Quantitative variation and chromosomal location of satellite DNAs

WOLFGANG STEPHAN*

Institut für Physikalische Chemie, Technische Hochschule Darmstadt, Petersenstr. 20, D-6100 Darmstadt

(*Received 24 November 1986 and in revised form 4 March 1987*)

Summary

A model of the evolutionary accumulation of highly repeated DNA (HRDNA) is proposed. The accumulation of HRDNA sequences, which are organized largely in tandem arrays and whose functional significance is obscure, is explained here as a consequence of the action of the forces of amplification (promoting increase in copy numbers) and unequal crossing over, random drift and natural selection (controlling copy numbers). This model provides a general framework (i) to study the chromosomal location of satellite DNAs present in the genomes of all higher eukaryotes, and (ii) to explain the significant variation in the amounts of satellites which is frequently found among closely related species, but only rarely within a species. A review of the relevant data is included and open questions are identified.

1. Introduction

Satellite DNA forms long clusters of tandemly repeated sequences of 10^4 to 10^7 copies, and is found in the chromosomes of all higher organisms. In particular, in mammals satellites comprise anywhere from a few percent (human) to over 50% (kangaroo rat) of total DNA. Apart from this, the most striking property of satellite DNA is its great variation in sequence, organization and quantity (reviewed by John & Miklos (1979) and Singer (1982)). Other important characteristics include (i) association with constitutive heterochromatin, and (ii) lack of measurable transcription.

A function of satellite DNA has not been identified so far, though a great deal of information has been gathered in the past 25 years from a more and more detailed analysis of structure and changes in structure. In two recent publications, we have opened up another way of looking at the problem of function by including population genetics considerations (Charlesworth, Langley & Stephan, 1986; Stephan, 1986*a*). By means of mathematical modelling it has been shown there that the distribution of HRDNA along the chromosome arms can most readily be explained as a consequence of the interaction of the forces of amplification (increasing copy numbers) and selection and recombination (controlling copy numbers). Accordingly, satellite DNA is likely to accumulate in those chromosomal regions in which recombination is heavily suppressed.

These results have been obtained in a rather indirect way, calculating the expected persistence time of HRDNA under the joint action of unequal crossing over, random genetic drift and natural selection, but in the absence of amplification (Stephan, 1986*a*). Assuming no amplification, the mean persistence time of an array starting with a certain amount of copies can be determined, since HRDNA will eventually be lost from the population. However, it is of much more biological interest to obtain non-trivial equilibrium distributions of array sizes. Therefore, this paper extends our original model by including an explicit mechanism for amplification. Furthermore, two different selection schemes are introduced, an additive and a truncation selection model. Both models are based on the assumption that satellites are generally functionless, such that they are neutral at low amounts and deleterious at too high ones (truncation selection), or that the reduction in fitness caused by a single extra copy is constant, i.e. independent of array size (additive selection). Obtaining analytical expressions for the stationary copy number distributions allows us to study the chromosomal distribution of satellites as well as the variation of copy numbers within and between populations or closely related species. The theoretical results are compared with relevant data. An attempt is made to systematize the information available on intra- and interspecific copy number variability and to draw attention to questions which must remain open in our interpretation of the evolution of satellite DNA, due mostly to a lack of experimental data; in particular, data on intraspecies variation are scarce.

* Present address: Laboratory of Genetics, P.O. Box 12233, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA.

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

2. The model

It is widely recognized that the mechanisms for the changes of satellites during evolution are unequal crossing over and saltatory amplification. Whereas the amounts of HRDNAs increase by saltatory events, unequal exchange and natural selection are thought to be forces controlling copy numbers (Charlesworth *et al.* 1986). In a given generation, the following processes are allowed to modify the copy numbers of HRDNAs: amplification, selection and sampling of individuals, and recombination among the chromosomes carried by the sampled and surviving individuals. The details of these processes are as follows.

Amplification. This term stands here for processes that lead to increases in the copy numbers of DNA sequences and occur locally in the genomes; that is, potentially long-range forces like transposition are not thought to play an important role for tandemly arrayed multigene families, except for the spread of HRDNA to non-homologous chromosomes which is not considered here. Genomic sequences can be amplified. This has been demonstrated for several systems, for instance, globin genes (Fritsch, Lawn & Maniatis, 1980), the gene for dihydrofolic acid reductase (Alt, Kellems, Bertino & Schimke, 1978), etc. In at least one case (mouse satellite), amplification of satellite DNA has been measured directly, due to being accompanied by the amplification of the dihydrofolate reductase gene (Bostock & Clark, 1980). The precise mechanisms of amplification are unknown but it has been proposed to occur by different sorts of replicational processes, a 'disproportionate replication' mechanism (Schimke, 1984), a 'rolling circle' mechanism (Hourcade, Dressler & Wolfson, 1973) or biased 'slippage replication' (Wells *et al.* 1967). Contrary to what seems to be widely believed, unequal crossing over itself, i.e. in the absence of other forces, does not work as an amplification mechanism, since it does not change the mean copy number of HRDNAs in a given population (Stephan, 1986*a*).

For our problem of copy number regulation in HRDNA families it is not necessary to go into the molecular details of the above amplification mechanisms. Instead we summarize their properties by the following two assumptions: (i) The number of new copies produced in a given amplification event depends on the current size, i, of the HRDNA cluster, and (ii), the number of new copies is randomly and uniformly distributed between 1 and i. Formally, the probability that an amplification even produces new copies leading from array size i to j is then given by

$$A_{ij} = \begin{cases} \frac{1}{i}, & i+1 \leqslant j \leqslant 2i \\ 0, & \text{else.} \end{cases} \tag{1}$$

This model contains the multiplicative (saltatory) aspect of amplification. Moreover, since the newly generated copies are assumed to be placed adjacent to the old ones, it follows a tandem duplication mechanism, as originally proposed by Southern (1970). Let μ denote the rate per chromosome per generation at which amplification events occur. Since the number of new copies is a uniformly distributed random number between 1 and i, the average duplication rate of the sequences is given by $\frac{1}{2}\mu(1+1/i)$. This is approximately equal to $\frac{1}{2}\mu$, as i is usually large.

Natural selection. We consider a sexual haploid species and assume that the fitness, w_i, of the individuals in a given population of size $2N$ is a simple function of copy number, i (Charlesworth *et al.* 1986). Examining the consequences of our hypothesis that satellites are generally functionless, we may assume that w_i is a constant (in the neutral case) or a decreasing function of i, if satellites are (slightly) deleterious. Two selection schemes are contrasted with each other, the additive selection model of our previous paper (Stephan, 1986*a*)

$$w_i = 1-s(i-1), \quad s>0 \tag{2a}$$

and a truncation model

$$w_i = \begin{cases} 1, & i \leqslant \Omega \\ 0, & i > \Omega. \end{cases} \tag{2b}$$

In the additive model, each copy of an HRDNA family imposes a burden s upon an individual, no matter what the current array size is. In contrast, the truncation model takes into account that satellites present in small amounts may well be considered neutral. The following analysis indicates that a realistic selection schemes lies between these two extreme models. For comparison we put $s^{-1} = \Omega$. This is the upper limit of copy number which is tolerable for the organisms in both models.

The existence of such a limit is certainly debatable, and so is its value. Although it is biologically plausible that selection prevents array from becoming arbitrarily large, there is no direct evidence for this effect. The fact that the copy numbers of different satellite families vary from 10^4 to 10^7 does not support our selection schemes either, at least at first sight. However, a quantity can be identified which is relatively constant across different (satellite) families and phyla of higher eukaryotes encompassing amphibians, insects and mammals. This is the product of copy number × repeat length. There is a correlation such that copy numbers of arrays are the larger the shorter the repeat units. For instance, for the alphoid satellite family of primates lying in the class of HRDNAs at the lower end of copy number ($1{\cdot}5 \times 10^4$ to $1{\cdot}3 \times 10^5$ copies per chromosome) (Pike, Carlisle, Newell, Hong & Musich, 1986), the length of a monomeric repeat is around 170 base pairs (bp), whereas for large families (5×10^6 to 5×10^7 copies) a repeat length around 10 bp is usually found (e.g. the three satellites of *Drosophila virilis* have a repeat length of 7 bp (Gall & Atherton, 1974)). This relative constancy of net satellite amounts

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

in eukaryotes (despite significant fluctuations in copy numbers within a particular genus) may indicate that there is a general force involved in copy number control (and thus in the regulation of genome size and of concomitant cellular and developmental processes, e.g. cell division (see Cavalier-Smith, 1985)) rather than specific molecular mechanisms only, and this force acts on satellites via quantity (number of copies or nucleotides) rather than sequence.

Unequal crossing over. As in our previous paper (Stephan, 1986*a*), we assume the following model of unequal meiotic exchange

$$Q_{ijk} \sim 1 - \left|\frac{2i}{j+k} - 1\right|. \qquad (3)$$

Q_{ijk} denotes the probability that an exchange between chromosomes with j and k copies, respectively, yields a daughter with i copies (conditional on an exchange having occurred). The rate of exchange, γ, is given per array and per generation.

3. Copy number distribution under additive selection

Analytical treatment of the model is possible only in asymptotic parameter ranges; these are (i) $N\gamma \ll 1$, $N\mu \ll 1$ and (ii) $N\gamma \gg 1$, $N\mu \gg 1$. In our previous papers (Charlesworth *et al.* 1986, Stephan, 1986*a*), we have identified the former as more relevant for the accumulation of HRDNA. Therefore, analysis of the model will be given only in this range. Possible extensions to intermediate parameters will be discussed later (section 5).

If $N\mu \ll 1$ and $N\gamma \ll 1$, the population is usually fixed for a single gamete type (Stephan, 1986*a*). Therefore, the process can first be studied on the time scale of successive fixation events, and its dynamics described by a finite Markov chain. Let p_{ij} be the probability of transition from state E_i $(i = 1, \ldots, \Omega)$ where all individuals of the population carry chromosomes with exactly i copies, to state $\mathrm{E_j}$. $\mathrm{p_{ij}}$ is obtained by including the amplification process in Stephan's (1986*a*) equation (3):

$$p_{ij} = \left(\frac{\gamma}{\gamma+2\mu} Q_{jii} + \frac{2\mu}{\gamma+2\mu} A_{ij}\right) r_{ij}, \quad (i \neq j). \qquad (4)$$

The factor 2 in this equation expresses the fact that, on the time scale of generations, recombination and amplification events altering array sizes of a population of $2N$ chromosomes occur with rate $(\gamma+2\mu)N$. r_{ij} is the probability of fixation of a variant chromosome with j copies in a population originally fixed for i copies. r_{ij} is given by (Stephan, 1986*a*)

$$r_{ij} \approx \frac{1-\exp(-2s(i-j))}{1-\exp(-4Ns(i-j))}. \qquad (5)$$

Following the methods of our previous paper, we replace the Markov chain, as defined by equations (4) and (5), bv a diffusion in the variable $x \equiv s(i-1)$, $x \in [0, 1]$. It is also convenient to approximate r_{ij} by a simpler function such that

$$\left.\begin{aligned} r_{ij} &= \frac{1}{2N} + s(i-j), & |i-j| &\leqslant \frac{1}{4Ns} \\ r_{ij} &= \begin{cases} 2s(i-j), & i-j \geqslant \dfrac{3}{4Ns} \\ 2s(j-i)\exp(-4Ns(j-i)), & i-j \leqslant -\dfrac{3}{4Ns}. \end{cases} \end{aligned}\right\} \qquad (6)$$

These approximations follow immediately from equation (5), when $|i-j|$ is sufficiently small (large) (see also Crow & Kimura (1970), p. 426). In the intermediate intervals, r_{ij} is approximated by functions which are linear in $s(i-j)$ and coincide with the above functions at the endpoints.

We are now ready to calculate the mean, $a(x)$, and variance, $b(x)$, of the rate of change in x between successive generations. Let $Y(\tau)$ indicate the state of the Markov chain at time τ (τ, time on the scale of fixation events). Then the first two moments of the change of Y are given by

$$E\{(Y(\tau+1) - Y(\tau))^n \mid Y(\tau) = i\} = \sum_{j=1}^{2i} p_{ij}(j-i)^n, \quad (n = 1, 2). \qquad (7)$$

Using (4) and (6), the evaluation of the sums in equation (7) is straightforward. After tedious calculations we obtain explicit formulae for the diffusion and drift coefficients by rescaling state space as $X(t) \equiv s(Y(t)-1)$ and time as $t \equiv ((\gamma+2\mu)N)^{-1}\tau$. According to (6), the formulae are derived for the following three domains of x:

$$x \leqslant \frac{1}{4N}, \quad \frac{1}{4N} < x < \frac{3}{4N} \quad \text{and} \quad x \geqslant \frac{3}{4N}.$$

The results are lengthy (see Appendix), and the distribution can in general be obtained only by numerical integration.

In Figs. 1–3, numerical results are shown for several sets of parameters. In order to examine the validity of the diffusion approximation, the theoretical results are compared with simulations (Figs. 1*a*, 1*b*). (The simulation technique is described in Stephan (1986*a*), and the data are taken from Stephan (1986*b*).) Apart from the fact that the theoretical results agree reasonably well with the simulations, several interesting points can be made:

First, the copy number distribution depends only on the ratio μ/γ (see Fig. 1*a*). This property follows immediately from (4) and the fact that $(\gamma+2\mu)N\{p_{ij}\}$ determines the process. (A consequence of this is that in the Figs. 2, 3, μ/γ is taken as the relevant variable on the x-axis.)

Secondly, there is a remarkable effect of population size on the amount of HRDNAs (see Figs. 1*b*, 2). The

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

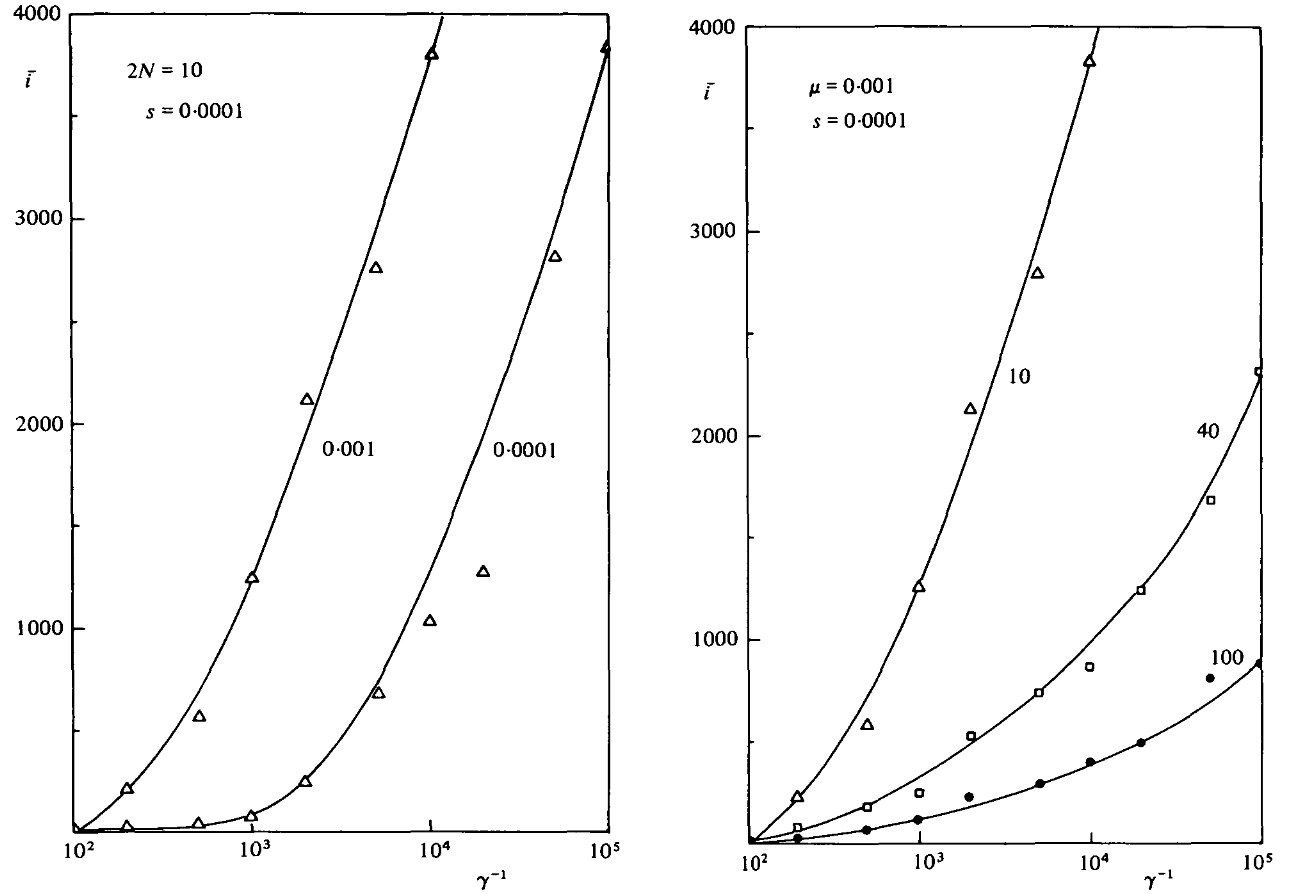

Fig. 1(*a*). Mean copy number of HRDNA *vs* inverse recombination rate under additive selection. The simulations were done with the selection coefficient $s = 0{\cdot}0001$ and population size $2N = 10$ (see Stephan, 1986*b*). The numbers by the curves are the values for μ. The solid lines represent the analytical results. (*b*) Mean copy number of HRDNA *vs* inverse recombination rate under additive selection. Here the parameters μ and s are fixed while the numbers by the curves are values for different population sizes.

model predicts that small populations have on the average more repetitive DNA sequences than large ones (conditional on the other parameters being comparable). This is a consequence of our assumption $N\mu \ll 1$ and $N\gamma \ll 1$. Since under these conditions a given population is usually fixed for a certain chromosome type, drift does not play a role in getting rid of HRDNA. For large μ/γ and sufficiently small population sizes the relation between population size and copy number reverses. Furthermore, we have $\bar{\imath} < \Omega$, since the variance (which is relatively large for small populations) keeps the mean away from the boundary Ω. This effect disappears, such that $\bar{\imath} \approx \Omega$, when population size becomes larger.

Thirdly, Fig. 3 displays the coefficient of variation *vs* μ/γ for various population sizes. For such small populations, $\sigma_i/\bar{\imath}$ is well above 0·25 for reasonable mean copy numbers $\bar{\imath}$, say $\bar{\imath} \lesssim 0{\cdot}5\Omega$, so that the model seems to predict large interpopulational variances. (In Fig. 3, the ranges of μ/γ underlined are those for which $0{\cdot}1\ \Omega \leqslant \bar{\imath} \leqslant 0{\cdot}5\ \Omega$; that is, parameter ranges for which considerable amounts of HRDNAs can be found.) Furthermore, we note that for intermediate values of μ/γ, $\sigma_t/\bar{\imath}$ is nearly independent of N. That means that σ_t, like $\bar{\imath}$, decreases with increasing population size.

In summary, the results look quite satisfactory for very small populations. The model shows that HRDNA is likely to accumulate in those regions of the chromosomes that have low recombination rates, and it predicts that closely related species may have extensive quantitative variation in their satellite DNA contents, whereas intraspecific variation should be very low. (Assuming no gene flow between populations, the words 'population' and 'species' are used interchangeably throughout this paper.) In fact, our assumptions that $N\gamma \ll 1$ and $N\mu \ll 1$ *a priori* imply that intraspecific copy number variation is near zero. But how does the model behave for larger population sizes? The apparent strong dependence of the mean on population size hints at possible limitations of the model for larger populations, since in this case the accumulation of HRDNAs seems possible only when μ/γ becomes sufficiently large. In the remainder of this section, we study this problem in more detail.

For larger populations, the analytical expressions for the diffusion and drift coefficient can be simplified such that (see formulae (A 5) and (A 6) of the Appendix)

$$a(x) \approx -\tfrac{1}{6}N\gamma x^2 + \tfrac{3}{20}\mu N^{-2}x^{-1} \qquad (8)$$

$$b(x) \approx \tfrac{1}{10}N\gamma x^3 + \tfrac{1}{10}\mu N^{-3}x^{-1},$$

where $3/4N \leqslant x \leqslant 1$. With these formulae, an explicit

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

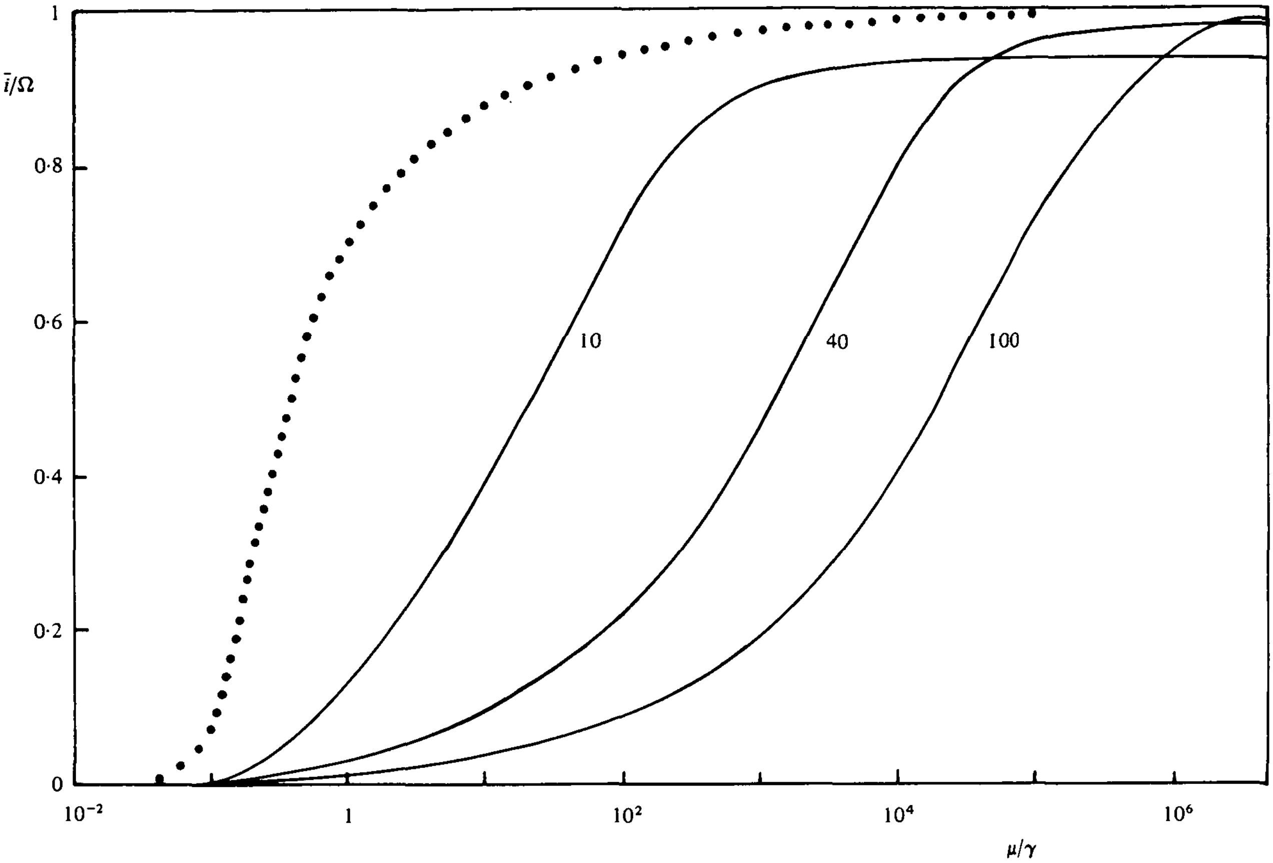

Fig. 2. Mean copy number of HRDNA *vs* μ/γ under additive selection (solid lines) and truncation selection (dotted line).

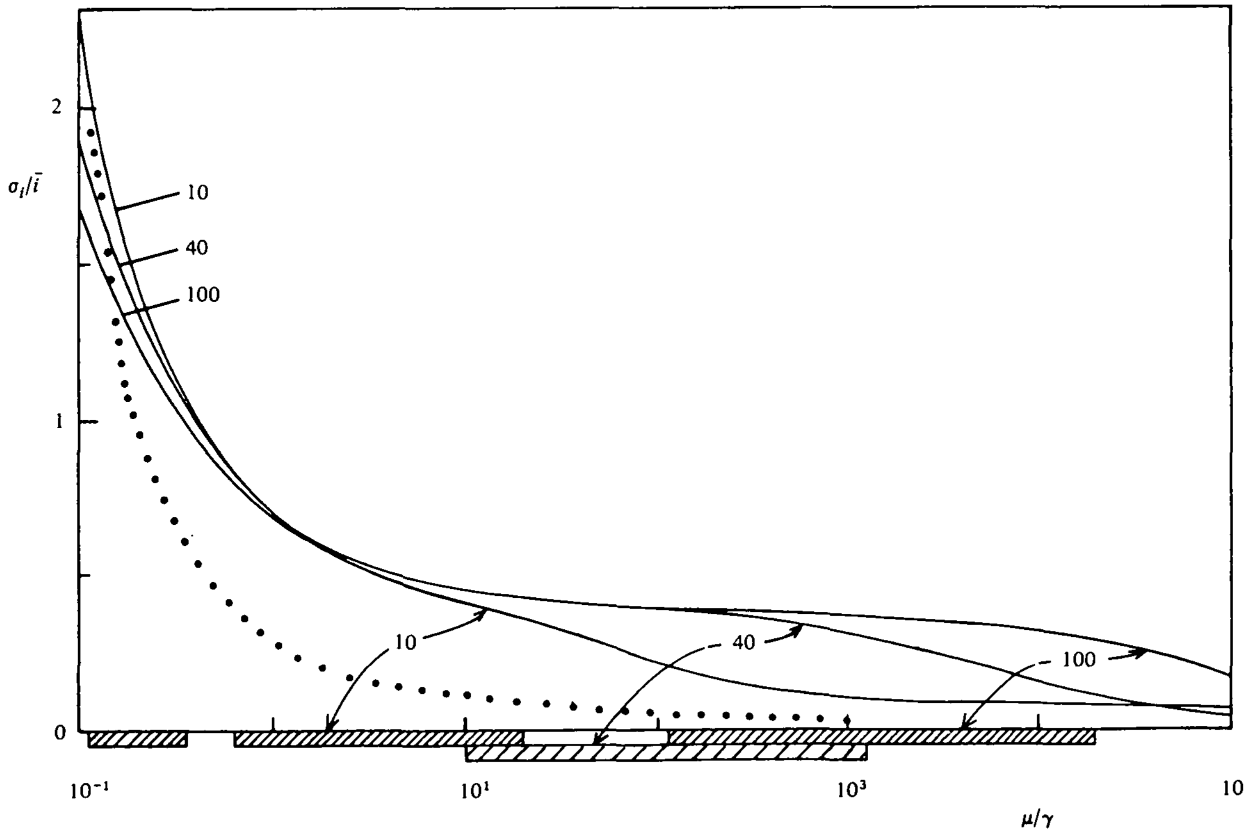

Fig. 3. Coefficient of variation *vs* μ/γ under both selection schemes. Underlined are the parameter domains for which mean copy numbers range from 0·1 Ω to 0·5 Ω (see text).

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

expression for the distribution f of copy numbers can be derived (Ewens (1979), p. 125). Introducing $z \equiv Nx/\alpha$ with $\alpha \equiv (\mu/\gamma)^{\frac{1}{4}}$, we obtain

$$f(z) \sim z(z^4+1)^{-\frac{11}{6}} \left\{\frac{z^2+\sqrt{2}z+1}{z^2-\sqrt{2}z+1}\right\}^{\frac{3\alpha}{4\sqrt{2}}} \times \exp\left\{\frac{3\alpha}{2\sqrt{2}} \arctan\left(\frac{\sqrt{2}z}{1-z^2}\right)\right\}, \qquad (9)$$

where $\frac{3}{4}\alpha^{-1} \leqslant z \leqslant N\alpha^{-1}$. For population size sufficiently large we may replace the lower boundary of the interval (corresponding to $x = 3/4N$) by 0. Then we obtain the mean and variance in copy numbers as

$$\bar{\imath} \approx \frac{\alpha}{Ns}\int_0^{N\alpha^{-1}} zf(z)\,dz, \qquad (10a)$$

$$\sigma_i^2 \approx \left(\frac{\alpha}{Ns}\right)^2 \int_0^{N\alpha^{-1}} (z-\bar{z})^2 f(z)\,dz. \qquad (10b)$$

The integrals in equations (10) cannot be evaluated explicitly. None the less, insight into the behaviour of the mean and variance can be obtained from these equations in a rather general way. We have to distinguish two cases, (i) $N\alpha^{-1} > 1$, and (ii) $N\alpha^{-1} < 1$. In the parameter range $N\alpha^{-1} > 1$, the copy number distribution f is dominated by the function

$$\left\{\frac{z^2+\sqrt{2}z+1}{z^2-\sqrt{2}z+1}\right\}^{\frac{3}{4\sqrt{2}}}, \qquad (11)$$

especially when $\alpha \gg 1$. (11) has the form of a resonance curve, with its resonance point (maximum) at $z = 1$. It follows that the integrals in (10) are only weakly dependent on both the upper bound of the integrals, $N\alpha^{-1}$, and also on α itself, when α is large enough. Numerical integration indeed shows that the integral in equation (10*a*) assumes values between 0·95 and 0·98, when $\alpha \geqslant 10$. Thus, unless population size is very small we obtain the following approximate formula for the mean

$$\bar{\imath} \approx \frac{1}{Ns}\left(\frac{\mu}{\gamma}\right)^{\frac{1}{4}}. \qquad (12)$$

On the other hand, when α is sufficiently large that $N\alpha^{-1}$ becomes smaller than 1, mean copy number reaches its asymptotic value $\bar{\imath} = \Omega$.

The $1/N-$ behaviour of $\bar{\imath}$ (and, similarly, σ_i) makes the model unlikely to apply under additive selection for larger populations, i.e. most natural populations. When populations sizes assume values around 10^4, unrealistically high values of μ/γ are required to obtain considerable amounts of HRDNAs and high interpopulational variances, as observed. Although the rate of recombination in heterochromatin is not known and probably very low, it becomes clear from independent calculations on sequence divergence in satellite DNA clusters that the recombination rate cannot be much smaller than the mutation rate for these DNAs (see Discussion).

4. Copy number distribution under truncation selection

Under truncation selection, the probability of fixation, r_{ij}, is given by

$$r_{ij} = \begin{cases} \dfrac{1}{2N}, & j \leqslant \Omega \\ 0, & j > \Omega, \end{cases} \qquad (13)$$

and the Markov chain is now defined by equations (4) and (13). Following the lines of the previous section and employing a diffusion approximation of the Markov chain, we have to evaluate the sums on the right-hand sides of equations (7). With (13) instead of (6), mean and variance in the rate of change in $X(t)$ per generation are obtained as

$$a(x) = \begin{cases} \frac{1}{2}\mu(x+2s), & x < \frac{1}{2} \\ \frac{1}{2}\mu\dfrac{(1-x)^2}{x} - \frac{1}{12}\gamma(2-x)\left(\dfrac{2x-1}{x}\right)^2, & x \geqslant \frac{1}{2} \end{cases} \qquad (14a)$$

and

$$b(x) = \begin{cases} \frac{1}{6}\mu(x+2s)(2x+3s) + \frac{1}{12}\gamma x(x+2s), & x < \frac{1}{2} \\ \frac{1}{3}\mu\dfrac{(1-x)^3}{x} + \frac{1}{12}\gamma x^2 - \\ \frac{1}{2}\gamma(2x-1)^2\left(\dfrac{1}{4}\left(\dfrac{2x-1}{x}\right)^2 - \dfrac{2}{3}\dfrac{2x-1}{x} + \frac{1}{2}\right), & x \geqslant \frac{1}{2} \end{cases} \qquad (14b)$$

Although these formulae are too complex to derive the distribution of copy numbers analytically, it becomes immediately evident that this distribution does not depend on population size. The reason is that, under truncation selection, the process is neutral until high copy numbers are reached, and the fixation of variant chromosomes is therefore inversely proportional to population size. On the other hand, the rate at which array size is altered by recombination or amplification is proportional to N, so that population size cancels out.

The result of the numerical integration is displayed in Figs. 2, 3 (dotted lines). Three aspects of this are important.

First, the accumulation of HRDNA occurs at values of μ/γ which are much smaller than in the case of additive selection, even for very small population sizes (Fig. 2).

Secondly, under truncation selection mean copy number changes relatively abruptly with μ/γ. Within an order of magnitude (that is, for $0{\cdot}01 \leqslant \mu/\gamma \leqslant 0{\cdot}1$), the mean increases by more than a factor 100 from $\bar{\imath} = 0{\cdot}0007\ \Omega$ to $0{\cdot}075\ \Omega$, and by roughly a factor 10

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

from $\bar{\imath} = 0{\cdot}075\ \Omega$ to $0{\cdot}71\ \Omega$ in the interval $0{\cdot}1 \leqslant \mu/\gamma \leqslant 1{\cdot}0$.

Thirdly, since satellite DNAs are neutral under truncation selection if copy numbers are not too large, variation between populations is markedly higher than in the additive selection scheme (Fig. 3). For instance, under additive selection $\sigma_i/\bar{\imath}$ never exceeds 1, when repetitive DNAs are present in considerable amounts, say $\bar{\imath} \geqslant 0{\cdot}1\ \Omega$, whereas under truncation selection values around 2 are reached for $\bar{\imath} \approx 0{\cdot}1\ \Omega$. (Note that the comparisons are made here for array sizes varying in the (underlined) range $0{\cdot}1\ \Omega \leqslant \bar{\imath} \leqslant 0{\cdot}5\ \Omega$; see Fig. 3.) In accordance with the abrupt change of the mean, the coefficient of variation decreases rapidly with increasing μ/γ. But the model predicts extensive variation even when copy number is fairly high. Note that for $\bar{\imath} \approx \frac{1}{2}\Omega$, σ_i is as large as $0{\cdot}6\ \bar{\imath}$.

5. Discussion

(i) *Evaluation of the models*

We first examine the conditions under which our models predict accumulation of HRDNAs. We have analysed the models of copy number control in the parameter range $N\gamma \ll 1$ and $N\mu \ll 1$. Under these conditions, the population is usually fixed for a single chromosome type, and the variance in copy number between individuals in a given population is therefore very small. This assumption is based on the notion that quantitative variation of HRDNAs within species is very low relative to interspecies differences (Dover & Flavell, 1982). However, the data on intraspecific variation are scarce (see below). I therefore conducted also some computer simulations for intermediate parameter ranges and found that the intrapopulational variance can be as high as the interpopulational one (e.g. for additive selection, population size $2N = 10$ and $N\gamma = N\mu = 1$, we have $\bar{\imath} \approx 0{\cdot}2\ \Omega$, $\sigma_i \approx 0{\cdot}08\ \Omega$). In the following, however, we focus on the case $N\gamma \ll 1$, $N\mu \ll 1$.

Under these conditions, we found for additive selection and realistic population sizes that $\bar{\imath} \approx 1/Ns\ (\mu/\gamma)^{\frac{1}{4}}$. The $1/N$-behaviour of the mean rules the additive model out, unless population size is very small. This is because of the difficulty of generating high expected copy numbers with additive selection. For instance, assume a population of size $2N = 2 \times 10^4$ and an amplification rate $\mu = 10^{-4}$ (see below), then the recombination rate must be very low to obtain HRDNAs. For $\bar{\imath} > 0{\cdot}01\ \Omega$, we find $\gamma < 10^{-12}$, and for $\bar{\imath} > 0{\cdot}1\ \Omega$, γ must be even smaller than 10^{-16}. At present, experimental evidence does not rule out such low recombination rates. Direct genetic measurements of recombination rates in heterochromatin are of course not possible, and the values of recombination frequencies measured between the markers which are closest to the centric heterochromatin provide only an upper estimate of the recombination rate within an array of HRDNA sequences. (A value of 10^{-3} is found for meiotic recombination in *Drosophila* X chromosomes (Carpenter & Baker, 1982).) However, there is evidence based on theoretical grounds that recombination rates in satellite DNA clusters must not be too low. Satellites exist as periodic structures in the genome. Assuming that unequal crossing over is acting on an array, as proposed above, it can be shown that periodicity can be created and maintained only when recombination rate is sufficiently large. I found by computer simulations that the recombination rate per unit, $\hat{\gamma}$, cannot be much lower than about $\frac{1}{100}\hat{u}$ ($\hat{u}$, mutation rate per unit). Otherwise, repetitive structures such as satellite DNAs would not exist (to be published).

Since mutation rates are relatively well known and there is also some evidence on the magnitude of the rates of gene amplification, we may now ask if our alternative selection scheme leads to reasonable results. Under truncation selection we found considerable amounts of HRDNAs, say $\bar{\imath} \geqslant 0{\cdot}1\ \Omega$, when $\mu/\gamma > 0{\cdot}1$ (see Fig. 2). To relate γ measuring the recombination frequency per array) to $\hat{\gamma}$, we put $\hat{\gamma} = \gamma/i$ for an array of size i. Thus, $i \geqslant 0{\cdot}1\ \Omega$ when $\hat{\gamma} \leqslant 10\ \mu/i$. Together with the results on periodicity, as mentioned above, we obtain the following condition to be satisfied by recombination rates for HRDNAs likely to accumulate (that is, to exceed a value of mean copy number of, say $\bar{\imath} = 0{\cdot}1\ \Omega$):

$$\tfrac{1}{100}\hat{u} \lesssim \hat{\gamma} \lesssim 10\ \mu/\bar{\imath}. \tag{15}$$

To test whether this condition is met in nature we consider two examples: (i) For satellites of a short repeat length ($\approx$ 10 bp), $\bar{\imath} = 0{\cdot}1\ \Omega$ may correspond to roughly 10^6 copies per chromosome. The mutation rate per unit may be around 10^{-8} per generation. For the amplification rate μ we may assume values around 10^{-4}. Stable duplications of a given genomic region occur at a frequency of 10^{-4} per cell generation at several loci in cultured mammalian cells kept under selection conditions (Schimke, 1984). A duplication rate of 10^{-5} was found for the *rosy* region of *Drosophila* (Gelbart & Chovnick, 1979). Noting that in our model duplication rate is given by $\frac{1}{2}\mu$, relation (15) reads: $10^{-10} \lesssim \hat{\gamma} \lesssim 2 \times 10^{-9}$, when duplication rate is 10^{-4}, or $10^{-10} \lesssim \hat{\gamma} \lesssim 2 \times 10^{-10}$, when duplication rate is 10^{-5}. (ii) For complex satellites (repeat length > 100 bp) with $\bar{\imath} \approx 10^5$, $\hat{u} \approx 10^{-7}$, we find a similar relation: $10^{-9} \lesssim \hat{\gamma} \lesssim 2 \times 10^{-8}$. Thus, in both cases, when amplification rate is around 10^{-4}, there is a 'window' for $\hat{\gamma}$ of approximately an order of magnitude where HRDNAs can accumulate. And, as we see in Fig. 2, this is about enough for satellite DNA to build up.

The width of this 'window' depends on the strength of the force controlling copy number. We assumed an unequal crossing over model in which unequal exchanges involve a large number of repeats, possibly the entire array size. Such exchanges can generate large variances and spread out the copy number distribution rapidly, so allowing selection to be very

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

effective in the elimination of individuals with high copy numbers. Constraints on unequal crossing over (e.g. caused by sterical hindrance in the misalignment of chromosomes) may lower the efficiency of the control mechanism so that the above 'window' becomes wider. Gene conversion promoting similarity of the members of an array may have a similar effect.

These considerations strongly favour the truncation model over the additive selection scheme. The essential difference between the two models is that under truncation selection satellites are essentially neutral, unless high copy numbers are reached. An immediate consequence of neutrality is that the accumulation of satellites does not depend on population size. The truncation model with its sharp transition from neutrality to zero fitness is of course meant only as a crude approximation. A realistic selection scheme might have a smoother (none the less steep) decay at $i \lesssim \Omega$, or even an asymptotic tail. However, it is not relevant to know the exact shape of the fitness curve at high copy numbers. It is only important for this model to work that the fitness function curves downwards to prevent array size from becoming arbitrarily large (Charlesworth *et al.* 1986). According to our results on the additive selection scheme, a negative slope of the fitness function at high copy numbers would lead to a (second order) population size effect such that smaller populations are likely to have larger amounts of HRDNAs in their genomes. However, it is not known at present whether this effect is met in nature.

Meanwhile a purely neutral model of copy number regulation has been proposed by Walsh (1986). In this model the exchange occurring within a chromatid between homologous units of a tandem array is thought to act as a force controlling copy number. A first analysis of the model revealed characteristics resembling those of the truncation selection model.

In the remainder of this section we discuss several applications of the truncation selection model.

(ii) *Chromosomal location of satellite DNA*

We have argued in previous papers (Charlesworth *et al.* 1986; Stephan, 1986*a*) that the accumulation of HRDNA sequences is a consequence of synergistic effects of the forces of amplification, unequal crossing over, drift and natural selection in regions of reduced rates of recombination. The results of this paper may be used to explore this hypothesis more accurately. According to (15), accumulation of HRDNAs is likely to occur in a rather defined range of γ. Lower recombination rates would also lead to the accumulation of repetitive sequences. However, through ongoing mutations those sequences would lose their periodic structure. As independent calculations show (to be published), the control mechanism is then no longer efficient enough to prevent array size from becoming arbitrarily large, i.e. reaching the truncation boundary. (Remember that copy number control is based on recombination which requires a certain degree of homology between the units of an array to work.) In contrast, chromosomal regions with crossing over frequencies one or two orders of magnitude larger than determined in relation (15) should be highly deficient in, if not completely devoid of, repetitive DNA sequences. This overall picture corresponds roughly to the partition of the chromosome into euchromatic and heterochromatic regions, with the heterochromatic, recombinationally inert blocks usually located around the centromere and telomeres (Charlesworth *et al.* 1986).

Apart from this rather crude characterization of the chromosomal location of HRDNAs, the model describes to some extent the distribution of tandem arrays along the entire chromosome too (i.e. regions of interstitial heterochromatin and less intense C banding). The model predicts that accumulation of repetitive sequences does not need extremely low recombination rates and that the copy numbers of the sequences accumulating depend sensitively on the value of μ/γ. (For instance, consider the parameter range of μ/γ around 0·1: In $0{\cdot}05 \leqslant \mu/\gamma \leqslant 0{\cdot}1$, $\bar{\imath}$ changes from 0·011 Ω by a factor 6·5 to 0·075 Ω and, in the adjacent interval $0{\cdot}1 \leqslant \mu/\gamma \leqslant 0{\cdot}2$, by a factor 4 to 0·3 Ω. That means $\bar{\imath}$ increases here more than linearly with increasing μ/γ.) Given the rather fine tuning of regional rates of crossing over along the chromosome arms (Szauter, 1984; Charlesworth, Mori & Charlesworth, 1985), it is therefore no surprise to find a whole hierarchy of tandemly repeated sequences in the genomes, ranging from families with low copy numbers like 'minisatellites' (Jeffreys, Wilson & Thein, 1985) or simple sequence DNAs (i.e. tracts of poly (A), poly (GT), etc. (Tautz & Renz, 1984; Walmsley *et al.* 1984)) to HRDNAs. Although the present model applies principally to all these sequence arrays, we will focus on middle repetitive sequences with copy numbers from 10^3 to 10^4, besides HRDNAs, because we have explored the model only in the parameter range $N\gamma \ll 1$, $N\mu \ll 1$.

Yamamoto (1979) and subsequently John & Miklos (1979) provide a list of species of both plants and animals showing interstitial C band structures with varying degrees of intensity. Although the distribution of repetitive DNA sequences need not parallel the distribution of constitutive heterochromatin as defined by C banding, it is clear that tandemly arrayed families are present at many chromosomal sites, in particular in plant cultivars (Jones & Flavell, 1982*a*, *b*). In rye, the portion of HRDNAs (with copy numbers from 6×10^5 to 6×10^6 (Rankejar, Lafontaine & Pallotta, 1974) corresponding to 8–12% of the genome) is principally located in the major blocks of heterochromatin (here, telomeres). But there is an (even higher) portion of sequence families, each with copy numbers of a factor 20 to 200 less (middle

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

repetitive DNA), which are accommodated by the interstitial heterochromatin. Interstitial heterochromatin is generally more variable in location and size and shows less intense C banding than the major heterochromatic blocks, suggesting that recombination rates are here relatively high.

A similar example is reported from *Caledia captiva*, a grasshopper (Arnold & Shaw, 1985). A family of a 144 bp repeat unit located around the centromere shows copy numbers of 10^5 and 2×10^6, respectively, for two different taxa, whereas a 168 bp family restricted to the interstitial heterochromatin has copy numbers of only $3{\cdot}5 \times 10^4$ or 4×10^3. The same tendency is known from newts. In the centromeric DNA from *Triturus cristatus karelinii* two abundant satellite families have been identified amounting to about $7{\cdot}5 \times 10^7$ copies (33 bp repeat) and $3{\cdot}7 \times 10^6$ (68 bp repeat), respectively. These satellites are also present in the genome from e.g. *T. c. cristatus* but in substantially lower copy numbers. *In situ* hybridization has shown that in the latter both satellite families are not located close to the centromeres but at pericentric sites (Baldwin & Macgregor, 1985).

The interpretation of these examples should be viewed with some *caveats*, since our theoretical results predict high interspecific variation in copy numbers (at a given recombination rate). At present, this makes it impossible to exclude the alternative explanation that the lower amount of repetitive DNA sequences in interstitial heterochromatin is due simply to such variation and not due to an increased rate of recombination in these regions. There is a need for more data on the distribution of tandemly arrayed sequences along the chromosomes.

(iii) *Quantitative variation of satellite DNAs*

Satellite DNA undergoes rapid changes in evolution relative to most, if not all, coding multigene families. This is manifested by the great variation in quantity between closely related species. One notion for satellite DNA is that the average amount of a particular family varies to a much greater extent between species than within a given species (Dover & Flavell, 1982). In the following we discuss the evidence for this notion and compare it with the predictions of our model.

Unfortunately, a direct comparison between intra- and interspecific variation in satellite DNA amounts cannot be made, since there is no species group, for which both quantitites have been measured. In particular, there is a considerable lack of information on intraspecific variability for our model assumptions to be tested. The above analysis has been given in the parameter ranges $N\gamma \ll 1$ and $N\mu \ll 1$. However, it is not clear at present to what extent these conditions are met in nature when population sizes are sufficiently large. For population size greater than 10^4, $N\mu$ may well be around 1 or even larger, so that intraspecies variation in copy numbers should be found, as our simulations indicate. There are a few examples indicating that copy number differences among individuals within a population do occur, but there are not enough data for a quantitative evaluation. Examining a satellite located in the centromeres of the African green monkey chromosomes, deca-satellite, Maresca, Singer & Lee (1984) report that the number of chromosomes with sufficient deca-satellite to give grains in *in situ* hybridization vary from one monkey to another. Moreover, quantitative variations in centric heterochromatin (and possibly also in the satellites accommodated) are known from mice (Forejt, 1973) and man (Kurnit, 1979).

With respect to interspecies data the situation is better, though not satisfying either. We will therefore focus on only one consequence of the model. This predicts a very high interspecific variance in copy numbers. For considerable amounts of HRDNAs, say $0{\cdot}1\,\Omega \leqslant \bar{\imath} \leqslant 0{\cdot}5\,\Omega$, we find under truncation selection values for $\sigma_i/\bar{\imath}$ between 2 and 0·6 (see Fig. 3). An immediate consequence of this would be that the satellites present in the genomes of the members of a particular species group may frequently not be shared by all members of this group. That means some satellites may be lost in a particular species since radiation of the group occurred and others will have been generated *de novo*. We examine this prediction for various examples of major satellite families.

(a) *Drosophila*

The situation in the *virilis* species group of *Drosophila* has been summarized by Throckmorton (1982). In *D. virilis* itself there are three major satellites which are related to one another by single base substitutions (Gall & Atherton, 1974). Satellite I is also found in *D. americana*, *D. texana*, *D. novamexicana* and *D. lumei* but neither of the other two satellites can be demonstrated elsewhere in the phylad. Together, these satellites constitute more than 40% of total DNA in *D. virilis*, whereas less than 25% have been found in the other species. In contrast, species in the *montana* phylad of *D. virilis* have low amounts of satellite DNA (< 10%). Some have apparently none at all, like *D. ezoana*. No single species shares the three major satellites of *D. virilis*. (A possible exception is *D. kanekoi*.)

A similar situation is reported from the *melanogaster* species group. Together, the amounts of the satellites present in this group vary from approximately 4% (*D. mauritiana*) to 9% (*D. melanogaster*) of total nuclear DNA (see Fig. 8 of Barnes, Webb & Dover, 1978). However, the interspecific fluctuations of a particular satellite are much higher. For instance, the most abundant satellite of *D. melanogaster* (4·5% of total DNA) is not found in the other species of this group, except *D. yakuba*. Furthermore, there is no single satellite of all nine present in the group showing a

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

distribution which is uniform across species, and none is shared by all species.

(b) *Rodents*

In kangaroo rats (genus *Dipodomys*), great variation in the amounts of the three (largely centromeric) satellite DNAs (HS-α, HS-β and MS) has been found in several closely related species (Hatch *et al.* 1976). For instance, *D. merriami* has only two components of these, HS-α and MS (each comprising about 15% of nuclear DNA), whereas another member of the *merriami* group, *D. nitradoides*, has only the MS type. Considerable differences in the banding profiles in CsCl and in reassociation rates within closely related species are also reported from other rodent families (e.g. Cricetidae & Muridae; Hennig & Walker, 1970). But the differences are here apparently less dramatic than in kangaroo rats and the average amounts are less than 10% (Singer, 1982).

(c) *Carnivora*

Data on copy number variation are available from the family Felidae. Fanning & O'Brien (1986) examined the distribution of a major centromeric satellite (cloned out from the genome of the domestic cat (*Felis catus*)) in 27 species of this family. The average amount of this satellite makes up 1–2% in the genomes of many cat species but some (e.g. puma, kodkod) contain reduced amounts. Of particular interest is the ocelot group within which the amount of this particular satellite varies by several orders of magnitude. Since it is likely that the ocelot group radiated within the last 1·5–2 million years, these data suggest that satellite sequences can change very rapidly in their amounts and disappear within several million years (or, in this case, 10^5 to 10^6 generations) or less.

(d) *Primates*

Most abundant in the genomes of several Old World monkeys are alphoid satellite DNAs which are clustered principally in the centromeric and telomeric regions of chromosomes (Kurnit & Maio, 1973). The alphoid satellite of the African green monkey appears to have the most simple structural organization. Its basic repeat consists of about 170 bp, whereas the predominant repeat unit of the other members (human, baboon, rhesus and Bonnet monkeys) is dimeric (340 bp). Alphoid satellite DNA represents 24% of the genome of African green monkeys, with $1{\cdot}3 \times 10^5$ copies per chromosome. It is less abundant in rhesus monkey and baboon genomes (8–10%), and in humans its amount is even more reduced (2%) (Musich, Brown & Maio, 1980). Similar differences have been found in the chromosomes of the great apes. The classical human satellite II, which is also present in gorilla and orang-utan cells has not been detected in chimpanzees (Gosden *et al.* 1977).

(e) *Plants*

Especially in higher plants, severalfold variation in genome size does occur (Walbot & Goldberg, 1979). The quantitative variation between species is due mostly to the variation in repeated DNAs, especially satellites. However, the situation is here less clear than in animals, mostly because only a few satellite sequences have been characterized. Data are available from rye (genus *Secale*) (Jones & Flavell, 1982*a*, *b*). In this genus most species can be distinguished from closely related species by at least one major satellite family. For instance, in *S. silvestre* three of the four major *S. cereale* HRDNAs are present in low copy number or absent. In the genome (telomeres) of *S. cereale* these four sequences makes up 8–12%. A similar situation is reported from cucumber (*Cucumis sativus*) where three major satellites represent about 20–30% of the total nuclear DNA. These sequences, however, are not homologous to the main satellite of the closely related melon species *Cucumis melo* (Ganal, Riede & Hemleben, 1986).

6. Conclusions

Turning around the usual assumption that there must be a function for satellite DNA, we started from the hypothesis that satellites are generally functionless in an evolutionary sense, i.e. neutral or deleterious, and studied its consequences under simple selection schemes. The predictions of our model which, apart from selection, includes gene amplification, unequal crossing over and drift are in qualitative agreement with experimental data on (i) intra- and interspecific copy number variation and (ii) the chromosomal location of HRDNAs. Quantitative analysis, however, is somewhat limited, due to a general lack of experimental data. Open questions concern (i) the values of the parameters $N\mu$ and $N\gamma$. For population sizes larger than 10^4, $N\gamma$ (like $N\mu$) could be around 1 or larger, even in the recombinationally inert heterochromatin. That is, for outbred populations our model would predict measurable intraspecific variances which may limit the generality of the notion that the average amount of a satellite DNA family varies to a much greater extent between closely related species than within a species. (ii) According to our results, there should be a hierarchy of tandem arrays along the chromosome arms, ranging from HRDNAs of copy number 10^6 or 10^7, associated with the major heterochromatic blocks around centromeres and/or telomeres to families with low copy numbers, like 'minisatellites' or simple sequence DNAs (10–100 copies). The latter are likely to be located in regions of normal or high recombination rates, so that $N\gamma \ll 1$ is not true, unless population size is very small. That means, in the genomes of most species small families should be (highly) polymorphic in repeat copy number, as has been demonstrated for human 'minisatellites' (Jeffreys *et al.* 1985). The present

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

model allows to study such families but analysis has then to be extended to parameter values $N\gamma \gtrsim 1$. Middle repetitive sequences of 10^3 to 10^4 copies, however, are described by this analysis, unless population size is too large. These sequences are likely to be accommodated by interstitial C banded regions showing presumably increased rates of recombination relative to the major heterochromatin. The data reported here are in support of this but more information on regional differences in both crossing over frequencies and amounts of tandemly repeated DNA sequences is needed.

I want to thank B. Charlesworth, C. H. Langley and J. B. Walsh for many valuable discussions and their comments on the manuscript.

References

Alt, W. F., Kellems, R. D., Bertino, J. R.& Schimke, R. T. (1978). Selective multiplication of dihydrofolate reductase genes in methotrexate-resistant variants of cultured murine cells. *Journal of Biological Chemistry* **253**, 1357–1370.

Arnold, M. L. & Shaw, D. D. (1985). The heterochromatin of grasshoppers from the *Caledia captiva* species complex. II. Cytological organisation of tandemly repeated DNA sequences. *Chromosoma* (*Berl.*) **93**, 183–190.

Baldwin, L. & Macgregor, H. C. (1985). Centromeric satellite DNA in the newt *Triturus cristatus karelinii* and related species: Its distribution and transcription on lamphbrush chromosomes. *Chromosoma* (*Berl.*) **92**, 100–107.

Barnes, S. R., Webb, D. A. & Dover, G. (1978). The distribution of satellite and main-band components in the *melanogaster* species subgroup of *Drosophila*. I. Fractionation of DNA in actinomycin D and distamycin A density gradients. *Chromosoma* (*Berl*). **67**, 341–363.

Bostock, C. J. & Clark, E. M. (1980). Satellite DNA in large marker chromosomes of methotrexate-resistant mouse cells. *Cell* **19**, 709–715.

Carpenter, A. T. C. & Baker, B. S. (1982). On the control of the distribution of meiotic exchange in *Drosophila melanogaster*. *Genetics* **101**, 81–84.

Cavalier-Smith, T. (ed.) (1985). *The Evolution of Genome Size*. Chichester: John Wiley.

Charlesworth, B., Langley, C. H. & Stephan, W. (1986). The evolution of restricted recombination and the accumulation of repeated DNA sequences. *Genetics* **112**, 947–962.

Charlesworth, B., Mori, I. & Charlesworth, D. (1985). Genetic variation in recombination in *Drosophila*. III. Regional effects on crossing over and effects on non-disjunction. *Heredity* **55**, 209–222.

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

Dover, G. A. & Flavell, R. B. (eds.) (1982). *Genome Evolution*. London: Academic Press.

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

Fanning, T. G. & O'Brien, S. J. (1986). Rapid evolution of satellite DNA loci in the family Felidae. *Genetics* **113** (Supplement), 40.

Forejt, J. (1973). Centromeric heterochromatin polymorphism in the house mouse. Evidence from inbred strains and natural populations. *Chromosoma* (*Berl.*) **43**, 187–201.

Fritsch, E. F., Lawn, R. M. & Maniatis, T. (1980). Molecular cloning and characterization of the human β-like globin gene cluster. *Cell* **19**, 959–972.

Gall, J. G. & Atherton, D. D. (1974). Satellite DNA sequences in *Drosophila virilis*. *Journal of Molecular Biology* **85**, 633–664.

Ganal, M., Riede, I. & Hemleben, V. (1986). Organization and sequence analysis of two related satellite DNAs in cucumber (*Cucumis sativus* L.). *Journal of Molecular Evolution* **23**, 23–30.

Gelbart, W. M. & Chovnick, A. (1979). Spontaneous unequal exchange in the *rosy* region of *Drosophila melanogaster*. *Genetics* **92**, 849–859.

Gosden, J. R., Mitchell, A. R., Seuanez, H. N. & Gosden, C. M. (1977). The distribution of sequences complementary to human satellite DNAs I, II and IV in the chromosomes of chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*) and orang utan (*Pongo pygmaeus*). *Chromosoma* (*Berl.*) **63**, 253–271.

Hatch, F. T., Bodner, A. J., Mazrimas, J. A. & Moore, D. H. (1976). Satellite DNA and cytogenetic evolution. DNA quantity, satellite DNA and karyotypic variations in kangaroo rats (genus *Dipodomys*). *Chromosoma* (*Berl.*) **58**, 155–168.

Hennig, W. & Walker, P. M. B. (1970). Variations in the DNA from two rodent families (Cricetidae and Muridae). *Nature* **225**, 915–919.

Hourcade, D., Dressler, D. & Wolfson, J. (1973). The amplification of ribosomal RNA genes involving a rolling circle intermediate. *Proceedings National Academy of Sciences USA* **70**, 2926–2930.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**, 67–73.

John, B. & Miklos, G. L. G. (1979). Functional aspects of satellite DNA and heterochromatin. *International Review of Cytology* **58**, 1–114.

Jones, J. D. G. & Flavell, R. B. (1982*a*). The mapping of highly-repeated DNA families and their relationship to C-bands in chromosomes of *Secale cereale*. *Chromosoma* (*Berl.*) **86**, 595–612.

Jones, J. D. G. & Flavell, R. B. (1982*b*). The structure, amount and chromosomal localisation of defined repeated DNA sequences in species of the genus *Secale*. *Chromosoma* (*Berl.*) **86**, 613–641.

Kurnit, D. M. (1979). Satellite DNA and heterochromatin variants: The case for unequal mitotic crossing over. *Human Genetics* **47**, 169–186.

Kurnit, D. M. & Maio, J. J. (1973). Subnuclear redistribution of DNA species in confluent and growing mammalian cells. *Chromosoma* (*Berl.*) **2**, 23–36.

Maresca, A., Singer, M. F. & Lee, T. N. H. (1984). Continuous reorganization leads to extensive polymorphism in a monkey centromeric satellite. *Journal of Molecular Biology* **179**, 629–649.

Musich, P. R., Brown, F. L. & Maio, J. J. (1980). Highly repetitive component alpha and related alphoid DNAs in man and monkeys. *Chromosoma* (*Berl.*) **80**, 331–348.

Pike, L. M., Carlisle, A., Newell, C., Hong, S.-B. & Musich, P. R. (1986). Sequence and evolution of rhesus monkey alphoid DNA. *Journal of Molecular Evolution* **23**, 127–137.

Rankejar, P. K., Lafontaine, J. G. & Pallotta, D. (1974). Characterization of repetitive DNA in rye (*Secale cereale*). *Chromosoma* (*Berl.*) **48**, 427–440.

Schimke, R. T. (1984). Gene amplification in cultured animal cells. *Cell* **37**, 705–713.

Singer, M. F. (1982). Highly repeated sequences in mamma-

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press

lian genomes. *International Review of Cytology* **76**, 67–112.

Southern, E. M. (1970). Base sequence and evolution of guinea-pig alpha-satellite DNA. *Nature* **227**, 794–798.

Stephan, W. (1986*a*). Recombination and the evolution of satellite DNA. *Genetical Research* **47**, 167–174.

Stephan, W. (1986*b*). Nonlinear phenomena in the evolution of satellite DNA. *Berichte der Bunsengesellschaft für Physikalische Chemie* **90**, 1029–1034.

Szauter, P. (1984). An analysis of regional constraints on exchange in *Drosophila melanogaster* using recombination-defective meiotic mutants. *Genetics* **106**, 45–71.

Tautz, D. & Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* **12**, 4127–4138.

Throckmorton, L. H. (1982). The *virilis* species group. In *The Genetics and Biology of Drosophila*, vol. 3b (ed. M. Ashburner, H. L. Carson and, J. N. Thompson). pp. 227–296 London: Academic Press.

Walbot, V. & Goldberg, R. B. (1979). Plant genome organization and its relationship to classical plant genetics. In *Nucleic Acids of Plants* (eds. T. C. Hall and J. W. Davies). Boca Raton, Florida: CRC Press.

Walmsley, R. W., Chan, C. S. M., Tye, B.-K. & Petes, T. D. (1984). Unusual DNA sequences associated with the ends of yeast chromosomes. *Nature* **310**, 157–160.

Walsh, J. B. (1986). Persistence of tandemly arrayed multigene families. Implications for satellite and simple-sequence DNAs *Genetics* (in press).

Wells, R. D., Büchi, H., Kössel, H., Ohtsuka, E. & Khorana, H. G. (1967). Studies on polynucleotides. LXX. Synthetic deoxyribopolynucleotides as templates for the DNA polymerase of *Escherichia coli*: DNA-like polymers containing repeating tetranucleotide sequences. *Journal of Molecular Biology* **27**, 265–272.

Yamamoto, M. (1979). Cytological studies of heterochromatin function in the *Drosophila melanogaster* male: Autosomal meiotic pairing, *Chromosoma* (*Berl.*) **72**, 293–328.

Appendix

Using (4), (6) and (7), explicit expressions for the mean, $a(x)$, and the variance, $b(x)$, of the rate of change in x between successive generations have been derived for additive selection. The formulae obtained in the three domains of x are as follows.

(i) $0 \leqslant x \leqslant 1/4N$

$$a(x) \approx -\tfrac{1}{6}\gamma Nx(x+2s)+\mu(x+2s)\left[\tfrac{1}{2}-\tfrac{1}{3}N(2x+3s)\right], \tag{A 1}$$

$$b(x) \approx \tfrac{1}{12}\gamma x(x+2s)+\mu(x+2s)\left[\tfrac{1}{6}(2x+3s) - \tfrac{1}{2}N(x+s)(x+2s)\right]. \tag{A 2}$$

(ii) $1/4N \leqslant x \leqslant 3/4N$

Writing $y \equiv Nx$, we obtain

$$a(y) \approx \gamma N^{-1}\left\{-\tfrac{1}{20480}y^{-2}+\tfrac{1}{2560}y^{-1}-\tfrac{1}{160}y-\tfrac{37}{240}y^{2}\right\} + \mu N^{-1}\left\{\tfrac{13}{3840}y^{-1}+\tfrac{27}{80}y-\tfrac{7}{30}y^{2}\right\}, \tag{A 3}$$

$$b(y) \approx \gamma N^{-2}\left\{-\tfrac{23}{409600}y^{-2}+\tfrac{23}{61440}y^{-1}+\tfrac{19}{320}y^{2}+\tfrac{23}{400}y^{3}\right\} + \mu N^{-2}\left\{\tfrac{13}{30720}y^{-1}+\tfrac{9}{40}y^{2}-\tfrac{7}{40}y^{3}\right\}. \tag{A 4}$$

(iii) $3/4N \leqslant x \leqslant 1$

$$a(y) \approx \gamma N^{-1}\left\{-\tfrac{33}{5120}y^{-2}+\tfrac{3}{256}y^{-1}-\tfrac{1}{6}y^{2}\right\} + \mu N^{-1}\left\{\tfrac{3}{20}y^{-1}-e^{-4y}\left(\tfrac{1}{8}y^{-1}+\tfrac{1}{2}+y\right)\right\}, \tag{A 5}$$

$$b(y) \approx \gamma N^{-2}\left\{-\tfrac{33}{2048}y^{-2}+\tfrac{17}{512}y^{-1}+\tfrac{1}{10}y^{3}\right\} + \mu N^{-2}\left\{\tfrac{1}{10}y^{-1}-e^{-4y}\left(\tfrac{3}{32}y^{-1}+\tfrac{3}{8}+\tfrac{3}{4}y+y^{2}\right)\right\}. \tag{A 6}$$

https://doi.org/10.1017/S0016672300023326 Published online by Cambridge University Press