*Genet. Res., Camb.* (1986), **47**, *pp.* 167–174 *With* 1 *text-figure Printed in Great Britain* 

# Recombination and the evolution of satellite DNA

WOLFGANG STEPHAN*

*Institute of Animal Genetics, University of Edinburgh, West Mains Road, Edinburgh EH9 3JN, U.K.*

(*Received 11 September 1985 and in revised form 6 December 1985*)

**Summary**

In eukaryotic chromosomes, large blocks of satellite DNA are associated with regions of reduced meiotic recombination. No function of highly repeated, tandemly arranged DNA sequences has been identified so far at the cellular level, though the structural properties of satellite DNA are relatively well known. In studying the joint action of meiotic recombination, genetic drift and natural selection on the copy number of a family of highly repeated DNA (HRDNA), this paper looks at the structure–function debate for satellite DNA from the standpoint of molecular population genetics. It is shown that (i) HRDNA accumulates most probably in regions of near zero crossing over (heterochromatin), and that (ii), due to random genetic drift the effect of unequal crossover on copy numbers is stronger, the smaller the population size. As a consequence, highly repeated sequences are likely to persist longest (over evolutionary times) in small populations. The results are based on a fairly general class of models of unequal crossing over and natural selection which have been treated both analytically and by computer simulation.

## 1. Introduction

Eukaryotic chromosomes contain nucleotide sequences of lengths from about 10 to several hundred base pairs that are repeated thousands to millions of times per haploid genome. HRDNA is arranged largely in long tandem arrays which are associated with the heterochromatic regions of chromosomes. Satellite DNA is associated with the bulk of constitutive heterochromatin but it does not need to parallel the distribution of heterochromatin (John & Miklos, 1979; Brutlag, 1980). Thus in humans four satellites which have been identified make up about 3% of the genome, whereas the C-banding material amounts to approximately 20%. Similar situations are reported from rye, *Secale cereale* (Peacock *et al.* 1977; Flavell, 1982), and from the Chinese hamster, *Cricetulus griseus* (see John & Miklos, 1979). Heterochromatin and HRDNAs are distributed along the chromosome arms. Though heterochromatic segments are mostly found in centromeres and telomeres, it is now apparent that chromosomes in many organisms also show interstitial C-bands which are highly variable in their locations (reviewed by John & Miklos, 1979).

Whereas very much is known about the structure, variability and location of satellite DNA, the problem of the function of these DNAs has been the subject of major controversies during the past two decades. So far no function of highly repeated, tandemly arranged DNA sequences has been clearly identified. It is largely the association between heterochromatin and satellite DNA that gave rise to speculations about the function of HRDNA. Numerous attempts have been made to resolve the question of function from a more and more detailed analysis of structure and changes in structure. But it is an assumption that important answers will be found exclusively at the cellular level (Miklos, 1982).

It was proposed in an earlier paper (Charlesworth, Langley & Stephan, 1986) that the distribution of HRDNA along the chromosome is a consequence of an evolutionary equilibrium between genetic drift, natural section and mutation pressure (amplification) in regions of restricted recombination and is not a property of HRDNA *per se*. The first step in the development of this hypothesis is to understand why certain regions of the chromosome should have intrinsically low recombination rates. The treatment of this question is the subject of our previous paper (Charlesworth *et al.* 1986). Several observations reviewed there suggest that the suppression of crossing over in regions such as centromeres and telomeres is not a direct physical property of HRDNA, but is a consequence of the long-range effects of centro-

Present address: Physikalische Chemie I, Technische Hochschule Darmstadt, Petersenstr. 20, D-6100 Darmstadt, F.R.G.

meric and telomeric factors. A second important step is to study the association between satellite DNA sequences and regions of restricted crossing over. Thus I will present here mathematical models which may explain why HRDNA accumulates preferentially in chromosomal regions where virtually no meiotic recombination occurs (heterochromatin), but is hardly found in euchromatin.

## 2. The model

It is clear from data on species comparisons (John & Miklos, 1979, and references therein) that changes in heterochromatin and satellite content are commonplace during evolution, and it is widely recognized that the mechanisms for their changes are unequal crossing over and saltatory amplification. Whereas HRDNA is formed by saltatory events, unequal exchange is considered as a secondary randomization process which leads to homogenization of the sequences within an HRDNA array or to variation of cluster size, but is not used as a direct means of amplification (see the definition of unequal exchange, below). Before outlining details of these processes, let me briefly describe the biological background which the model is based on. For simplicity, I assume a sexual haploid species, in which transient diploid zygotes are formed by random mating. Meiosis follows immediately to produce the haploid phase, during which the development of the organisms occurs and the individuals formed from these gametes may be exposed to selection. Thus 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 gametes produced by the sampled and surviving individuals. For exploring the basic question of this paper, the association between HRDNA and regions of restricted recombination, it is not necessary to assume a particular mechanism of sequence amplification. It is only important that HRDNA is somehow generated. The model of recombination and the selection scheme, however, do have to be made explicit (Charlesworth *et al.* 1986).

*Natural selection.* Suppose a given chromosome carries a tandemly repeated DNA sequence, and there is variation in the number of copies of members of this sequence between different representatives of this chromosome. (I consider here only one cluster of HRDNA and neglect the fact that there is often sequence variation within the cluster.) It is assumed that selection acts on the individuals through the copy number of the chromosomes which they carry. In a haploid species, the fitness of the individuals is then solely a function of the copy number, $i$ (Ohta, 1983). As in our previous paper, I shall assume that the fitness, $w_i$, is a decreasing function of $i$ and is zero beyond a certain threshold, $\Omega$. This is because I am not considering multigene families with specific functions such as histone or ribosomal RNA genes which presumably have optimal copy numbers of repeats. On the other hand, an upper limit to copy number seems likely, since cells with large amounts of HRDNA must have substantially altered properties such as long division times (John & Miklos, 1979). A possible form of $w_i$ is given by

$$w_i = 1 - s(i-1), \tag{1}$$

which is mostly used in the following and referred to as the additive selection model.

*Unequal exchange.* Little is known at present about the distribution of equal and unequal meiotic exchanges in HRDNA. However, large changes in the amount of heterochromatin in specific chromosomal regions have been observed in humans from one generation to the next (Craig-Holmes, Moore & Shaw, 1975; Seabright, Gregson & Mould, 1976). This suggests that unequal exchanges can involve a large number of repeats. Various models of unequal exchange have been proposed (Krüger & Vogel, 1975; Perelson & Bell, 1977; Ohta & Kimura, 1981). The present analysis is based on the model of Takahata (1981), which seems most general. It only assumes that the exchange is symmetric, i.e. the probability that an exchange between two chromosomes with copy numbers $j$ and $k$ results in a daughter chromosome with copy number $i$ ($1 \leqslant i < j+k$) is equal to the probability of production of a daughter chromosome with $j+k-i$ copies. It follows from this assumption that the mean copy number is not changed by unequal crossing over from generation to generation. (Accordingly, unequal crossing over itself must not be viewed as an amplification mechanism.)

Let $Q_{ijk}$ denote 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 probability of an exchange occurring is denoted by $\gamma Q_{ijk}$, where $\gamma$ is the rate of exchange per cluster and generation for a certain pair of chromosomes. I choose the following explicit model of $Q_{ijk}$ (Takahata, 1981):

$$Q_{ijk} = c\left\{1 - \left|\frac{2i}{j+k} - 1\right|\right\}, \tag{2a}$$

where the normalization constant is given by

$$c = \begin{cases} \dfrac{2}{j+k}, & j+k \quad \text{even} \\ \dfrac{2(j+k)}{(j+k)^2-1}, & j+k \quad \text{odd}. \end{cases} \tag{2b}$$

The next two sections study the joint action of selection and unequal crossing over, as exemplified by the above models, on the distribution of HRDNA in a finite population of $2N$ haploid individuals. Analytic approximations are possible for asymptotic parameter ranges, i.e. $2N\gamma \ll 1$ and $N\gamma \gg 1$. Given the above selection scheme, it can be shown (Charlesworth *et al.* 1986) that, having initially been

https://doi.org/10.1017/S0016672300023089 Published online by Cambridge University Press

accumulated, HRDNA will ultimately be lost from the population in the absence of (further) amplification or migration. In the following I give some asymptotic formulae for the mean time to loss of HRDNA. The effect of recombination on the rate of loss is of particular interest.

## 3. Asymptotic theory

### (i) *Mean time to loss for small $N\gamma$*

I assume here that population size $2N$ and recombination rate $\gamma$ are sufficiently small that the population is usually fixed for a single gamete type. In the absence of selection, this requires $2N\gamma \ll 1$, so that no new type is produced by unequal crossing over while a given one is on its way to fixation (a process which takes in the neutral case an average of $4N$ generations while the time between successive recombination events amounts to approximately $1/N\gamma$ generations; since, assuming neutrality, only $1/2N$ recombinant gametes get fixed, the present model requires $2N\gamma \ll 1$). A similar approach has been adopted by Walsh (1985) in describing the evolution of multigene families under gene conversion. It should be noted that the following formal treatment is also valid for sister-strand exchange.

It is convenient to study the process first on the time scale of successive fixation events. Since I assumed an upper limit, $\Omega = s^{-1}$, to copy numbers, the system occupies $\Omega$ discrete states $E_i$ ($i = 1, \ldots, \Omega$), numbered after the copy numbers of the individuals, and its dynamics can be described by a finite Markov chain. Let $p_{ij}$ be the probability of transition from state $E_i$, to $E_j$ between times $\tau$ and $\tau+1$ on the scale of fixation events. $p_{ij}$ is related to the unequal crossing-over transition probabilities by

$$\left.\begin{aligned} p_{ij} &= Q_{jii} r_{ij}, \quad i \neq j \\ p_{ii} &= \sum_{j \neq i} Q_{jii}(1-r_{ij}) + Q_{iii} \end{aligned}\right\}, \tag{3}$$

where $r_{ij}$ is the probability of fixation of a variant with $j$ copies in a population originally fixed for $i$ copies. The calculation of the probability of fixation, $r_{ij}$, is difficult unless we assume that a variant chromosome with $j$ copies (say $j < i$) has an overall selective advantage over all other individuals of the population (note that apart from a chromosome with $j$ copies, a variant with $2i-j$ copies is also produced by unequal crossing over between two individuals both carrying $i$ copies). In the additive selection model the overall advantage is given by $s(i-j)$. This can be inserted into the result by Moran (1962, chapter 5) for haploid populations to obtain an approximate expression for the probability of fixation at state $E_j$, coming from $E_i$:

$$r_{ij} \approx \frac{1-e^{-2s(i-j)}}{1-e^{-4Ns(i-j)}}. \tag{4}$$

A simpler expression for $r_{ij}$ can be obtained by assuming that the system is always far away from the boundary $\Omega$. This may be realistic for HRDNA since selection is very weak and thus $\Omega = s^{-1}$ must be very large; hence,

$$r_{ij} \approx \frac{1}{2N} + s(i-j). \tag{5}$$

Given equations (4) or (5), the Markov chain (3) is fully determined. To calculate the expected time until HRDNA is lost from a given population, it is impracticable to use the method of our previous paper by considering the eigenvalues of the matrix $\{p_{ij}\}$. Instead, we resort to a diffusion approximation of the Markov chain model (Ewens, 1979; chapter 4).

Let $Y(\tau)$ indicate the state of the Markov chain at time $\tau$. The expected change in $Y$ between $\tau$ and $\tau+1$ can easily be calculated using equations (2) and (5):

$$\begin{aligned} E\{Y(\tau+1)-Y(\tau) \mid Y(\tau)=i\} &= \sum_{j=1}^{2i} p_{ij}(j-i) \\ &\approx -s \sum_{j=1}^{2i} Q_{jii}(j-i)^2. \end{aligned} \tag{6}$$

According to the above assumption I neglected boundary effects here too, so that $2i < \Omega$ always holds. The term $\sum_{j=1}^{2i} Q_{jii}(j-i)$ vanishes because it is a sum over a product of a symmetric and an asymmetric function. Similarly,

$$E\{(Y(\tau+1)-Y(\tau))^2 \mid Y(\tau)=i\} \approx \frac{1}{2N} \sum_{j=1}^{2i} Q_{jii}(j-i)^2. \tag{7}$$

Since $s$ is assumed to be very small this is approximately equal to the variance in the change of $Y$ between $\tau$ and $\tau+1$. The sums on the right-hand sides of equations (6) and (7) can elementarily be evaluated to be

$$\sum_{j=1}^{2i} Q_{jii}(j-i)^2 = \tfrac{1}{6}(i^2-1). \tag{8}$$

In order to obtain the operator for the diffusion equation, space is rescaled by introducing $X(t) = (Y(t)-1)/\Omega$ and time as $t = (N\tilde{\gamma})^{-1}\tau$. Since in the additive selection model $s = \Omega^{-1}$, the mean $a(x)$ and the variance $b(x)$ in the rate of change in $X(t) = x$ per generation are then obtained as

$$a(x) \approx -\tfrac{1}{6}x(x+2s), \tag{9a}$$

$$b(x) \approx \frac{1}{2N}\tfrac{1}{6}x(x+2s). \tag{9b}$$

As shown in our previous paper, the process starting at a certain copy number will ultimately get absorbed in state $E_1$, where each chromosome carries only a single copy. If the initial copy number of the repeated sequences is $i_o$, the expected time to absorption at copy number 1 is in the diffusion approximation given by

$$\bar{t}(i_o) \approx \frac{6}{N\gamma}\left\{\int_0^{x_0} \frac{1-e^{-4Nx}}{x(x+2s)}dx + e^{\frac{-4Nx}{x(x+2s)}} \int_{x_0}^{1} dx\right\}$$

$$\text{with} \quad x_0 \equiv \frac{(i_0-1)}{\Omega}. \tag{10}$$

https://doi.org/10.1017/S0016672300023089 Published online by Cambridge University Press

If population size is small, such that $N \leqslant \Omega/i_o$ and if $1 \ll i_o \ll \Omega$, as it might be realistic for HRDNA, the integrals in equation (10) can be evaluated approximately to obtain the following simple results:

$$\bar{t}(i_o) \approx \frac{24}{\gamma} \ln \frac{i_o}{2}, \quad i_o \leqslant \frac{1}{4Ns}, \tag{11a}$$

$$\bar{t}(i_o) \approx \frac{24}{\gamma} \ln \frac{1}{8Ns}, \quad \frac{1}{4Ns} \leqslant i_o \leqslant \Omega/N. \tag{11b}$$

At the end of this section let me briefly consider the effect of unequal crossing over on HRDNA in the absence of selection ($s = 0$, neutral case). As outlined above, the mean copy number of sequences is unaffected by unequal exchanges in this case, but their distribution might well be changing until a non-trivial stationary distribution of copies is ultimately approached. None the less, interest centres in this case also on the expected time to absorption at $E_1$, but conditional on this state having been reached. By regarding the diffusion as an approximation to the (modified) Markov chain, the expected absorption time can again be calculated. Since mean copy number does not change per generation, one has $a(x) = 0$, instead of equation (9*a*), whereas $E\{(Y(t+1) - Y(t))^2 \mid Y(t) = i\}$ is again given by equations (7) and (8). It follows that the variance is obtained as

$$b(x) \approx \frac{1}{2N} \tfrac{1}{6} x(x + 2M^{-1}), \tag{12}$$

where $M$ is an arbitrary but large number used to map the process on to the unit interval [0, 1] ($M$ is an auxiliary parameter which must not appear in the end results). It follows from standard diffusion theory (Ewens, 1979; chapters 4, 6) that when the initial copy number of the population is $i_o$, such that $1 \ll i_o$, the conditional expected time to absorption at copy number 1 is given by

$$\bar{t}(i_o) \approx \frac{24}{\gamma} \ln \frac{i_o}{2}. \tag{13}$$

This is the same result as in the case of additive selection, when the initial copy number is small enough for the process to be scarcely affected by selection (see equation (11*a*)). The results will be discussed below, together with the simulations.

### (ii) *Mean time to loss for large $N\gamma$*

In this case population size $2N$ and recombination rate $\gamma$ are supposed to be sufficiently large for the following argument to be applied. Since selection is weak relative to recombination, the mean copy number changes slowly, relative to the rapid establishment of a quasi-equilibrium distribution of copy number, after a recombination event had occurred. (Sampling drift can be neglected, because $N$ is assumed to be large.) The dynamics of the process is thus a function of the current mean, $\bar{\imath}$, only and will be approximated by a diffusion in the single variable $\bar{\imath}$. Since population size is large, the change in the mean can be calculated approximately by considering a population of infinite size. Sampling is incorporated subsequently to obtain the diffusion coefficient.

*Infinite population.* Let $x_i$ be the frequency of individuals of a sexual haploid species with $i$ copies of the family, measured after selection in a given generation. After allowing copy number to be modified by unequal exchange, the new frequencies before selection in the next generation are given by (Takahata, 1981)

$$\bar{x}_i = (1-\gamma)x_i + \gamma \sum_{j,k} Q_{ijk}\, x_j x_k, \quad i = 1, 2, \ldots \tag{14}$$

The frequencies after selection are

$$x_i' = \bar{x}_i\, w_i/\bar{w}, \tag{15a}$$

where

$$\bar{w} \equiv \sum \bar{x}_i w_i. \tag{15b}$$

I now again make the assumption that $\Omega$ is sufficiently large that the system is always far away from the boundary $\Omega$. Given an explicit fitness function of the form $w_i = 1 - s(i-1)$, one is able to develop equation (14) further and arrives at the following equation:

$$x_i' = x_i + \gamma \left\{ \sum_{j,k-1}^{\Omega} Q_{ijk}\, x_j x_k - x_i \right\} + s x_i \{E\{i\} - i\} + O(s\gamma), \quad i = 1, \ldots, \Omega, \tag{16}$$

where $E\{\ldots\}$ denotes some moments of the distribution $\{x_i\}$, e.g.

$$E\{i^n\} \equiv \sum_{i-1}^{\Omega} i^n x_i, \quad n \geqslant 1.$$

In the derivation of equation (16) only the first-order terms in $\gamma$ and $s$ have been kept. The next higher-order term $O(s\gamma)$ plays no role, even when $\gamma \lesssim 1$.

Progress in treating equation (16) can be made by employing a moment expansion of the distribution $\{x_i\}$, a standard procedure used, for instance, to study master equation systems (Van Kampen, 1975). Upon use of the distribution for $Q_{ijk}$ (equation (2)), a straightforward calculation leads to the following equations for the three lowest-order moments (with $\bar{\imath} \equiv E\{i\}$):

$$\Delta\bar{\imath} = \bar{\imath}' - \bar{\imath} \approx -sE\{(i-\bar{\imath})^2\}, \tag{17a}$$

$$\Delta E\{(i-\bar{\imath})^2\} \approx \gamma(\tfrac{1}{6}\bar{\imath}^2 - \tfrac{1}{6} - \tfrac{5}{12}E\{(i-\bar{\imath})^2\}) - sE\{(i-\bar{\imath})^3\}, \tag{17b}$$

$$\Delta E\{(i-\bar{\imath})^3\} \approx \gamma(\tfrac{1}{2}\bar{\imath}E\{(i-\bar{\imath})^2\} - \tfrac{5}{8}E\{(i-\bar{\imath})^3\}) + s(3E\{(i-\bar{\imath})^2\}^2 - E\{(i-\bar{\imath})^4\}). \tag{17c}$$

Clearly, these equations would still contain frequency terms without some further but minor approximations. I approximated in equations (17*b*) and (17*c*)

$$\tfrac{1}{6} - \tfrac{1}{24} \sum_{\substack{j,k-1 \\ j+k \text{ odd}}}^{\Omega} x_j x_k$$

by $\frac{1}{6}$, and similarly,

$$\tfrac{1}{2}\bar{\imath} - \tfrac{1}{16} \sum_{\substack{j,k-1 \\ j+k \text{ odd}}}^{\Omega} (j+k) x_j x_k$$

https://doi.org/10.1017/S0016672300023089 Published online by Cambridge University Press

by $\frac{1}{2}\bar{\imath}$. Some insight into the validity of these approximations can be obtained by considering the following two extreme cases. If copy numbers are equally distributed, the first term is given by $\frac{1}{6}-\frac{1}{48}$ and the second by $\frac{1}{2}\bar{\imath}-\frac{1}{64}$. On the other hand, if the system is in the neighbourhood of the absorption state, both sums will vanish, as they should do.

Since the equation for a given moment always depends on the next higher moment, the moment expansion does not stop, unless one breaks it off at a certain point. This is done after the third moment because the term containing the fourth moment in equation (17*c*) can be neglected in the kind of approximation I pursue (see below). To solve the moment equations I employed an adiabatic approximation procedure (Haken, 1977, chapter 7), noting that unequal crossing over does not affect mean copy number in the absence of selection, and using the argument from the outset of this section. For our case of very weak selection it is thus possible to conclude that the variance and third moment change rapidly on the time scale of the very slowly moving mean, so that, after a short transient phase, one has

$$\left.\begin{aligned}\Delta E\{(i-\bar{\imath})^2\} &\approx 0,\\ \Delta E\{(i-\bar{\imath})^3\} &\approx 0.\end{aligned}\right\} \tag{18}$$

In dealing with the fourth moment, it is realistic to assume that it is of the order of $E\{(i-\bar{\imath})^2\}^2$ or smaller. (This is fulfilled for most of the known distributions.) In the following, I consider only the case $E\{(i-\bar{\imath})^4\} \leqslant 6E\{(i-\bar{\imath})^2\}^2$. (Relaxing this condition a bit does not change the main result (equation (21)) and has only slight influence on the domain of validity of this result.) Under these circumstances, $|3E\{(i-\bar{\imath})^2\}^2 - E\{(i-\bar{\imath})^4\}|$ (see equation (17*c*)) can be estimated by $3E\{(i-\bar{\imath})^2\}^2$. Using this approximation and equation (18), equations (17*b*) and (17*c*) can be combined to give the following equation, which is quadratic in the second moment:

$$\frac{8}{5}\left(\frac{s}{\gamma}\right)^2 E\{(i-\bar{\imath})^2\}^2+\left(\frac{4}{5}\frac{s}{\gamma}\bar{\imath}+\frac{5}{12}\right)E\{(i-\bar{\imath})^2\}-\tfrac{1}{6}(\bar{\imath}^2-1)\approx 0. \tag{19}$$

The first term of equation (19), which is proportional to the square of the second moment and which is basically an upper bound of the fourth moment can, in turn, be neglected if

$$\frac{\frac{16}{15}\left(\frac{s}{\gamma}\right)^2(\bar{\imath}^2-1)}{\left(\frac{4}{5}\frac{s}{\gamma}\bar{\imath}+\frac{5}{12}\right)^2} \ll 1, \tag{20a}$$

or, to be more specific and requiring the left-hand side of (20*a*) to be smaller than 0·1, if

$$\bar{\imath} \leqslant 0{\cdot}17\frac{\gamma}{s}. \tag{20b}$$

Under these conditions one is allowed to approximate system (14) of equations by their three lowest order moments. It follows then from equation (19) that the variance is given by

$$E\{(i-\bar{\imath})^2\} \approx \frac{10(\bar{\imath}^2-1)}{48\frac{s}{\gamma}\bar{\imath}+25}, \tag{21}$$

provided $\bar{\imath}$ is sufficiently small. However, since selection is presumably very weak, say $s \lesssim 10^{-10}-10^{-8}$, and recombination rate is assumed to be rather high, conditions (20) do not imply any greater restraints on the present analysis. Note that the dynamics becomes independent of recombination, if $\bar{\imath}$ is sufficiently small, say $\bar{\imath} \lesssim 0{\cdot}05\ \gamma/s$ (see (21)).

*Finite population.* The deterministic model, presented above, studies the change in copy number under unequal crossing over and selection only. In finite populations sampling has to be taken into account. This occurs in a given generation after selection to form new zygotes. Let $\{x_i\}$ again be the frequency distribution of copy number in a given generation after selection. A possibly different frequency distribution $\{\hat{x}_i\}$ is produced by $N$ male and $N$ female gametes when taken as random samples from $\{x_i\}$. The expected change of the mean, $\hat{E}\{\Delta\bar{\imath}\}$, per generation of the distributions $\{\hat{x}_i'\}$ and $\{\hat{x}_i\}$ can be calculated from standard multinomial formulae (Crow & Kimura, 1970, p. 330):

$$\hat{E}\{\Delta\bar{\imath}\} = \Delta\bar{\imath} \tag{22a}$$

and the variance of the change of the mean is given by

$$\hat{E}\{(\Delta\bar{\imath}-\hat{E}\{\Delta\bar{\imath}\})^2\} = \frac{1}{2N}E\{(i-\bar{\imath}')^2\} \tag{22b}$$

where $\hat{E}\{\ldots\}$ indicates that expectation has to be taken over the distributions $\{\hat{x}_i\}$ and $\{\hat{x}_i'\}$, respectively.

Formulae (22) constitute the relationship between the deterministic model and finite populations. They allow one to employ a diffusion approximation, with the mean copy number as the diffusing variable. The combination of equations (17*a*) and (22*a*) and, similarly, of equations (21) and (22*b*) yields the mean, $a(x)$, and the variance, $b(x)$, in the change in $x$ per generation:

$$a(x) \approx -s\frac{10x(x+2)}{48\frac{s}{\gamma}x+25} \tag{23a}$$

$$b(x) \approx \frac{1}{2N}\frac{10x(x+2)}{48\frac{s}{\gamma}x+25}. \tag{23b}$$

If the initial mean copy number is $\bar{\imath}_0$, the expected time to absorption at copy number 1 is approximately given by:

$$\bar{t}(\bar{\imath}_0) \approx \frac{1}{s}\int_0^{x_0} dx\,\frac{48\frac{s}{\gamma}(x+1)+25}{10x(x+2)}(1-e^{-4Nsx}) + \frac{1}{s}(e^{4Ns\bar{x}_0}-1)\int_{\bar{x}_0}^{z} dx\,\frac{48\frac{s}{\gamma}(x+1)+25}{10x(x+2)}e^{-4Nsx}, \tag{24}$$

https://doi.org/10.1017/S0016672300023089 Published online by Cambridge University Press

where $z$ is the upper boundary determined by conditions (20). For evaluating the integrals, $z$ can be replaced by $\infty$. The integrals can approximately be carried out if $\bar{\imath}_o$ is sufficiently large. The mean time to loss is then given by

$$\bar{t}(\bar{\imath}_o) \approx 10N \ln \frac{\bar{\imath}_o}{2}, \quad \bar{\imath}_o \leqslant \frac{1}{4Ns} \tag{25a}$$

and

$$\bar{t}(\bar{\imath}_o) \approx 10N \ln \frac{1}{8Ns} + \frac{1}{\gamma}\frac{24}{5}(1 + \ln(4Ns\bar{\imath}_o)), \quad \bar{\imath}_o \geqslant \frac{1}{4Ns}. \tag{25b}$$

These results may again be compared with those of the corresponding neutral case (conditional on the absorption state $E_1$ having been reached). In this case the mean copy number does not change, so that $\Delta\bar{\imath} = 0$, whereas equation (21) has to be replaced by

$$E\{(i-\bar{\imath})^2\} \approx \tfrac{2}{5}(\bar{\imath}^2 - 1). \tag{26}$$

The expected time to loss of HRDNA, conditional on $E_1$ having been reached, is then given by

$$\bar{t}(\bar{\imath}_o) \approx 10N \ln \frac{\bar{\imath}_o}{2}, \tag{27}$$

provided the process started with a high copy number $\bar{\imath}_o$.

## 4. The simulations

To obtain the expected time to loss of HRDNA from a given population for intermediate values of the recombination rate $\gamma$ and to examine the validity of the analytic results, a simulation study was performed. A haploid species of population size $2N$ was considered. In each generation, the following processes were allowed to modify copy numbers: sampling and selection of gametes, and recombination among the sampled and survived individuals. Sampling and selection were done simultaneously. Individuals were sampled at random from the population, with replacement. Their survival was determined according to the additive selection scheme of equation (1). This process was repeated until population size became $2N$. Recombination was simulated as follows. Two gametes were chosen randomly from the pool of the sampled and survived individuals and were paired up according to the probability distribution (equation (2)) to generate recombinants. The number of crossovers between the parental chromosomes had a Poisson distribution with mean $N\gamma$.

The simulations were started with an initial copy number $i_o = 50$ (for each gamete) and were run for $5 \times 10^6$ generations, or until absorption at copy number 1 occurred. For populations of sizes 10 and 40, the simulations were repeated 50 times, for sizes 100 and 400 25 times. The upper limit to copy number was $\Omega = 10^4$. The results of the simulations are summarized in Fig. 1. Terminating runs at $5 \times 10^6$ generations had almost no effect on the estimates of the times to loss since, for the given set of parameter values, the boundary has hardly been reached. However, a slight artifact has arisen through the way in which the simulations were realized on the computer. Since each batch job could run for only a limited amount of time, longer runs remained unfinished (and could not be counted) more often than shorter ones. The times to loss are therefore slightly underestimated, in particular, for the larger populations ($2N = 100$ and 400) and small $\gamma$s.

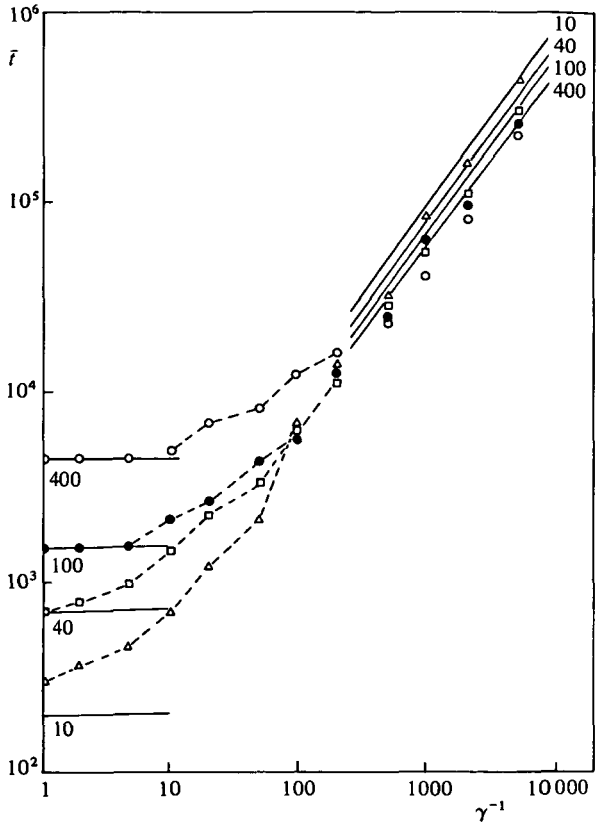


Fig. 1. Mean time to loss of satellite DNA *vs* inverse recombination rate. The simulations were done with the following sets of parameter values: selection coefficient, $s = 0{\cdot}0001$; initial copy number, $i_o = 50$; and population sizes, $2N = 10$ (△), 40 (□), 100 (●) and 400 (○). The theoretical curves are represented by solid lines, with the corresponding population size by the curve.

In Fig. 1 the simulation results are compared with those predicted by the model, as obtained by numerical integration of equations (10) and (24), respectively. The asymptotic results agree reasonably well with the simulations for large $\gamma$s. The agreement is the better the larger the recombination rates and population sizes are. The deviations for very low population sizes, $2N = 10$, are due to the semi-deterministic approach used in the analysis. On the other hand, the adiabatic approximation technique leading to equation (18) is responsible for the deviations of the theoretical curve for smaller $\gamma$s ($\gamma < 1$). In this approximation, $\bar{t}$ is nearly independent of $\gamma$ since, apart from the change in the mean (see equation (17*a*)), the vari-

https://doi.org/10.1017/S0016672300023089 Published online by Cambridge University Press

ance $E\{(i-\bar{i})^2\}$ is hardly affected by unequal crossing over, due to the weakness of selection (see equation (21)). On the whole, the comparison between theoretical and simulation results suggests that for realistic population sizes the mean time to loss of HRDNA from a population is correctly described by the model over a wide range of recombination rates including those of euchromatic chromosomal regions.

More important for the evolution of HRDNA and heterochromatin where virtually no meiotic exchange occurs (Szauter, 1984) is the other asymptotic case $2N\gamma \ll 1$. The model predicted an inverse linear relationship between $\bar{t}$ and recombination rate for small $\gamma$s (see equation (10)). This character of $\bar{t}$ is found in the simulations over an even wider range of $\gamma$ than required by the condition $2N\gamma \ll 1$, a fact which suggests that the model could also be used as an approximation in circumstances in which copy number is known to vary from individual to individual within a given population (for instance, deca-satellite in the African green monkey genome; Maresca, Singer & Lee, 1984). The values obtained from equation (10) are slightly higher than those found by the simulations. To some extent, this is due to the time limitations of the batch jobs, mentioned above.

Two remarkable consequences of the model, which have been obtained analytically, are confirmed by the simulations. First, the expected times to loss of HRDNA are greater in small populations. This result is not intuitively obvious and is in contrast to the large $N\gamma$-case. It follows from this effect, which is a consequence of sampling drift, that the amount of HRDNA should be higher in small populations. Fig. 1 indicates that the effect is not large. But a final answer cannot be given without including the amplification process explicitly. Secondly, the simulations confirmed the possibility of the accumulation of HRDNA in chromosomal regions where recombination is suppressed. In such regions the mean time to loss of HRDNA increases by several orders of magnitude relative to segments of high recombination rates, i.e. $N\gamma \geqslant 1$. The interpretation of this result is as follows. Frequent unequal exchanges tend to spread out the distribution of copy number, so allowing selection to be very effective in the elimination of individuals with high copy numbers. Selection itself, being not supported by unequal crossing over, appears to be rather weak in preventing the spread of tandemly repeated DNA sequences.

## 5. Conclusions

Meiotic exchanges occur non-uniformly along the chromosomes of most higher eukaryotes. There are at least two aspects of this. First, there is virtually no crossing over in heterochromatin. Secondly, there are fewer cross-overs per unit of physical length in regions near the centric, telomeric and apparently also near the interstitial heterochromatin than in the remainder of euchromatin (Szauter, 1984, and references cited therein). Apart from the fact that we do not know whether small amounts of satellite DNA have gone undetected at other regions throughout the lengths of chromosomes, there is a strong association between heterochromatin and satellite DNA. As a contribution to the structure-function debate about HRDNA, the possibility of this association has been investigated in this paper on the basis of quantitative modelling.

To this end, the time to loss of HRDNA from a given population has been calculated, both analytically and by computer simulations. To understand why HRDNAs are likely to persist longest in regions of suppressed recombination, it was not necessary to introduce an explicit model of amplification. The evolutionary dynamics of HRDNA can also be understood in the following way. At each time, when new repeated sequences are regenerated by an amplification event the (unidirectional) process of loss of HRDNA is restarted. Since the rate of producing new copies of DNA sequences is presumably very low (Schimke, 1984), interaction between selection (operating through unequal exchange) and amplification thus leads to a mutation–selection balance, i.e. a steady-state probability distribution of copy numbers in evolutionary times. Sufficiently strong amplification could certainly generate limit cycles or chaotic behaviour, but this possibility seems somewhat remote. Given the results presented above, the steady-state distribution of copy numbers will be such that mean copy number is highest in segments of chromosomes with low values of $\gamma$. In fact, the greatest effect of meiotic recombination on HRDNA is in the neighbourhood of zero crossing over (see Fig. 1).

Thus the present analysis questions the often-used argument that the localization of HRDNAs along chromosomes gave hints to understanding the functional aspects of satellite DNA. Many arguments concerning satellite DNA function are based on the centric or telomeric localization of these DNAs, but it is now apparent that heterochromatin is distributed throughout the entire chromosome arms (see Introduction) and may persist everywhere in the genome where recombination is restricted. Under these circumstances our earlier paper (Charlesworth *et al.* 1986) attempted to answer the question why recombination is suppressed in some chromosomal regions separately and to treat the accumulation of HRDNAs as a pure consequence of such restriction. Two (independent) observations, the non-uniform distribution of meiotic exchanges along the chromosomes and the distribution of satellites, are thus brought together to constitute a causal relationship. We presented evidence that this relationship is essentially established by long-range effects of centromeric and telomeric factors in controlling recombination.

A more complete picture of the accumulation of satellite DNA can be obtained, considering possible

counter-forces of the mechanisms controlling recombination. Recent data on minisatellites (Jeffreys, Wilson & Thein, 1985) show that repeated, non-transcribed DNA sequences are not inherently incapable of recombination. These observations suggest, that, in order to behave as recombinationally inert, satellite sequences must obey certain constraints with respect to their DNA structure. Further evidence pointing in this direction has been reported by Strauss & Varshavsky (1984). They found a non-histone protein that binds to α-satellites from African green monkeys at three specific sites per repeat. Due to their apparent nucleosome-positioning activity, such proteins may underlie the highly regular nucleosome arrangements in regions of repetitive DNA and so stabilize repeated DNA sequences against strand breakage and subsequent recombination events. Accordingly, we advanced the following hypothesis on the evolutionary accumulation of satellite DNA. There is a preferential accumulation of repeated sequences in the neighbourhood of centromeres and telomeres, due to these regions having been selected for restricted rates of crossing over, and the HRDNA sequences which accumulate are those with low rates of exchange.

This research has been carried out at the Universities of Sussex and Edinburgh. I am considerably indebted to Brian Charlesworth and Charles H. Langley for numerous discussions and their comments on this paper, and John Maynard Smith for his continuing help and advice. Moreover, I thank the members of the Department of Genetics at Edinburgh University for their hospitality. My work has been financially supported by a fellowship from the Deutsche Forschungsgemeinschaft.

## References

Brutlag, D. L. (1980). Molecular arrangement and evolution of heterochromatic DNA. *Annual Review of Genetics* **14**, 121–144.

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

Craig-Holmes, A. P., Moore, F. B. & Shaw, M. W. (1975). Polymorphism of human C-band heterochromatin. II. Family studies with suggestive evidence of somatic crossing over. *American Journal of Human Genetics* **27**, 178–189.

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

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

Flavell, R. B. (1982). Sequence amplification, deletion and rearrangement: major sources of variation during species divergence. In *Genome Evolution* (eds. G. A. Dover and R. B. Flavell). London: Academic Press.

Haken, H. (1977). *Synergetics. an Introduction*. Berlin: Springer.

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.

Krüger, J. & Vogel, F. (1975). Population genetics of unequal crossing over. *Journal of Molecular Evolution* **4**, 201–247.

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.

Miklos, G. L. G. (1982). Sequencing and manipulating highly repeated DNA. In *Genome Evolution* (eds. G. A. Dover and R. B. Flavell). London: Academic Press.

Moran, P. A. P. (1962). *The Statistical Processes of Evolutionary Theory*. Oxford: Clarendon Press.

Ohta, T. (1983). Theoretical study on the accumulation of selfish DNA. *Genetical Research* **41**, 1–15.

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

Peacock, W. J., Lohe, A. R., Gerlach, W. L., Dunsmuir, P., Dennis, E. S. & Appels, R. (1977). Fine structure and evolution of DNA in heterochromatin. *Cold Spring Harbor Symposium on Quantitative Biology* **42**, 1121–1135.

Perelson, A. S. & Bell, G. I. (1977). Mathematical models for the evolution of multigene families by unequal crossing over. *Nature* **265**, 304–310.

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

Seabright, M., Gregson, N. & Mould, S. (1976). Trisomy 9 associated with an enlarged segment in a liveborn. *Human Genetics* **34**, 323–325.

Strauss, F. & Varshavsky, A. (1984). A protein binds to a satellite DNA repeat at three specific sites that would be brought into mutual proximity by DNA folding in the nucleosome. *Cell* **37**, 889–901.

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

Takahata, N. (1981). Mathematical study of the distribution of the number of repeated genes per chromosomes. *Genetical Research* **38**, 97–102.

Van Kampen, N. G. (1975). The expansion of the master equation. In *Advances in Chemical Physics* (eds. I. Prigogine and S. A. Rice). New York: Wiley.

Walsh, J. B. (1985). Interaction of selection and biased gene conversion in a multigene family. *Proceedings National Academy Sciences, USA* **82**, 153–157.