THE MAINTENANCE OF ALLELES BY MUTATION[1]

W. J. EWENS[2]

Department of Mathematics, Stanford University, Stanford, California[3]

Received June 15, 1964

IN a recent paper, KIMURA and CROW (1964) have investigated quantitatively the possibility that the wild-type allele is not a single entity, but rather a population of different isoalleles that are indistinguishable from each other by any ordinary procedure. The reasons for investigating this possibility are outlined sufficiently by KIMURA and CROW and are not repeated here.

As the most extreme case it may be assumed that all new alleles which arise by mutation are different from any allele which exists or has existed in the population. Thus the only way in which two alleles may be identical is that they be identical by descent. This is the case considered by KIMURA and CROW and is also the case considered in this paper. When the process settles down to equilibrium, there will exist a variable number of different alleles in the population, maintained by a balance between loss of alleles by mutation and drift and a creation of new alleles by mutation.

The quantity of interest to KIMURA and CROW was called the "effective number" n of alleles maintained in the population, defined as the reciprocal of the probability that an individual selected at random in the population be homozygous, or alternatively as the reciprocal of the equilibrium inbreeding coefficent. KIMURA and CROW have shown that the relatonship between n, the mutation rate u to new alleles, and the population size N_e, is given, in the case of selectively neutral genes, by the formula

$$n = 4N_e u + 1 \tag{1}$$

if terms of order u^2 are ignored. In this paper we consider not the effective number of alleles but the actual number of alleles, or more exactly we consider the mean value $\overline{n}$ of the actual number. We consider in turn the case of selectively neutral isoalleles and then the extension to heterotic alleles.

We shall maintain throughout the same notation as KIMURA and CROW. Further, we shall assume that the process has been continuing for a sufficiently long time to suppose that the equilibrium state has been reached.

Selectively Neutral Alleles

Since the population under consideration is diploid and of size N_e, we expect on the average, $2N_e u$ new alleles to arise per generation by the mutation rate u.

[1] Research supported by National Institutes of Health, Contract GM 10452-01A1.

[2] On leave from Statistics Department, Australian National University, Canberra, Australia.

[3] See footnote 2 for present address.

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024

At equilibrium, these new alleles must be balanced by a similar number of "old" alleles being lost by drift and/or mutation. If we suppose that at equilibrium there are $\bar{n}$ different alleles, on the average, in each generation, and that the mean number of generations that any new allele exists in the population before being lost is $\bar{t}$, then the relation

$$2N_e u = \bar{n}/\bar{t} \tag{2}$$

will express the required balance between new alleles being formed and "old" alleles being lost. Thus if an expression can be found for $\bar{t}$, the value of $\bar{n}$ follows immediately.

In order to find $\bar{t}$, it is necessary to set up a model to describe the behavior of any newly formed allele. The model considered here is a particular case of that due to WRIGHT (1931). If in any generation the number of genes of a particular allele is i, then we expect that in the next generation the number of genes of the same allele will be $i(1-u)$, the decrease $-iu$ being due to mutation to new alleles. Thus the model is that the probability p_{ij} that the number of genes of the allele in question changes from i to j in consecutive generations is given by

$$p_{ij} = \binom{2N_e}{j}\left(\frac{i(1-u)}{2N_e}\right)^j \left(\frac{2N_e - i(1-u)}{2N_e}\right)^{2N_e - j} \tag{3}$$

Since any new allele occurs initially exactly once, the Markov chain characterized by (3) and the initial condition $i_0 = 1$ are sufficient in principle to determine the mean time $\bar{t}$ until the allele in question is lost forever from the population, an event which happens eventually with probability unity. In practice, however, $\bar{t}$ seems to be very difficult to find by using (3) and some approximation is necessary. The approximation used here is to replace the discrete process (3) by a continuous diffusion process, a procedure which is valid, for all practical purposes, whenever u is of order $(N_e)^{-1}$.

Details of the diffusion process approximating (3) have been given by the present author (1964). For our purposes, the result is that a close approximation to $\bar{t}$ is given by the expression

$$\bar{t}^* = 2\int_{(2N_e)^{-1}}^{1} x^{-1}(1-x)^{4N_e u - 1}\, dx \quad \text{generations} \tag{4}$$

Thus, using (2), a close approximation to the mean number of alleles maintained in a diploid population of size N_e with mutation rate u to entirely new alleles, in the case u of order $(N_e)^{-1}$, is given by

$$\bar{n} = 4N_e u \int_{(2N_e)^{-1}}^{1} x^{-1}(1-x)^{4N_e u - 1}\, dx \tag{5}$$

It is interesting to compare the number $\bar{n}$ obtained from this formula with the effective number n defined by KIMURA and CROW. This is done in Table 1 for various values of N_e and u. The values given in this table have been chosen to make evaluation of (5) comparatively simple; for wider values of N_e and u, evaluation of (5) becomes tedious and numerical methods may be required.

As expected by KIMURA and CROW, the mean number exceeds the effective number in all cases. We shall examine in detail why the excess is as large as it

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024

TABLE 1

Mean number ($\bar{n}$) *and effective number* (n) *of alleles maintained for various* N_e *and* u

$u = \frac{1}{4} \times 10^{-6}$	N_e			
	10^6	2×10^6	3×10^6	4×10^6
$\bar{n}$	13.8	27.0	40.2	53.5
n	2	3	4	5
$N_e = 250{,}000$	u			
	10^{-6}	2×10^{-6}	3×10^{-6}	4×10^{-6}
$\bar{n}$	12.4	22.9	32.8	42.4
n	2	3	4	5

is in the next section. For the moment we make some qualitative statements by examining equation (5). We may say immediately that for fixed N_eu (i.e., fixed expected number of new alleles per generation), the mean number $\bar{n}$ of different alleles increases slowly with N_e (since N_e occurs, other than in a product N_eu, only in the lower terminal of the integral). This behaviour is to be expected, since increasing u and decreasing N_e (with N_eu fixed), while maintaining a fixed number of newly formed alleles, will lead to a decreasing total number of alleles since those alleles already in existence will tend to disappear faster with the higher mutation. Another way of noting the same phenomenon is to observe, from Table 1, that for fixed N_e, $\bar{n}$ increases somewhat less than linearly with u.

The distribution of allele frequency: The original derivation of (4) allows a much more detailed examination of the distribution of allele frequency to be made. The approximation (4) is a particular case of a more general approximation, which is that the mean number of generations for which a given allele has a frequency in any range (x_1,x_2) $[(2N_e)^{-1} \leq x_1 < x_2 \leq 1]$ before being lost is

$$2\int_{x_1}^{x_2} x^{-1}\,(1-x)^{4N_eu-1}\,dx \qquad \text{generations.} \tag{6}$$

In fact (4) is obtained by putting $x_1 = (2N_e)^{-1}$, $x_2 = 1$. In the present case we may use the above formula to derive a more general formula than (4). This is that in any generation, the mean number of alleles in the population which have frequency between x_1 and x_2 is

$$4N_eu\int_{x_1}^{x} x^{-1}(1-x)^{4N_eu-1}\,dx \qquad \text{alleles.} \tag{7}$$

It is therefore useful to discuss the function

$$f(x) = 4N_eu\,x^{-1}(1-x)^{4N_eu-1} \tag{8}$$

For fixed N_eu, this function increases as x approaches $(2N_e)^{-1}$. This indicates that on the average, a large number of alleles will occur which have only a very small frequency. Among these will probably be many of those alleles only recently formed in the population. On the other hand, if $4N_eu$ is small enough, $f(x)$ increases also at $x = 1$, indicating that for very small mutation rates, the

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024

most likely situation is that where some allele has become temporarily fixed, or almost so, in the population. This agrees with what would be expected on common-sense grounds. Thus the curve of $f(x)$ is either J-shaped or U-shaped, indicating that many alleles occur with small frequency, alleles with moderate frequency occur rarely, while for small enough mutation rate, a single allele will entirely, or almost entirely, occupy most of the population.

This indicates why the numerical values of n and $\bar{n}$ in Table 1 differ as much as they do. The large number of alleles occurring with small frequency contribute a correspondingly large amount to $\bar{n}$, but very little to n, since it is very unlikely that an individual chosen at random from the population will be homozygous for one of the rare alleles.

The total mean frequency of all alleles, which must be unity, should be derived by multiplying any frequency by the mean number of alleles having that frequency, and adding over all possible frequencies. For the continuous diffusion approximation, this quantity is

$$\int_{(2N_e)^{-1}}^{1} xf(x)\,dx = 1 - 0(N_e^{-1}) \tag{9}$$

The small deviation from unity in this formula may be shown to be due entirely to approximations made in passing from (3) to (4). Furthermore, the coefficient of inbreeding, or the probability that an individual chosen at random be homozygous, will be

$$\int_{(2N_e)^{-1}}^{1} x^2 f(x)\,dx \tag{10}$$

which reduces to

$$\frac{1}{4N_e u + 1} + 0(N_e^{-1}) \tag{11}$$

in agreement with the result of KIMURA and CROW. This derivation, in fact, brings out in an interesting way the difference between n and $\bar{n}$. If we define $g(x) = xf(x)$, then $g(x)$ is a density function (ignoring the small error in (9)). In fact the probability that a gene chosen at random in the population comes from an allele having frequency in the population between x and $x + dx$ is $g(x)\,dx$. Then from (5), (10) and (11),

$$n = \left[\int_{(2N_e)^{-1}}^{1} xg(x)\,dx \right]^{-1}$$

$$\bar{n} = \int_{2N_e}^{1} x^{-1} g(x)\,dx.$$

Thus the difference between n and $\bar{n}$ is essentially the difference between the reciprocal of a mean and the mean of a reciprocal.

Heterotic Alleles

The analysis for the case of selectively neutral alleles may be extended immediately to the case of heterotic alleles. Equation (2) will still hold, and all that is necessary is to replace formula (4) for $\bar{t}^*$. This is done by setting up a model analogous to (3), incorporating extra terms allowing for the heterosis. Once

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024

more it is necessary to use a diffusion approximation. This approximation is found by considering the mean $M(\delta x)$ and the variance $V(\delta x)$ of the change δx in the frequency x of any allele in consecutive generations. KIMURA and CROW (1964) have shown that to a sufficiently close approximation,

$$(12) \qquad \begin{cases} M(\delta x) = -ux - sx(x-F) \\ V(\delta x) = x(1-x)/2N_e \end{cases}$$

Here s is the selective advantage of heterozygotes, and $F = \sum x_i^2$ is the sum of squares of the frequencies of all alleles currently in the population. Following KIMURA and CROW, we make the approximation that F may be replaced by its mean value, which is now denoted F.

In order to justify the use of the diffusion methods used to approximate $\bar{t}$, it is assumed that both u and s are of order N_e^{-1}. Thus F will be a sufficiently close approximation to the coefficient of inbreeding. Using an immediate extension of the formula employed for selectively neutral alleles, it is found that the diffusion approximation for $\bar{t}$ is

$$(13) \qquad \int_{(2N_e)^{-1}}^{1} n(x)\, dx \qquad \text{generations,}$$

where

$$(14) \qquad n(x) = 2x^{-1}(1-x)^{4M+4S(1-F)-1}\, exp(4Sx)$$

and $S = N_e s$ and $M = N_e u$, both being of order unity.

Combining equations (2) and (13), we find that to a close approximation, the mean number of alleles present in the population is

$$(15) \qquad \bar{n} = 2M \int_{(2N_e)^{-1}}^{1} n(x)\, dx$$

Evaluation of F: Once an expression for F has been found, we can evaluate the right-hand side in (15) for any S and M and hence find $\bar{n}$. We evaluate F, which is a function of S and M, as follows. In the same way that (6) extends equation (4), we may say that the mean number of generations for which the frequency of any allele lies in an arbitrary range (x_1,x_2), where $(2N_e)^{-1} \leq x_1 < x_2 \leq 1$, is

$$(16) \qquad \int_{x_1}^{x_2} n(x)\, dx \qquad \text{generations.}$$

Thus we may extend (15) and state that in any generation, the mean number of alleles having frequency between x_1 and x_2 is

$$(17) \qquad 2M \int_{x_1}^{x_2} n(x)\, dx \qquad \text{alleles.}$$

Since we require total mean allele frequency to be unity, we must have

$$(18) \qquad 2M \int_{0}^{1} x\, n(x)\, dx = 1,$$

where now a negligible error is introduced by replacing the terminal $(2N_e)^{-1}$ by zero. Equation (18) determines F implicitly in terms of S and M. As $S \to 0$, the solution for F approaches $(4M+1)^{-1}$, the solution obtained by KIMURA and

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024

CROW for selectively neutral alleles. Further, with F defined by (18), it follows readily that the equation

$$2M \int_0^1 x^2 n(x)\, dx = F \tag{19}$$

holds. This is as we expect, as the left-hand side in (19) is the expected value of $\sum x_i^2$.

Thus for any fixed M and S, we find $\bar{n}$ by first solving either (18) or (19) for F, and then insert the value obtained in (15). Clearly, unless M and S take special values, this process will require numerical methods.

The quantity F was the parameter of interest to KIMURA and CROW. In order to obtain a useable formula, they replaced the value of $V(\delta x)$ in (2) by $x/2N_e$, and thus derived their equations (16) and (19) for F. It may be shown that if the formula $x(1-x)/2N_e$ is retained, then their (new) equations corresponding to (16) and (19) agree with our equation (18). In the present paper, we shall use the more accurate formula (18) for F because our primary interest lies in equation (15), and the resemblance of the integrals in (15) and (18) can be utilized when calculating $\bar{n}$.

The effect of heterosis: We are now able to examine the effect of heterosis on the distribution of the frequencies of alleles, and on the mean and effective number of alleles. In doing this we may consider both positive and negative values of S; that is, we may also consider the case where the heterozygote has a selective disadvantage.

Without heterosis, equation (17) indicates that the mean number of alleles having frequency between x_1 and x_2 in any generation is

$$4M \int_{x_1}^{x_2} x^{-1}(1-x)^{4M-1}\, dx \qquad \text{alleles.} \tag{20}$$

When heterosis operates, the integrand must be multiplied by the modifying factor

$$(1-x)^{4S(1-F)}\, exp(4Sx). \tag{21}$$

For positive S, this modifying factor increases from unity at $x = 0$, reaches a maximum somewhere in (0,1), and then decreases to zero at $x = 1$. This indicates that with positive S, the mean number of alleles occurring with low frequency is somewhat higher than in the case $S = 0$. Also, the mean number of alleles occurring with high frequency is decreased. This will lead to an increase in both n and $\bar{n}$, although no statement can be made about the comparative proportionate rates of increase of n and $\bar{n}$.

For negative S, the modifying factor (21) decreases from unity at $x = 0$, reaches a minimum in the interior of (0,1), and then increases sharply as x approaches unity. This shows that for negative S, the mean number of low frequency alleles is slightly diminished compared to the case $S = 0$, the mean number of moderate frequency alleles will be diminished even more while the mean number of high frequency alleles increases. This leads to a decrease in both n and $\bar{n}$. Clearly the effect of heterosis corresponds to what is expected on intuitive grounds.

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024

We illustrate the behaviour of n and $\bar{n}$ with a numerical example. Let $2N_e = 10^6$ and $4M = (e-2)^{-1}$. Then for $4S = 0$ we get $n = 2.392$, $\bar{n} = 18.5$ (approximately). For $4S = 1$ we have $n = 2.550$, $\bar{n} = 18.68$. Both n and $\bar{n}$ have increased with increasing S, as expected. The proportionate increase in $\bar{n}$ is .066, while the proportionate increase in n is only about .01. The numerical values suggest that increasing S tends to make n and $\bar{n}$ less unequal, an effect expected by KIMURA and CROW. However, it is clear that very large values of S would be necessary before anything like equality is reached. Approximate equality would possibly be reached in the case of self-sterility alleles, corresponding to infinite S, but this case cannot be considered by the present analysis which requires that S be of order unity.

As a second example we let $2N_e = 10^6$ $M = \frac{1}{2}$. For $S = 0$ we have $n = 3.0000$, $\bar{n} = 25.63$. For $S = -.3985$ we have $n = 2.684$, $\bar{n} = 25.38$. The proportionate decrease in n is .11, while the proportionate decrease in $\bar{n}$ is only .01. It appears that n is much more sensitive to the value of S, whether positive or negative, than is $\bar{n}$. This happens essentially because the modifying factor (21) has more effect on large values of x than on small values, and it is the large values which primarily determine n.

A Note on Population Size

The numerical examples given in this paper refer to populations of the order of 10^6 individuals. Laboratory populations will, of course, be much smaller than this, and it may be asked whether the diffusion approximations used (for example equation (4)) will hold reasonably well for smaller values of N. Numerical results (EWENS 1963) suggest that this is in fact the case, and that for populations as small as 30 or 40, the approximation is surprisingly accurate. It is, of course, necessary that whatever the value of N, use of diffusion methods requires that selective advantages and mutation rates be of no larger order of magnitude than N^{-1}.

The author has benefited greatly from several discussions with PROFESSOR J. F. CROW.

SUMMARY

A mathematical analysis has been made of the number of different alleles in a population when it is assumed that all new alleles which arise by mutation are entirely new types. Because of a very skew distribution of allele frequencies, it is shown that this number will differ considerably from the effective number of alleles, defined as the reciprocal of the probability that an individual chosen at random in the population is homozygous. The effect of heterosis is also considered; it is shown that the effective number of alleles is more sensitive to changes in heterosis than is the actual number of alleles.

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024

LITERATURE CITED

EWENS, W. J., 1963 Numerical results and diffusions approximations in a genetic process. Biometrika **50**: 241–249. —— 1964 Correcting diffusion approximations in finite genetic models. Stanford University Technical Report 4, Contract NIH GM 10452–01A1 (Department of Mathematics), April 1.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49**: 725–738.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16**: 97–159.

Downloaded from https://academic.oup.com/genetics/article/50/5/891/6033066 by guest on 20 April 2024