no new formulas. One can also consider configuration distributions defined by (12) with $\beta(x)/k$ in place of $x\beta(x)/r$, where k is the number of distinct types represented in the sample. This does lead to a consistent family of sampling formulas which can be obtained in closed form. However, one can show that none of them are of the form $P(r, \beta, F)$ for any distribution F(dx). If the $\{Z_k\}$ are constant (i.e., if F (dx) is concentrated at one point) then the frequencies in (1) or (3) are geometric, and $P(r, \beta, F)$ describes configurations resulting from multinomial sampling from a fixed geometric distribution. In that case, expressing $P(r, \beta, F)$ in closed-form for an arbitrary k seems to be a difficult combinatorial problem.

3. Species abundance and residual allocation models

There are a number of models in the literature with type frequencies of the form (3). McCloskey (1965) has a "species intensity" model which was shown, under natural assumptions, to have Poisson–Dirichlet type frequencies. McCloskey proved, among other things, the following result. For $k = 1, 2, \ldots$, let M_k be the *population* frequency of the kth distinct new type discovered in a large sequential sample from a source with Poisson–Dirichlet type frequencies. Then the frequencies $\{M_k\}$ are of the form (3) where the random variables $\{Z_k\}$ are independent with the distribution (7). An algorithm for simulating sampling formulas for fixed θ was derived which was similar to that suggested in remark (ii) above (§2.1). In addition, goodness of fit methods were used to show that the distribution of 15,609 individuals among 233 species of *Macrolepidoptera* was consistent with Poisson–Dirichlet frequencies, but that the distribution of 4112 species among 826 genera of *Orthoptera* was not.

Engen (1975) considered type frequencies $\{\pi_k\}$ defined by (3) where the random variables $\{Z_k\}$ are independent with distribution (7). While it was not proven (or claimed) that the $\{\pi_n\}$ were the Poisson–Dirichlet, calculations were made showing that the *frequency spectrum* (Ewens 1972) was the same as that of the comparable Poisson–Dirichlet. Patil and Taillie (1977) defined a similar model:

> "Let the total resource be represented by the unit interval .... A first species preempts a fraction Q_1 of the total, then a second species preempts a fraction Q_2 of the residual $1-\pi_1$, then a third species preempts a fraction Q_3 of the new residual $1-\pi_1-\pi_2$, etc. The random community is called a *residual allocation model* (or a *preemption model*) when the residual fractions are independently distributed ...."

(Patil and Taillie 1977, p. 507). If the types in a population are ordered by their appearance in a sample (as in McCloskey's result quoted above), this is called a *size-biased permutation* of the model. Patil and Taillie continue:

> "It is well-known that the Dirichlet with parameters s and k is a residual allocation model with $Q_i \approx \text{Beta}[k, (s-i)k]$. The size-biased permutation of the Dirichlet is also a residual allocation model but with $Q_i \approx \text{Beta}[k+1, (s-i)k]$."

Using this statement and standard results (see e.g. Watterson 1976), Patil and Taillie quickly conclude that the frequencies in Engen's model are the Poisson–Dirichlet. Since the type frequencies $\{\pi_n\}$ are statistically the same as (3), this leads to an alternate proof of theorem 1.

Note that the residual allocation model is similar to the local selection model (1)–(3) discussed above, but not identical. In the model of (1)–(3), the nth type preempts the same fraction of everything previous allocated, and so is, in this sense, much less polite than the nth type in Patil and Taillie's model. Donnelly and Tavare (1985) have a detailed description of age structure in the infinite alleles model, and remark that their age ordering is consistent with Engen's model but is the reverse of that implied by (1)–(3).

4. Discussion

As mentioned earlier, the Ewens distribution (6) is the distribution expected in samples of selectively neutral alleles, but it also arises in a class of selection models in diploids with temporally varying selection coefficients (Gillespie 1977), as well as in other contexts in population genetics (Rothman and Templeton 1980). It is remarkable that the same sampling distribution also holds for *haploid* populations undergoing the type of local selection described here. This has relevance to the interpretation of the widespread electrophoretic variation found in natural populations of *E. coli* (Milkman 1973; Selander and Levin 1980). Dr. David Haymer of our (D.H.) laboratory has carried out the homozygosity test of Watterson (1978) to determine whether the sample configurations in *E. coli* data provided by Selander and Levin conform to what would be expected from the Ewens distribution. The results are as yet unpublished, but tests of 20 loci sampled in 109 clones revealed only one locus deviating significantly at the 4·4% level, which is about what one would expect were all the alleles selectively neutral. The simplest interpretation of this finding is that the electrophoretic variants are selectively neutral or nearly neutral, which has been shown directly in a number of instances (Dykhuizen and Hartl 1980, 1983; Dykhuizen *et al* 1984; Hartl and Dykhuizen 1981, 1984). However, the results of statistical tests based on sample configurations must be interpreted with caution, as the sort of local selection considered here can also lead to the Ewens sampling formula.

There is a species abundance interpretation of the local selection model which is congruent with the genetic interpretation. In this interpretation a rich habitat, as, for example, on an island, is successively invaded by a series of closely related species from a mainland, each newly arriving species replacing a random fraction of each of the residents. On a much longer time scale, one may consider a larger geographical area with each invading species being newly evolved. In either case the sampling distribution of species abundances will be governed by the fundamental recursion (4), and in the case when F is given by the power law (7), sampling will again conform to a Ewens distribution. Along this line, we remark that McCloskey (1965) found that the distribution of 15,609 individuals among 233 species of *Macrolepidoptera* was consistent with Poisson–Dirichlet frequencies.

The local selection model presented here can also model age structure under appropriate conditions. Suppose that organisms in a forest or colony are initially all the same age. In the kth year thereafter, a fixed proportion Z_k of each age class dies and is replaced by organisms of age zero. After n years, the proportions of the n youngest age cohorts will be given by (2). In particular, if the $\{Z_k\}$ are independent random variables with the power law distribution (7), the equilibrium distribution of a sample of organisms among various age cohorts will be the Ewens distribution.

Note that after a new type has been established, it has exactly the same selective advantage or disadvantage against future favorable mutations as any other pre-existing type. Thus local selection is a model of a kind of "impulsive" selection in which a favorable mutation has an initial favorable advantage, but is thereafter selectively neutral with respect to older types. While this may be difficult to reconcile with classical selection models, it may not be unreasonable in the context of multiple microhabitats or the age-structured model described above.

Acknowledgements

We wish to thank Dr Warren J Ewens for his encouragement and helpful comments on the manuscript. This work was partially supported by NSF grant MCS 82-02858 to S.S. and by NIH grants GM 33741 and GM 30351 to D.H.

Appendix A

We prove a result that is slightly stronger than is needed for the converse of theorem 1:

*Theorem A*1. *The limiting distribution $P(r, \beta, F)$ determines $F(\mathrm{d}x)$ uniquely. That is, if*

$$P(r, \beta, F_1) = P(r, \beta, F_2) \quad \textit{for all } \beta$$

for two possible fitness distributions F_1 and F_2, then $F_1 = F_2$.

In particular, if $F(\mathrm{d}x)$ is such that $P(r, \beta, F)$ is the Ewens distribution (6), then $F(\mathrm{d}x)$ must be (7).

First, let β be the configuration with r singlets; i.e., $\beta(1) = r$ and $\beta(x) = 0$ for $x > 1$. Set $a(r) = P(r, \beta, F_1) = P(r, \beta, F_2)$. Since β has only singlets, there is only one term in (5), and

$$a(r) = \frac{rE[Z(1-Z)^{r-1}]}{1-E[(1-Z)^r]}\, a(r-1) = r\,\frac{E(Y^{r-1})-E(Y^r)}{1-E[(1-Z)^r]}\, a(r-1)$$

where $Z = 1 - Y$ has the distribution either F_1 or F_2, and

$$\frac{1-E(Y^{r-1})}{1-E(Y^r)} = 1 - \frac{a(r)}{ra(r-1)} = h(r)^{-1}.$$

Hence for $r \geqslant 2$

$$1-E(Y^r) = [1-E(Y)]h(r)h(r-1)\ldots h(2) = E(Z)Q(r)$$

where $Q(r)$ depends only on $P(r, \beta, F_i)$. If $C_i = E(Z_i) = \int xF_i(\mathrm{d}x)$, where $Z_i = 1 - Y_i$ has the distribution F_i, then

$$C_1[1-E(Y_2^r)] = C_2[1-E(Y_1^r)] = C_1C_2Q(r) = E(Z_1)E(Z_2)Q(r)$$
$$= C_1[1-\int(1-x)^r F_2(\mathrm{d}x)] = C_2[1-\int(1-x)^r F_1(\mathrm{d}x)]$$

for $r \geqslant 0$. Thus the measures $C_1[\delta(\mathrm{d}x) - F_2(\mathrm{d}x)] = C_2[\delta(\mathrm{d}x) - F_1(\mathrm{d}x)]$, where $\delta(\mathrm{d}x)$ puts unit mass at zero. Since fitness distributions have $Z > 0$ by assumption, we conclude $F_1 = F_2$.

References

Donnelly P and Tavare S 1985 *Adv. Appl. Probab.* (in press)

Dykhuizen D E and Hartl D L 1980 *Genetics* 96: 801–817

Dykhuizen D E and Hartl D L 1983 *Genetics* 105: 1–18

Dykhuizen D E, de Framond J and Hartl D L 1984 *Mol. Biol. Evol.* 1: 162–170

Engen S 1975 *Biometrika* 62: 697–699

Ewens W 1972 *Theor. Popul. Biol.* 3: 87–112

Ewens W 1979a *Theor. Popul. Biol.* 15: 205–216

Ewens W 1979b *Mathematical population genetics. Biomathematics* (New York: Springer-Verlag) Vol 9

Gillespie J 1977 *Nature* (*London*) 266: 443–445

Hartl D L and Dykhuizen D E 1981 *Proc. Natl. Acad. Sci. USA* 78: 6344–6348

Hartl D L and Dykhuizen D E 1984 *Annu. Rev. Genet.* 18: 31–68

Kingman J F C 1975 *J. R. Stat. Soc. B*32: 1–22

Kingman J F C 1978 *Proc. R. Soc. London A*361: 1–20

McCloskey J W 1965 *A model for the distribution of individuals by species in an environment.* Ph.D. Thesis, Michigan State University

Milkman R 1973 *Science* 182: 1024–1026

Patil G P and Taillie C 1977 *Bull. Int. Stat. Inst.* 47 Book 2: 497–515

Rothman E and Templeton A 1980 *Theor. Popul. Biol.* 18: 135–150

Selander R K and Levin B R 1980 *Science* 210: 545–547

Stewart F M 1977 *Genetics* 86: 482–483

Watterson G 1976 *J. Appl. Probab.* 13: 639–651

Watterson G 1978 *Genetics* 88: 405–417