Persistence of Tandem Arrays: Implications for Satellite and Simple-Sequence DNAs

James Bruce Walsh

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

Manuscript received May 27, 1986
Revised copy accepted November 6, 1986

ABSTRACT

Recombination processes acting on tandem arrays are suggested here to have probable intrinsic biases, producing an expected net decrease in array size following each event, in contrast to previous models which assume no net change in array size. We examine the implications of this by modeling copy number dynamics in a tandem array under the joint interactions of sister-strand unequal crossing over (rate γ per generation per copy) and intrastrand recombination resulting in deletion (rate ε per generation per copy). Assuming no gene amplification or selection, the expected mean persistence time of an array starting with z excess copies (*i.e.*, array size $z + 1$) is $z(1 + \gamma/\varepsilon)$ recombinational events. Nontrivial equilibrium distributions of array sizes exist when gene amplification or certain forms of selection are considered. We characterize the equilibrium distribution for both a simple model of gene amplification and under the assumption that selection imposes a minimal array size, n. For the latter case, $n + 1/\alpha$ is an upper bound for mean array size under fairly general conditions, where $\alpha(=2\varepsilon/\gamma)$ is the scaled deletion rate. Further, the distribution of excess copies over n is bounded above by a geometric distribution with parameter $\alpha/(1 + \alpha)$. Tandem arrays are unlikely to be greatly expanded by unequal crossing over unless $\alpha \ll 1$, implying that other mechanisms, such as gene amplification, are likely important in the evolution of large arrays. Thus unequal crossing over, by itself, is likely insufficient to account for satellite DNA.

THE variety of gene families existing in tandem arrays is enormous, ranging from repeats of large, well-defined genetic units (such as rRNA and histone genes) to arrays whose individual units consist of as few as one or two bases. Tandemly arrayed sequences undergo both recombination and replication processes producing variation in array size. Mispairing of repeated sequences between different DNA molecules followed by recombination generates duplications and deletions. In an analogous fashion, mispairing between bases *within* a single DNA molecule followed by replication, a process known as replication slippage, also generates variation in array copy number (STREISINGER *et al.* 1966; JONES and KAFATOS 1982; MOORE 1983; SCHMID and SHEN 1985). Continual production of copy number variation results in homogenization of arrays (SMITH 1974, 1976; TARTOF 1974; BLACK and GIBSON 1974; OHTA 1980).

Aside from a few families producing well-defined products (*i.e.*, rRNA, histones), the function, if any, of most tandemly arrayed gene families is unknown. Such families may simply exist due to chance duplications/gene amplifications which are subsequently expanded by unequal crossing over and/or replication slippage to sizes sufficiently large to invite the attention of molecular biologists. Satellite DNAs, which comprise large fractions of many genomes, have been interpreted in this fashion (*e.g.*, MIKLOS 1982, 1985). An important issue is how long such arrays persist in the absence of any selection for their maintenance. If arrays persist for only a short evolutionary time, existing selectively neutral arrays within a genome reflect an equilibrium between processes (such as gene amplification) creating new arrays and recombinational/replicational processes removing them.

Existing models for array copy number dynamics in the absence of selection focus only on unequal crossing over (KRÜGER and VOGEL 1975; PERELSON and BELL 1977) and give the mistaken impression that some tandem arrays can persist for an infinite time in the absence of selection. These results, while mathematically correct, are based on assumptions that are biologically incomplete. Both models assume mean array size does not change from generation to generation. As discussed in detail below, drift invalidates this for KRÜGER and VOGEL's model, and, more important, failure to consider the full range of recombination events in arrays invalidates this assumption for both models. Specifically, while both models assume recombination between different chromatids, neither considers recombination *within* a chromatid. As shown in Figure 1, such events always result in deletions. Coupled with unequal crossovers which produce no expected change in array size, the net effect of both recombinational events is an expected net decrease in array size. Here we develop a simple model that takes

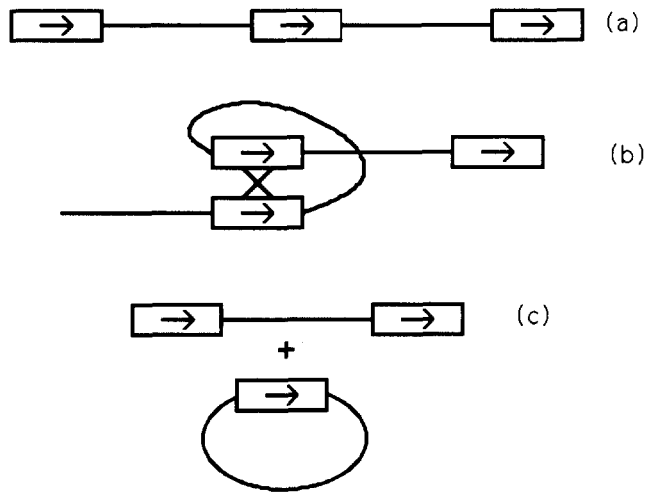

FIGURE 1.—Intrastrand exchange. Tandemly arrayed genes (a) can potentially pair with members of themselves on the same chromatid (b). A recombination event following such a pair produces a deletion in the chromosomal array and a plasmid carrying the deleted copies (c). Most plasmids are likely lost, but a few may be important in gene amplification events (see Figure 8).

this into account, examining expected persistence time for a selectively neutral array and the equilibrium distribution in array length under amplification/recombination balance. Our analysis complements those of CHARLESWORTH, LANGELY and STEPHAN (1986) and STEPHAN (1986) who examined persistence times and equilibrium distributions when selection limits the size of an array. We also examine the distribution of array sizes resulting when selection retains an array by requiring a minimal array size. Finally, replication slippage, which by its very nature is likely to be biased, is also examined.

EXISTING SELECTIVELY NEUTRAL MODELS

CHARLESWORTH, LANGELY and STEPHAN (1986) and STEPHAN (1986) have provided important models by considering the implications of selection limiting array size. To see how selection and/or recombinational biases influence array size, we need first to consider properties of selectively neutral models. Two previous models examined array copy number dynamics in the absence of selection. PERELSON and BELL (1977) assumed strictly intrachromosomal events, all crossovers being between duplicated sister strands of the same chromosome. Here is suffices to consider the dynamics of a single chromosomal lineage, with $X_k(t)$, the probability at time t that an array has exactly k copies, describing the process. KRÜGER and VOGEL (1975) assumed strictly interchromosomal crossing over, with $Y_k(t)$, the population frequency at time t of chromosomes containing arrays of size k, describing the process. Expected persistence times of an array are defined slightly differently for each model—the expected time for a chromosome lineage to reach array size one for the PERELSON-BELL model, versus the expected time for the population to become fixed for a chromosome with array size one of the KRÜGER-VOGEL model. In both models expected persistence time is infinite, but for different reasons. Under intrachromosomal crossovers, with probability one all arrays eventually become fixed at size one (*i.e.*, $X_k(t) \to 0$ for $k > 1$, $X_1(t) \to 1$ as $t \to \infty$), but the expected time to fixation is infinite, as a few arrays reach arbitrarily large size before becoming fixed at size one. This rather strange result follows directly from classic gambler's ruin problems in probability theory. Unequal crossing over produces no net change in array size (as duplications and deletions are equally likely). The array size at any point in time can thus be viewed as the current capital of a gambler playing a fair game against an adversary with infinite capital. In such cases, the gambler always loses (arrays become fixed), but along some sample paths, the gambler's acquired winnings (array size) become arbitarily large before finally being lost (FELLER 1968). Biologically, one would imagine that selection prevents arrays from becoming arbitarily large, although the array size above which selection begins to operate is certainly debatable. CHARLESWORTH, LANGLEY and STEPHAN (1986) and STEPHAN (1986) showed that when an upper limit is imposed by selection, arrays are lost in finite time. Imposing a selective upper limit on an array is equivalent to imposing a reflecting barrier at some finite value on a random walk. Such processes have finite persistence times.

For interchromosomal crossovers, a nontrival equilibrium distribution exists, and hence arrays persist for an infinite amount of time in the population. Specifically $Y_k(\infty) = \pi(1 - \pi)^{k-1}$, with π^{-1} the initial mean array size. The difference in equilibrium behavior between the two models results from differences for the class $k = 1$. Under strictly intrachromosomal crossovers, once a chromosome enters this class it remains there forever. With interchromosomal crossovers, a chromosome with array size one can expand by interacting with a chromosome with array size larger than 1. In the KRÜGER-VOGEL model, $k = 1$ is fixed if and only if $Y_1(t) = 1$. Newly introduced arrays can reach arbitary size before returning to size one, implying that chromosomes with more than one copy are always around (in an infinite population) to interact with chromosomes with array size one. However, the existence of this nontrival equilibrium distribution requires an infinite population. The KRÜGER-VOGEL distribution is neutrally stable—if mean array size changes, the distribution shifts to a new geometric distribution parameterized by $\pi = 1/(\text{new mean})$. In a finite population, sampling each generation continually changes the mean, and eventually results in all chromosomes in the population having $k = 1$. Thus, in a finite population, the equilibrium distribution has

probability one at $k = 1$. PERELSON and BELL's results are not affected by population size, as a single chromosome lineage is being followed.

Both models depend critically on the assumption that recombinational events do not change expected array size. Given the complementary nature of duplication/deletion produced by unequal crossing over, this is a reasonable assumption, *provided* no other events occur. However, if recombination occurs either between arrays on sister-strands or between different homologs, it is also likely to occur between members within a strand, a process we call *intrastrand exchange* (Figure 1), resulting in deletion of members on that array. Intra-strand exchange results in deletions being more frequent than duplications, giving (from classical gamblers ruin theory, see FELLER 1968, p. 349) a finite persistence time for such arrays.

FORMULATION

To examine the expected persistence time of a tandem array and the equilibrium distribution of array size under gene amplification counterbalanced by unequal crossing over, we consider the following simple model. A single chromosome lineage is followed through time, and only sister strand exchanges (which are unequal with probability γ per repeat per generation) and intrastrand deletion events (which occur with probability ε per repeat per generation) are allowed. We assume that any recombinational event alters array size by at most one copy—unequal crossovers produce duplications/deletions of a single copy, likewise intra-strand recombination deletes only a single copy. The latter assumption is almost certainly biologically incorrect, as many repeated units would be expected to loop out during an intra-strand recombination, resulting in the loss of multiple copies. Assuming only a single copy is deleted suggests our persistence times are best regarded as upper limits.

The state space is $\{1, 2, \ldots, \infty\}$, state i corresponding to a chromatid with array size i. In the absence of gene amplification events, $i = 1$ is an absorbing state, as unequal exchange is assumed not to operate between sister strands each containing only a single member. We focus first on persistence times, treating $i = 1$ as absorbing, and then subsequently allow gene duplication (transition from $i = 1$ to $i = 2$) to obtain a stationary distribution.

Array size is modeled as a birth and death process, specified by λ_i and μ_i, the instantaneous rates for transition from state i to state $i + 1$ and state $i - 1$, respectively. For a chromatid in state i, a recombination event altering array size occurs with rate $i(\gamma + \varepsilon)$. At most only one type of recombination event is assumed to occur per cycle (either sister strand or intrastrand). Following a recombinational event, one of the two daughters is chosen at random to continue the lineage. With probability $(1/2)\gamma/(\gamma + \varepsilon)$ the array gains a copy (moves from state i to state $i + 1$), and likewise loses a copy with probability $[(1/2)\gamma + \varepsilon)]/(\gamma + \varepsilon)$. Putting these together gives:

$$\lambda_i = i\gamma/2, \qquad \text{for } i \geq 2, \tag{1a}$$

$$\mu_i = i(\gamma/2)(1 + 2\varepsilon/\gamma), \quad \text{for } i \geq 2. \tag{1b}$$

To investigate the equilibrium between unequal exchange and gene amplification, we assume an array consisting of a single member is duplicated with rate υ per generation, implying:

$$\lambda_1 = \upsilon. \tag{2}$$

PERSISTENCE TIMES

Consider the imbedding of the above birth and death process into a discrete time Markov chain, each time step corresponding to a recombination event (sister-strand unequal exchange or intrastrand deletion). As above, state space is $\{1, 2, \ldots, \infty\}$, with state 1 being absorbing. Let $p_{i,j}$ be the probability of moving from state i to state j, observing for our imbedding that $p_{i,j} > 0$ for only $j = i + 1$ and $i - 1$. From (1):

$$p_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i} = \frac{1}{2(1 + \varepsilon/\gamma)} \tag{3a}$$

$$p_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i} = \frac{1 + 2\varepsilon/\gamma}{2(1 + \varepsilon/\gamma)}. \tag{3b}$$

There has been some discussion as to whether i or i^2 is a more appropriate scaling for recombination events (*e.g.*, PERELSON and BELL 1977; KOCH 1979; NAGYLAKI and PETES 1982). Since persistence times depend only on the ratio of λ_i/μ_i, either assumption gives the same persistence times (in terms of total number of recombination events). More generally, persistence times (measured in number of recombination events) are unchanged if we replace i in λ_i and μ_i by $g(i)$, an arbitrary (strictly positive) function of i.

From (3), $p_{i,i+1}$ and $p_{i,i-1}$ are constants independent of i, reducing our problem to a classic gambler's ruin, the expected duration of which is a standard result (FELLER 1968, p. 348). Let $T(z)$ be the expected number of recombinational events required for fixation starting with z excess copies (*i.e.*, $z + 1$ total members), then:

$$T(z) = z(1 + \gamma/\varepsilon). \tag{4}$$

For fixed γ, $T(z) \to \infty$ as $\varepsilon \to 0$, as expected from previous models (PERELSON and BELL, 1977). Persistence time depends linearly on starting position and on only the ratio of recombination parameters. If intra-sister deletion events are rare compared to sister-strand unequal exchanges (*i.e.*, $\varepsilon/\gamma \ll 1$),

$$T(z) \approx z(\gamma/\varepsilon) \quad \text{for } \varepsilon/\gamma \ll 1. \tag{5}$$

Likewise if intra-sister deletion events are much more common than sister-strand unequal exchanges ($\varepsilon/\gamma \gg 1$),

$$T(z) \approx z \quad \text{for } \varepsilon/\gamma \gg 1. \tag{6}$$

Since (4) measures persistence time in terms of number of recombinational events, not actual number of generations, arrays persist longer (in terms of absolute generations) in regions with reduced rates of recombination (both unequal crossing over and intrastrand deletion). CHARLESWORTH, LANGLEY and STEPHAN (1986) showed that when only unequal crossing over is considered, sequences persist longer in regions of reduced recombination. An instructive example (provided by W. STEPHAN, personal communication) is to consider $g(i) = 1$, that is, rates of recombination are independent of array size i. Under this assumption, the time between each recombination event has mean $1/(\gamma + \varepsilon)$, and hence the actual mean time for loss of an array is z/ε generations. Here the actual persistence time is independent of the rate of unequal crossing over γ, with arrays persisting longest in regions of reduced ε. However, given that both intrastrand deletion and unequal crossing over use homologous recombination, we expect that γ and ε for a given chromosomal region to be highly correlated, so that reduced rates for intrastrand deletions very likely mean reduced rates unequal crossing over and vice versa.

The exact distribution of fixation times is obtainable. Let $u_{z,\tau}$ be the probability of fixation on the τth recombinational event, starting with z excess copies. From FELLER (p. 368, problem 13):

$$u_{z,\tau} = (1 + \varepsilon/\gamma)^{-\tau}(1 + 2\varepsilon/\gamma)^{(\tau+z)/2} \cdot \int_0^1 \cos^{\tau-1}(\pi x) \sin(\pi x) \sin(\pi\tau x)dx. \tag{7}$$

The above analysis applies to a PERELSON-BELL model. The infinite population model of KRÜGER and VOGEL can also be extended to incorporate intrastrand deletions. Let $i\varepsilon$ be the rate at which a chromosome with array size $i > 1$ loses a single copy and let $m(t)$ be the population mean array size at generation t. It can be easily shown that the change in mean array size in generation t, $\Delta m(t) = -\varepsilon[m(t) - Y_1(t)]$, where $Y_1(t)$ is the population frequency (in generation t) of chromosomes carrying arrays of length one. The unique equilibrium is $m = 1$ (*i.e.*, $Y_1(t) = 1$), and the rate of approach to equilibrium is proportional to ε.

AMPLIFICATION-RECOMBINATION EQUILIBRIUM

In the absence of processes capable of expanding arrays of size one, such arrays remain fixed. Gene amplification (a variety of poorly characterized processes which amplify sequences) can create new multicopy arrays (SCHMIKE 1984), allowing for an equilibrium distribution of arrays sizes by balancing the removal of multicopy arrays by recombination with the input of new arrays. We model this by assuming an array of size one undergoes a gene duplication event to become an array of size two with probability υ per generation, *i.e.*, $\lambda_1 = \upsilon$.

We wish to compute π_i = probability of being in state i at equilibrium, with state space as before ($i \geq 1$). Define

$$\rho_i = \frac{\lambda_1\lambda_2 \cdots \lambda_{i-1}}{\mu_2\mu_3 \cdots \mu_i} \quad \text{for } i > 1. \tag{8a}$$

A trivial modification of standard results (*e.g.*, KARLIN and TAYLOR 1977, p. 135) gives:

$$\pi_i = \pi_1\rho_i. \tag{8b}$$

π_1 is determined from the condition that the π_i must sum to one, giving:

$$\pi_1 = 1 \Big/ \left(1 + \sum_{i=2}^{\infty} \rho_i\right). \tag{9}$$

Let α (the scaled deletion rate) $= 2\varepsilon/\gamma$, β (the scaled amplification rate) $= 2\upsilon/\gamma$, and $\theta = 1/(1 + \alpha)$. From (1) and (8):

$$\rho_i = \beta\theta^{i-1}(1/i). \tag{10}$$

Substitution of (10) into (9) and a slight rearrangement gives

$$\sum_{i=2}^{\infty} \rho_i = (\beta/\theta) \sum_{i=2}^{\infty} \theta^i/i. \tag{11}$$

Recalling $\theta^i/i = \int_0^\theta x^{i-1}dx$, interchanging order of integration and summation in (11) following this substitution, and using elementary identities for a geometric series gives:

$$\sum_{i=2}^{\infty} \rho_i = (\beta/\theta) \int_0^\theta x/(1 - x)dx, \tag{12}$$

which yields:

$$\sum_{i=2}^{\infty} \rho_i = \beta[(1 + \alpha)\ln\{(1 + \alpha)/\alpha\} - 1]. \tag{13}$$

From (9) and (13)

$$\pi_1 = [1 + \beta\{(1 + \alpha)\ln\{(1 + \alpha)/\alpha\} - 1\}]^{-1}. \tag{14}$$

From (8) and (10),

$$\pi_i/\pi_{i+1} = [(i + 1)/i](1 + \alpha) \quad \text{for } i > 1, \tag{15}$$

implying (since the right hand side of (15) > 1) that the distribution mode occurs at either array size one or two. From (8) and (10), $\pi_2/\pi_1 = \beta\theta/2$, giving $\pi_2 > \pi_1$ iff

$$\beta > 2(1 + \alpha), \tag{16a}$$

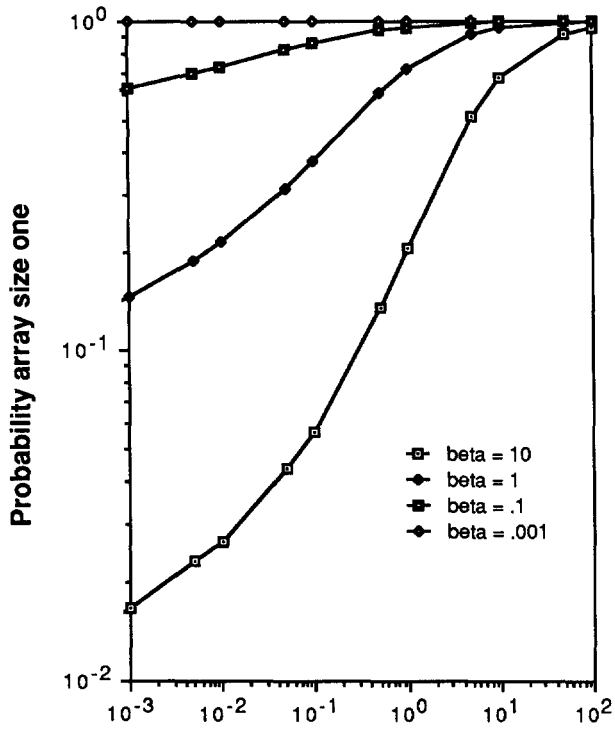

FIGURE 2.—Probability equilibrium array size is length one (π_1) as a function of the scaled deletion rate $\alpha(=2\varepsilon/\gamma)$ and scaled duplication rate $\beta(=2\upsilon/\gamma)$. The equilibrium is between recombination events (unequal crossing over, intra-strand exchange) removing array members and gene duplication creating new arrays by duplicating arrays of size one.

or equivalently,

$$\upsilon > \gamma + 2\varepsilon. \tag{16b}$$

A necessary condition for the distribution mode to be at array size two is that the rate of gene duplication must exceed both the rate of unequal crossing over and intrastrand deletion, else array size one is the commonest class.

Figure 2 plots π_1 for various values of α and β. If intrastrand deletions are rare relative to unequal crossing over, such that $-\ln(\alpha) \gg 1$,

$$\pi_1 \approx 1/[1 - \beta \ln(\alpha)] \tag{17a}$$

$$\approx 1 + \beta \ln(\alpha) \qquad \text{if } \beta|\ln(\alpha)| \ll 1, \tag{17b}$$

$$\approx -[\beta \ln(\alpha)]^{-1} \qquad \text{if } \beta|\ln(\alpha)| \gg 1. \tag{17c}$$

If intrastrand deletion is common relative to unequal crossing over, such that $\alpha \gg 1$,

$$\pi_1 \approx [1 + \beta/(2\alpha)]^{-1} \tag{18a}$$

$$\approx 1 - \beta/(2\alpha) \qquad \text{if } \beta/(2\alpha) \ll 1, \tag{18b}$$

$$\approx 2\alpha/\beta \qquad \text{if } \beta/(2\alpha) \gg 1. \tag{18c}$$

Using (7b), (10) and (14), the equilibrium mean array size, $E(x)$, can be obtained (see APPENDIX):

$$E(x) = \pi_1(1 + \upsilon/\varepsilon) \equiv \pi_1(1 + \beta/\alpha). \tag{19}$$

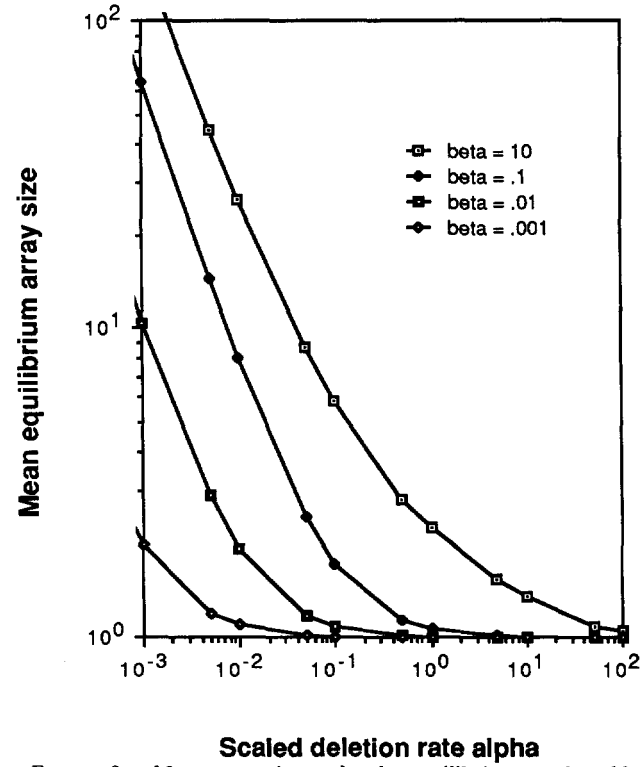

FIGURE 3.—Mean array size under the equilibrium produced by recombinational events balanced by duplication events (as in Figure 2).

Likewise, the equilibrium variance in array size is (see APPENDIX):

$$\mathrm{Var}(x) = \pi_1(\beta/\alpha^2)(1 + \alpha) - E(x)[E(x) - 1] \tag{20}$$

Figures 3 and 4 plot $E(x)$ and $\mathrm{Var}(x)$ for various values of α and β. For extreme values of α and β, asymptotic expressions for $E(x)$ and $\mathrm{Var}(x)$ are available. If $-\ln(\alpha) \gg 1$ (intrastrand deletions very rare relative to unequal crossovers) and $-\beta \ln(\alpha) \gg 1$,

$$E(x) \approx -[\alpha \ln(\alpha)]^{-1} \quad \text{and} \quad \mathrm{Var}(x) \approx -[\alpha^2 \ln(\alpha)]^{-1}. \tag{21}$$

If intrastrand deletions are rare and duplications are not too infrequent ($-\beta \ln(\alpha) \gg 1$), equilibrium mean and variance are essentially independent of β, and both are very large. If $\alpha \gg 1$,

$$E(x) \approx 1 + \beta/(2\alpha), \quad \text{and } \mathrm{Var}(x) \approx \beta/(2\alpha) \quad \text{if } \beta/\alpha \ll 1 \tag{22a}$$

$$E(x) \approx 2(1 - \alpha/\beta), \quad \text{and } \mathrm{Var}(x) \approx 2\alpha/\beta \quad \text{if } \alpha/\beta \ll 1, \tag{22b}$$

implying that when intrastrand deletions are common relative to unequal crossovers, mean array size at equilibrium is small, with a small variance.

EQUILIBRIUM ARRAY SIZE DISTRIBUTION UNDER SELECTION

A second scenario for the persistence of a tandem array is selection which restricts the array to be above

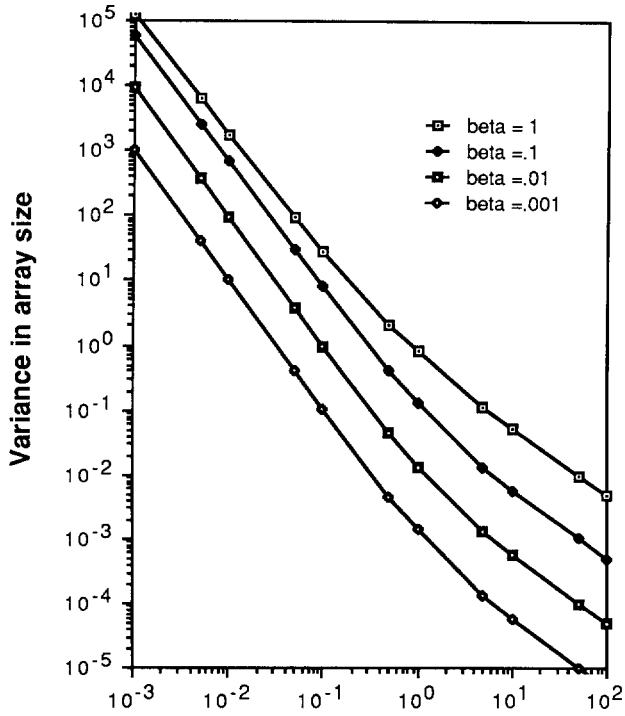

FIGURE 4.—Variance in equilibrium array size under duplication/recombination equilibrium (as in Figures 2 and 3).

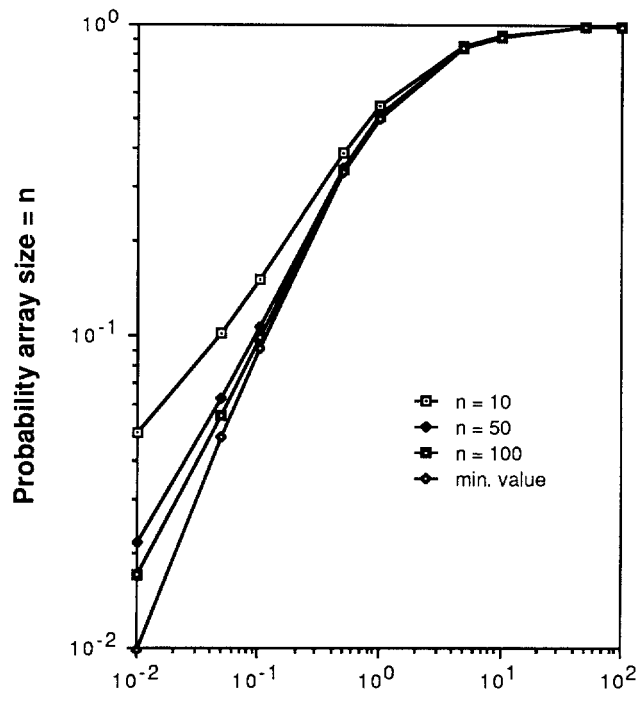

FIGURE 5.—Probability that equilibrium array size equals n (π_n) as a function of scaled deletion rate α and n, where selection constrains arrays to be no smaller than n. The minimum value of $\pi_n(=\alpha/[1-\alpha])$, obtained in the limiting geometric distribution is also plotted.

a critical size. One notion for satellite DNA is that a small fraction of such sequences are essential for some genomic function, with the remaining copies being a selectively neutral surplus generated by genomic forces acting upon the retained array (*e.g.*, LEWIN 1982). In our case, requiring a certain critical array size, $n(>1)$, allows unequal crossing over to expand that array above n.

A simple restriction on (1) allows us to model excess array size. When a chromatid has less than n array members, it is lethal. This is the simplest model of selection, and captures the essence of the problem: if we set a lower limit, how much do recombinational processes expand the array above that limit? Take λ_i and μ_i as in (1), with the restriction that λ_i is zero for $i < n$; μ_i is zero for $i \leq n$. The state space now becomes $\{n, n+1, \ldots \infty\}$, and define π_{n+i} as the equilibrium frequency of an array of length $n + i$ (*i.e.*, i excess copies over the minimum level set by selection). Analysis follows exactly as above. Define:

$$\rho_i = \frac{\lambda_n\lambda_{n+1} \cdots \lambda_{n+i-1}}{\mu_{n+1}\mu_{n+2} \cdots \mu_{n+i}} = n/(n+i)\theta^i, \quad \text{for } i > 1. \tag{23}$$

with $\theta = 1/(1+\alpha)$ as before. Further,

$$\pi_{n+i} = \pi_n\rho_i, \tag{24}$$

with π_n determined by

$$\pi_n = 1\Big/\left(1 + n\sum_{i=1}^{\infty}\theta^i/(n+i)\right). \tag{25}$$

From (23)–(25), mean array size at equilibrium, $E(x)$, is;

$$E(x) = n\pi_n(1 + 1/\alpha), \tag{26}$$

where, as above, $\alpha = 2\epsilon/\gamma$. The series in (25) is bounded above, term by term, by a geometric series. Thus the distribution of excess copies over n is bounded by a geometric distribution with parameter $(1-\theta)$, giving

$$\pi_n \geq 1 - \theta = \alpha/(1+\alpha), \tag{27a}$$

and

$$\text{Prob(array size} \geq n + \zeta;\ \zeta \geq 1) \leq \theta^{\zeta} = (1+\alpha)^{-\zeta}. \tag{27b}$$

Figure 5 plots π_n for various values of n and α as well as plotting the minimal value given by (27a). Assuming the limiting geometric series, the maximal increase in mean array size is bounded above by $1/\alpha$, *i.e.*, $E(x) \leq n = 1/\alpha$, and, likewise, $\text{Var}(x) \leq (1+\alpha)/\alpha^2$. These results, as well as (27), hold under fairly general conditions: we can replace i by $g(i)$ in (1), provided $g(n+i) \geq g(n)$ (see APPENDIX for proof). The increase in mean array size generated by unequal crossing over increases with n toward a maximal increase given by $1/\alpha$. For small n the shift may be dramatic, but becomes trival for n sufficiently large (Figure 6). If

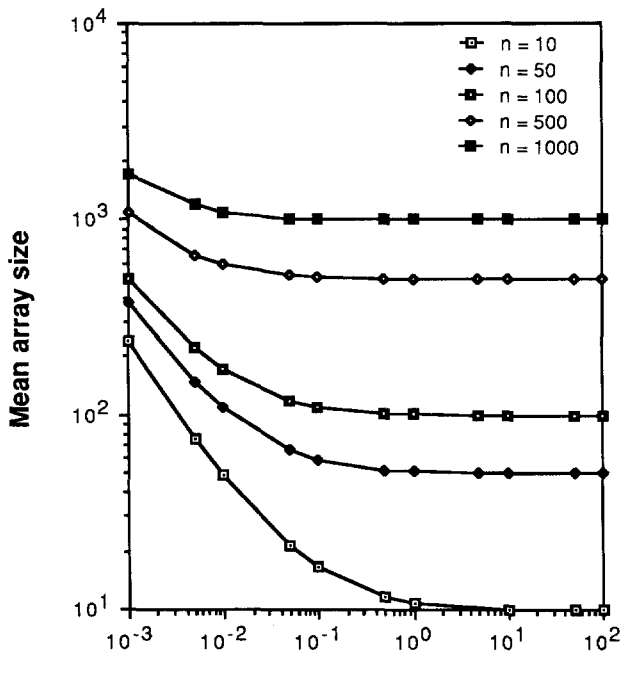

FIGURE 6.—Mean array size under selection for a minimum array size, n (as in Figure 5). An upper limit on mean size is given by $n + 1/\alpha$, obtained as the distribution approaches the limiting geometric.

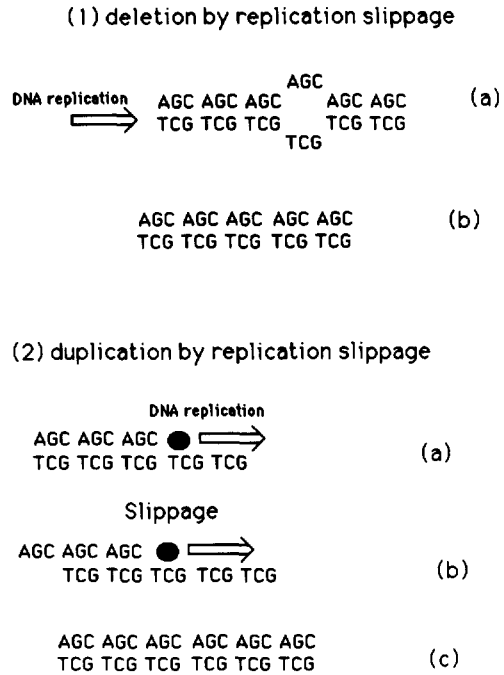

FIGURE 7.—Changes in tandem array size by replication slippage (other schemes are also likely). (1) Deletion: (a) repeat units may mispair within a single double helix. If a replication fork passes through while repeats are mispaired, the polymerase may skin over repeats, resulting in a deletion (b). (2) Duplication: DNA polymerase (indicated by the solid circle) is replicating through a tandem array (a), slippage among recently replicated repeats occurs (b), re-replicating repeats, giving a duplication (c).

$n\alpha \ll 1$, the percent increase in mean array size over n is quite considerable, but becomes trivial when $n\alpha \gg 1$. For example, consider $\alpha = 0.1$, which gives a limiting increase in array size of $1/\alpha = 10$. For $n = 10$, $E(x) \approx 17$; for $n = 100$, $E(x) \approx 109$ and for $n = 1000$, $E(x) \approx 1010$. Suggestions that unequal crossing over, by itself, generates huge satellite DNA array sizes as a by-product of selection for retention of a much smaller array are unlikely unless α is very small.

EXTENSION TO REPLICATION-SLIPPAGE

The individual elements comprising a tandem array are most often thought of as being reasonably large stretches of DNA (say ≥100 bp). However, arrays can exist on a much finer scale, such as simple sequences—stretches of DNA consisting of one or a few tandemly repeated nucleotides (TAUTZ and RENZ, 1984). Such short stretches are common features of eukaryotic genomes, and certain families (such as poly(dC-dA)) are widely distributed across evolutionarily diverse eukaryotes (HAMADA, PETRINO and KAKUNAGA 1982; HAMADA and KAKUNAGA 1982; ROGERS 1983; GEBHARD and ZACHAU 1983; SCHMID and SHEN 1985; YAVACHEV *et al.* 1986; MORRIS, KUSHNER and IVARIE 1986). Variation in array size of such simple sequences occurs (JONES and KAFATOS 1982; MOORE 1983; HAUSWIRTH *et al.* 1984). In particular, many human restriction length fragment polymorphisms appear to be copy number variation in tandem arrays of simple sequence (BELL, SELBY and RUTTER 1982; GOODBOURN *et al.* 1983; KRONTIRIS *et al.* 1985; JEFFREYS, WILSON and THEIN 1985). Several workers (*e.g.*, JONES and KAFATOS 1982; MOORE 1983; SCHMID and SHEN 1985; TAUTZ, TRICK and DOVER 1986) have proposed that variation in simple sequence arrays is generated by replication processes, rather than by (or perhaps in addition to) recombination processes. Just as individual chromatids may mispair allowing for unequal crossing over, likewise sections of individual DNA strands may occasionally mispair. If this occurs when a DNA replication fork passes through, duplications and deletions can result, a process called replication slippage (see Figure 7). STREISINGER *et al.* (1966) first proposed such a process to account for spontaneous frameshift mutants in bacteriophage T4.

Unequal crossing over may be restricted to arrays above a certain absolute length. Below this length, recombination may face topological constraints. Arrays whose repeat units are moderate to large require few copies to be sufficiently long, while arrays whose units are very small can nevertheless be subjected to unequal crossing over if they contain a sufficient number of repeats. Below this threshold (if it indeed exists), replication-slippage can generate variation in array length and both processes may operate simultaneously at intermediate lengths. Both processes are likely important for families whose primary repeat

units are themselves composed of smaller subrepeats, such as satellite DNA wherein the "primary" repeating unit usually consists of a collection of many smaller repeating units, often 5–6 bases in length (SMITH 1976; BRUTLAG 1980; SINGER 1982; MIKLOS and GILL 1982; MIKLOS 1985). Replication-slippage may build up larger units upon which unequal crossing over subsequently acts, and indeed may be an important source for the initiation of new satellite DNAs.

Unlike unequal crossing over, a single replication slippage event does not produce complementary duplications and deletions. Rather, only a duplication *or* a deletion occurs per event. Further, events producing duplications are different from events producing deletions, making it unlikely that the probability of a duplication exactly equals the probability of a deletion. A simple modification extends our model to replication slippage. Let Λ be the per copy per generation rate at which replication slippage occurs. With probability P such an event results in an increase in array size by one, and with probability $1 - P$ results in a decrease in array size by one. Thus:

$$\lambda_i = i\Lambda P, \qquad \text{for } i \geq 2, \tag{28a}$$

$$\mu_i = i\Lambda(1 - P) \quad \text{for } i \geq 2. \tag{28b}$$

If $P > \frac{1}{2}$, the array grows without bound in the absence of any other forces. STREISINGER and OWEN (1985) review data from a number of studies on the *rII* gene of T4, all of which show a bias toward deletions (*i.e.*, $P < \frac{1}{2}$). We can transform (28) into our previous results by setting $\alpha = (1 - 2P)/P$, provided $P < \frac{1}{2}$. Note that we are not assuming any intrastrand deletion, as replication slippage is by itself a biased process. While the *rII* data are the best currently available, it is possible that slippage may be intrinsically biased upward in a number of system (*i.e.*, $P > \frac{1}{2}$).

Arrays of size one are assumed not to experience replication slippage. Mutation events can duplicate nucleotides to initiate an array. Simple point mutants may initiate runs of a single nucleotide, such as poly(dA) (NUSSINOV, 1980), and more complex arrays may be initiated by such processes as integration of template-independent synthesized DNA. Such synthesis often produces defined runs of simple sequences (KORNBERG 1980, pp. 143–150; GREIDER and BLACKBURN 1985), which might account for the widespread conservation of certain simple sequence arrays (*e.g.*, LEVINSON *et al.* 1985). Finally, insertion and subsequent removal of certain transposable elements in maize leaves behind a tandem duplication of 4–6 bases (DORING and STARLINGER 1984; SAEDLER and NEVERS 1985; SCHWARZ-SOMMER *et al.* 1985). Such a mechanism may also initiate runs of simple sequences in other organisms. The distribution of simple sequence arrays within a genome may largely be a reflection of an equilibrium between these various mutation processes and replication slippage. Assuming that arrays of size one are duplicated with probability v per generation, we can use the equilibrium results obtained above by noting that $\beta = v/(\Lambda P)$.

DISCUSSION

Theoretical overview: Both recombination and replication processes potentially act on tandem arrays to generate length variation, and as a consequence, arrays may become lost (*i.e.*, reach size one). It is natural to ask both how long arrays persist and what expected array sizes are under various evolutionary forces. We have examined simple models to gain insight into both of these questions, by assuming either no selection or selection simply to maintain a minimal array size which is otherwise not selectively constrained.

Previous selectively neutral models (KRÜGER and VOGEL 1975; PERELSON and BELL 1977) assumed recombination events generating length variation produce no net change in array size. However, the entire spectrum of recombination events acting on a tandem array is very likely to be intrinsically biased. Although unequal crossing over, by itself, produces no net change in array copy number, intrastrand exchange produces only deletions, resulting in an overall net decrease in copy number when both recombination events are considered. Replication processes acting on arrays, such as replication slippage, are also expected to be intrinsically biased (see REPLICATION SLIPPAGE). Inclusion of the natural biases in these processes changes the behavior of tandem arrays compared with predictions from unbiased models.

First, in the absence of selection or gene amplification, all arrays are lost (reach size one) in finite time. For our simple model, the expected persistence time of an array starting with $z + 1$ copies is $z(1 + 2/\alpha)$ recombinational events (unequal crossover and intrastrand exchange), where $\alpha = 2\varepsilon/\gamma$ is the ratio of intrastrand exchanges to unequal crossovers.

If gene amplification occurs (expansion of an array by processes other than unequal crossovers), a nontrival equilibrium distribution of array sizes exists. Gene amplification creates new variation in array length, offsetting variation removed by recombination. Our model examined the simplest case—gene amplification acts only when the array is of size one, and acts by simply duplicating the array. This is the most conservative model for how amplification acts as we ignore generation of multiple copies and amplification from any array size other than one. If $\alpha \ll 1$, the distribution of array sizes is very broad, and large arrays can occur, while if $\alpha \geq 1$, the variance is very

small, and the distribution of array size is centered at size two if duplications are common relative to recombinational events, else it is centered at array size one. The basic implications of these results for more general models of amplification and recombination is that regardless of the strength of amplification, if $\alpha > 1$ the equilibrium mean array size is unlikely to be larger than the mean size of a newly amplified array. If $\alpha \ll 1$, mean array size can be considerably larger than the mean array size produced strictly by amplification events.

Selection can also maintain arrays. A number of models assume stablizing selection on array size (CROW and KIMURA 1970, pp. 294–296; KRÜGER and VOGEL 1975; KOCH 1979; HOOD, HUANG and HOOD 1980; OHTA 1981; TAKAHATA 1981), wherein arrays with copy number above or below a certain optimum are selected against. This type of model seems reasonable for certain arrays producing known gene products, such as rRNA. Not surprisingly, the equilibrium distribution in array size is centered near the optimum, with variance decreasing as selection increases. Another possibility, considered here, is that selection imposes a lower limit on array size, but does not impose an upper limit. We assume arrays below a critical size, n, are lethal, but that otherwise all arrays are selectively equivalent. This captures the spirit of suggestions that highly repetitive families, such as satellite DNA, result from selection from some (much smaller) minimal array size essential for genome function, with arrays being greatly expanded above this minimal size by unequal crossing over or other genomic forces. For this model, $n + 1/\alpha$ is an upper limit for mean array size. Unequal crossing over coupled with intra-strand exchanges produce *at most* an increase of $1/\alpha$ in mean array size. Provided $n\alpha \gg 1$, the percentage increase in mean array size is small (see Figure 6). Small arrays can be quite sensitive to recombination events (as seen in our amplification model, which becomes the selection model with $n = 2$ as $\beta \rightarrow \infty$), but larger arrays are less sensitive. This places constraints on the notion that most satellite DNA is a byproduct of unequal crossing over greatly expanding some smaller required array, but still leaves open the possibility that gene amplification, rather than unequal crossing over, could generate a vast excess in array size.

Evidence for intrastrand deletions in tandem arrays: There is considerable direct and indirect evidence that unequal crossovers occur in tandem arrays. Direct evidence comes from ribosomal gene clusters in yeast (PETES 1980; SZOSTAK and WU 1980), Drosophila (TARTOF 1974; RAE 1982) and primates (ARNHEIM *et al.* 1980), while various observations suggests unequal crossovers occur in satellite DNA (reviewed in SOUTHERN 1975; SMITH 1976; KURNIT 1979; JOHN and MIKLOS 1979; BRUTLAG 1980). It is difficult to imagine that unequal crossing over could occur without intra-strand exchanges also occurring.

There are numerous examples of intrastrand deletion between duplicated DNA segments, for example the loss of yeast *Ty* elements by recombination between their flanking δ sequences (ROEDER and FINK 1983). There is also indirect evidence for intrastrand deletion acting on tandem arrays. Several studies show a net loss in tandem arrays over time, as is predicted if intrastrand deletions occur in addition to unequal crossovers. Some families of satellite sequences cloned on plasmids show a strong net loss during passage through *Escherichia coli* (*e.g.*, BRUTLAG *et al.* 1977; LOHE and BRUTLAG 1986), which has plagued attempts to clone certain satellite families. This might be explained by unequal crossing over and subsequent selection for plasmids of smaller size. However, the finding of a few plasmids expanding the amount of cloned satellite sequences complicates this interpretation. Human fibroblast cells show a strong net decrease in the amount of human alphoid satellite during serial passage (SHMOOKLER, REIS and GOLDSTEIN 1980), losing roughly one fourth of this satellite before the cultures senescence. Polymona virus often integrates as multiple copies in the form of a tandem array, and can be excised from this array by homologous recombination (LANIA, BOAST and FRIED 1982). Chromosome diminution, the loss of repetitive sequences from the soma of certain animals (reviewed in AMMERMANN 1985), is a striking natural example of net loss through time. Significantly, diminution in Ascaria involves only the loss of satellite sequences (ROTH and MORITZ 1981). Finally, the most direct evidence is the finding of circular plasmids consisting entirely of tandemly repeated sequences. Intrastrand exchanges generate circular molecules containing the deleted copies (Figure 1), while unequal crossing over does not. Certain cloned human HeLa small polydispersed circular DNAs are composed entirely of human alphoid satellite repeats (JONES and POTTER 1985). Yeast rDNA has been isolated on circular plasmids (MEYERINK, KLOOTWIKJ and PLANTA 1979). Some caution is required on the interpretation of circular plasmids in that they can be generated by gene amplification models based on overreplication and subsequent recombination (SCHIMKE 1984; SCHIMKE *et al.* 1986). Thus, while no study has shown directly that intrastrand deletions occur in tandem arrays, a number of observations suggest both that it is reasonable, and, in fact, likely.

Our models depend critically on the magnitude of α. Unfortunately, only PETES (1980) provides data for estimation of α, but this study suffers from small sample size. PETES used tetrad analysis to examine yeast rDNA copy variation (based on an inserted *LEU2* marker) generated at meiosis. Unequal crossover gives tetrads containing complementary deletions and du-

plications, while intrastrand gives tetrads with only deletions. Ten out of 12 tetrads showing copy number changes are unequal crossover events by the above criteria, while the remaining two each contain two deleted spores. These could be either (1) two intrastrand deletions, (2) an unequal crossing over in a previous mitotic division, or (3) a single intrastrand deletion in a previous mitotic division. Assuming the worst case as far as large values of α are concerned (case 2), no intrastrand deletion events are scored in 12 trials, giving a 95% upper confidence limit for ε/γ of 0.28, or $\alpha \leq 0.56$. To place $\varepsilon/\gamma \leq 0.01$ with 95% confidence, 302 tetrads showing copy number changes must be scored. More generally, to place a 95% limit for $\varepsilon/\gamma \leq \delta$ ($\ll 1$) requires a sample size greater than $3/\delta$.

Implications for satellite DNA: Unequal crossing over has generally been assumed to be the primary evolutionary force acting on satellite sequences—satellites are created, homogenized and ultimately lost entirely through the actions of unequal crossovers. However, our models suggest a modified view, with other processes perhaps being more important. Given intrastrand exchange, the operation of recombinational processes (unequal crossovers and intrastrand exchanges) on an array leads to a net reduction in array size. As a consequence, the expansion of an array to very large size by cumulative effects of numerous unequal crossovers becomes difficult, if not impossible. Further, the life span of an array experiencing only recombinational events can be short on an evolutionary time scale. We suggest that gene amplification plays a major role in satellite evolution. Here, gene amplification is a loose term covering a variety of poorly characterized processes that expand sequences in a single event, and can be either a few large events (i.e. expanding array size by orders of magnitude) or a series of smaller events. Unequal crossing over, by itself, can not offset the net decrease in array size caused by intrastrand exchanges. If intrastrand exchanges occur at moderate rate, it is unlikely that a small array can be expanded up to the size of existing satellite arrays by unequal crossing over without some assistence from gene amplification. We suggest the following evolutionary scenario for most satellite DNAs: an initial array, created perhaps by both replication slippage and unequal crossing over, is subsequently expanded greatly by gene amplification to create a new satellite. Recombinational processes acting on the new satellite tend to remove it unless counterbalanced by recurrent gene amplifications. Selection for retention of satellite sequences modifies the latter suggestion, but no evidence of selection has been forthcoming. Direct experimental manipulations on amounts of satellite sequences have failed to demonstrate any phenotypic effects, other than alterations in recombination rates (JOHN and MIKLOS 1979; BOSTOCK 1980; MIKLOS 1985). Indirect molecular evidence for function has been suggested by the finding of specific proteins that bind to certain satellite sequences (STRAUSS and VARSHAVSKY 1984; LINXWEILIER and HORZ 1985). However, recent evidence is that at least some of these proteins bind to any sufficiently long (>5 bases) region of A·T base pairs (SOLOMON, STRAUSS and VARSHAVSKY 1986), and sequencing data shows that specific structures thought to be involved in protein binding are very poorly conserved between repeats within a species (MIKLOS 1985). Furthermore, since satellite sequences persist longer in regions of low recombination [CHARLESWORTH, LANGELY and STEPHAN 1986; our equation (4)], the "function" of satellite-specific proteins may simply be to reduce recombination.

Satellite sequences often show rapid turnover between even closely related species—shared satellites are lost and new satellites can be created over very short evolutionary times (*e.g.*, GILLESPIE, DONEHOWEVER and STRAYER 1982; MIKLOS 1985), a pattern consistent with biased recombinational processes removing satellites and gene amplification creating new ones. Under this view, the relative homogenization of satellite members can be accounted for simply by recent common ancestory, rather than any correction mechanism. Unequal crossing over can spread a variant through an array, but is likely to greatly reduce (if not remove) the array in the process unless new length variation is generated by gene amplification.

What are the possible mechanisms of gene amplification which might be acting on satellite sequences? First, it is important to note that gene amplification, especially in somatic cells, is not uncommon (SCHIMKE 1984; FOX 1984). Several models (*e.g.*, SCHIMKE 1984; SCHIMKE *et al.* 1986) invoke recombination within certain replication structures. Reports of nonrandom distribution of some satellites in the higher-order domains of chromatin structure (SMALL, NELKIN and VOGELSTEIN 1982) are intriguing, given that replication is thought by some to be intimately involved with such structures (VOGELSTEIN, PARDOLL and COFFEY 1980; WILKINS 1981; DIJKWEL, WENINK and PODDIGHE 1986; JACKSON and COOK 1986; but see MIRKOVITCH, MIRAULT and LAEMMLI 1984; RATTNER and LIN 1985; ZAKIAN 1985 for possible complications). Another candidate mechanism, which is attractive in that it relates intra-strand exchange events to gene amplification, is shown in Figure 8. Here, circular plasmids created by intrastrand exchange integrate (at some low level) into arrays by homologous recombination, either on the same chromosome or possibly on nonhomologous chromosomes. Foreign DNA introduced by calcium phosphate in mouse L cell cultures integrates into repetitive DNA with high pref-

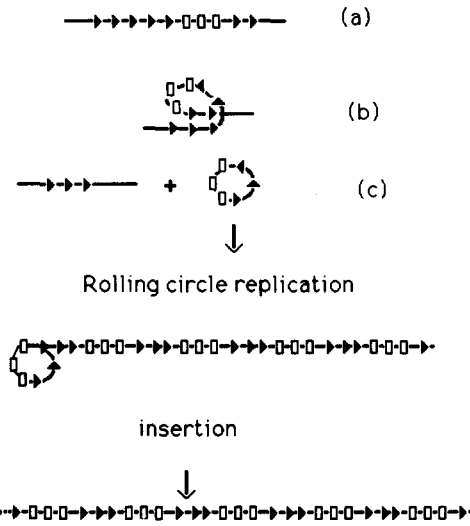

FIGURE 8.—Gene amplification resulting from plasmids generated by intrastrand exchanges. (a) Tandem array undergoes a sister-strand exchange, resulting (b) in deletion of some array members and the formation of a circular plasmid carrying copies of the array (c). Some of these plasmids may undergo rolling circle replication which amplifies a small section of the array before integration into an array. Periodicities can be generated by this process.

erence (KATO, ANDERSON and CAMERINI-OTERO 1986), suggesting that nonhomologous recombination may target sequences into existing satellites even in the absence of sequence homology. The finding of both circular plasmids containing monomers of the basic 170-bp repeat unit of human alphoid satellite sequences (JONES and POTTER 1985), and the subsequent characterization of human alphoid satellites composed of circularly permutated 170-bp monomers (DEVILEE *et al.* 1986) strongly suggests such a process occurs. Further, if the circular plasmids contain replication origins, rolling circle replication can greatly expand a short array sequence on a plasmid, allowing for rapid amplification (HOURCADE, DRESSLER and WOLFSON 1973; FLAVELL 1982). Recent reports of ARS sequences in Drosophilia satellites suggest this is plausible (MARUNOUCHI and HOSOYA 1984). WONG, ABRAHAMSON and NAZAR (1984) report an unusual tandemly arrayed cluster of cytoplasmic 5S RNA pseudogenes in the fungus *Thermomyces lanuginosus*, which they suggest is the result of a rolling circle-like gene amplification. Satellite periodicities can be generated by such an amplification process (Figure 8). Such observed periodicities are often taken as direct evidence for unequal crossing over, may but simply reflect the nature of the amplification mechanism. Amplification can also expand a region which has been acted upon by unequal crossing over (or replication slippage), generating hierarchical periodicities.

In summary, we suggest that unequal crossing over, by itself, is insufficient to account for very large tandem arrays, such as satellite DNAs. Rather, consideration of a wider range of recombinational, replicational, and gene amplification mechanisms is required. The amount of bias in the processes acting on satellite DNAs remains an unresolved issue. The nature of possible biases acting on tandem arrays is likely to be a complex function of array size, for example it might change sign as array size increases. It might be that intrastrand exchanges occur only after the array has exceeded a certain size. Likewise, gene amplification biases may also be a function of array size. Empirical elucidation of these possible biases is an important, but poorly explored, area of genome evolution.

Implications for simple sequence DNA: Replication slippage has been assumed to operate on tandem arrays of simple sequence DNAs (see REPLICATION SLIPPAGE above). Such a process is likely intrinsically biased given that duplications and deletions are produced by different pathways. The r131 site in the *rII* gene of T4 shows a bias in favor of deletions ($\alpha \approx 1$), while another site in the same gene (r117) does not (STREISINGER and OWEN, 1985). If eukaryotic simple sequence arrays show the same bias as r131, our results predict arrays should be lost after a moderate number of replication slippage events. High degrees of length polymorphism occur in some human simple sequence arrays, and estimates of per generation rates of length variants are on the order of 10^{-4} per kilobase of array (JEFFREYS, WILSON and THEIN 1985), suggesting that slippage is reasonably frequent. Studies on the persistence of simple sequence arrays are few. A comparison of human and chimpanzee $\zeta 1$ globin introns finds two shared blocks of simple sequences which have persisted since at least the human-chimp split (WILLARD *et al.* 1985; SAWADA *et al.* 1985). A 12-base sequence is repeated 39 times in both human and chimp, the structure of these repeats strongly suggests numerous events have occurred on these arrays since their divergence. A 5-base repeat occurs 52 times in human and 26 times in chimp and also shows evidence of multiple slippage events. If replication slippage is acting on these arrays, and has the same bias as seen in the *rII* genes (*i.e.*, $\alpha \approx 1$), each array should persist for $\approx 3z$ replication slippage events. Assuming the larger repeat size represents the ancestoral array, $3z = 114$ for the 12-base repeat and $3z = 153$ for the 5-base repeat. While it is difficult to estimate how many replication events have occurred on each array, the $\zeta 1$ data may not meet our expectations, especially the 12-base repeats. One immediate explanation is that slippage bias is different from the *rII* estimate. Assuming the bias estimate is correct, then either recurrent gene amplification or selection may be required to explain the persistence of these arrays.

Although simple sequence DNAs are often common in eukaryotes, they are usually rare to absent in eubacteria and archaebacteria (MORRIS, KUSHNER and

IVARIE 1986). These differences could reflect increased selection for a streamlined genome in prokaryotes (DOOLITTLE 1978), differences in underlying mutational processes or differences in bias. Given that the *rll* data show a very strong negative bias, the absence of most simple sequence DNAs in prokaryotes could be entirely do to bias. If this is true, we expect that biases in replication slippage in eukaryotes will be much less extreme.

There are examples of simple sequences having functional roles. Several simple sequence arrays are known to be involved in gene expression, either by directly binding proteins (BUSBY and REEDER 1983; DYNAN and TJIAN 1985; WEBER and SCHAFFNER 1985; TREISMAN and MANIATIS 1985), altering mRNA stability (SHAW and KAMEN 1986) or having less well defined roles (STUART *et al.* 1984; STRUHL 1985; MELTON *et al.* 1986). Other sequences are involved in aspects of chromatin structure (BONVEN, GOCKE and WESTERGAARD 1985; LILLEY 1986; KOO, WU and CROTHERS 1986) and recombination (SLIGHTOM *et al.* 1985). Just what fraction of simple sequences have function roles remains an open question. The view we favor is that the majority of such sequences are simply byproducts of DNA metabolism. Estimates of the inherent biases (or lack thereof) for eukaryotic simple sequences of different composition are desperately needed, and would provide an important step toward a fuller understanding of the evolutionary forces experienced by such sequences.

I thank MICHAEL TURELLI, WOLFGANG STEPHAN, CHUNG-I WU and JOE FELSENSTEIN for useul comments and criticisms. Special thanks to WOLFGANG STEPHAN for clarifying an important point.

LITERATURE CITED

AMMERMANN, D., 1985 Chromatin diminution and chromosome elimination: mechanisms and adaptive significance. pp. 427–442. In: *The Evolution of Genome Size*, Edited by T. CAVALIER-SMITH. John Wiley & Sons, New York.

ARNHEIM, N., M. KRYSTAL, R. SCHMICKEL, G. WILSON, O. RYDER and E. ZIMMER, 1980 Molecular evidence for genetic exchanges among ribosomal genes on nonhomologous chromosomes in man and apes. Proc. Natl. Acad. Sci. USA **77:** 7323–7327.

BELL, G. I., M. J. SELBY and W. J. RUTTER,1982 The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. Nature **295:** 31–35.

BLACK, J. A. and D. GIBSON, 1974 Neutral evolution and immunoglobulin diversity. Nature **250:** 327–328.

BONVEN, B. J., E. GOCKE and O. WESTERGAARD, 1985 A high affinity topoisomerase I binding sequence is clustered at DNAase I hypersensitive sites in *Tetrahymena* R-chromatin. Cell **41:** 541–551.

BOSTOCK, C., 1980 A function for satellite DNA? Trends Biochem. Sci. **5:** 117–119.

BRUTLAG, D. L., 1980 Molecular arrangement and evolution of heterochromatic DNA. Annu. Rev. Genet. **14:** 121–144.

BRUTLAG, D., M. CARLSON, K. FRY and T. S. HSIEH, 1977 Synthesis of hybrid bacterial plasmids containing highly repeated satellite DNA. Cell **10:** 509–519.

BUSBY, S. J. and R. H. REEDER, 1983 Spacer sequences regulate transcription of ribosomal gene plasmids injected into *Xenopus* embryos. Cell **34:** 989–996.

CHARLESWORTH, B., C. H. LANGELY and W. STEPHAN, 1986 The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics **112:** 947–962.

CROW, J. F. and M. KIMURA, 1970 *An Introduction to Population Genetics Theory.* Harper & Row, New York.

DEVILEE, P., P. SLAGBOOM, C. J. CORNELISSE and P. L. PEARSON, 1986 Sequence heterogeneity within the human alphoid repetitive DNA family. Nucleic Acids Res. **14:** 2059–2072.

DIJKWEL, P. A., P. W. WENINK and J. PODDIGHE, 1986 Permanent attachment of replication origins to the nuclear matrix in BHK-cells. Nucleic Acids Res. **14:** 3241–3249.

DOOLITTLE, W. F., 1978 Genes in pieces: were they ever together? Nature **273:** 581–582.

DORING, H.-P. and P. L. STARLINGER, 1984 Barbara McClintock's controlling elements: now at the DNA level. Cell **71:** 621–630.

DYNAN, W. S. and R. TJIAN, 1985 Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature **316:** 774–778.

FELLER, W., 1968 *An Introduction to Probability Theory and Its Applications*, Vol. 1, Ed. 3. John Wiley & Sons, New York.

FLAVELL, R. B., 1982 Sequence amplification, deletion and rearrangement: major sources of variation during species divergence. pp. 301–323. In: *Genome Evolution*, Edited by G. A. DOVER and R. B. FLAVELL. Academic Press, New York.

FOX, M., 1984 Gene amplification and drug resistance. Nature **307:** 212–213.

GEBHARD, W. and H. G. ZACHAU, 1983 Simple DNA sequences and dispersed repetitive elements in the vicinity of mouse immunoglobulin κ light chain genes. J. Mol. Biol. **170:** 567–573.

GILLESPIE, D., L. DONEHOWEVER and D. STRAYER, 1982 Evolution of primate DNA organization. pp. 113–133. In: *Genome Evolution*, Edited by G. A. DOVER and R. B. FLAVELL. Academic Press, New York.

GOODBOURN, S. E. Y., D. R. HIGGS, J. B. CLEGG and D. J. WEATHERALL, 1983 Molecular basis of length polymorphism in the human ζ-globin gene complex. Proc. Natl. Acad. Sci. USA **80:** 5022–5026.

GREIDER, C. W. and E. H. BLACKBURN, 1985 Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. Cell **43:** 405–413.

HAMADA, H. and T. KAKUNGA,1982 Potential Z-DNA forming sequences are highly dispersed in the human genome. Nature **298:** 396–398.

HAMADA, H., M. G. PETRINO, and T. KAKUNAGA, 1982 A novel repeated element with Z-DNA-forming potential is widely found in evolutionary diverse eukaryotic genomes. Proc. Natl. Acad. Sci. USA **79:** 6464–6469.

HAUSWIRTH, W. W., M. J. VAN DE WALLE, P. J. LAIPIS and P. D. OLIVO, 1984 Heterogeneous mitochondrial DNA D-loop sequences in bovine tissue. Cell **37:** 1001–1007.

HOOD, J. M., H. V. HUANG and L. HOOD, 1980 A computer simulation of evolutionary forces controlling the size of a multigene family. J. Mol. Evol. **15:** 181–196.

HOURCADE, D., D. DRESSLER and J. WOLFSON, 1973 The amplification of ribosomal RNA genes involving a rolling circle intermediate. Proc. Natl. Acad. Sci. USA **70:** 2926–2930.

JACKSON, D. A. and P. R. COOK, 1986 Replication occurs at a nucleoskeleton. EMBO **5:** 1403–1410.

JEFFREYS, A. J., V. WILSON and S. L. THEIN, 1985 Hypervariable "minisatellite" regions in human DNA. Nature **314:** 67–73.

JOHN, B. and G. L. G. MIKLOS, 1979 Functional aspects of satellite DNA and heterochromatin. Int. Rev. Cytol. **58:** 1–119.

JONES, C. W. and F. C. KAFATOS, 1982 Accepted mutations in a

gene family: evolutionary diversification of duplicated DNA. J. Mol. Evol. **19:** 87–103.

JONES, R. S. and S. S. POTTER, 1985 Characterization of cloned human alphoid satellite with an unusual monomeric construction: evidence for enrichment in HeLa small polydispersed circular DNA. Nucleic Acids Res. **13:** 1027–1042.

KARLIN, S. and H. M. TAYLOR, 1975 *A First Course in Stochastic Processes*, Ed. 2. Academic Press, New York.

KATO, S., R. A. ANDERSON and R. D. CAMERINI-OTERO, 1986 Foreign DNA introduced by calcium phosphate is integrated into repetitive DNA elements of the mouse L cell genome. Mol. Cell. Biol. **6:** 1787–1795.

KOCH, A. L., 1979 Selection and recombination in populations containing tandem multiplet genes. J. Mol. Evol. **14:** 273–285.

KOO, H.-S., H.-M. WU and D. M. CROTHERS, 1986 DNA bending at adenine-thymine tracts. Nature **320:** 501–506.

KORNBERG, A., 1980 *DNA Replication*. W. H. Freeman, San Francisco.

KRONTIRIS, T. G., N. A. DIMARTINO, M. COLB and D. R. PARKINSON, 1985 Unique allelic restriction fragments of the human Ha-*ras* locus in leukocyte and tumor DNAs of cancer patients. Nature **313:** 369–374.

KRÜGER, J. and F. VOGEL, 1975 Population genetics of unequal crossing over. J. Mol. Evol. **4:** 201–247.

KURNIT, D. M., 1979 Satellite DNA and heterochromatin variants: The case for unequal mitotic crossing over. Hum. Genet. **47:** 169–186.

LANIA, L., S. BOAST and M. FRIED, 1982 Excision of polyoma virus genomes from chromosomal DNA by homologous recombination. Nature **295:** 349–351.

LEVINSON, G., J. L. MARSH, J. T. EPPLEN and G. A. GUTMAN. 1985 Cross-hybridizing snake satellite, *Drosophila*, and mouse DNA sequences may have arisen independently. Mol. Biol. Evol. **2:** 494–504.

LEWIN, R., 1982 Repeated DNA still in search of a function. Science **217:** 621–623.

LILLEY, D., 1986 Bent molecules—how and why? Nature **320:** 487–488.

LINXWEILIER, W. and W. HORZ, 1985 Reconstitution experiments show that sequence specific histone-DNA interactions are the basis for nucleosome phasing on mouse satellite DNA. Cell **42:** 281–290.

LOHE, A. R. and D. L. BRUTLAG, 1986 Multiplicity of satellite DNA sequences in *Drosophilia melanogaster*. Proc. Natl. Acad. USA **83:** 696–700.

MARUNONUCHI, T. and H. HOSOYA, 1984 Isolation of an autonomously replicating sequence (ARS) from satellite DNA of *Drosophila melanogaster*. Mol. Gen. Genet. **196:** 258–265.

MELTON, D. W., C. MCEWAN, A. B. MCKIE and A. M. REID, 1986 Expression of the mouse HPRT gene: deletional analysis of the promoter region of an X-chromosome linked housekeeping gene. Cell **44:** 319–328.

MEYERINK, J. H., J. KLOOTWIKJ and R. J. PLANTA, 1979 Extrachromosomal circular ribosomal DNA in the yeast *Saccharomyces carlsbergensis*. Nucl. Acids Res. **7:** 69–76.

MIKLOS, G. L. G., 1982 Sequencing and manipulating highly repeated DNA. pp. 41–68. In: *Genome Evolution*, Edited by G. A. DOVER and R. B. FLAVELL. Academic Press, London.

MIKLOS, G. L. G., 1985 Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. pp. 241–321. In: *Molecular Evolutionary Genetics*, Edited by R. J. MACINTYRE. Plenum, New York.

MIKLOS, G. L. G. and A. C. GILL, 1982 Nucleotide sequences of highly repeated DNAs: compilation and comments. Genet. Res. Camb. **39:** 1–30.

MIRKOLVITCH, J., M.-E. MIRAULT and U. K. LAEMMLI, 1984 Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. Cell **39:** 223–232.

MOORE, G. P., 1983 Slipped-mispairing and the evolution of introns. Trends Biochem. Sci. **8:** 411–414.

MORRIS, J., S. R. KUSHNER and R. IVARIE, 1986 The simple repeat Poly(dT-dG)·Poly(dC-dA) common to eukaryotes is absent from eubacteria and archaebacteria and rare in protozoans. Mol. Biol. Evol. **3:** 343–355.

NAGYLAKI, T. and T. D. PETES, 1982 Intrachromosomal gene conversion and maintenance of sequence homogeneity among repeated genes. Genetics **100:** 315–337.

NUSSINOV, R., 1980 Strong adenine clustering in nucleotide sequences. J. Theor. Biol. **85:** 285–291.

OHTA, T., 1980 *Evolution and Variation of Multigene Families*. Springer-Verlag, New York.

OHTA, T. 1981 Genetic variation in small multigene families. Genet. Res. Camb. **37:** 133–149.

PERELSON, A. S. and G. I. BELL, 1977 Mathematical models for the evolution of multigene families by unequal crossing over. Nature **265:** 304–310.

PETES, T., 1980 Unequal meiotic recombination within tandem arrays of yeast ribosomal DNA genes. Cell **19:** 765–774.

RAE, P. M. M., 1982 Unequal crossing-over accounts for the organization of *Drosophila virilis* rDNA insertions and the integrity of flanking 28S gene. Nature **296:** 579–581.

RATTNER, J. B. and C. C. LIN, 1985 Radial loops and helical coils coexist in metaphase chromosomes. Cell **42:** 291–296.

ROEDER, G. S. and G. R. FINK, 1983 Transposable elements in yeast. pp. 300–328. In: *Mobile Genetic Elements*, Edited by J. A. SHAPIRO. Academic Press, New York.

ROGERS, J., 1983 CACA sequences—the ends and the means? Nature **305:** 101–102.

ROTH, G. E. and K. B. MORITZ,1981 Restriction enzyme analysis of the germ line limited DNA of *Ascaris suum*. Chromosoma **83:** 169–190.

SAEDLER, H. and P. NEVERS, 1985 Transposition in plants: a molecular model. EMBO **4:** 585–590.

SAWADA, I., C. WILLARD, C.-K. J. SHEN, B. CHAPMAN, A. C. WILSON and C. W. SCHMID, 1985 Evolution of ALU family repeats since the divergence of human and chimpanzee. J. Mol. Evol. **22:** 316–322.

SCHIMD, C. W. and C.-K. J. SHEN. 1985 The evolution of interspersed repetitive DNA sequences in mammals and other vertebrates. pp. 323–358. In: *Molecular Evolutionary Genetics*, Edited by R. J. MACINTYRE. Plenum, New York.

SCHIMKE, R. T., 1984 Gene amplification in cultured animal cells. Cell **37:** 705–713.

SCHIMKE, R. T., S. W. SHERWOOD, A. B. HILL and R. H. JOHNSON, 1986 Overreplication and recombination of DNA in higher eukaryotes: Potential consequences and biological implications. Proc. Natl. Acad. Sci. USA **83:** 2157–2161.

SCHWARZ-SOMMER, Z., A. GIERL, H. CUYPERS, P. A. PETERSON and H. SAEDLER, 1985 Plant transposable elements generate the DNA sequence diversity needed in evolution. EMBO **4:** 591–597.

SHAW, G. and R. KAMEN. 1986 A conserved AU sequence from the 3′ untranslated region of GM-CSF mRNA mediates selective mRNA degradation. Cell **46:** 659–667.

SHMOOKLER REIS, R. J. and S. GOLDSTEIN, 1980 Loss of reiterated DNA sequences during serial passage of human diploid fibroblast. Cell **21:** 739–749.

SINGER, M. F. 1982 Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. **76:** 67–112.

SLIGHTOM, J. L., L.-Y. E. CHANG, B. F. KOOP and M. GOODMAN, 1985 Chimpanzee fetal $^{G}\gamma$ and $^{A}\gamma$ globin gene nucleotide sequences provide further evidence of gene conversions in Hominine evolution. Mol. Biol. Evol. **2:** 370–389.

SMALL, D., B. NELKIN and B. VOGELSTEIN, 1982 Nonrandom

distribution of repeated DNA sequences with respect to supercoiled loops and the nuclear matrix. Proc. Natl. Acad. Sci. USA **79:** 5911–5915.

SMITH, G. P., 1974 Unequal crossover and the evolution of multigene families. Proc. Cold Spring Harbor Symp. Quant. Biol. **38:** 507–513.

SMITH, G. P., 1976 Evolution of repeated DNA sequences by unequal crossover. Science **191:** 528–535.

SOLOMON, M. J., F. STRAUSS and A. VARSHAVSKY, 1986 A mammalian high mobility group protein recognizes any stretch of six A·T base pairs in duplex DNA. Proc. Natl. Acad. Sci. USA **83:** 1276–1280.

SOUTHERN, E. M., 1975 Long range periodicities in mouse satellite DNA. J. Mol. Biol. **94:** 51–69.

STEPHAN, W., 1986 Recombination and the evolution of satellite DNA. Genet. Res. **47:** 167–174.

STRAUSS, F. and A. VARSHAVSKY, 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.

STREISINGER, G., Y. OKADA, J. EMRICH, J. NEWTON, A. TSUGITA, E. TERZAGNI and M. INOUYE, 1966 Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. **31:** 77–84.

STREISINGER, G. and J. E. OWEN. 1985 Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. Genetics **109:** 633–659.

STUART, G. W., P. F. SEARLE, H. Y. CHEN, R. L. BRINSTER and R. D. PALMITTER, 1984 A 12-base-pair DNA motif that is repeated several times in metallothionein gene promoters confers metal regulation to a heterologous gene. Proc. Natl. Acad. Sci. USA **81:** 7318–7322.

STRUHL, K., 1985 Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. Proc. Natl. Acad. Sci. USA **82:** 8419–8423.

SZOSTAK, J. W. and R. WU, 1980 Unequal crossing over in the ribosomal DNA of *Saccharomyces cerevisiae.* Nature **284:** 426–430.

TAKAHATA, N. 1981 A mathematical study on the distribution of the number of repeated genes per chromosome. Genet. Res. Camb. **38:** 97–102.

TARTOF, K. D., 1974 Unequal mitotic sister chromatid exchange and disproportionate replication as mechanisms regulating ribosomal RNA gene redundancy. Proc. Cold Spring Harbor Symp. Quant. Biol. **38:** 491–500.

TAUTZ, D. and M. RENZ, 1984 Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucl. Acids Res. **12:** 4127–4138.

TAUTZ, D., M. TRICK and G. A. DOVER. 1986 Cryptic simplicity in DNA is a major source of genetic variation. Nature **322:** 652–656.

TREISMAN, R. and T. MANIATIS,1985 Simian virus 40 enhancer increases number of RNA polymerase II molecules on linked DNA. *Nature* **315:** 72–75.

VOGELSTEIN, B., D. M. PARDOLL and D. S. COFFEY, 1980 Supercoiled loops and eucaryotic DNA replication. Cell **22:** 79–85.

WEBER, F. and W. SCHAFFNER, 1985 Simian virus 40 enhancer increases RNA polymerase density within the linked gene. Nature **315:** 75–77.

WILKINS, A. S., 1981 Eukaryotic chromosome replication and the radial loop model. J. Theor. Biol. **89:** 715–717.

WILLARD, C., E. WONG, J. F. HESS, C.-K. J. SHEN, B. CHAPMAN, A. C. WILSON and C. W. SCHMID, 1985 Comparison of human and chimpanzee ζ1 globin genes. J. Mol. Evol. **22:** 309–315.

WONG, W. M., J. L. A. ABRAHAMSON and R. N. NAZAR, 1984 Are DNA spacers relics of gene amplification events? Proc. Natl. Acad. Sci. USA **81:** 1768–1770.

YAVACHEV, L. P., O. J. GEORGIEV, E. A. BRAGA, T. A. AVDONINA, A. E. BOGOMOLOVA, V. B. ZHURKIN, V. V. NOSIKOV and A. A. HADJIOLOV, 1986 Nucleotide sequence analysis of the spacer regions flanking the rat rRNA transcription unit and identification of repetitive elements. Nucleic Acids Res. **14:** 2799–2809.

ZAKIAN, V. A., 1985 Nuclear structure: taken with a grain of salt. Nature **314:** 223–224.

Communicating editor: D. CHARLESWORTH

APPENDIX

Mean under amplification/recombination equilibrium

By definition,

$$E(x) = \sum_{i=1}^{\infty} i\pi_i = \pi_1 + \sum_{i=2}^{\infty} i\pi_i = \pi_1\left(1 + \sum_{i=2}^{\infty} i\rho_i\right), \quad \text{(A1)}$$

where the last equality follows from [8b]. From [10]:

$$\sum_{i=2}^{\infty} i\rho_i = \beta \sum_{i=1}^{\infty} [1/(1+\alpha)]^i = \beta/\alpha, \quad \text{(A2)}$$

which follows from basic identities for a geometric series. Combining (A1) and (A2), we obtain

$$E(x) = \pi_1(1 + \beta/\alpha) = \pi_1(1 + \upsilon/\varepsilon). \quad \text{(A3)}$$

Variance under amplification/recombination equilibrium

We proceed by computing $E(x^2)$. Following along the lines of (A1),

$$E(x^2) = \sum_{i=1}^{\infty} i^2\pi_i = \pi_1 + \sum_{i=2}^{\infty} i^2\pi_i = \pi_1\left(1 + \sum_{i=2}^{\infty} i^2\rho_i\right), \quad \text{(A4)}$$

where from (10) and defining $\theta = 1/(1+\alpha)$,

$$\sum_{i=2}^{\infty} i^2\rho_i = (\beta/\theta)\sum_{i=2}^{\infty} i\theta^i = (\beta/\theta)\theta^2(2-\theta)/(1-\theta)^2, \quad \text{(A5)}$$

the last equality follows from the identity $\sum_{i=0}^{\infty} ix^i = x/(1-x)^2$.

Combining (A4) with (A5) and simple factoring gives

$$E(x^2) = \pi_1[1 + (\beta/\alpha^2)(1 + 2\alpha)]. \quad \text{(A6)}$$

The equilibrium variance, $\mathrm{Var}(x) = E(x^2) - [E(x)]^2$. From (A3) and (A6),

$$\begin{aligned}\mathrm{Var}(x) &= \pi_1[1 + (\beta/\alpha^2)(1+2\alpha)] - [\pi_1(1+\beta/\alpha)]^2 \\ &= \pi_1[(1+\beta/\alpha) + (\beta/\alpha^2)(1+\alpha)] \\ &\quad - [\pi_1(1+\beta/\alpha)]^2 \\ &= E(x) + \pi_1(\beta/\alpha^2)(1+\alpha) - [E(x)]^2,\end{aligned} \quad \text{(A7)}$$

which rearranges slightly to give (20):

$$\mathrm{Var}(x) = \pi_1(\beta/\alpha^2)(1+\alpha) - E(x)[E(x) - 1].$$

Limiting distribution under selection

Consider the general form of our model:

$$\lambda_i = g(i)\gamma/2, \qquad \text{for } i \geq n,$$

$$\mu_i = g(i)(\gamma/2)(1 + 2\varepsilon/\gamma) \quad \text{for } i > n,$$

which gives

$$\rho_i = \theta^i g(n)/g(n+i) \leq \theta^i. \quad \text{(A8)}$$

provided that $g(n) \leq g(n+i)\ \forall\ i \geq 1$, which holds if rates of

recombination do not decrease as array size increases, a very reasonable biological assumption. From (A8),

$$1 + \sum_{i=1}^{\infty} \rho_i \leq \sum_{i=0}^{\infty} \theta^i = 1/(1 - \theta). \tag{A9}$$

(A9) in conjunction with (25) gives (27). From (A8) and (24),

$$\pi_{n+i}/\pi_n \leq \theta^i \quad \text{for } i \geq 1, \tag{A10}$$

implying the probability an array has i (≥ 1) excess members is bounded above by a geometric distribution (with parameter $1 - \theta$), giving (27b). Assuming the limiting geometric gives $E(x) \leq n + 1/\alpha$ and $\text{Var}(x) \leq (1 + \alpha)/\alpha^2$.