Genet. Res., Camb. (1983), **42**, *pp.* 1–27
With 4 text-figures
Printed in Great Britain

The population dynamics of transposable elements

BY BRIAN CHARLESWORTH AND DEBORAH CHARLESWORTH
School of Biological Sciences, University of Sussex, Brighton BN1 9QG

(*Received 19 August 1982 and in revised form 14 February 1983*)

SUMMARY

This paper describes analytical and simulation models of the population dynamics of transposable elements in randomly mating populations. The models assume a finite number of chromosomal sites that are occupable by members of a given family of elements. Element frequencies can change as a result of replicative transposition, loss of elements from occupied sites, selection on copy number per individual, and genetic drift. It is shown that, in an infinite population, an equilibrium can be set up such that not all sites in all individuals are occupied, allowing variation between individuals in both copy number and identity of occupied sites, as has been observed for several element families in *Drosophila melanogaster*. Such an equilibrium requires either regulation of transposition rate in response to copy number per genome, a sufficiently strongly downwardly curved dependence of individual fitness on copy number, or both. The probability distributions of element frequencies, generated by the effects of finite population size, are derived on the assumption of independence between different loci, and compared with simulation results. Despite some discrepancies due to violation of the independence assumption, the general pattern seen in the simulations agrees quite well with theory.

Data from *Drosophila* population studies are compared with the theoretical models, and methods of estimating the relevant parameters are discussed.

1. INTRODUCTION

It is now well established that the genomes of eukaryotes consist to a significant extent of families of dispersed, repeated sequences of DNA (middle repetitive DNA). The structure of members of these families is similar to that of the well-characterized transposable genetic elements of prokaryotes (Kleckner, 1981), and it seems increasingly clear that they are capable of replication and transposition to new sites within the genome, just like the prokaryote transposons and insertion sequencies (Doolittle, 1982; Finnegan, Will, Bayev, Bowcock & Brown, 1982). Although in most cases there is only indirect evidence for mobility of middle repetitive DNA, there is direct genetic evidence for transposable elements in both *Drosophila* (Green, 1980) and maize (McClintock, 1956). Furthermore, it has been shown that the phenomenon of hybrid dysgenesis in *D. melanogaster* (Kidwell, Kidwell & Sved, 1977; Engels, 1981) is due to the mobilization and integration

at new chromosomal sites of a transposable element, the P factor (Rubin, Kidwell & Bingham, 1982; Bingham, Kidwell & Rubin, 1982).

The likelihood that middle repetitive DNA consists of families of transposable elements generated the 'selfish DNA' hypothesis of Doolittle & Sapienza (1980) and Orgel & Crick (1980), which asserts that the maintenance and distribution of such DNA in natural populations can be understood in terms of the selective advantage to the elements themselves of their power of multiplication within the genome, and not to any advantages to the individuals who carry them. The known role of transposable elements in bacteria, yeast and *Drosophila* in inducing mutations as a result of insertion into new sites suggests that they are likely to have an adverse effect on individual fitness, if any. A stable distribution of number of copies per individual of a given class of element could be set up in a sexual population as a result of a balance between increase in copy number by replicative transposition, and elimination by selection (Brookfield, 1982, 1983; Charlesworth, 1983). Another possibility for maintaining stable copy numbers is that replication is regulated, so that the transposition rate per element declines with the number of elements of the same family present in the same genome. There is evidence for such regulation in prokaryotes (Kleckner, 1981; Kitts, Lamond & Sherratt, 1982; Reed, Shibuya & Steitz, 1982), and the properties of the P factor system in *D. melanogaster* suggests that this can occur in eukaryotes as well.

These considerations indicate that transposable elements may be involved in a novel class of population process. Despite the extensive verbal discussions of selfish DNA (reviewed by Doolittle, 1982, and Dover, 1982), there have been few attempts to provide well-worked out population genetics models of their evolutionary dynamics. In view of the possible role of population studies in testing the hypotheses concerning the significance of transposable elements, it seems important to develop such models. Various types of model are described by Ohta (1981, 1983), Ohta & Kimura (1981), Brookfield (1982, 1983), Hickey (1982), Barrett (1982), Langley, Brookfield & Kaplan (1983) and Kaplan & Brookfield (1983*b*).

The purpose of this paper is to present some models of the population genetics of a transposable element in diploid, sexually reproducing populations mating at random, in the hope of generating predictions that could be tested by comparison with the distributional properties of such elements in natural populations. Our models assume that increase in the number of copies of the element occurs by means of transposition of replicates of pre-existing elements to new genomic sites. We also allow for the possibility that an element may be lost from a site, independently of transposition. There is good evidence for these processes in prokaryotes (Kleckner, 1981), but the situation in eukaryotes is less clear. We consider both regulated transposition and selection against individuals carrying elements, by means of analytical models and computer simulations.

2. THE COMPUTER MODEL

Two independently segregating chromosome pairs were assumed, each with up to 31 sites at which the transposable element could integrate; most of our simulations assumed 31 sites per chromosome. Recombination between loci on the same chromosome was modelled by the method described below. Each of the four

chromosomes in a genome was stored as one computer word, using zeros to denote non-occupied sites, and 1's for loci at which the element had integrated.

The initial population for each run was set up by specifying an expected number of copies of the element per individual (usually 10). The number of individuals in the population (N) was also specified. Each of the two gametes constituting an individual was set up locus by locus, using the appropriate probability of occupation of a site derived from the total number of loci and the specified expected number of elements per individual, and choosing a random number to decide whether or not a given locus was occupied. This initial population formed the basis for a single run, with the following sequence of events, which assumes that the starting population consists of adult individuals about to reproduce.

Two parent individuals were taken at random from the initial set. Each parent in turn was used to generate a gamete containing two non-homologous chromosomes. This step involves modelling the recombination process within each chromosome. The number of crossover events for a chromosome was determined by a random number, using a Poisson distribution whose mean was the total map distance between the two extreme loci. This distance was usually 90 map units, corresponding to a recombination frequency of 0·03 between adjacent loci. Their location was determined by random numbers, assuming a uniform distribution of crossovers along the chromosome. The reciprocal recombinant chromosomes were then generated from the appropriate pair of homologous chromosomes of the parent in question, using the usual masking procedure (Franklin & Lewontin, 1970), and one of them was chosen at random for inclusion in the gamete. When both parents had generated gametes by this process, the gametes were combined to form a zygote.

In runs where selection was being modelled, the fitness of the zygote was assumed to be a function of the number of elements in its genome (see Section 4(i) for details). The number of elements was therefore counted for the zygote in question, and a random number was generated in order to decide whether or not to accept it as viable. The whole procedure of generating new zygotes was repeated from the beginning until N viable zygotes had been produced.

Finally, the population was subjected to processes of transposition and loss of elements. The probability of transposition of a given element was assumed to be a decreasing function of n, the number of elements in the genome in which the element is situated (see Section 3(i) for details). The probability of loss was treated as constant. The probability of transposition or loss per element was multiplied by n for the individual in question, and the result gave the mean of a Poisson distribution for the total number of transposition or loss events. In order to save computer time, low probabilities of loss or transposition were assumed, so that the Poisson distributions could be approximated by their first two terms (i.e. at most one of each type of event occurred in a given individual in a given generation).

Loss of an element from a locus of an individual was modelled by first choosing a random number to decide whether an element was to be lost from that individual, or whether all elements were to be retained. If an element was to be lost, the site of loss was chosen at random from a list of all occupied loci for the given individual, assuming an equal probability of loss for each occupied locus. The appropriate bit of the computer words specifying the genome in question was then altered.

The occurrence of a transposition event in an individual was determined in the same way as for loss events. The process of transposition was assumed to involve the appearance of the element at a site that was previously unoccupied, without any simultaneous loss of the element from occupied sites. In order to decide at which site a new element would appear, we chose a locus at random from the entire genome; if unoccupied, it was altered to the occupied state, but if it proved to be occupied, another site was chosen at random, and so on until an unoccupied site was found.

When all N zygotes had undergone transposition and loss, the original individuals present at the start of the generation were discarded and replaced by their progeny. The entire generation cycle of events was repeated for a number of generations (usually 1000). Replicate runs of the same parameter set were performed using the final random number of the preceding run for determining the state of the first locus of the first individual in the initial population.

3. REGULATED TRANSPOSITION

In this section, we shall develop some analytical models for the case with no selection, but when the probability of transposition of an element in a diploid genome containing n elements is a decreasing function u_n of n. The probability of loss of an element is assumed to be a constant, v. (Parallel results can be obtained for the case when v is an increasing function of n, but we shall not pursue this further.) The analytical models are then compared with simulation results. Throughout the rest of the paper, we assume that there is a finite number of chromosomal sites that may be occupied by an element, which we write as T for a diploid individual. The number of occupable chromosomal loci is thus $T/2$, and these will be treated in the same way as conventional gene loci in the population genetics models.

(i) *Infinite population size*

Let the frequency of the element at the ith locus be x_i, and let $\bar{n}$ be the mean number of elements per individual in the population. We have

$$\bar{n} = 2\sum_i x_i, \tag{1}$$

where the summation is taken over all occupable loci. The change per generation is $\bar{n}$ is given by

$$\Delta\bar{n} = E\{nu_n\} - \bar{n}v, \tag{2}$$

where the expectation is taken over all individuals in the population. Expanding this around $\bar{n}$, we can write

$$\Delta\bar{n} \approx \bar{n}(u_{\bar{n}} - v) + \frac{V_n}{2}\left[2\frac{\partial u_{\bar{n}}}{\partial\bar{n}} + \bar{n}\frac{\partial^2 u_{\bar{n}}}{\partial\bar{n}^2}\right], \tag{3}$$

where V_n is the variance in copy number between individuals.

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

We have (Bulmer, 1980, p. 158)

$$V_n = 2\sum_i x_i(1-x_i)+4\sum_{i<j} D_{ij}$$
$$= \bar{n}\left(1-\frac{\bar{n}}{T}\right)-T\sigma_x^2+4\sum_{i<j} D_{ij}, \tag{4}$$

where σ_x^2 is the variance in x_i between loci, and D_{ij} is the coefficient of linkage disequilibrium between loci i and j. If linkage disequilibrium effects are small, as seems likely unless linkage is very tight, the terms in D_{ij} can be ignored. (The validity of this is examined in Section 3(iv) below.) Furthermore, in an infinite population, the processes of transposition and loss will soon equalize the frequencies of the element at all loci, so that $\sigma_x^2 \to 0$ (see Appendix 1). Copy number per individual will then follow a binomial distribution with mean $\bar{n}$ and variance $\bar{n}(1-\bar{n}/T)$, and equation (3) becomes

$$\Delta\bar{n} \approx \bar{n}(u_{\bar{n}}-v)+\frac{\bar{n}}{2}\left(1-\frac{\bar{n}}{T}\right)\left[2\frac{\partial u_{\bar{n}}}{\partial\bar{n}}+\bar{n}\frac{\partial^2 u_{\bar{n}}}{\partial\bar{n}^2}\right]. \tag{5}$$

Provided that transposition is not too strongly regulated, the second term on the right-hand side of equation (5) may be neglected in most cases. $\bar{n}$ will then tend to an equilibrium value $\hat{n}$ given by the approximate expression

$$u_{\hat{n}} = v. \tag{6a}$$

In the simulation work, we used the function $u_n = u_0/(1+kn)$. The validity of the above approximations then depends on k being small ($\ll 1$), and the solution of equation (6a) is

$$\hat{n} = (u_0-v)/kv. \tag{6b}$$

(ii) *Finite population size: an analytical model*

Consider a population of N breeding individuals, of effective size N_e. In such a population, the assumption of equal frequencies of the element at all loci will no longer hold, nor will the mean number of copies per individual stabilize at the level corresponding to the infinite population case. If linkage disequilibrium effects are ignored, however, the change in frequency per generation of the element can be determined for an individual locus i. For simplicity, the subscript i will be dropped from the frequency of the element at locus i, so that x will be used instead of x_i.

If the mean number of copies per individual in the present generation is $\bar{n}$, the mean number of transposition events is, by the above results, approximately $\bar{n}u_{\bar{n}}$. New elements can insert themselves only at unoccupied sites, to which locus i contributes a fraction $2(1-x)/(T-\bar{n})$. The mean increase in copy number at this locus is thus approximately $2\bar{n}u_{\bar{n}}(1-x)/(T-\bar{n})$, and the change in x due to transposition and loss is given by

$$\Delta x \approx \mu_{\bar{n}}(1-x)-vx, \tag{7}$$

where

$$\mu_{\bar{n}} = \bar{n}u_{\bar{n}}/(T-\bar{n}).$$

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

As a result of sampling, there will be a probability distribution of x, whose density function at time t may be written as $\phi(x,t)$. This will obey the usual diffusion approximation, provided that $\mu_{\bar{n}}$, v and $1/N_e$ are sufficiently small. Hence

$$\frac{\partial\phi}{\partial t} = \frac{1}{2}\frac{\partial^2(\phi V_{\delta x})}{\partial x^2} - \frac{\partial(\phi M_{\delta x})}{\partial x}, \tag{8}$$

where $V_{\delta x} = x(1-x)/2N_e$ and $M_{\delta x}$ are the variance and expectation of the change in x per generation (e.g. Crow & Kimura, 1970, p. 372). From equation (7) the latter is given by

$$M_{\delta x} \approx E\{\mu_{\bar{n}}\}(1-x) - vx, \tag{9}$$

where the expectation is taken over the probability distribution of $\bar{n}$. This assumes independence between the distributions for different loci.

In order to utilize equations (8) and (9), it is obviously necessary to evaluate $E\{\mu_{\bar{n}}\}$. This can be done if we assume that the probability distributions for all loci tend to a steady-state, such that the expectation of $\bar{n}$ approximates the infinite-population equilibrium value $\hat{n}$ given by equations (5) and (6). Formally, this is impossible, since the state of loss of the element from all loci of each member of the population is the absorbing boundary of the process. But in practice the chance of loss in a given generation of all copies from a reasonably large population with even a moderate value of $\bar{n}$ is a very low (approximately $\exp -\bar{n}N_e$). With $\bar{n} = 20$ and $N_e = 50$, for example, this probability is about 10^{-434}. Provided a state with a sufficiently high value of $E\{\bar{n}\}$ is approached, the rate of absorption into the state of loss of all copies will be negligible in practice. Computer simulations show that there is indeed fairly rapid convergence to an approximate steady-state in which $E\{\bar{n}\} \approx \hat{n}$. (See Section 3(iv) for details.)

For sufficiently large t, we can therefore write

$$E\{\mu_{\bar{n}}\} \approx \mu_{\hat{n}} + \tfrac{1}{2}\hat{\sigma}^2_{\bar{n}}\frac{\partial^2\mu_{\hat{n}}}{\partial\hat{n}^2} \approx \mu_{\hat{n}} + \frac{T\hat{\sigma}^2_x}{2(T-\hat{n})}\left\{\hat{n}\frac{\partial^2 u_{\hat{n}}}{\partial\hat{n}^2} - \frac{2T}{(T-\hat{n})}\frac{\partial u_{\hat{n}}}{\partial\hat{n}} + \frac{T^2}{(T-\hat{n})^2}u_{\hat{n}}\right\}, \tag{10}$$

where $\hat{\sigma}^2_{\bar{n}}$ and $\hat{\sigma}^2_x$ are the steady-state variances in mean copy number and element frequency respectively. Provided that $\hat{\sigma}^2_x$ is small and $\hat{n}$ is small compared with T, only the $\mu_{\hat{n}}$ term need be employed.

It is therefore legitimate to solve equation (8) for the steady-state probability distribution $\phi(x)$, satisfying $\partial\phi/\partial t = 0$, by substituting $\mu_{\hat{n}}$ for $E\{\mu_{\bar{n}}\}$ is equation (9), which thereby becomes identical in form with the usual equation for a single locus with reversible mutation between two alleles (Crow & Kimura, 1970, p. 442). $\phi(x)$ therefore follows a beta distribution

$$\phi(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \tag{11}$$

where $\alpha = 4N_e\mu_{\hat{n}}$ and $\beta = 4N_e v$. The mean and variance of this distribution are

$$\hat{x} = \alpha/(\alpha+\beta) = \hat{n}/T \tag{12a}$$

and

$$\hat{\sigma}^2_x = \hat{x}(1-\hat{x})/(1+\alpha+\beta). \tag{12b}$$

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

(iii) *Methods of comparison with simulation results and data*

If the assumption of independence of the distributions between different loci is correct, then $\phi(x)$ can be interpreted either as the distribution of frequencies among different loci in the same genome, or as the distribution at a single locus over separate populations. Data or simulation results on the frequencies of the element at different sites can then be pooled over loci and populations to provide a distribution of frequencies to compare with $\phi(x)$. The probability that a locus has a frequency in the range x_1 to x_2 $(0 < x_1 < x_2 < 1)$ is given by integrating over this range. The proportions of loci at which the element is lost or fixed are given by the formulae (Crow & Kimura, 1970, p. 441)

$$\text{loss:} \qquad P_0 = \int_0^{1/2N} \phi(x)\,dx \tag{13}$$

$$\text{fixation:} \quad P_1 = \int_{1-(1/2N)}^{1} \phi(x)\,dx. \tag{14}$$

(These formulae are not necessarily accurate for small population sizes, and more adequate approximations to P_0 and P_1 can be obtained [Ewens, 1979, p. 158].)

Since we do not know the total number of occupable sites in the genome for any family of transposable elements in eukaryotes, data from populations must be restricted to information on frequencies of the element at loci where it has been identified as occurring at least once in the sample. The observed distribution of frequencies can therefore be compared with the distribution conditioned on non-loss, i.e.

$$\phi^*(x) = \phi(x)/(1-P_0). \tag{15}$$

An alternative method has been suggested by Langley, Brookfield & Kaplan (1983) for the case when T is very large, using an analogy with the infinite-alleles model of standard neutral mutation theory (Crow & Kimura, 1970, p. 455). Their method can be placed in the present framework as follows. The expected number of loci in a genome with the element in the frequency range x to $x+dx$ can be written

$$\Phi(x)\,dx = \tfrac{1}{2}T\phi(x)\,dx. \tag{16}$$

As $T \to \infty$, $\alpha \to 0$ in equation (11), and $\Gamma(\alpha) \to \alpha^{-1}$. For a large number of occupable sites, therefore, we have (using the fact that $T\alpha \to 4N_e u_{\hat{n}}\hat{n} = 4N_e v\hat{n} = \beta\hat{n}$)

$$\Phi(x) = \tfrac{1}{2}\hat{n}\beta x^{-1}(1-x)^{\beta-1}. \tag{17}$$

This is similar in form to equation (4) of Langley *et al.* (1983).

(iv) *Simulation results*

The adequacy of the approximate theory developed above was checked by the simulation procedure of Section 2. All the runs described here were done with a system of two independent chromosomes each with 31 loci ($T = 124$). In most cases, the recombination fraction between adjacent loci was 0·03. This provides an approximation to what would be expected for sites distributed uniformly along

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

the major autosomes of *Drosophila melanogaster*, although the total number of sites per chromosome is probably unrealistically small. Each run was started by generating independent individuals with an expected number of ten copies of the element, using the procedures described in Section 2. The random sampling method used to form each new generation ensures that $N_e = N$. Each run was terminated after 1000 generations, or after loss of all copies of the element from the population, if this happened first.

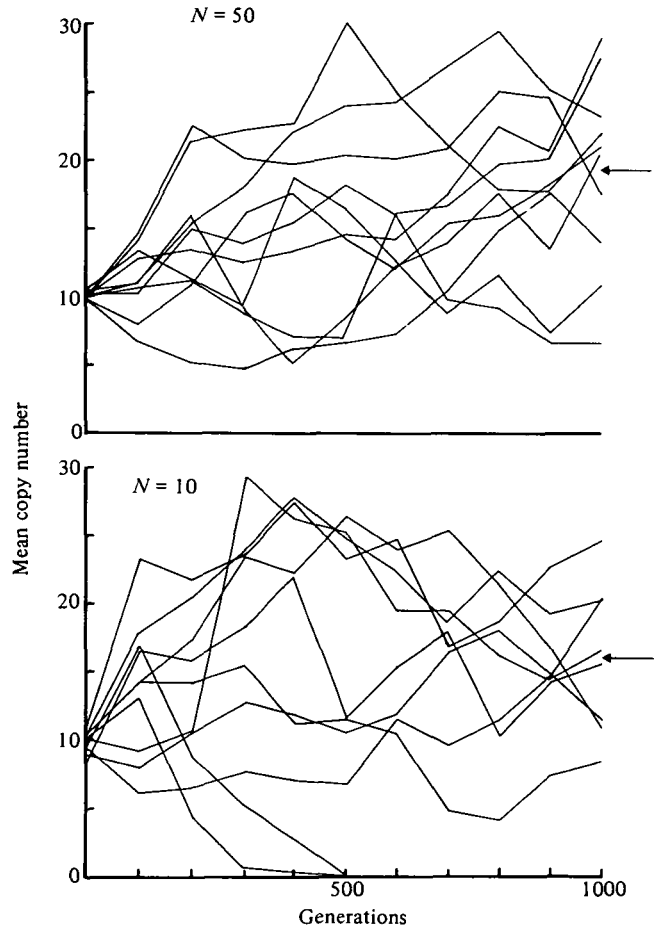

Fig. 1. Changes in time of the mean copy number ($\bar{n}$) of ten simulated populations for parameter sets 1 and 2 of Table 1 ($u_0 = 0{\cdot}01$, $v = 0{\cdot}005$, $k = 0{\cdot}05$, and $N = 10$ and 50 respectively). The arrows indicate the means of $\bar{n}$ for those populations that have not lost all copies of the element. (The expected value is 20 in both cases.)

As shown in Fig. 1, even populations of moderate size tend to move fairly quickly towards a state in which the expected copy number per individual approaches that predicted by equations (6), with only the occasional case of loss of all copies from the population. Loss will obviously be more frequent, the lower the initial mean copy number, but the assumption of convergence to a steady-state value of expected copy number close to $\hat{n}$ seems to be justified for those populations that survive early loss (see also Table 1).

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

The results of simulations done with four different parameters set are shown in Table 1, where cases of loss have been discarded. Fig. 2 displays histograms of the conditional distributions of element frequency for these parameter sets. The distributions of element frequencies were compiled by pooling the distribution of frequencies across loci at generation 1000 of each run. The estimates of mean copy

Table 1. *Parameter sets used in the computer simulations with regulated transposition, and a summary of the results*

Parameter set	1	2	3	4
k	0·05	0·05	0·05	0·0838
N	10 ($\alpha = 0{\cdot}038$, $\beta = 0{\cdot}2$)	50 ($\alpha = 0{\cdot}192$, $\beta = 1$)	100 ($\alpha = 0{\cdot}385$, $\beta = 2$)	50 ($\alpha = 0{\cdot}106$, $\beta = 1$)
Number of runs	17	23	13	20
Mean copy number per individual				
Theoretical	20	20	20	11·93
Simulated	18·59**	19·59**	19·01**	11·16**
Variance in element frequency between loci (σ_x^2)				
Theoretical	0·1092	0·0617	0·0400	0·0413
Simulated	0·1041	0·0655	0·0438	0·0421
Fraction of loci with zero frequency (P_0)				
Theoretical	0·7596	0·4125	0·1802	0·6124
Simulated	0·7666	0·4390*	0·2246**	0·6548**
Variance in copy number between individuals within populations (V_n)				
Theoretical	2·789	8·309	10·674	5·522
Simulated	2·192	9·211*	10·521	5·042

$u_0 = 0{\cdot}01$ and $v = 0{\cdot}005$ in each case, and $u_n = u_0/(1+kn)$.
* and ** indicate deviations from the theoretical values at $p < 0{\cdot}05$ and $p < 0{\cdot}01$ respectively.

number per individual, σ_x^2, and P_0 were obtained directly from these distributions, and compared with the formulae derived above. The existence of any systematic linkage disequilibrium effects (i.e. effects consistent across different runs) was tested by comparing the mean over runs of V_n (the variance between individuals in copy number) at generation 1000 with the expected value given by equation (4) on the hypothesis of no linkage disequilibrium, $\bar{n}(1-\bar{n}/T)-\sigma_x^2\,(T+1)$, using the simulated values of n and σ_x^2.

Table 1 shows that simulated and theoretical values are close for mean copy number per individual and σ_x^2, although in every case the expected values of $\bar{n}$ are significantly higher than the observed. Similarly, the theoretical values of P_0 tend to be somewhat higher than the simulated. This may partly reflect an inadequacy in the approximations involved in equation (13), but more probably is due to a movement of the distribution towards the absorbing boundary (see Section 5(iv)). The simulated conditional distributions in Fig. 2 show good agreement with theory, with no significant discrepancies except possibly for the case of $N = 10$, where the diffusion approximation is expected to be unreliable anyway.

The comparatively good agreement between simulated and theoretical values of these measures suggest that the assumption of no linkage disequilibrium and independent distributions for different loci is valid. This is largely confirmed by the comparison of simulated and theroetical values of V_n except for the case of $N = 10$. Linkage effects were also tested for by simulations of parameter set 2 with

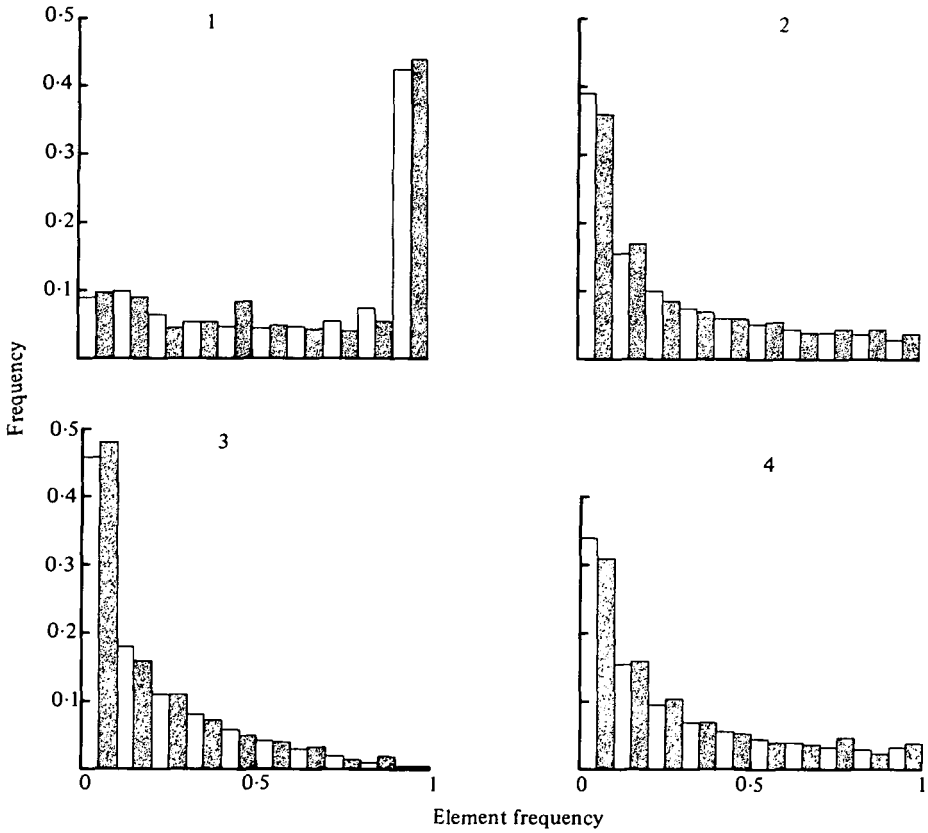

Fig. 2. Histograms of the conditional distributions of element frequencies for the parameter sets of Table 1 (identified by the numbers above each histogram). Ten per cent intervals of element frequency were used. The clear columns are the theoretical frequencies and the shaded columns are the frequencies obtained in the simulations.

a recombination fraction of 0·003 between adjacent loci. Ten runs were carried out; the simulated mean copy number was 14·99 instead of the expected 20, σ_x^2 was 0·0548 compared with an expected 0·0617, and P_0 was 0·5710 compared with the expected 0·4125. The differences in mean and P_0 are highly significant; the difference in variance has $p < 0·05$. The conditional distribution of copy number does not differ significantly from the expected, however, with χ^2_{19} of 25·29 ($p < 0·10$). This indicates that the main effect of close linkage is to increase the probability of loss of copies. There is no evidence that close linkage causes any systematic linkage disequilibrium (simulated and theoretical values of V_n were 6·350 and 6·328 respectively). Evidence for some lack of independence of the distributions at different loci is presented in Section 5(iv), however.

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

4. TRANSPOSITION AND SELECTION

In this Section we analyse models in which selection against individuals carrying the transposable elements opposes its spread. We assume that the probabilities of both transposition and loss are constants (u and v), independent of copy number. The fitness of an individual carrying n copies is assumed to be a decreasing function of n, w_n. Otherwise, the notation and assumptions are the same as in Section 3. We first consider the case of an infinitely large population, and then examine the properties of the distribution of element frequencies in a finite population.

(i) *Infinite population size*

Combining equation (7) and Section 3(ii) with standard selection theory, and assuming no linkage disequilibrium, we obtain the following equation for the change in element frequency at a given locus i.

$$\Delta x_i = \frac{x_i(1-x_i)}{2\bar{w}}\frac{\partial \bar{w}}{\partial x_i} + \frac{u\bar{n}(1-x_i)}{(T-\bar{n})} - vx_i, \tag{18a}$$

where $\bar{w}$ is the mean fitness of the population, $E(w_n)$. Approximating $\bar{w}$ by $w_{\bar{n}}$ and noting that $\partial\bar{n}/\partial x_i = 2$, we obtain

$$\Delta x_i = x_i(1-x_i)\frac{\partial \ln w_{\bar{n}}}{\partial \bar{n}} + \frac{u\bar{n}(1-x_i)}{(T-\bar{n})} - vx_i. \tag{18b}$$

If element frequencies have the same value, x, at all loci, as would be expected to be true eventually in an infinite population (see Appendix 1 for the proof of this), then $\bar{n} = Tx$ and equation (18b) becomes

$$\Delta x = x(1-x)\frac{\partial \ln w_{\bar{n}}}{\delta \bar{n}} + x(u-v), \tag{19a}$$

and

$$\Delta \bar{n} = \bar{n}\left(1-\frac{\bar{n}}{T}\right)\frac{\partial \ln w_{\bar{n}}}{\partial \bar{n}} + \bar{n}(u-v). \tag{19b}$$

The direction of change in $\bar{n}$ is given by the sign of

$$u-v-f(\bar{n}) \tag{20a}$$

where

$$f(\bar{n}) = \left(1-\frac{\bar{n}}{T}\right)\left|\frac{\partial \ln w_{\bar{n}}}{\partial \bar{n}}\right| \tag{20b}$$

$$\approx \left|\frac{\partial \ln w_{\bar{n}}}{\partial \bar{n}}\right| \quad (T \text{ large}). \tag{20c}$$

An equilibrium $\hat{n}$ in mean copy number therefore exists when $f(\hat{n}) = 0$, subject to the constraint $0 < \hat{n} < T$. For mean copy number to increase from zero, we require $f(0) < u-v$. It follows that $\partial^2 \ln w_n/\partial n^2 < 0$ is necessary for there to be a biologically meaningful value of $\hat{n}$, in addition to the condition on $f(0)$. Such an

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

equilibrium is locally stable by the usual criteria when

$$-1 < \hat{n}\left\{\left(1-\frac{\hat{n}}{T}\right)\frac{\partial^2 \ln w_{\hat{n}}}{\partial \hat{n}^2}-\frac{1}{T}\frac{\partial \ln w_{\hat{n}}}{\partial \hat{n}}\right\} < 0. \tag{21a}$$

For large T, this reduces to

$$\frac{-1}{\hat{n}} < \frac{\partial^2 \ln w_{\hat{n}}}{\partial \hat{n}^2} < 0. \tag{21b}$$

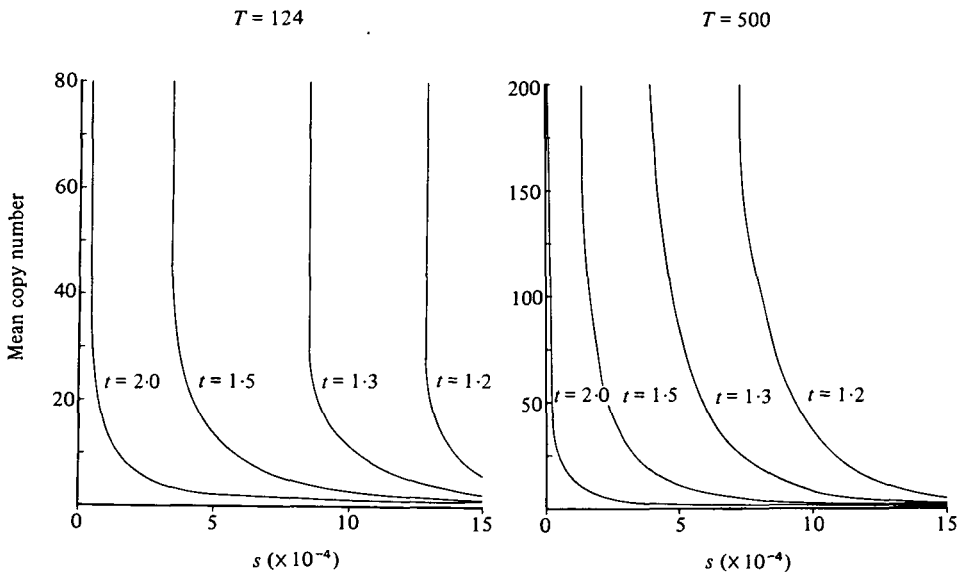

Fig. 3. Equilibrium values of mean copy number ($\hat{n}$) for an infinite population, with the fitness function $w_n = 1-sn^t$. The left-hand graphs are for the case with $T = 124$ (used in the computer simulations), and the right-hand graphs for $T = 500$. $u-v$ is 0·0025 in both cases. The asymptotes of the curves correspond to the s values below which saturation of all T sites occurs for the given value of t.

These results imply that fitness must fall off more steeply with n than does a multiplicative function $w_n = (1-s)^n$ corresponding to independent effects s of each element on fitness. This conclusion has also been reached by Brookfield (1983) using a different method. Too fast a rate of fall-off could yield a limit cycle about $\hat{n}$ rather than a stable equilibrium, but this possibility seems somewhat unlikely. It is theoretically possible that a linear fitness function $w_n = 1-ns$ could yield a stable equilibrium in $\bar{n}$. The equilibrium copy number with linearity is

$$\hat{n} = \frac{1}{s}\left\{\frac{u-v-s}{u-v-1/T}\right\} \quad (s < u-v). \tag{22}$$

This requires $T > 1/u-v$, which means a very large number of occupable sites, in view of the probable low values of u and v. Furthermore, the value of s must be closely adjusted to the other parameters in order to obtain a realistic value of $\hat{n}$. A linear fitness function thus seems unlikely; a more strongly convex function

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

of n is less tightly constrained. In our simulation work we used functions of the form

$$w_n = 1 - sn^t. \tag{23}$$

Fig. 3 illustrates the dependence of $\hat{n}$ on s, t, and $u-v$ with this model.

(ii) *Finite population size*

Selection with finite population size can be studied in much the same way as for regulated transposition (Section 3(ii)). There is, however, the complication that it is no longer true that the expected value of $\bar{n}$ is usually close to $\hat{n}$ for the infinite population case, given by equation (20a). We denote the expected value of $\bar{n}$ by n^*, and write

$$\gamma = 4N_e \left| \frac{\partial \ln w_n^*}{\partial n^*} \right|. \tag{24}$$

Using this in equation (18b), it follows from standard theory (Crow & Kimura, 1970, p. 442) that the steady-state distribution of element frequency is given approximately by

$$\phi(x) = Ce^{-\gamma x} x^{\alpha-1} (1-x)^{\beta-1}, \tag{25a}$$

$$C = \Gamma(\alpha+\beta)/\Gamma(\alpha)\,\Gamma(\beta) \left\{ 1 + \sum_{j=1}^{\infty} \frac{(-\gamma)^j}{j!} \frac{(\alpha+j-1)(\alpha+j-2)\ldots\alpha}{(\alpha+\beta+j-1)(\alpha+\beta+j-2)\ldots(\alpha+\beta)} \right\}, \tag{25b}$$

where $\alpha = 4N_e un^*/(T-n^*)$ and $\beta = 4N_e v$.

The expectation of this distribution is $\hat{x} = n^*/T$, which is given by the equation

$$C^{-1}\hat{x} = \Gamma(\alpha+\beta+1)/\Gamma(\alpha+1)\,\Gamma(\beta) \left\{ 1 + \sum_{j=1}^{\infty} \frac{(-\gamma)^j}{j!} \frac{(\alpha+j)(\alpha+j-1)\ldots\alpha}{(\alpha+\beta+j)(\alpha+\beta+j-1)\ldots(\alpha+\beta)} \right\}. \tag{26}$$

The value of $\hat{x}$ (and hence n^*, C, etc.) can be obtained by eliminating C between equations (25b) and (26), and iterating the resulting expression in $\hat{x}$.

The variance in element frequency, $\hat{\sigma}_x^2$, can be obtained by the method of Kimura & Ohta (1971, p. 185) as

$$\hat{\sigma}_x^2 \approx \hat{x} \left\{ 1 - \hat{x} - \frac{4N_e(u-v)}{\gamma} \right\}. \tag{27}$$

If N_e is sufficiently large in comparison with the force of selection that most values of x are close to $\hat{x}$, equation (18b) can be linearized about $\hat{x}$. With large T, so that $\hat{x}$ is close to zero, the linearized equation is approximately

$$\Delta x_i = \frac{u\hat{n}(1-x_i)}{(T-\hat{n})} - x_i(v + s_{\hat{n}}), \tag{28}$$

where $-s_{\hat{n}}$ is the derivative of $\ln w_n$ at $\hat{n}$, $\hat{n}$ being the equilibrium solution of equation (20c). The steady-state distribution is approximated by a beta distribution

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

as in equation (11), where α is now $4N_e u\hat{x}/(1-\hat{x})$ and $\beta = 4N_e(v+s_{\hat{n}})$ ($\hat{x} = \hat{n}/T$). This approximation is probably adequate for most natural populations, where N_e is large, especially when migration is taken into account (see Section 5(iii)).

(iii) *Simulation results*

The distribution of equations (25) was compared with the results of computer simulations, using the selection model of equation (23) with $s = 0{\cdot}001$ and $t = 1{\cdot}5$,

Table 2. *Simulation results with selection*

	$N = 50$		$N = 250$	
Linkage	Loose	Tight	Loose	Tight
Number of runs	10	10	10	5
Mean copy number per individual				
Theoretical	29·50	29·50	15·39	15·39
Simulated	18·54**	17·30**	11·64**	11·47**
Variance in element frequency between loci (σ_x^2)				
Theoretical	0·0587	0·0587	0·0096	0·0096
Simulated	0·0487**	0·0481**	0·0074**	0·0084
Fraction of loci with zero frequency (P_0)				
Theoretical	0·1012	0·1012	0·0034	0·0034
Simulated	0·3371**	0·3613**	0·0500**	0·0419**
Variance in copy number between individuals within populations (V_n)				
Theoretical	9·679	10·345	9·623	9·359
Simulated	8·872	8·660**	9·716	9·020

$u = 0{\cdot}01$, $v = 0{\cdot}005$ and $w_n = 1{-}0{\cdot}001\ n^{1{\cdot}5}$ in each case. Loose and tight linkage correspond to map lengths of 90 and 9 units respectively.

and with $u = 0{\cdot}01$, $v = 0{\cdot}005$ and $T = 124$. In an infinite population, this model applied to equations (20) gives an equilibrium mean copy number per individual of 12·56, corresponding to an element frequency per locus of 0·1013. The expected values of mean copy number and element frequency (n^* and $\hat{x}$) were obtained from numerical solution of equations (25*b*) and (26). The simulations were carried out with recombination fractions of 0·03 and 0·003 between adjacent loci ('loose' and 'tight' linkage, respectively), with population sizes of 50 and 250. The results are summarized in Table 2. Fig. 4 displays the theoretical and simulated conditional distributions with loose linkage; the simulated distributions with tight linkage are very similar to those with loose linkage.

It will be seen that there is substantial disagreement between the theoretical and simulated distributions, even with the larger population size. The main source of this disagreement lies in the much greater frequency of the zero-frequency class in the simulated, compared with the theoretical distributions. Fig. 4 shows that the simulated conditional distributions are of similar form to the theoretical ones,

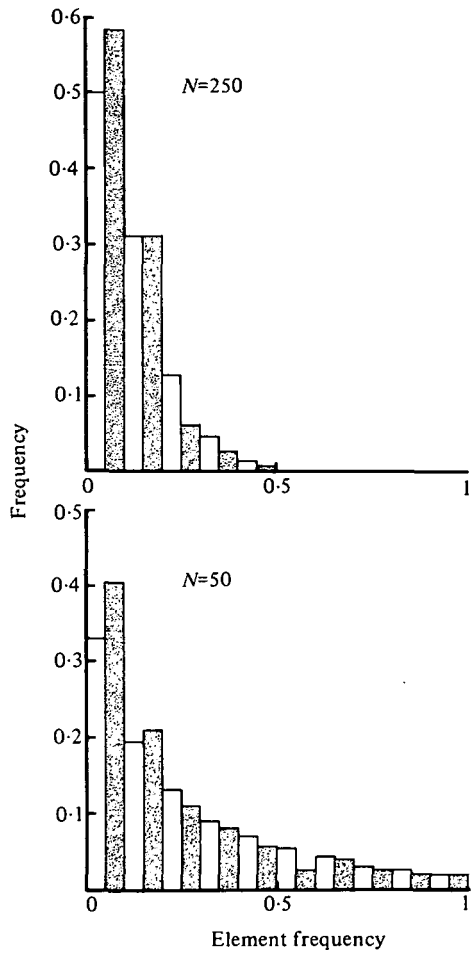

Fig. 4. Conditional distributions of element frequencies for the parameter sets of Table 2 with loose linkage. Other details as for Fig. 2.

although they tend to show an excess at the left-hand end of the distribution. This tendency is slightly but significantly exaggerated with tight linkage with $N = 50$, but not with $N = 250$.

(iv) *Linkage disequilibrium*

At first sight, linkage disequilibrium generated by selection provides an obvious explanation for the discrepancies between the simulations and theoretical expectations, which are much more marked than in the case of no selection described in Section 3(iv). Several lines of evidence rule this explanation out, however. In the first place, there is little effect of tighter linkage on the goodness of fit to the theoretical distribution. Similarly, reducing the intensity of selection does not appreciably improve goodness of fit, as was shown by a set of runs with $s = 0{\cdot}0005$, $u = 0{\cdot}005$, $v = 0{\cdot}0025$, $N = 250$ and loose linkage. Furthermore, of the 4 cases

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

shown in Table 2, only that with $N = 50$ and tight linkage yields a significant result on the test for linkage disequilibrium based on between-individual variance in copy number, V_n. The combined χ^2 over all 4 cases for the comparison of V_n with its expectation with no linkage disequilibrium is 5302 for 5548 D.F. ($p < 0{\cdot}02$), suggesting some negative linkage disequilibrium (smaller V_n than expected). Such negative linkage disequilibrium is expected from the fact that copy number per individual can be regarded as a quantitative character (with heritability of one) under directional selection (cf. Bulmer, 1980, p. 166). As shown in Appendix 2, however, the degree of linkage disequilibrium generated by the rather weak selection employed in the simulations is very small, even for adjacent loci.

5. DISCUSSION

(i) *Maintenance of a stable copy number*

The results derived in Sections 3 (i) and 4 (i) yield conditions for the maintenance of a stable mean number of copies per individual of a transposable element, such that not all possible genomic sites are occupied, resulting in variation between individuals in both copy number and identity of occupied sites. Such variation has been observed for several families of elements in *D. melanogaster* (Ananiev *et al.* 1978; Strobel, Dunsmuir & Rubin, 1979; Young, 1979; Ilyin *et al.* 1980; Pierce & Lucchesi, 1981; Langley & Montgomery, 1983). But it is important to note that these conditions may not necessarily be satisfied if the pressure of selection against individuals carrying the element, the degree of regulation of transposition rate, or the total number of occupable sites in a diploid genome (T), are too small. Under such circumstances, all available sites in all individuals may become saturated with the element, removing any possibility for variation. Different families of elements within the same species, or the same family in different species, might well behave differently with respect to the level of saturation of occupable sites (cf. Fig. 3) and to mean copy numbers, as a result of differences in these parameters. Orgel & Crick (1980) have discussed ways in which the strength of selection could vary with life-history.

In the case of selection without regulation of transposition rate, the results of Section 4 (i) indicate that individual fitness must be sufficiently downwardly curved as a function of copy number in order for saturation to be prevented. Brookfield (1982) has come independently to a similar conclusion on the basis of computer simulation results, but thinks that the equilibrium mean copy number per individual, $\bar{n}$, must be so high in relation to the intensity of selection that many individuals suffer lethality. But, as our results show, this is not in general the case. For example, with the parameters used in Table 2, the mean fitness of an infinite population at equilibrium is expected to be approximately 95·5 % of that of copy-free individuals; the fitness of individuals with copy number 2 standard deviations above the mean is 91·3 %. This is due to the fact that copy number stabilizes at a level far below that at which a substantial loss in fitness occurs, with this model. If $\bar{n}$ is small compared with T, equation (20*c*) shows that at equilibrium

$$\left|\frac{\partial \ln w_{\bar{n}}}{\partial \bar{n}}\right| \approx u - v. \tag{29}$$

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

This is analogous to Haldane's (1937) result for the genetic load with mutation-selection balance, and shows that, with small values of u and v (as seemed realistic to assume), a small decrement in fitness is sufficient to balance the increase in copy number by transposition. The reason for this disagreement is not clear, but Brookfield's conclusion that selection cannot plausibly maintain stable copy numbers does not seem to be generally valid. Selection can, of course, occur jointly with regulated transposition, but we have not analysed this in detail.

This model of selection is, of course, not the only possible one. It is appropriate for a situation in which individuals carrying elements are at a selective disadvantage, due to the induction of semi-dominant deleterious mutations by insertion of the elements, provided that element frequencies at each locus are sufficiently small that elements rarely become homozygous. In this situation, n would refer to the number of heterozygous elements carried by an individual. It is known that deleterious mutants generally have heterozygous effects on viability in *Drosophila* (Simmons & Crow, 1977), so that this model has some biological plausibility. There is also evidence for a downwardly curved relationship between viability and the number of mutants accumulated experimentally on *Drosophila* chromosomes (Simmons & Crow, 1977), which fits the above condition for maintenance of a stable copy number.

(ii) *Variation in copy number between individuals*

If an equilibrium is maintained without saturation of available sites, then the results of the simulations shown in Tables 1 and 2 (together with the considerations of Appendix 2) suggest strongly that linkage disequilibrium effects can be neglected for practical purposes. Furthermore, in a large population, variation in element frequencies between loci can be neglected. Equation (4) implies that, under these conditions, variation in copy number between individuals should be approximately binomial in form, with variance $\bar{n}(1-\bar{n}/T)$; with large T, a Poisson distribution with variance $\bar{n}$ is a sufficiently good approximation for most purposes. Unfortunately, there are too little population data available at present to test this with any precision. Langley & Montgomery (1983) have sereened 20 X chromosomes extracted from a natural population of *D. melanogaster* in North Carolina, and established the salivary chromosome locations of members of the families *copia*, *297* and *412* by means of *in situ* hybridization. The means and variances of copy number per chromosome are 1·60 and 1·94 respectively for *copia*, 3·80 and 4·17 for *297*, and 2·50 and 2·37 for *412*. The distributions of copy number fit Poisson expectations in the case of *copia* and *412*, but deviate significantly ($\chi^2_2 = 10{\cdot}88$) for *297*. (The discrepancy is mainly due to an excess of chromosomes with 3 copies of *297*.) As far as these limited data go, they suggest that observed and expected distributions of copy number agree reasonably well, except for *297* in North Carolina.

Dover (1982) has suggested that transposition of the sort discussed here can cause evolutionary change in the absence of significant variation between individuals in numbers of copies of the genetic elements concerned. The models described above, and the data just quoted, do not support this contention. Indeed, unless

all occupable sites are close to saturation, it is difficult to see how variation in copy number of an approximately binomial form could be avoided.

As is obvious from equations (6) and (20), and Fig. 3, the equilibrium copy number of an element is strongly dependent on the relative rates of transposition and loss; if selection is involved in its maintenance, equation (29) shows that the strength of selection against additional copies is of the order of $u-v$ at equilibrium. The available evidence for *Drosophila* suggests that transposition and loss typically occur at rates of the order of 10^{-4} to 10^{-5} per generation (Rasmuson *et al.* 1981; Ising & Block, 1981). Any hope of directly detecting selection effects by comparing the fitnesses of individuals with different copy numbers seems to be futile. The well-documented increase in copy number of several families in cell culture lines (Potter *et al.* 1979; Tchurikov *et al.* 1981) is, however, consistent with a role for selection, although other explanations are possible.

A possible exception to this is provided by the P factor involved in hybrid dysgenesis. The frequency of new insertions of the P element in dysgenic crosses seems to be very high (Bingham *et al.* 1982), and there is evidence that *copia* may also be mobilized simultaneously (Rubin *et al.* 1982). It has been suggested that the absence of P factors from most laboratory stocks of *D. melanogaster*, in contrast to the situation in wild populations, is due to its having only recently spread (Kidwell, Novy & Feeley, 1981). If this is the case, then the rate of transposition involved in its spread may have been much higher than that characteristic of non-dysgenic crosses. The dynamics of this process would be rather different from that envisaged in the present models, although these might still be relevant to the interpretation of copy number distributions in natural populations.

(iii) *Samples from natural populations*

As mentioned in Section (ii), data on the distributions of numbers of copies and chromosomal locations of transposable elements are becoming available for samples of chromosomes drawn from natural populations of *Drosophila*, using *in situ* hybridization of probes to polytene chromosomes. In order to test the adequacy of the fit of the distributions derived earlier to these data, it is clearly necessary to have a theory of the statistical properties of samples drawn from such distributions. Work on this problem is only in its infancy. Kaplan & Brookfield (1983*a*, *b*) have derived some results that give the expected occupancy profile for a set of haploid genomes taken from a population, i.e. the numbers of chromosomal sites (identified by virtue of being occupied by the given element in at least one of the sampled genomes) that are occupied in 1, 2, 3,...independent genomes in the sample. Their methods assume that the total number of occupable sites is effectively infinite, so that α in equations such as (11) can be taken as zero. The distribution is then characterized by the single parameter β, which they estimate from an equation relating β (their θ) to the total number of chromosomal sites identified in the sample.

A related procedure can be developed for the case of general values of α and β, assuming that the probability distribution of element frequencies fits a beta distribution, as in equation (11). The goodness of fit of the observed occupancy

profile to the expected one for any given α, β pair can be computed by a χ^2-like statistic ζ (Appendix 3). The value of ζ for the α, β pair that minimize it can be compared with other assumed values. Two useful extreme alternatives to contrast with the case of best fit are:

(*a*) The case of $\alpha = 0$, with β obtained by the method of Kaplan & Brookfield (1983*a*). In this case, all multiple occupancy must be due to drift of element frequencies in a finite population.

(*b*) The case of equal element frequencies at all loci (α and β infinite), described in Appendix 3. In this case, multiple occupancy is due entirely to the fact that element frequencies at each site are non-zero, because of the finite number of occupable sites.

Table 3. *Analysis of the data of Langley & Montgomery (1983) on the distribution of* 297 *over 20 X chromosomes from a N. Carolina population of* D. melanogaster

No. of chromosomes occupied at a site	1	2	3	4	5	> 5	ζ	Estimated $\frac{1}{2}T$
No. of sites observed with given occupancy	35	11	2	2	1	0	—	—
No. of sites expected with given occupancy								
(*a*)	35·57	9·73	3·46	1·35	0·54	0·33	1·85	∞
(*b*)	27·55	14·36	4·73	1·10	0·19	0·03	8·50	73
(*c*)	34·32	9·84	3·59	1·42	0·58	0·40	1·79	1340

(*a*) are the expectations using the method of Kaplan & Brookfield (1983*a*), who assume $\alpha = 0$ and estimate β as 16·72; (*b*) are the estimates obtained assuming infinite α and β and estimating the value of $\hat{x}$ that minimizes ζ on this assumption ($\hat{x} = 0{\cdot}052$); (*c*) uses the joint estimation of α and β ($\alpha = 0{\cdot}048$, $\beta = 16{\cdot}70$) by minimization of ζ.

As an illustration of this method, which is computationally very simple, Table 3 shows the analysis of the data of Langley & Montgomery (1983) on the distribution of *297* in the sample of X chromosomes mentioned earlier. The observed and expected occupancy profiles are very similar for case (*a*) ($\alpha = 0$), and fit is barely improved by joint estimation of α and β. Case (*b*), with infinite α and β, gives a noticeably worse fit, but this is not statistically significant. The estimates of $\frac{1}{2}T$, the number of occupable sites in a haploid X chromosome genome, are 1340 for the joint estimation of α and β, and 73 for case (*b*).

Similar calculations can be done for the data on *copia* and *412* for the same sample. In the case of *412*, the best fit is obtained with $\alpha = 0$ and $\beta = 30$; with *copia*, infinite α and β, with an element frequency of 0·02 per site ($\frac{1}{2}T = 82$), give the best fit. As in the case of *297*, the differences in goodness of fit are not large for the three alternatives considered, although case (*b*) again gives the worst fit for *412*.

It is clearly not easy to discriminate between alternative models that can generate an adequate fit to the same set of data, particularly as the *in situ* hybridization methods can only locate the positions of elements to polytene chromosome bands (at best). Apparent multiple occupancy of the same site may thus reflect occupancy of different, but neighbouring, sites, as discussed by Kaplan

& Brookfield (1983). The estimated number of sites is clearly biased downwards for this reason, and β will tend to be underestimated.

Nevertheless, it is interesting to note that the data of Langley & Montgomery (1983) suggest high β values (ranging from 16·7 for *297* to infinity for *copia*) when the joint estimation procedure for α and β is applied. This does not necessarily mean that the loss rate v that appears in the product $\beta = 4N_e v$ in equation (11) is very high. Natural populations of *Drosophila* are not closed entities, as this formula assumes, and the evidence from allozyme variation in *Drosophila* suggests that there is usually little genetic differentiation between local populations (Lewontin, 1974). The problem of obtaining the form of gene frequency distributions for general models of migration and genetic drift has not been solved (Felsenstein, 1976); the only case where a solution is known is when a fraction m of the genes of a local population are derived from the gene pool of the species at large, and $(1-m)$ are of local origin. In this case, α and β in equation (11) include the additional terms $4N_e m\hat{x}$ and $4N_e m(1-\hat{x})$ respectively, where $\hat{x}$ is the expected element frequency given by the same equation as previously ($\hat{x} = \mu_{\hat{n}}/[\mu_{\hat{n}} + v]$) (cf. Crow & Kimura, 1970, p. 437). As mentioned in Section 4(ii), the effects of weak selection can also be incorporated into β (equation [28]), so that there is some hope that the beta distribution can be used as a basis for inference about natural populations. Large values of β can clearly arise because $\hat{x}$ is close to zero (if T is large) and $4N_e m$ is substantial, even if $4N_e v$ is small.

It would clearly be of great interest to have data on geographical patterns of variation in element frequencies, to compare with the patterns observed at conventional gene loci, although the lack of resolution of *in situ* hybridization techniques may hinder accurate analysis of such patterns. The data of Pierce & Lucchesi (1981) on *Dm25* do not suggest any obvious geographical differentiation, but they are very limited in quantity.

(iv) *Discrepancies between simulation results and theory*

As described in Sections 3(iii) and 4(ii), there are some discrepancies between the theoretical distributions of element frequency in finite populations and those obtained in the simulations. In particular, there is a consistent tendency in the simulations for more loci to have lost the element than is expected, particularly with selection. As shown in Appendix 2, linkage disequilibrium generated by selection does not seem to be responsible for these effects.

A factor which probably does play a role is a considerable excess in variance of the mean copy number between different runs of the same parameter set, over the variance that is expected on the assumption of independence between the probability distributions of element frequencies at different loci. This excess indicates a positive correlation in frequencies between loci from the same population compared with loci from different populations, and is analysed in detail in Appendix 4.

Whatever the source of the excess variance, it would be expected to cause an increased rate of loss of the element, since there will be more populations with low values of $\bar{n}$, which are the most susceptible to the loss of segregating loci.

Furthermore, the relative reduction with small n in the net rate of generation of new copies (nu_n) of the element by transposition is smaller with regulated transposition, because of the compensating effect of the increase in u as n declines. The rate of escape of loci from the state of loss or low frequency of the element is therefore relatively higher with regulated transposition. This may explain the greater proportional discrepancies in the cases with selection and no regulation.

(v) *Conclusions*

Despite these discrepancies, there is remarkably good general agreement between the simulation results and the distributions expected on the simplest assumptions concerning independence between loci. It would seem that classical population genetics models are easily adapted to the analysis of the dynamics of transposable elements with sexual reproduction. Although our models have assumed diploidy, the general results are equally applicable to sexual species where there is a predominantly haploid life-history. As pointed out by Hickey (1982), however, the occurrence of sexual reproduction and segregation (but not recombination) is crucial for replicative transposition to cause an increase in copy number within a large population.

We thank W. R. Engels, J. Haigh, D. A. Hickey, J. Maynard Smith, T. Ohta and A. Robertson for their comments on this paper, and J. F. Y. Brookfield, C. H. Langley, E. A. Montgomery and N. L. Kaplan for providing copies of their unpublished manuscripts, and for helpful discussions. Peter Croyden, of the University of Sussex Computer Centre, provided valuable assistance with the simulations.

APPENDIX 1. EQUALIZATION OF ELEMENT FREQUENCIES OVER LOCI IN AN INFINITE POPULATION

Using equation (18a), and writing $\bar{x}$ and σ_x^2 for the mean and variance of element frequency across loci, we have

$$\Delta\bar{x} \approx [\bar{x}(1-\bar{x})-\sigma_x^2]\frac{\partial \ln \bar{w}}{\partial \bar{n}}+(u-v)\bar{x}. \tag{A 1}$$

Writing $\delta x_i = x_i - \bar{x}$, we also have

$$\Delta(\delta x_i) \approx \delta x_i[(1-2\bar{x})+\sigma_x^2-\delta x_i]\frac{\partial \ln \bar{w}}{\partial \bar{n}}-\delta x_i(\mu_\pi+v). \tag{A 2}$$

Noting that $\sigma_x^2 = 2\Sigma\delta x_i^2/T$, so that $\frac{1}{2}T\Delta\sigma_x^2 \approx 2\Sigma\delta x_i\Delta(\delta x_i)$, and neglecting terms in δx_i^3, we obtain from equation (A 2)

$$\Delta\sigma_x^2 \approx 2\sigma_x^2\left[(1-2\bar{x})\frac{\partial \ln \bar{w}}{\partial \bar{n}}-(\mu_\pi+v)\right]. \tag{A 3}$$

Assuming $\partial \ln \bar{w}/\partial \bar{n} \leqslant 0$ and $\bar{x} \leqslant \frac{1}{2}$, which is the case in most biologically realistic situations, equation (A 3) implies that $\Delta\sigma_x^2 \leqslant 0$, the equality holding only when $\sigma_x^2 = 0$, i.e. when element frequencies have been equalized. This result is true even in the absence of selection.

APPENDIX 2. LINKAGE DISEQUILIBRIUM BETWEEN TRANSPOSABLE ELEMENTS GENERATED BY SELECTION

The theoretical estimate of $\delta = 4\Sigma_{i<j} D_{ij}$, the contribution of linkage disequilibrium to V_n in equation (4), can be obtained using the method of Bulmer (1980, pp. 158–160). Assuming that copy number per individual, n, is normally distributed within a population, Bulmer's equation (9·45) shows that selection in an infinite population causes δ to converge to the value given by

$$\delta = \Delta V_n / 2H, \tag{A 4}$$

Table 4. *Linkage disequilibrium effects with selection*

	$N = 50$		$N = 250$	
Linkage	Loose	Tight	Loose	Tight
Expected δ	−0·029	−0·250	−0·042	−0·321
Simulated δ	−0·807	−1·685	+0·093	−0·339
Expected D for adjacent loci	$-3{\cdot}70 \times 10^{-6}$	$-3{\cdot}70 \times 10^{-5}$	$-4{\cdot}90 \times 10^{-6}$	$-4{\cdot}58 \times 10^{-5}$

See text for further explanation.

where H is the harmonic mean of the recombination fractions over all pairs of loci involved, and ΔV_n is the change in variance in copy number induced by selection within a single generation. If $f(n)$ is the frequency of individuals with n copies, then

$$V_n + \Delta V_n = \frac{1}{\bar{w}} \Sigma w_n (n - \bar{n})^2 f(n). \tag{A 5}$$

Assuming normality, and approximating w_n by the first two terms of a Taylor's expansion about $\bar{n}$, this yields the expression

$$\Delta V_n \approx \frac{V_n^2}{2\bar{w}} \frac{\partial^2 w_{\bar{n}}}{\partial \bar{n}^2}. \tag{A 6}$$

Using the means of $\bar{n}$ and V_n found in the simulations, it is easy to calculate δ from equation (A 4), for a given value of H. In the case of 2 chromosomes each carrying m loci at intervals of r map units, it is easily found from Haldane's mapping function that

$$H = m(2m-1) \Big/ \left\{ 2m^2 + 4 \sum_{i=1}^{m-1} i/(1 - e^{-2r(m-i)}) \right\}. \tag{A 7}$$

For $m = 31$ ($T = 124$), H takes values of 0·1277 and 0·0146 for $r = 0{\cdot}03$ and 0·003 respectively, corresponding to the cases of loose and tight linkage in the simulations.

Table 4 shows values of δ calculated from these formulae, and compares them with values found in the simulations in Table 2. The expected D values for adjacent loci, calculated from Bulmer's equation (9.44), are also shown. It is clear that linkage disequilibrium effects are expected to be slight, even with tight linkage, with this intensity of selection. Because of the variances in D expected

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

with small population size, the general lack of good quantitative agreement between theoretical and simulated values of δ is not surprising; there is, however, close agreement in the case with tight linkage and $N = 250$, which is the most favourable case for detecting selectively generated linkage disequilibrium. It therefore seems most unlikely that linkage disequilibrium due to selection can explain the discrepancies between the theoretical and simulated distributions.

APPENDIX 3. ANALYSIS OF DATA FROM NATURAL POPULATIONS

It is assumed that we have a set of m haploid genomes, drawn independently from a natural population. The chromosomal sites at which a probe for a given family of elements hybridizes are assumed to be known for each sampled genome, enabling the occupancy profile to be determined for the sample. Let n_i be the number of chromosomal sites that are occupied by the element in i genomes ($i = 1, 2, \ldots, m$). The mean number of sites per haploid genome ($\frac{1}{2}\hat{n}$) is estimated by $\Sigma i n_i/m$. The expected value of n_i is, from equation (16), given by

$$E\{n_i\} = \tfrac{1}{2}T\binom{m}{i}\int_0^1 x^i(1-x)^{m-i}\,\phi(x)\,dx. \tag{A 8}$$

If we assume the beta distribution of equation (11) for $\phi(x)$, this reduces to

$$E\{n_i\} = \tfrac{1}{2}\hat{n}\frac{(\alpha+1)(\alpha+2)\ldots(\alpha+i-1)(\beta+1)\ldots(\beta+m-i-1)}{(\alpha+\beta+1)(\alpha+\beta+2)\ldots(\alpha+\beta+m-1)}\binom{m}{i}, \tag{A 9}$$

where $\hat{n} = T\alpha/(\alpha+\beta)$.

A simple procedure for estimating α and β is to substitute the estimator of $\frac{1}{2}\hat{n}$ into equation (A 9), and then minimize the goodness-of-fit statistic

$$\xi = \sum_i \frac{[n_i - E\{n_i\}]^2}{E\{n_i\}}, \tag{A 10}$$

Values of ξ obtained by this method can be compared with the values generated by the substitution of alternative estimates of α and β into equation (A 9), such as the estimates of β used by Kaplan & Brookfield (1983*a*) who assume $\alpha = 0$. Another alternative, mentioned in the text, is to assume infinite population size, so that α and β are both infinite and element frequencies are equal to $\hat{x} = \alpha/(\alpha+\beta)$ at each locus. In this case equation (A 8) reduces to

$$E\{n_i\} = \tfrac{1}{2}\hat{n}\hat{x}^{i-1}(1-\hat{x})^{m-i}, \tag{A 11}$$

ζ can then be minimized with respect to $\hat{x}$. As pointed out to us by Dr A. W. F. Edwards, this case is equivalent to the problem studied by Lewontin & Prout (1956), who provided a maximum likelihood estimator of $\frac{1}{2}T$, and hence $\hat{x}$. Dr Edwards' calculations on the data analysed in the text show that their method yields consistently lower estimates of $\frac{1}{2}T$ than the minimization of ζ, although the estimates are of similar magnitude.

The significance of the fit of the observed and expected distributions can be tested by pooling classes to avoid expectations < 5, in the usual way. The resulting goodness-of-fit statistic will be distributed approximately as χ^2, with degrees of

freedom equal to the number of cells minus the number of parameters estimated from the data (2 in the case of methods [*a*] and [*b*] of Table 3, and 3 for method [*c*]). Unfortunately, most cells have such small numbers that all the d.f. are used up, except for the case of methods (*a*) and (*b*) applied to *297*. The χ^2_1 values are 0·26 and 2·98 respectively, showing no significant deviations from expected.

Table 5. *Variation in mean copy number between runs and between-locus correlations in the simulations of Table 1 and 2*

	(*a*) Regulated transposition with $k = 0{\cdot}05$			
Population size	10	50		100
Linkage	Loose	Loose	Tight	Loose
No. of runs analysed	10	10	10	10
$\sigma^2_{\bar{n}}$ (simulated)	69·09**	50·30**	103·11**	49·12**
$\sigma^2_{\bar{n}}$ (expected, if no correlations)	25·82	16·24	13·59	10·86
Covariance	0·0029	0·0029	0·0059	0·0025
Correlation	0·0275	0·0274	0·1080	0·0577

	(*b*) Selection			
Population size	50		250	
Linkage	Loose	Tight	Loose	Tight
No. of runs analysed	10	10	5	5
$\sigma^2_{\bar{n}}$ (simulated)	64·80**	50·66**	6·54**	2·33
$\sigma^2_{\bar{n}}$ (expected, if no correlations)	12·08	11·93	1·84	2·08
Covariance	0·0035	0·0026	0·0003	0·0002
Correlation	0·0716	0·0532	0·0420	0·0164

** Significant excess over the expected variance at the 1 % level.

APPENDIX 4. EXCESS VARIANCE IN MEAN COPY NUMBER BETWEEN RUNS

The expected variance in $\bar{n}$ between replicate runs of the same parameter set can be obtained as follows, assuming that the between-run variance in element frequency takes the same value σ^2 at each locus. From equation (1) we have

$$\sigma^2_{\bar{n}} = 2T\sigma^2 + 8 \sum_{i>j} \mathrm{cov}\,(x_i, x_j), \tag{A 12}$$

where cov (x_i, x_j) is the between-run covariance in element frequency between loci i and j. If $l(= \frac{1}{2}T)$ is the total number of loci (usually 62), the mean covariance and correlation between locus pairs are given by

$$\mathrm{cov} = (\sigma^2_{\bar{n}} - 2T\sigma^2)/4l(l-1), \tag{A 13}$$

$$\rho = \mathrm{cov}/\sigma^2. \tag{A 14}$$

For a given set of runs, $\sigma^2_{\bar{n}}$ and σ^2 can be estimated directly, and cov and ρ obtained from these formulae. Table 5 displays the results for the simulations of Table 1 ($k = 0{\cdot}05$) and Table 2. There are significant effects in all but one case; ρ is usually of the order of a few per cent, which is sufficient to generate a large excess variance because of the large number of locus pairs (1891 with 62 loci). It

https://doi.org/10.1017/S0016672300021455 Published online by Cambridge University Press

is clear that the assumption of independence between the probability distributions for different loci is to some extent violated in the simulations.

Two factors may contribute to this effect. Firstly, all loci from the same populations share a common value of $\bar{n}$, so that the contributions of transposition and selection to change in element frequency (which both involve terms in $\bar{n}$) are correlated across loci. Calculations indicate that this effect should be larger with selection than with regulated transposition, but this is not apparent from Table 5. Secondly, linkage disequilibrium generated by drift may contribute to the correlations between loci. Substantial D values may be generated with the population sizes and recombination fractions used in the simulations (Hill, 1976). They can contribute a large component of the variance in copy number (equation [4]), and tend to persist from generation to generation (Bulmer, 1980, pp. 226–232). A reduction in variance due to net negative linkage disequilibrium will tend to reduce the efficiency of selection at all loci, but would increase the net effect of regulated transposition, as can be seen from equation (3). An increase in variance due to positive linkage disequilibrium has the opposite effects. This factor would be expected to be most important with tight linkage and small population size. There are no consistent effects of these parameters on the correlations in Table 5, but the sampling errors are rather large and may obscure any real trends.

REFERENCES

Ananiev, E. V., Gvozdev, V. A., Ilyin, Y. V., Tchurikov, N. A. & Georgiev, G. P. (1978). Reiterated genes with varying location in intercalary heterochromatin regions of *Drosophila melanogaster*. *Chromosoma* **70**, 1–17.

Barrett, J. A. (1982). Junk DNA and the 'parasite paradigm': neo-Darwinism revisited. (Unpublished MS.)

Bingham, P. M., Kidwell, M. G. & Rubin, G. M. (1982). The molecular basis of *P-M* hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell* **29**, 995–1004.

Brookfield, J. F. Y. (1982). Interspersed repetitive DNA sequences are unlikely to be parasitic. *Journal of Theoretical Biology* **94**, 281–299.

Brookfield, J. F. Y. (1983). A simple analytical model for the spread of independent transposable elements of equal effect. *Journal of Theoretical Biology* (submitted).

Bulmer, M. G. (1980). *The Mathematical Theory of Quantitative Genetics*. Oxford: Oxford University Press.

Charlesworth, B. (1983). Recombination, genome size and chromosome number. In *DNA and Evolution: Natural Selection and Genome Size* (ed. T. Cavalier-Smith). Chichester: Wiley. (In press.)

Crow, J. F. & Kimura, M. (1970). *An Introduction to Population Genetics Theory*. New York: Harper & Row.

Doolittle, W. F. (1982). Selfish DNA after fourteen months. In *Genome Evolution* (ed. G. A. Dover and R. B. Flavell), pp. 3–28. London: Systematics Association and Academic Press.

Doolittle, W. F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature* **272**, 123–124.

Dover, G. A. (1982). Molecular drive: a cohesive mode of species evolution. *Nature* **299**, 111–117.

Engels, W. R. (1981). Hybrid dysgenesis in *Drosophila* and the stochastic loss hypothesis. *Cold Spring Harbor Symposium on Quantitative Biology* **45**, 561–566.

Ewens, W. J. (1979). *Mathematical Population Genetics*. Berlin: Springer-Verlag.

Felsenstein, J. (1976). The theoretical population genetics of variable selection and migration. *Annual Review of Genetics* **10**, 253–280.

FINNEGAN, D. J., WILL, B. H., BAYEV, A. A., BOWCOCK, A. M. & BROWN, L. (1982). Transposable DNA sequences in eukaryotes. In *Genome Evolution* (ed. G. A. Dover and R. B. Flavell), pp. 29–40. London: Systematics Association and Academic Press.

FRANKLIN, I. & LEWONTIN, R. C. (1970). Is the gene the unit of selection? *Genetics* **65**, 701–734.

GREEN, M. M. (1980). Transposable genetic elements in *Drosophila* and other Diptera. *Annual Review of Genetics* **14**, 109–120.

HALDANE, J. B. S. (1937). The effect of variation on fitness. *American Naturalist* **71**, 337–349.

HICKEY, D. A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics* **101**, 519–531.

HILL, W. G. (1976). Non-random associations of neutral linked genes in finite populations. In *Population Genetics and Ecology* (eds S. Karlin and E. Nevo), pp. 339–376. New York: Academic Press.

ILYIN, Y., CHMELIAUSKAITE, V., ANANIEV, E. V. & GEORGIEV, G. P. (1980). Isolation and characterisation of a new family of mobile dispersed genetic elements, *mdg3*, in *Drosophila melanogaster*. *Chromosoma* **81**, 27–53.

ISING, G. & BLOCK, K. (1981). Derivation-dependent distribution of insertion sites for a *Drosophila* transposon. *Cold Spring Harbor Symposium on Quantitative Biology* **45**, 527–544.

KAPLAN, N. L. & BROOKFIELD, J. F. Y. (1983*a*). Transposable elements in Mendelian populations. III. Statistical results. *Genetics*. (In the Press.)

KAPLAN, N. L. & BROOKFIELD, J. F. Y. (1983*b*). The effect on homozygosity of selective differences between sites of transposable elements. *Theoretical Population Biology*. (In the Press.)

KIDWELL, M. G., KIDWELL, J. F. & SVED, J. A. (1977). Hybrid dysgenesis in *Drosophila melanogaster*: a syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics* **86**, 813–833.

KIDWELL, M. G., NOVY, J. B. & FEELEY, S. M. (1981). Rapid unidirectional change of hybrid dysgenesis potential in *Drosophila*. *Journal of Heredity* **72**, 32–38.

KIMURA, M. & OHTA, T. (1971). *Theoretical Aspects of Population Genetics*. Princeton: Princeton University Press.

KITTS, P. A., LAMOND, A. & SHERRATT, D. J. (1982). Inter-replicon transposition of *Tn1/3* occurs in two sequential genetically separable steps. *Nature* **295**, 626–628.

KLECKNER, N. (1981). Transposable elements in prokaryotes. *Annual Review of Genetics* **15**, 341–404.

LANGLEY, C. H. & MONTGOMERY, E. A. (1983). Transposable elements in Mendelian populations. II. Distribution of three *copia*-like elements in a natural population of *Drosophila melanogaster*. *Genetics*. (In the Press.)

LANGLEY, C. H., BROOKFIELD, J. F. Y. & KAPLAN, N. L. (1983). Transposable elements in Mendelian populations. I. A theory. *Genetics*. (In the Press).

LEWONTIN, R. C. (1974). *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press.

LEWONTIN, R. C. & PROUT, T. (1956). Estimation of the number of different classes in a population. *Biometrics* **12**, 211–223.

MCCLINTOCK, B. (1956). Controlling elements and the gene. *Cold Spring Harbor Symposium on Quantitative Biology* **21**, 197–216.

OHTA, T. (1981). Population genetics of selfish DNA. *Nature* **292**, 648–649.

OHTA, T. (1983). Theoretical study on the accumulation of selfish DNA. *Genetical Research*. **41**, 1–16.

OHTA, T. & KIMURA, M. (1981). Some calculations on the amount of selfish DNA. *Proceedings of the National Academy of Sciences, USA* **78**, 1129–1132.

ORGEL, L. E. & CRICK, F. H. C. (1980). Selfish DNA: the ultimate parasite. *Nature* **284**, 606–607.

PIERCE, D. A. & LUCCHESI, J. C. (1981). Analysis of a dispersed repetitive DNA sequence in isogenic lines of *Drosophila*. *Chromosoma* **82**, 471–492.

POTTER, S., BROREIN, W. J., DUNSMUIR, P. & RUBIN, G.M. (1979). Transposition of elements of the *412*, *copia*, and *297* dispersed repeated gene families in *Drosophila*. *Cell* **17**, 415–427.

RASMUSON, R., WESTERBERG, B. M., RASMUSON, A., GVOZDEV, V. A., BELYAEVA, E. S. & ILYIN, Y. V. (1981). Transposition, mutable genes, and the dispersed gene family *Dm225* in *Drosophila melanogaster*. *Cold Spring Habor Symposium on Quantitative Biology* **45**, 545–551.

REED, R. R., SHIBUYA, G. I. & STEITZ, J. A. (1982). Nucleotide sequence of $\gamma\delta$ resolvase gene and demonstration that its gene product acts as a repressor of transcription. *Nature* **300**, 381–383.

RUBIN, G. M., KIDWELL, M. G. & BINGHAM, P. M. (1982). The molecular nature of P-M hybrid dysgenesis: the nature of induced mutations. *Cell* **29**, 987–994.

SIMMONS, M. J. & CROW, J. F. (1977). Mutations affecting fitness in *Drosophila* populations. *Annual Review of Genetics* **11**, 49–78.

STROBEL, E., DUNSMUIR, P. & RUBIN, G. M. (1979). Polymorphisms in the chromosomal locations of elements of the *412*, *copia*, and *297* dispersed repeated gene families in *Drosophila*. *Cell* **17**, 429–439.

TCHURIKOV, N. A., ILYIN, Y. V., SKRYABIN, K. G., ANANIEV, E. V., BAYEV, A. A., KRAYEV, A. S., ZELENTSOVA, E. S., KULGUSKIN, V. V., LYOBOMIRSKAYA, N. V. & GEORGIEV, S. P. (1981). General properties of mobile dispersed genetic elements in *Drosophila melanogaster*. *Cold Spring Harbor Symposium on Quantitative Biology* **45**, 655–671.

YOUNG, M. W. (1979). Middle repetitive DNA: a fluid component of the *Drosophila* genome. *Proceedings of the National Academy of Sciences, U.S.A.* **76**, 6274–6278.