Proc. Natl. Acad. Sci. USA
Vol. 78, No. 2, pp. 1129–1132, February 1981
Genetics

Some calculations on the amount of selfish DNA*

(molecular evolution/junk DNA/transposable element/repetitive sequence)

TOMOKO OHTA AND MOTOO KIMURA

National Institute of Genetics, Mishima 411, Japan

Contributed by Motoo Kimura, October 27, 1980

ABSTRACT A population genetical theory is developed to treat the amount of selfish DNA in a genome. We assume that the selfish DNA consists of replicating units and that it evolves by multiplication within a genome, exchange between genomes, and random genetic drift at reproduction. Special reference is made to the mean and variance of the number of replicating units per genome in the population. Under the assumption of no systematic evolutionary pressure, the number of units changes randomly with time, and its variance increases by replication process. Although under certain circumstances the variance increases also by exchange process, under ordinary circumstances this process tends to decrease the variance. Random genetic drift also reduces the variance. The relationship between the mean and variance at equilibrium of the number of replicating units per genome in the population was derived. The results obtained will be useful in understanding various observations on repeated DNA which presumably does not contain genetic information and which is likely to be selectively neutral.

Doolittle and Sapienza (1) and Orgel and Crick (2) discussed the evolution of what they call "selfish DNA." This is a piece of DNA that has little or no phenotypic effect yet spreads in the species because of its rapid replication within the genome. Highly and moderately repetitive nucleotide sequences in eukaryotes are considered to belong to this class of DNA. Their papers suggest a need for a treatment, based on population genetical' theory, of the problem of DNA segments spreading within a genome and subsequently spreading within a population. On the other hand, the repeated gene families such as transfer and ribosomal RNA, histone, and immunoglobulin genes are called multigene families (3). Population genetical consequences of such genetic systems have been worked out (4, 5) by assuming that the gene family size (i.e., the number of repeated genes per family) stays fairly constant over the generations because of its functional requirements. Under such an assumption, the nature of gene diversity was investigated. Also, the family size and proportion of defective genes were examined by Monte Carlo simulation experiments by Hood *et al.* (6).

For a proper treatment of the evolution of selfish DNA, however, we have to take into account the possibility that its total amount changes with time. The purpose of this note is to present a theoretical treatment of the amount of selfish DNA based on population genetics.

BASIC THEORY

Let us assume that selfish DNA consists of replicating units and let n_i be the number of such replicating units in the ith genome in the population. (n_i may increase or decrease by unequal crossing-over as in multigene families (7) or by an integration mechanism as in transposons and insertion sequences of bacteria (2). Let N be the number of breeding individuals in the population. We consider the process of selfish DNA evolving under replication within a genome and under random genetic drift at reproduction within a population. In addition, we assume that exchange of replicating units occasionally occurs at reproduction between the two genomes in a diploid individual—i.e., we consider a population of sexually reproducing diploid organisms. To simplify the treatment, we first neglect the effect of natural selection, although its effect will be discussed later.

Let $\bar{n}$ and σ_n^2 be the mean and variance, respectively, of the amount of selfish DNA per individual genome in the population so that

$$\bar{n} = \sum_{i=1}^{2N} n_i/2N \qquad [1]$$

and

$$\sigma_n^2 = \sum_i n_i^2/2\mathrm{N} - \bar{n}^2. \qquad [2]$$

We now investigate the rates of change of the mean and variance due to various forces.

Multiplication of replicating units may occur through the following two processes: unequal crossing-over as in multigene families (3, 7), and integration as in mobile genetic elements (2, 8). For our purpose, however, we consider the following two types of replication.

(*i*) Duplication or deletion which occurs independently for each unit; each unit of selfish DNA has a constant probability α_1 of either duplicating or being deleted in a genome in each generation. This may be called single replication.

(*ii*) A certain number of units are simultaneously duplicated or deleted in a genome (say, the ith genome) and this number depends on the total number n_i. This may be called cluster replication.

In the first type, single replication (sr), the mean and variance of the change of n_i for a given i are

$$M_{\mathrm{sr}}(\Delta n_i|i) = 0$$

and [3]

$$V_{\mathrm{sr}}(\Delta n_i|i) = \alpha_1 n_i,$$

where M and V denote operators for taking the mean and the variance, respectively, and the subscript sr means that these operations refer to single replication. Δn_i is the change of n_i in one generation. Note that these are conditional expectations for a given i.

In the second type of process, cluster replication (cr), we assume that the mean number of simultaneously duplicated or deleted units is equal to an_i. Also, let α_2 be the probability of such a duplication or deletion occurring in each generation.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

*This paper is contribution no. 1341 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411, Japan.

Then the mean and variance of the change of n_i due to cluster replication are

$$M_{\mathrm{cr}}(\Delta n_i|i) = 0$$

and [4]

$$V_{\mathrm{cr}}(\Delta n_i|i) = \alpha_2 a^2 n_i^2,$$

where the subscript cr refers to cluster replication. Therefore, the expected value of $\bar{n}$ does not change whereas σ_n^2 increases on the average through replications in one generation as follows.

$$\begin{aligned} E_{\mathrm{rep}}(\Delta\sigma_n^2) &= \frac{1}{2N}\sum_i(\alpha_1 n_i + \alpha_2 a^2 n_i^2) \\ &= \alpha_1\bar{n} + \alpha_2 a^2(\bar{n}^2 + \sigma_n^2), \end{aligned} \qquad [5]$$

where E stands for taking the expectation and the subscript rep refers to the replication process.

Next we consider the change of n_i by intergenome exchange such as interchromosomal crossing-over at meiosis. It is convenient to treat the following two processes separately. The first is the treatment of the selfish DNA which is dispersed in the whole genome (8–12), and the second is that of the clustered gene families. Let us call the former the dispersed exchange process, and the latter the clustered exchange process. We shall first formulate the dispersed exchange process. At meiosis, the dispersed replicating units are likely to be equally distributed to the two daughter cells. We assume that, with the rate β_1 per generation, the genomes exchange their selfish DNA with each other, and the dispersed units are equally divided. Then, when the genomes with n_i and n_j units exchange their selfish DNA, the expected number of the units per daughter genome is $(n_i + n_j)/2$.

In order to derive the variance of the change of n_i, we first ask what fraction of the units in the two genomes share homologous positions on the chromosome. If the two units share the same position on homologous chromosomes, they are distributed one/one to the daughter genomes, whereas if they occupy different positions and are thus hemizygous, they would be distributed two/zero, one/one, and zero/two in the ratio 1:2:1—that is, they follow the binomial distribution. We consider the state in which the fraction of homologous units are held in equilibrium between transposition of the units (which decreases homozygosity) and random genetic drift (which increases it). It is assumed that each unit has a probability α_1 of either duplicating or deleting itself each with equal probability of $\alpha_1/2$ in one generation. When duplication occurs, it is assumed that one of the two units jumps to a different position while the other unit remains in the same site. Thus, with probability $u = \alpha_1/2$, a transposable unit occupies a new position. In other words, $\alpha_1/2$ is the probability of a new hemizygous unit being created in a new position. This site later may become homozygous by random drift of chromosomes. Let h be the probability that one unit of a randomly chosen genome has a homozygous partner in another randomly chosen genome. Following the method used by Kimura and Crow (13) for deriving the probability of allele identity, we can show that, in one generation of random drift and jumping of units, h changes to h' according to the following equation.

$$h' = \left(1 - \frac{1}{2N}\right)\left(1 - \frac{\alpha_1}{2}\right)^2 h + \frac{1}{2N}.$$

The value of h at equilibrium, which we denote by $\hat{h}$, may be obtained by putting $h = h'$ in the above formula, and it becomes,

$$\hat{h} = 1/(1 + 2N\alpha_1). \qquad [6]$$

In terms of this fraction, and by letting β be the rate of dispersed exchange process, the variance of the change of n_i due to segregation is

$$\begin{aligned} V_{\mathrm{ds.ex.}}(\Delta n_i|i,j) &= \frac{1}{4}\beta(1 - \hat{h})(n_i + n_j) \\ &= \frac{N\alpha_1\beta(n_i + n_j)}{2(1 + 2N\alpha_1)}, \end{aligned} \qquad [7]$$

where the subscript ds.ex. denotes the dispersed exchange process. On the other hand, the variance of n between genomes is halved by recombination so that, the mean change of σ_n^2 in one generation is

$$E_{\mathrm{ds.ex.}}(\Delta\sigma_n^2) = -\frac{\beta}{2}\sigma_n^2 + \frac{N\alpha_1\beta\bar{n}}{(1 + 2N\alpha_1)}. \qquad [8]$$

The clustered exchange process through crossing-over is more complicated. Let us assume that the two genomes involved in the exchange have segments with n_i and n_j units, respectively, and that $n_j > n_i$ as in Fig. 1. Let ℓ be the number of unpaired units on the longer segment that lie to the left of the left end of the shorter segment (point O in the figure), taking negative values when the pairing is so skewed that point O lies to the left of the left end of the longer segment. Thus, ℓ takes values between $-n_i$ and $+n_j$. The upper diagram shows the probability density of point O. This probability density function implies that the pairing is equally likely for the region $0 \le \ell < n_j - n_i$ and it becomes less frequent for $\ell < 0$ or $\ell > n_j - n_i$. Let us denote these three regions as I, II, and III, calling the middle region region I and the left and the right ones regions II and III. Let β_{I} be the crossing-over frequency per generation in region I ($0 \le \ell \le n_j - n_i$), and let β_{II} and β_{III} be the frequencies of regions II ($-n_i < \ell < 0$) and III ($n_j - n_i < \ell < n_j$). Thus, the total rate of the clustered exchange is $\beta_{\mathrm{I}} + \beta_{\mathrm{II}} + \beta_{\mathrm{III}}$. After recombination as in Fig. 1, the resulting chromosomes have $n_i + \ell$ and $n_j - \ell$ units. In order to calculate the changes of the mean ($\bar{n}$) and variance (σ_n^2) of n_is, let us first evaluate the mean and mean square of $n_i + \ell$.

The mean and the mean square of $n_i' = n_i + \ell$ under the condition that one crossing-over occurs while point O lies in region I ($0 \le \ell \le n_j - n_i$) following the uniform probability distribution are

$$\begin{aligned} E_{\mathrm{I}}(n_i'|n_i,n_j) &= \frac{n_i + n_j}{2} \quad \text{and} \\ E_{\mathrm{I}}(n_i'^2|n_i,n_j) &= \frac{1}{3}(n_i^2 + n_j^2 + n_i n_j). \end{aligned} \qquad [9]$$

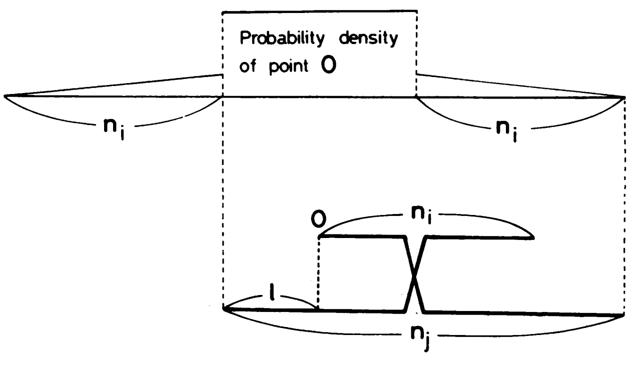

FIG. 1. Diagram showing the model of exchange process of the clustered repeating DNA. Upper diagram illustrates the probability function of the point O; lower figure shows the crossing-over between the DNA segments with n_i and n_j repeating units.

These are conditional expectations, given that one crossing-over occurs while O lies in region I. Next, when one crossing-over occurs while O lies in region II or in region III, we assume that the probability density function of ℓ follows the distribution $(n_i + \ell)/(n_i^2)$ for region II and $(n_j - \ell)/(n_i^2)$ for region III, so that the total probability is unity for the regions II and III. By using these density functions, the mean and the mean square become,

$$E_{\mathrm{II,III}}(n_i'|n_i,n_j) = \frac{n_i + n_j}{2}$$

and [10]

$$E_{\mathrm{II,III}}(n_i'^2|n_i,n_j) = \frac{1}{3}n_i^2 + \frac{1}{2}n_j^2 + \frac{1}{3}n_in_j.$$

These are conditional expectations given that one crossing-over occurs while O lies in region II or III. Let β_{I}' $(= \beta_{\mathrm{II}} + \beta_{\mathrm{III}})$ be the rate of crossing-over for the regions II and III per generation. Taking the expectations of the right hand sides of formulae **9** and **10**, the mean and the mean square of the number of units per genome after one generation of interchromosomal recombination are

$$E_{\mathrm{c}\ell.\mathrm{ex.}}(n_i') = \bar{n}$$

and [11]

$$E_{\mathrm{c}\ell.\mathrm{ex.}}(n_i'^2) = \bar{n}^2 + \sigma_n^2 + \frac{\beta_{\mathrm{I}}'}{6}\bar{n}^2 - \left(\frac{\beta_{\mathrm{I}}}{3} + \frac{\beta_{\mathrm{I}}'}{6}\right)\sigma_n^2,$$

where the subscript cℓ.ex. denotes the clustered exchange process and β_{I} and β_{I}' are assumed to be much less than unity. Therefore, the mean number of the units does not change, but the variance of the number changes on the average by the amount,

$$E_{\mathrm{c}\ell.\mathrm{ex.}}(\Delta\sigma_n^2) = \frac{\beta_{\mathrm{I}}'}{6}\bar{n}^2 - \left(\frac{\beta_{\mathrm{I}}}{3} + \frac{\beta_{\mathrm{I}}'}{6}\right)\sigma_n^2. \qquad [12]$$

It is expected that, in reality $\beta_{\mathrm{I}} >> \beta_{\mathrm{I}}'$—that is, highly skewed chromosomal pairings ($\ell < 0$ or $\ell > n_j - n_i$) occur disproportionately less as compared with more symmetric pairings. From this, we expect that the variance of the unit number decreases by the exchange process.

By sampling of gametes at reproduction, the mean of n_is does not change but its variance decreases. This may be seen by writing the expected value of the variance as

$$\begin{aligned} E\left\{\frac{\sum_i n_i^2}{2N} - \bar{n}^2\right\} &= E\left\{\frac{\sum_i n_i^2}{2N} - \frac{\sum_i n_i^2}{(2N)^2}\right\} \\ &= \frac{1}{(2N)^2}E\left\{(2N-1)\sum_i n_i^2 \right. \\ &\qquad \left. - \sum_i\sum_{j\neq i} n_in_j\right\} \\ &\approx E(n_i^2) - \underset{j\neq i}{\mathrm{E}}(n_in_j). \end{aligned}$$

Through sampling, the fraction, $1/(2N)$ of $E_{j\neq i}(n_in_j)$ becomes equal to $E(n_i^2)$, because the two randomly sampled chromosomes happen to be identical with probability $1/(2N)$. Since the expected value of n_i^2 does not change by sampling, we have

$$E_{\mathrm{smp}}(\Delta\sigma_n^2) = -\frac{\sigma_n^2}{2N}, \qquad [13]$$

where the subscript smp denotes the sampling process. The variance, σ_n^2, is decreased by $1/(2N)$.

Taking all the preceding factors into account, the variance among n_is at equilibrium may be obtained by putting

$$E_{\mathrm{rep}}(\Delta\sigma_n^2) + E_{\mathrm{ds.ex.}}(\Delta\sigma_n^2) + E_{\mathrm{c}\ell.\mathrm{ex.}}(\Delta\sigma_n^2) + E_{\mathrm{smp}}(\Delta\sigma_n^2) = 0,$$

which leads to

$$\hat{\sigma}_n^2 = \frac{6\alpha_1\bar{n}\{1 + N\beta(1 + 2N\alpha_1)^{-1}\} + \bar{n}^2(6\alpha_2a^2 + \beta_{\mathrm{I}}')}{3\beta + 2\beta_{\mathrm{I}} + \beta_{\mathrm{I}}' + 3N^{-1} - 6\alpha_2a^2}, \qquad [14]$$

where the circumflex on σ_n^2 denotes that it refers to the equilibrium value. One has to understand that there is no stable equilibrium in our model and $\bar{n}$ (the mean number in a particular population) actually varies with time. Nevertheless, the relationship between the mean and variance at a given moment may be predicted by the above equation. In particular, it is expected that, when the exchange rates β, β_{I}, and β_{I}' are small, the variance becomes large compared with the mean. Under extreme situations, these quantities are so small that $6\alpha_2a^2 \geq 3\beta + 2\beta_{\mathrm{I}} + \beta_{\mathrm{I}}' + 3N^{-1}$. In such cases, the denominator becomes zero or negative, and the above formula no longer holds. Theoretically, this means that the variance increases indefinitely with time.

In applying Eq. **14** to actual examples, we must choose different parameter values depending on the kind of selfish DNA. For dispersed repeated DNA, α_2, β_{I}, and β_{I}' are likely to be zero. Then, the last term of the numerator vanishes and σ_n^2 may become quite small. On the other hand, for highly repeated sequences, gene exchange may be very limited (14), and β, β_{I}, and β_{I}' may take very small values and therefore σ_n^2 may become large.

For example, let us consider a dispersed repeated DNA with $\bar{n} = 20$. If $\alpha_2 = \beta_{\mathrm{I}} = \beta_{\mathrm{I}}' = 0$ and $\alpha_1 = 10^{-4}$, $\beta = 10^{-3}$, and $N = 10^4$, σ_n^2 becomes 15.8. If α_1 is smaller and is 10^{-5}, with the same value for other parameters, σ_n^2 becomes 3.4. Compared with this, a larger variance would be associated with the clustered repeated sequence. Let $\bar{n} = 10^2$, $\alpha_1 = \beta = 0$, $a = 0.1$, $\alpha_2 = 10^{-3}$, $\beta_{\mathrm{I}} = 10^{-2}$, $\beta_{\mathrm{I}}' = 10^{-3}$, and $N = 10^4$. Then $\hat{\sigma}_n^2$ becomes 500. If α_2 is 10 times larger (i.e., $\alpha_2 = 10^{-2}$) with the same values for the other parameters, σ_n^2 becomes 773. Also, negative natural selection may be responsible for eliminating the genomes having too large an amount of selfish DNA, and its effect awaits future investigation.

The present result may be compared with that of Crow and Kimura (ref. 15, pp. 295–296) where they treated the distribution of the number of repeated units under equilibrium between *stabilizing selection* and the generation of new units by unequal crossing-over. In their model, it is assumed that the increase or decrease of the number of replicating units per genome per generation is independent of n_i and follows a normal distribution and that selection is of the centripetal type such that the fitness of a genome with n_i units is $1 - s(n_i - n_{\mathrm{op}})^2$, where n_{op} is the optimum number of units per genome and s is the selection coefficient which is assumed to be positive. Then, at equilibrium, the mean and variance of n_i become (15),

$$\bar{n} = n_{\mathrm{op}}$$

and [15]

$$\sigma_n^2 = \sqrt{\frac{ux^2}{2s}},$$

where u is the rate at which a chromosome with a certain number of repeats is converted into something else and x^2 is the variance of the change in the number of repeating units due to duplication or deletion in one generation. It is assumed that the mean change of the number of units is zero. In our model, the variance is a function of n_i and ux^2 corresponds to $\alpha_1\bar{n} + \alpha_2 a^2(\bar{n}^2 + {\sigma_n}^2)$ (Eq. 5). The model used by Crow and Kimura is presumably better suited than the present model for treating multigene families having some important function such as ribosomal RNA or histone gene families. Selfish DNA is more likely to be selectively neutral or nearly neutral unless the amount becomes too large (2), and the present model may be more suitable to treat it because, in this model, random genetic drift, intergenome exchange, and the independently replicating property of each unit are taken into account.

If n_i is assumed to be normally distributed in the population, it is possible to combine the two models. This assumption is valid for the steady-state distribution investigated by Crow and Kimura but may not hold in our case because the variance of Δn_i depends upon n_i. In the following, we tentatively assume that n_i is normally distributed at equilibrium and we consider the mean and variance of the distribution. Then, it can be shown that $\bar{n} = n_{\mathrm{op}}$ and the variance ${\sigma_n}^2$ is reduced through selection by the amount $2s{\sigma_n}^4$. Therefore, at equilibrium, we have

$$E_{\mathrm{rep}}(\Delta{\sigma_n}^2) + E_{\mathrm{ds.ex.}}(\Delta{\sigma_n}^2) + E_{\mathrm{c\ell.ex.}}(\Delta{\sigma_n}^2) + E_{\mathrm{smp}}(\Delta{\sigma_n}^2) - 2s{\sigma_n}^4 = 0$$

Solving the above equation, we get, as the equilibrium variance,

$${\sigma_n}^2 = \frac{1}{4s}\{\sqrt{B^2 + 8sA} - B\}, \qquad [16]$$

where $A = \alpha_1\bar{n}\{1 + [N\beta/(1 + 2N\alpha_1)]\} + \bar{n}^2(\alpha_2 a^2 + [\beta_{\mathrm{I}}'/6])$ and $B = (\beta/2) + (\beta_{\mathrm{I}}/3) + (\beta_{\mathrm{I}}'/6) + (1/2N) - \alpha_2 a^2$. When $B = 0$ $[\beta = \beta_{\mathrm{I}} = \beta_{\mathrm{I}}' = 1/(2N) = \alpha_2 = 0]$, formula **16** reduces to $\sqrt{\alpha_1\bar{n}/(2s)}$. In other words, when random drift, exchange process, and cluster replication process are negligible, the result agrees with Eq. **15** because $\alpha_1\bar{n}$ corresponds to ux^2 of Crow and Kimura (15) under such an assumption. Eq. **16** may be applied to treat observed polymorphisms of the repeat length of some gene families. An interesting example is the length polymorphism of the silk fibroin gene which possesses an internally repetitive structure (16). It is likely that unequal crossing-over, random drift, and natural selection are responsible for the length heterogeneity.

DISCUSSION

In applying the present results to real observations, one may choose parameter values depending upon the kind of selfish DNA. As already mentioned, the parameters are likely to be quite different for the clustered and the dispersed repetitive DNA. For the former, the "cluster" type replication and exchange processes would prevail ($\alpha_2 \gg \alpha_1$ and $\beta_{\mathrm{I}} + \beta_{\mathrm{I}}' \gg \beta$), whereas for the dispersed repeating DNA, the single replication and the "dispersed" type exchange would predominate ($\alpha_1 \gg \alpha_2$ and $\beta \gg \beta_{\mathrm{I}} + \beta_{\mathrm{I}}'$).

Our theory is limited to the cases in which the selfish DNA segment has no tendency for systematic increase or decrease. In other words, we assumed that the mean change per generation of the number of replicating units is zero, and therefore the course of change in the amount of selfish DNA is left to chance. There may exist truly selfish DNA in the sense that it tends to increase deterministically, such as the B chromosomes observed in some plant species. A notable example is a supernumerary or B chromosome called f_ℓ in the lily *Lilium callosum* (17). This is a large telocentric chromosome which appears to show no phenotypic effect except that when more than one copy exists in an individual, pollen and seed fertility of the plant are reduced. However, it has a tendency to increase in number due to preferential segregation in embryosac mother cells in a plant with one f_ℓ chromosome, so that the f_ℓ chromosome is included in the egg cell in about 80% of the cases. No such preferential segregation occurs in pollen formation. In natural populations of this lily, this chromosome is contained in nearly 70% of individuals due to the balance between preferential segregation and negative natural selection. An analogous situation occurs with the SD factor in *Drosophila melanogaster* (18). Different formulation is needed in order to understand evolution of this kind.

A problem related to the amount of selfish DNA is the so called "C value paradox"—i.e., the problem that too much DNA exists in a cell as compared with the estimated gene number in higher organisms (19). At the moment, we do not know which of the following concepts is applicable to this paradox: (*i*) polyploidization and subsequent degeneration (20), (*ii*) truly selfish DNA such as B chromosomes, (*iii*) highly and middle repetitive sequences with no phenotypic effects as considered in this paper, and (*iv*) pseudogenes such as found in 5S ribosomal DNA (21) and in hemoglobin α loci (22, 23). Quantitative analyses based on population genetics theory would be required for a correct understanding of their relative importance. Also, in the future, the effect of natural selection in eliminating genomes with too large an amount of selfish DNA has to be investigated.

We thank Dr. K. Aoki for carefully going over the first draft of this paper and for making many useful suggestions to improve the presentation.

1. Doolittle, W. F. & Sapienza, C. (1980) *Nature (London)* **284**, 601–603.
2. Orgel, L. E. & Crick, F. H. C. (1980) *Nature (London)* **284**, 604–607.
3. Hood, L., Campbell, J. H. & Elgin, S. C. R. (1975) *Annu. Rev. Genet.* **9**, 305–353.
4. Ohta, T. (1978) *Genet. Res.* **31**, 13–28.
5. Ohta, T. (1980) *Evolution and Variation of Multigene Families*, Lecture Notes in Biomathematics (Springer, New York), Vol. 37.
6. Hood, J. M., Huang, H. V. & Hood, L. (1980) *J. Mol. Evol.* **15**, 181–196.
7. Smith, G. P. (1974) *Cold Spring Harbor Symp. Quant. Biol.* **38**, 507–513.
8. Calos, M. P. & Miller, J. H. (1980) *Cell* **20**, 579–595.
9. Young, M. W. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 6274–6278.
10. Cameron, J. R., Loh, E. Y. & Davis, R. W. (1979) *Cell* **16**, 739–751.
11. Finnegan, D. J., Rubin, G. M., Young, M. W. & Hogness, D. S. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1053–1063.
12. Ilyin, Y. V., Tchurikov, N. A., Ananiev, E. V., Ryskov, A. P., Yenikolopov, G. N., Limborska, S. A., Maleeva, N. E., Gvozdev, V. A. & Georgiev, G. P. (1978) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 959–969.
13. Kimura, M. & Crow, J. F. (1964) *Genetics* **49**, 725–738.
14. Yamamoto, M. & Miklos, G. L. G. (1978) *Chromosoma* **66**, 71–98.
15. Crow, J. F. & Kimura, M. (1970) *An Introduction to Population Genetics Theory* (Harper & Row, New York).
16. Sprague, K. U., Roth, M. B., Manning, R. F. & Gage, L. P. (1979) *Cell* **17**, 407–413.
17. Kimura, M. & Kayano, H. (1961) *Genetics* **46**, 1699–1712.
18. Crow, J. F. (1979) *Sci. Am.* **240** (2), 134–146.
19. Muller, H. J. (1967) in *Heritage from Mendel*, ed. Bring, R. A. (Univ. Wisconsin Press, Madison, WI), pp. 419–447.
20. Ohno, S. (1970) *Evolution by Gene Duplication* (Springer, Berlin).
21. Fedoroff, N. V. (1979) *Cell* **16**, 697–710.
22. Nishioka, Y., Leder, A. & Leder, P. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2806–2809.
23. Vanin, E. F., Goldberg, G. I., Tucker, P. W. & Smithies, O. (1980) *Nature (London)* **286**, 222–226.