THE FREQUENCY DISTRIBUTION OF LETHAL CHROMOSOMES IN FINITE POPULATIONS

By Masatoshi Nei

DIVISION OF GENETICS, NATIONAL INSTITUTE OF RADIOLOGICAL SCIENCES, CHIBA, JAPAN

The statistical techniques for studying the population dynamics of recessive lethal genes were developed mainly by Wright,[1–3] who studied the distribution of gene frequencies in populations. In reality, however, the frequencies of individual lethal genes cannot be observed but the frequency of lethal-bearing chromosomes is determined. In large populations the random fluctuation of the frequency of lethal chromosomes is negligibly small, but it becomes appreciably large in intermediate or small populations. Thus, it is desirable to find the frequency distribution of *lethal chromosomes* rather than *lethal genes*.

To obtain a general distribution function of the chromosome frequencies is not easy, but the author obtained several formulas for some important special cases. As we shall see, these formulas give some useful relations among mutation rate, population size, allelic rate, etc., which may be used for the study of genetic structure of populations or the heterozygous effects of lethal genes. In recent years much attention has been focused on the latter problem in view of the fact that the genetic load due to these genes is profoundly affected by a slight positive or negative effect in the heterozygotes.[4, 5]

Partially Recessive Lethals.—Consider a randomly mating population of size N, in which a lethal gene a and its allelic normal gene A are segregating. Let the frequency of a be q and the fitnesses of the three possible genotypes AA, Aa, and aa be 1, $1 - h$, and $1 - s$, respectively, where s is 1 for a completely lethal gene and $0.5 < s < 1$ for a semilethal gene. The mean fitness, $\overline{W}$, then becomes $1 - 2hq - (s - 2h)q^2$. We assume that the gene A mutates to a at the rate of u per generation, the reverse mutation being negligible. Inserting these quantities into Wright's[1] general formula, we find the frequency distribution of lethal gene in an equilibrium population as follows:

$$\phi(q) \sim e^{-4Nhq-2N(s-2h)q^2}q^{4Nu-1}(1 - q)^{-1}. \tag{1}$$

Noting that q is generally very small compared to unity and that $2N(s - 2h)q^2$ is negligibly small compared to $4Nhq$ if h is much larger than $\sqrt{su}$, the distribution $\phi(q)$ is approximated by the following gamma distribution.

$$\phi(q) = [(4Nh)^{4Nu}/\Gamma(4Nu)]e^{-4Nhq}q^{4Nu-1}. \tag{2}$$

The mean, $\bar{q}$, and variance, σ_q^2, of the gene frequency become

$$\bar{q} = u/h, \qquad \sigma_q^2 = \bar{q}/(4Nh). \tag{3}$$

The expression for $\bar{q}$ indicates that the mean gene frequency is independent of the population size and equal to the value for an infinitely large population. This contrasts with the case of completely recessive genes in which the mean gene frequency decreases as the population size decreases (cf. (15)).
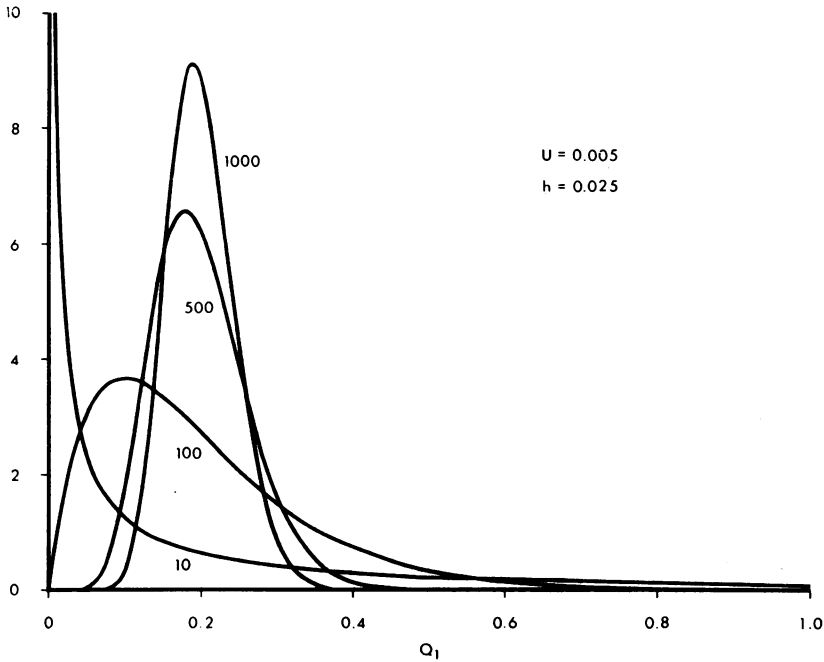
FIG. 1.—Frequency distributions of lethal chromosomes for various sizes of populations. It is assumed that the fitness of the heterozygote for a lethal gene is reduced by 2.5% and the total mutation rate per chromosome (U) is 0.005. The number given to each curve represents the population size.

In order to obtain the distribution of lethal chromosome frequency, we assume that a chromosome carries n loci in which lethal genes may occur and the selection coefficient h is the same for all loci. The mutation rate may vary with loci, and let u_i be the mutation rate at the ith locus.

The frequency of lethal chromosomes, Q, is not a simple sum of the frequencies of lethal genes at the individual loci, so that it is difficult to obtain the distribution of the lethal chromosome frequency itself. However, if we make the following transformation

$$Q_1 = -\log_e(1 - Q), \tag{4}$$

then Q_1 becomes a sum of the individual gene frequencies, i.e., $\sum_{i=1}^{n} q_i$, where q_i denotes the lethal gene frequency at the ith locus. Since a sum of gamma variates is distributed again as a gamma variate, the distribution of Q_1 is easily obtained and becomes

$$\phi(Q_1) = [(4Nh)^{4NU}/\Gamma(4NU)] \, e^{-4NhQ_1} Q_1^{4NU-1}, \tag{5}$$

where $U = \sum u_i$ or nu, u being the mean mutation rate. Some of the forms of this distribution are given in Figure 1.

The mean, $\bar{Q}_1$, and variance, V_{Q_1}, of Q_1 are then given by

$$\bar{Q}_1 = U/h \text{ or } n\bar{q}, \qquad V_{Q_1} = n\bar{q}/(4Nh), \tag{6}$$

where $\bar{q}$ now represents the mean gene frequency over all loci, i.e., $\sum \bar{q}_i/n$. It is seen from the expression for V_{Q_1} that the variance increases with increase of n and with decrease of N or h.

From expression (6) a new method for estimating the value of h is obtained. Namely, if we know the effective population size, h is estimated by

$$\hat{h} = \hat{\bar{Q}}_1/(4N\hat{V}_{Q_1}), \tag{7}$$

where $\hat{\bar{Q}}_1$ and $\hat{V}_{Q_1}$ are the estimates of $\bar{Q}_1$ and V_{Q_1}, respectively. On the other hand, if h is known, the effective size is estimated by

$$\hat{N} = \hat{\bar{Q}}_1/(4h\hat{V}_{Q_1}). \tag{8}$$

In the study of lethal genes, a quantity called *allelic rate* or *frequency of allelism of lethal genes* plays an important role.[2, 3] This quantity is defined as

$$I_g = \sum_{i=1}^{n} q_i^2/Q_1^2 \tag{9}$$

and can be estimated by $-\log_e(1 - I_c Q^2)/[\log_e(1 - Q)]^2$, where I_c stands for the allelic rate of lethal chromosomes. Assuming that the number of lethal gene loci is large, the expectation of I_g becomes

$$E(I_g) = E(\sum q_i^2)/E(Q_1^2) + \sum \text{cov}(q_i^2, 1/Q_1^2) \approx E(\sum q_i^2)/E(Q_1^2)$$
$$= (\bar{q}^2 + \sigma_{q(d)}^2 + \sigma_q^2)/(n\bar{q}^2 + \sigma_{q(d)}^2 + \sigma_q^2), \tag{10}$$

where $\sigma_{q(d)}^2$ is the variance of q due to differences in u_i (if h is variable, the variance due to this factor is also included in this term). Wright's equivalent formula does not have the terms $\sigma_{q(d)}^2$ and σ_q^2 in the denominator.

Substituting σ_q^2 by $\bar{q}/(4Nh)$, $E(I_g)$ reduces to $[4N(\bar{q}^2 + \sigma_{q(d)}^2) + \bar{q}/h]/[4N(n\bar{q}^2 + \sigma_{q(d)}^2) + \bar{q}/h]$. This indicates that the allelic rate increases as the population size decreases. On the other hand, the allelic rate in an infinitely large population or between lethal genes of two completely isolated populations will be $(\bar{q}^2 + \sigma_{q(d)}^2)/(n\bar{q}^2 + \sigma_{q(d)}^2)$. Thus, $E(I_g)$ is roughly $1/n$ when $\sigma_{q(d)}^2$ is small. This property was first used by Dobzhansky and Wright[2] for estimating the minimum number of lethal gene loci. The expected allelic rates for various values of N when $n = 500$, $u = 10^{-5}$, and $\sigma_{q(d)}^2 = 0$ are as follows:

N	10	50	10^2	10^3	10^4	∞
$E(I_g)$	0.834	0.501	0.335	0.050	0.007	0.002

If $\sigma_{q(d)}^2$ is 0 and u and U are known, the effective population size can be estimated by

$$\hat{N} = (1 - \hat{I}_g)/[4(\hat{I}_g U - u)]. \tag{11}$$

This estimator, however, appears to be less efficient than formula (8) or Wright's formula[3] since it depends on the assumption that both h and u are the same for all loci.

Another parameter (h_e) that is often used is a quantity devised to estimate the value of h.   Namely,

$$h_e = (U - I_c Q^2)/Q_1. \tag{12}$$

This formula is due to Crow and Temin[4] and roughly similar to Wright's.[2]   In the present case the expectation of this parameter is approximately given by $h - [4Nh(\bar{q}^2 + \sigma_{q(d)}^2) + \bar{q}]/(4Nh\bar{q})$, which, if $\sigma_{q(d)}^2$ is negligible, reduces to $h[1 - 1/(4Nh^2)]$, approximately.   Therefore, h_e gives an erroneous estimate in small populations; particularly, when $4Nh^2 < 1$, h_e may become negative even if h is positive.   For example, if $h = 0.05$, h_e may be negative for N smaller than 100.

Apparently, a better estimate of h is

$$\hat{h} = \hat{U}/\hat{\bar{Q}}_1, \tag{13}$$

if h is sufficiently large.

Completely Recessive Lethals.—In this case $h = 0$, and making use of the transformation $t = q^2$ and omitting the term $(1 - q)^{-1}$ in (1), we have

$$\phi(t) = [(2Ns)^{2Nu}/\Gamma(2Nu)]e^{-2Nst}t^{2Nu-1}. \tag{14}$$

The mean and variance of q approximately become

$$\bar{q} = \Gamma(2Nu + {}^1\!/_2)/(\sqrt{2Ns}\,\Gamma(2Nu)), \qquad \sigma_q^2 = u/s - \bar{q}^2. \tag{15}$$

When $s = 1$, these expressions are identical to those of Wright.[1, 2]   If $2Nu$ is larger than unity, $\bar{q}$ is approximately $\sqrt{u/s}$, and if $2Nu$ is much smaller than unity, this becomes $u\sqrt{2\pi N/s}$, approximately.

Now consider the quantity $T = \sum_{i=1}^{n} q_i^2$, which can be estimated by $-\log_e(1 - I_c Q^2)$.   Assuming the same value of $s(> 0.5)$ for all loci, the distribution of T is given by

$$\phi(T) = [(2Ns)^{2NU}/\Gamma(2NU)]e^{-2NsT}T^{2NU-1}, \tag{16}$$

and the mean, $\bar{T}$, and variance, V_T, of T by

$$\bar{T} = U/s, \qquad V_T = U/(2Ns^2). \tag{17}$$

Therefore, if lethal genes are completely recessive, then $T = \sum q_i^2$ rather than $Q_1 = \sum q_i$ is distributed as a gamma variate.   This property may be used for studying the heterozygous effect of lethal genes.

The expectation of the allelic rate can be obtained by substituting (15) into (10).   It becomes

$$E(I_q) = \frac{2Ns(u/s + \sigma_{q(d)}^2)\{\Gamma(2Nu)\}^2}{2Ns(u/s + \sigma_{q(d)}^2)\{\Gamma(2Nu)\}^2 + (n - 1)\{\Gamma(2Nu + {}^1\!/_2)\}^2}, \tag{18}$$

which is close to $(u/s + \sigma_{q(d)}^2)/(nu/s + \sigma_{q(d)}^2)$ if $2Nu$ is larger than unity.   On the other hand, if Nu is small, it approximates to $(u + s\sigma_{q(d)}^2)/(u + s\sigma_{q(d)}^2 + 2(n - 1)\pi Nu^2)$.   Thus, with decrease of N, $E(I_q)$ again increases.   If $n = 500$,

$u = 10^{-5}$, $s = 1$, and $\sigma^2_{q(d)} = 0$, the expected allelic rates for various values of N become as follows:

N	10	50	10^2	10^3	10^4	∞
$E(I_g)$	0.762	0.390	0.242	0.031	0.003	0.002

The expectation of h_e is 0 if $s = 1$, as it should be, but if $s < 1$, i.e., for semilethal genes, it becomes negative, since the expectation of $I_e Q^2$ is larger than U.

Overdominant Lethals.—It is not easy to obtain a general formula for this case and the author has been able to find one only for a special case of $4Nu(1 - \tilde{q}) + 2\tilde{q} = 1$ or approximately $4Nu = 1$, where $\tilde{q} = k/(1 + k)$, k being the selective advantage of Aa compared to AA (the fitness of aa is assumed to be 0). If k is much larger than $2u$, the distribution of gene frequency for this special case is given by a normal distribution with mean $\tilde{q} = k/(1 + k)$ and variance $\sigma^2_q = 1/[4N(1 + k) - (1 + k)^3/k]$ or approximately $\sigma^2_q = 1/[4N(1 + k)]$. Therefore, $Q_1 = \sum q_i$ is again distributed as a normal variate with mean $\sum \bar{q}_i$ and variance $\sum \sigma^2_{q_i}$.

Effects of Migration.—The effect of migration is roughly similar to increasing population size. Let m be the rate of migration per generation. Then, if $4Nm$ is much larger than unity, the distribution of gene frequency for partially recessive lethals can be obtained from Wright's general formula and approximately becomes

$$\phi(q) = Ce^{-4N[m(1 - \tilde{q}) + h]q} q^{4N(m\tilde{q} + u) - 1}, \tag{19}$$

where $C = [4N\{m(1 - \tilde{q}) + h\}]^{4N(m\tilde{q} + u)}/\Gamma\{4N(m\tilde{q} + u)\}$ and $\tilde{q}$ is the gene frequency of immigrating individuals or the mean gene frequency in Wright's island model. The mean ($\bar{q}$) and variance (σ^2_q) are then given by $(m\tilde{q} + u)/[m(1 - \tilde{q}) + h]$ and $\bar{q}/[4N\{m(1 - q) + h\}]$, respectively. If $\tilde{q} = u/h$, $\bar{q}$ and σ^2_q approximate to u/h and $\bar{q}/[4N(m + h)]$, respectively. Therefore, $Q_1 = \sum q_i$ is distributed as a gamma variate with the following mean and variance.

$$\bar{Q}_1 = n\bar{q}, \qquad V_{Q_1} = n\bar{q}/[4N(m + h)]. \tag{20}$$

Thus, the effect of migration in this case is simply to replace h by $(m + h)$ in the denominator of variance. This is in effect to decrease the variance for a given value of N.

The distribution of the frequency of completely recessive gene can be obtained in the same manner, if we assume $\overline{W} = 1$, as done by Wright.[2] It becomes a gamma distribution with mean $\bar{q} = \sqrt{u/s}$ and variance $\bar{q}/(4Nm)$ approximately, where $4Nm$ is assumed to be much larger than unity and also $m\bar{q} \gg u$. The mean and variance are roughly equal to those given by Wright. With these approximations $Q_1 = \sum q_i$ is distributed as a gamma variate with mean $n\bar{q}$ and variance $(n\bar{q}/(4Nm))$.

With overdominance the frequency distribution of lethal chromosomes can be obtained only in a special case, i.e., the case of $4Nu = 1$, as before. In this case the gene frequency is distributed as a normal variate with mean $\tilde{q} = k/(1 + k)$

TABLE 1. *Frequencies of the third chromosomes that are lethal or semilethal in American populations of Drosophila pseudoobscura.*

Location	n_0	Q	$-\text{Log}_e$ $(1 - Q)$	Reference
(1) Lida	55	0.255	0.294	(2)
(2) Cottonwood Mts.	93	0.183	0.201	(2)
(3) Grapevine Mts.	56	0.125	0.134	(2)
(4) Cose Range	124	0.177	0.195	(2)
(5) Panamint Mts.	105	0.114	0.121	(2)
(6) Kingstone Range	101	0.119	0.127	(2)
(7) Charleston Mts.	103	0.146	0.158	(2)
(8) Sheep Range	90	0.122	0.130	(2)
(9) Providence Mts.	99	0.152	0.165	(2)
(10) Andreas A	343	0.120	0.128	(3)
(11) Andreas B	317	0.126	0.135	(3)
(12) Pinon A	296	0.139	0.150	(3)
(13) Pinon B	231	0.087	0.091	(3)
(14) Keen A	169	0.172	0.189	(3)
(15) Keen B	139	0.130	0.139	(3)
(16) Keen C	120	0.158	0.172	(3)
(17) Keen D	149	0.181	0.200	(3)
(18) Keen E	130	0.085	0.089	(3)
(19) Wildrose A	102	0.147	0.159	(3)
(20) Wildrose B	94	0.202	0.226	(3)
(21) Mexico-Guatemala	120	0.300	0.357	(3)
(22) Mather (1951)	116	0.250	0.288	(6)
(23) Mather (1957)	140	0.229	0.260	(7)
(24) Austin	153	0.203	0.227	(7)
(25) Colombia	252	0.167	0.183	(8)

n_0 = number of chromosomes examined.
Q = frequency of lethal chromosomes among n_0.

and variance $\sigma_q^2 = 1/[4N\{1 + k + m\bar{q}^{-1}(1 - \bar{q})^{-1}\}]$, approximately. Therefore, the mean and variance of Q_1 are $n\bar{q}$ and $\sum \sigma_{qi}^2$, respectively.

Data on Lethal Chromosomes in Drosophila pseudoobscura.—There are a large number of data on lethal chromosomes in *Drosophila* populations. However, they have been obtained by different authors often with different techniques in various places in the world, so that they are not necessarily suited for the analysis of their frequency distribution. The only data that can be used are those obtained by Dobzhansky and his co-workers in various populations of *Drosophila pseudoobscura* in America, although even these data are never free from the effects of environmental differences among the locations in which samples were taken (the lethal chromosome frequency in Death Valley region (1–9) in Table 1 is slightly but significantly larger than that in San Jacinto region (10–20)[3]). The data on the third chromosomes that are lethal or semilethal are presented in Table 1.

The mean and variance of observed values of Q_1 for these data are 0.181 and 0.004312, respectively. The mean is the estimate of $\bar{Q}_1$ but the variance includes the sampling variance in addition to V_{Q_1}. This sampling variance, V_s, is obtained by $Q/[(1 - Q)n_0]$, where n_0 is the number of chromosomes observed. Since the average sampling variance amounts to 0.00175 in the present case, the estimate of V_{Q_1} becomes 0.00256. The rate of lethal mutations of the third chromosome (U) has been estimated to be 0.0031 by Wright, Dobzhansky, and

Hovanitz.[3] These authors also showed that $U > I_eQ^2$, so that there must be selection against lethal genes in the heterozygous condition or some degree of inbreeding. Assuming no excessive consanguineous matings such as brother-sister mating but a limited size of population, the disadvantage of the heterozygotes is estimated to be $\hat{h} = U/\hat{Q}_1 = 0.0172$, which is close to the estimates by Wright, Dobzhansky, and Hovanitz[3] and Crow and Temin.[4] On the other hand, $N(m + h)$ is estimated by $\hat{\bar{Q}}_1/(4\hat{V}_{Q_1})$ and becomes 17.63. This is perhaps an underestimate but within the range of the estimate of Nm by Dobzhansky and Wright,[2] i.e., between 5 and 40.

Discussion.—Crow and Temin's formula for estimating the degree of disadvantage of the lethal heterozygotes (h_e) is based on the assumption that the random fluctuation of lethal chromosome or lethal gene frequency is negligibly small. In small populations, however, this assumption does not hold and, as shown previously, h_e may be smaller than the true value and even negative for a positive value of h. Nevertheless, the values of h_e so far obtained are mostly positive and suggest that lethal genes have a deleterious effect in the heterozygous condition. The only exception is the data obtained by Wallace (1966) in *Drosophila melanogaster*. In his data h_e is negative and he claims that it is due to an overdominant effect of lethal genes. However, as the allelic rate of his lethal chromosomes is very high, the effective size of the population from which his data were sampled appears to be small, and the negative value of h_e might be due to this small effective size.

Wallace's data includes both lethal and semilethal chromosomes. The frequency of these chromosomes amounts to 79.8 per cent for the second and third chromosomes together. The total mutation rate of *completely lethal genes* for the second and third chromosomes has been estimated to be 0.01. He does not mention the fraction of completely lethal chromosomes in his data. In view of the unusually high frequency of lethals plus semilethals, it appears that a considerable number of semilethals have been included. The fraction of semilethals in other data is variable but often as large as that of complete lethals.[9] If one third of his data are semilethals, then the total mutation rate to be used should be 0.015 instead of 0.01. Thus, from (13) h is estimated to be 0.01, approximately. Therefore, Wallace's data are never incompatible with the idea that lethal genes are slightly deleterious in the heterozygous condition.

The equilibrium mean frequency of an overdominant lethal gene in populations larger than $1/(4u)$ is given by $\bar{q} = k/(1 + k)$ approximately, and the expectation of Q_1 is $\sum q_i$. The number of lethal genes on the second chromosome in *D. melanogaster* has been estimated to be 400 ⌢ 500. Therefore, if all the loci show an overdominance with $k = 0.02$, $\bar{Q}_1$ must be 8 ⌢ 10 in a large population. If $k = 0.01$, $\bar{Q}_1$ must be 4.7 ⌢ 5.9. However, the weighted mean of Q_1, which is obtained from Crow and Temin's survey data (Table 3 in their paper), is 0.433, including semilethals. If the population size of *D. melanogaster* is sufficiently large, this indicates that the number of overdominant lethals must be limited if they exist. Of course, if the effective population size is small, the mean gene frequency could be smaller than $\bar{q} = k/(1 + k)$, and the above argument requires some modification.

So far most studies on the selection against lethal genes have been conducted with natural populations or with a single cage population. A much better experimental method would be to study the frequency distribution of lethal chromosomes in artificially controlled small populations. In these populations the effective size and migration rate can be determined almost exactly, so that the selection against lethal genes can be studied more accurately. This type of experiment has been initiated by M. Murata in our laboratory, using *D. melanogaster*, and his preliminary data indicate that the fitness of heterozygotes for lethal genes is reduced by several per cent.

Kimura, Maruyama, and Crow[10] showed that the genetic load due to deleterious mutations in small populations is much larger than that in large populations. They were, however, concerned with those genes which have rather small deleterious effects. With lethal genes the genetic load is almost independent of population size. The expected genetic load for a partially recessive lethal gene is

$$\bar{L} \approx \int_0^\infty 2hq \; \frac{(4Nh)^{4Nu}}{\Gamma(4Nu)} \; e^{-4Nhq}q^{4Nu-1}dq \; = \; 2u,$$

while the genetic load for a completely recessive lethal similarly becomes u. These values are the same as those for a large population.

Summary.—Formulas for the frequency distribution of chromosomes carrying recessive lethals are worked out, taking into account the heterozygous effect of lethal genes. The methods of estimating the heterozygous effect and the effective population size are also discussed. Analyses and interpretations of *Drosophila* data have indicated that lethal genes are on the average slightly deleterious in the heterozygous condition.

[1] Wright, S., these PROCEEDINGS, **23**, 307–320 (1937).

[2] Dobzhansky, Th., and S. Wright, *Genetics*, **26**, 23–54 (1941).

[3] Wright, S., Th. Dobzhansky, and W. Hovanitz, *Genetics*, **27**, 363–394 (1942).

[4] Crow, J. F., and R. G. Temin, *Am. Naturalist*, **98**, 21–33 (1964).

[5] Wallace, B., *Am. Naturalist*, **100**, 565–578 (1966).

[6] Dobzhansky, Th., and B. Spassky, *Genetics*, **38**, 471–484 (1953).

[7] Spassky, B., N. Spassky, O. Pavlovsky, M. G. Krimbas, C. Krimbas, and Th. Dobzhansky, *Genetics*, **45**, 723–740 (1960).

[8] Dobzhansky, Th., A. S. Hunter, O. Pavlovsky, B. Spassky, and B. Wallace, *Genetics*, **48**, 91–103 (1963).

[9] Oshima, C., and T. K. Watanabe, *Ann. Rept. Natl. Inst. Genet., Japan*, No. 17, 77–78 (1967).

[10] Kimura, M., T. Maruyama, and J. F. Crow, *Genetics*, **48**, 1303–1312 (1963).