Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 3967–3971, May 1991
Genetics

On the association of restriction fragment length polymorphisms across species boundaries

(mitochondrial DNA/gene genealogies/reproductive isolating barriers/gametic-phase disequilibrium)

YUN-XIN FU AND JONATHAN ARNOLD

Department of Genetics, University of Georgia, Athens, GA 30602

Communicated by Michael T. Clegg, January 22, 1991 (received for review October 1, 1990)

ABSTRACT **We study the expected values of gametic-phase disequilibrium in both nuclear and cytonuclear systems in a finite population composed of reproductively isolated subpopulations. Random drift alone within each subpopulation will generate permanent overall non-zero gametic-phase disequilibrium in both a two-locus nuclear system and a cytonuclear system unless the population has an overall initial disequilibrium of zero. We derive formulae for the expected overall disequilibrium when mutation is involved and demonstrate that these values decay to zero at an extremely slow rate compared with the decay of the expected disequilibria within each subpopulation.**

Restriction fragment length polymorphisms (RFLPs) in rapidly evolving nuclear and mitochondrial DNA provide a major tool by which to identify the shared ancestry between hybridizing taxa. In hybrid zones, where such taxa coexist, varying degrees of concordance in the distribution of molecular markers across taxonomic boundaries are observed (1). As a consequence, a number of authors (2–4) now make a distinction between the gene genealogies created by RFLPs on the one hand and the organismal pedigree through which genes descend on the other. A gene genealogy traces out only a portion of the organismal pedigree linking individuals to their ancestors on either side of a hybrid zone and may vary from gene to gene across the zone.

The genealogical concordance or discordance of molecular markers across a hybrid zone may ultimately determine the phylogenetic status of the hybridizing taxa (5). In fact, Avise and Ball (6) suggest that subspecies status should be applied only to two or more taxa when these interbreeding populations are separated concordantly by multiple gene genealogies. This suggestion is motivated in part by simulations (7) showing little association between gene genealogies in the absence of reproductive barriers between subpopulations. To resolve the phylogenetic status of hybridizing taxa, it is then necessary to know what associations can be expected between RFLPs among subpopulations once reproductive barriers are introduced. These barriers to dispersal can be either intrinsic, like the mating system (8), or extrinsic, like geographic barriers.

In this report we address three questions about the expected associations of RFLPs in a subdivided population with reproductive barriers, as measured by their gametic-phase disequilibrium. (*i*) Will genetic drift alone generate permanent gametic-phase disequilibrium across subpopulations? (*ii*) What effects do mutation and genetic drift have on the dynamical decay of gametic-phase disequilibrium in a subdivided population? (*iii*) How does the linkage of RFLPs affect the mean disequilibrium between these reproductively isolated subpopulations?

Model and Methods

Consider a population that is subdivided into n subpopulations or islands and that has discrete, nonoverlapping generations. For simplicity, we shall consider only the case in which each subpopulation is of the same size, N. These subpopulations are reproductively isolated either by intrinsic or extrinsic barriers to gene exchange, so they are evolving independently. The entire population is polymorphic at two restriction sites, A and B. The presence or absence of a site is denoted by A,a and B,b, respectively. The measure of allelic association between two restriction sites or *allelic disequilibrium* within each subpopulation is defined by

$$D_i = p_{i_1}p_{i_4} - p_{i_2}p_{i_3}, \qquad (i = 1, \ldots, n),$$

where $p_{i_1}, \ldots, p_{i_4}$ are the frequencies of the four gametic types (AB, Ab, aB, and ab) in the ith subpopulation. The four gametic frequencies in the ith subpopulation are random variables with joint density g_i. This density is conditional on the gametic frequencies in the previous generation. The gametic frequencies in different subpopulations are assumed to be stochastically independent; therefore, all expectations are with respect to the product density $g_1, g_2 \ldots g_n$. The overall frequencies of these gametes (AB, Ab, aB, ab) in the entire population are denoted by $P_1, \ldots, P_4$ and are defined by:

$$P_j = \left(\sum_{i=1}^{n} p_{ij}\right)/n, \qquad (j = 1, \ldots, 4),$$

assuming each subpopulation is of the same size.

Ohta (9) has considered several types of disequilibrium measures for subdivided populations. Following her definitions, the *overall disequilibrium* is defined by

$$D'_{\mathrm{st}} = P_1P_4 - P_2P_3.$$

We also introduce several expectations:

$$\delta_{\mathrm{st}} = E(P_1)E(P_4) - E(P_2)E(P_3),$$

$$\overline{D} = \sum_i \{E(p_{i_1}p_{i_4}) - E(p_{i_2}p_{i_3})\}/n,$$

$$\overline{\delta} = \sum_i \{E(p_{i_1})E(p_{i_4}) - E(p_{i_2})E(p_{i_3})\}/n.$$

It can easily be shown that

$$D_i = p_{i_1} - p_iq_i \quad \text{and} \quad D'_{\mathrm{st}} = P_1 - PQ,$$

where $p_i = p_{i_1} + p_{i_2}$ is the frequency of site A in the ith subpopulation and $q_i = p_{i_1} + p_{i_3}$ is the frequency of site B in the ith subpopulation. We use P and Q to denote the overall

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

Abbreviation: RFLP, restriction fragment length polymorphism.

frequencies of sites *A* and *B*, respectively. The frequencies of the four gametes and their relationship to the allelic disequilibrium and allelic frequencies on each island are shown in Table 1. A similar relationship holds for the overall frequencies.

Since each island is isolated, we have

$$E(D'_{\text{st}}) = E\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(p_{i_1}p_{j_4} - p_{i_2}p_{j_3})\right\}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{E(p_{i_1})E(p_{j_4}) - E(p_{i_2})E(p_{j_3})\}$$

$$+ \frac{1}{n^2}\sum_{i}\{E(p_{i_1}p_{i_4} - p_{i_2}p_{i_3})\}$$

$$- \frac{1}{n^2}\sum_{i}\{E(p_{i_1})E(p_{i_4}) - E(p_{i_2})E(p_{i_3})\}.$$

In the last two terms on the right-hand side, the cross terms have canceled each other because each subpopulation represents an independent evolutionary process [i.e., $E(XY) = E(X)E(Y)$]. It follows that

$$E(D'_{\text{st}}) = \delta_{\text{st}} + \frac{1}{n}\overline{D} - \frac{1}{n}\bar{\delta}.$$

When each colony has the same initial condition, we have for each generation, $E(P_i) = E(p_{ij}), j = 1, \ldots, n$. This implies $\delta_{\text{st}} = \bar{\delta}$, which reduces the previous equation to

$$E(D'_{\text{st}}) = \frac{1}{n}\overline{D} + \frac{n-1}{n}\bar{\delta}. \qquad \textbf{[1A]}$$

The first term on the right-hand side is the within population component of the expected disequilibrium, and the second term is the between population component.

The quantity $\overline{D}$ measures the average allelic disequilibrium within subpopulations. If we define $\overline{\text{cov}}(p, q) = 1/n\ \Sigma\ [E(p_iq_i) - E(p_i)E(q_i)]$, then the other measure $\bar{\delta}$ is related to the covariance (cov) in site frequencies averaged across subpopulations:

$$\bar{\delta} = \overline{D} + \overline{\text{cov}}(p, q).$$

The right-hand side can now be written in another informative way:

$$E(D'_{\text{st}}) = \overline{D} + \frac{n-1}{n}\cdot\overline{\text{cov}}(p, q). \qquad \textbf{[1]}$$

Notice this expression is different from that of Nei and Li (10) or Asmussen and Arnold (11), since they were considering the sampling mean and covariance in an infinite population.

Table 1. Gametic frequencies within each subpopulation

Site	Site *A*		
B	*A*	*a*	Total
B	$p_{i_1} = p_iq_i + D_i$	$p_{i_3} = (1 - p_i)q_i - D_i$	q_i
b	$p_{i_2} = p_i(1 - q_i) - D_i$	$p_{i_4} = (1 - p_i)(1 - q_i) + D_i$	$1 - q_i$
Total	p_i	$1 - p_i$	1

We shall now consider a nuclear and cytonuclear system. In the former system the two restriction sites *A* and *B* occur in nuclear DNA and are separated by a recombination fraction *r*. In the latter cytonuclear system, site *A* occurs in nuclear DNA, but site *B* is unlinked and occurs in cytoplasmic DNA. The random union of zygotes (RUZ) model will be used as developed by Kimura (12) and Watterson (13) for a nuclear system and will be developed here for a cytonuclear system. Another commonly used model, the random union of gametes (RUG) model (14, 15), is not considered because it is not suitable for a cytonuclear system (unpublished results). In the RUZ model, we assume that each individual of the next generation is formed by taking a pair of gametes randomly from the gametic pool generated by the *N* individuals of the current generation in the *i*th isolated subpopulation. Let X_{ifm} be the number of individuals in the *i*th subpopulation receiving gametes of type *f* from the father and gametes of type *m* from the mother (recall that gametes *AB, Ab, aB*, and *ab* are indexed 1, 2, 3, and 4). The constants α_{fmk} will denote the probability of producing a gamete of type *k* from such an individual with genotype (*f, m*). The values of α_{fmk} for both the nuclear and cytonuclear systems are given in Table 2. From this table gametic frequencies can be computed in generation *t*:

$$p_{ik}(t) = \left(\sum\sum X_{ifm}\alpha_{fmk}\right)/N, \qquad (k = 1, \ldots, 4).$$

While normally the genotypic counts X_{ifm} and hence $p_{ik}(t)$ are random, we condition on them in calculating the density of $p_{i_1}(t + 1), \ldots, p_{i_4}(t + 1)$ in the next generation. For the nuclear system, Watterson (13) showed that the joint moment-generating function of the gametic frequencies $[p_{i_1}(t + 1), p_{i_2}(t + 1), p_{i_3}(t + 1), p_{i_4}(t + 1)]$ in generation $t + 1$ is

$$M(\Theta_1,\Theta_2,\Theta_3,\Theta_4)$$

$$= E\{\exp[\Theta_1 p_{i_1}(t + 1) + \Theta_2 p_{i_2}(t + 1) + \Theta_3 p_{i_3}(t + 1)$$

$$+ \Theta_4 p_{i_4}(t + 1)] \mid p_{ik}(t)\},$$

$$= \left[\sum\sum p_{if}(t)p_{im}(t)\exp\left(\sum \Theta_k\alpha_{fmk}/N\right)\right]^N. \qquad \textbf{[2]}$$

This result is conditional on gametic frequencies in generation *t*. A similar result holds in a cytonuclear system. By

Table 2. Values of α_{fmk} for a nuclear and cytonuclear system

	k			
f, m	1	2	3	4
Nuclear				
1, 1	1	0	0	0
1, 2	1/2	1/2	0	0
1, 3	1/2	0	1/2	0
1, 4	(1 − *r*)/2	*r*/2	*r*/2	(1 − *r*)/2
2, 2	0	1	0	0
2, 3	*r*/2	(1 − *r*)/2	(1 − *r*)/2	*r*/2
2, 4	0	1/2	0	1/2
3, 3	0	0	1	0
3, 4	0	0	1/2	1/2
4, 4	0	0	0	1
Cytonuclear				
1, 1 (or 2, 1)*	1	0	0	0
1, 2 (or 2, 2)	0	1	0	0
1, 3 (or 2, 3)	1/2	0	1/2	0
1, 4 (or 2, 4)	0	1/2	0	1/2
3, 1 (or 4, 1)	1/2	0	1/2	0
3, 2 (or 4, 2)	0	1/2	0	1/2
3, 3 (or 4, 3)	0	0	1	0
3, 4 (or 4, 4)	0	0	0	1

*The paternal allele *f* is not passed onto offspring.

definition, we have:

$$M(\Theta_1,\Theta_2,\Theta_3,\Theta_4) = E\{\exp[\Sigma\Theta_k p_{ik}(t+1)]\},$$

$$= \Sigma\, pr\{X_{ifm}(t+1);$$

$$m, f = 1, \ldots, 4\}\exp[\Sigma\Theta k pik(t+1)];$$

$$= \Sigma\, pr\{X_{ifm}(t+1);$$

$$m, f = 1, \ldots, 4\}\exp\{\Sigma\Theta k[\Sigma\Sigma\, Xifm(t+1)\alpha fmk]/N\},$$

$$= [\Sigma\Sigma p_{if}(t)p_{im}(t)\exp(\Sigma\Theta_k\alpha_{fmk}/N)]^N.$$

This establishes Watterson's fundamental result in Eq. **2** for a cytonuclear system with only the constants α_{fmk} being different. Coefficients α_{fmk} are defined in Table 2. With Eq. **2**, various moments of the gametic frequencies can be studied, and we shall derive the recurrence formulae for the expected disequilibria in the following sections. Result **2** uniquely specifies the density g_i.

Expected Disequilibria Due to Random Drift

The behavior of the expected overall disequilibrium $E(D'_{\mathrm{st}})$ can be determined by the study of the expected disequilibrium on each island. In the following sections, we shall drop the subscript i for an island when the context is clear. Using the moment-generating function **2** for a nuclear system (13), it is easy to show on each island that

$$E\{p_k(t)\} = E\{p_k(t-1)\} + \eta_k r E[D(t-1)], \qquad k = 1, \ldots, 4,$$

$$E\{D(t)\} = \left(1 - \frac{1}{2N} - r\right)E\{D(t-1)\},$$

where $\eta_1 = \eta_4 = -1$ and $\eta_2 = \eta_3 = 1$. Repeating the process for generations $t-1, \ldots, 1$, we thus have

$$E\{p_k(t)\} = p_k(0) + \eta_k \cdot \frac{2rN}{2rN+1} \cdot D(0)\left\{1 - \left(1 - \frac{1}{2N} - r\right)^t\right\},$$

$$(k = 1, \ldots, 4),$$

$$E\{D(t)\} = \left(1 - \frac{1}{2N} - r\right)^t D(0). \qquad [3]$$

The corresponding cytonuclear expressions are obtained by letting $r = 1/2$ in the above equations. Although, in general, a cytonuclear system is different from a nuclear system, expressions considered in this paper happen to be obtainable by letting $r = 1/2$ in formulae for a nuclear system. Thus, we shall refer mainly to the nuclear system. From Eq. **3** we can establish that

$$E(p_1)E(p_4) - E(p_2)E(p_3)$$

$$= D(0) - \frac{2rN}{2rN+1} \cdot D(0)\left\{1 - \left(1 - \frac{1}{2N} - r\right)^t\right\}.$$

In the case of the same initial condition for each island, substituting Eq. **3** and the equality above into Eq. **1A** yields:

$$E\{D'_{\mathrm{st}}(t)\} = \left(1 - \frac{1}{2N} - r\right)^t \overline{D}(0)$$

$$+ \frac{n-1}{n} \cdot \frac{\overline{D}(0)}{2rN+1}\left\{1 - \left(1 - \frac{1}{2N} - r\right)^t\right\}. \qquad [4]$$

Comparing Eq. **4** with Eq. **1**, we obtain the solution for the average allelic disequilibrium ($\overline{D}$) within subpopulations and the covariance in site frequencies averaged across subpopulations. The within-subpopulation component $\overline{D}$ goes to zero quickly, but not the between-population component, $\overline{\mathrm{cov}}(p, q)$. The asymptotic value of the overall disequilibrium is

$$E\{D'_{\mathrm{st}}(\infty)\} = \frac{n-1}{n} \cdot \frac{\overline{D}(0)}{2rN+1}, \qquad [5]$$

and by letting $r = 1/2$, we have the corresponding limit for a cytonuclear system.

Expected Disequilibria When Mutation Exists

When mutation is involved, we need to consider the expected disequilibrium before and after mutation. Suppose that the mutation rates from A to a and from a to A are μ_1 and ν_1, respectively. Furthermore, suppose that the mutation rates from B to b and from b to B are μ_2 and ν_2, respectively. The frequencies of the four gametes after mutation but before sampling for both the nuclear and cytonuclear systems are

$$p'_1(t) = p_1(t)(1 - \mu_1 - \mu_2) + p_2\nu_2 + p_3\nu_1,$$

$$p'_2(t) = p_2(t)(1 - \nu_2 - \mu_1) + p_1\mu_2 + p_4\nu_1,$$

$$p'_3(t) = p_3(t)(1 - \nu_1 - \mu_2) + p_1\mu_1 + p_4\nu_2,$$

$$p'_4(t) = p_4(t)(1 - \nu_1 - \nu_2) + p_2\mu_1 + p_3\mu_2.$$

After mutation and before sampling we also have

$$D'(t) = (1 - U)D(t).$$

The constant $U = \mu_1 + \nu_1 + \mu_2 + \nu_2$ is the total mutation rate.

As in the case with drift above, we condition on the current frequencies and we use these in Eq. **2** to compute the moment-generating function of $p_1(t+1), \ldots, p_4(t+1)$. From the moment-generating function in successive generations, we can derive recursions for the moments of gametic frequencies and functions of them.

Since sampling takes place after mutation, we have by analogy to Eq. **3**:

$$E\{p_k(t)\} = E\{p'_k(t-1)\} + \eta_k r E\{D'(t-1)\}, \qquad k = 1, \ldots, 4,$$

$$E\{D(t)\} = \left(1 - \frac{1}{2N} - r\right)E\{D'(t-1)\}.$$

After substituting and simplifying, we obtain:

$$E\{D'_{\mathrm{st}}(t)\} = (1-U)^t\left\{\left(1 - \frac{1}{2N} - r\right)^t \overline{D}(0)\right.$$

$$\left. + \frac{n-1}{n} \cdot \frac{\overline{D}(0)}{2rN+1}\left[1 - \left(1 - \frac{1}{2N} - r\right)^t\right]\right\}. \qquad [6]$$

The corresponding formula for a cytonuclear system is obtained by letting $r = 1/2$. A comparison with Eq. **1** yields the solution for $\overline{D}$ and $\overline{\mathrm{cov}}(p, q)$. The asymptotic values in both systems are zero, but the between-population component can persist for thousands of generations.

Numerical Results

Examples of the dynamics of the expected values of the overall disequilibrium, D'_{st}, are shown in Figs. 1 and 2.

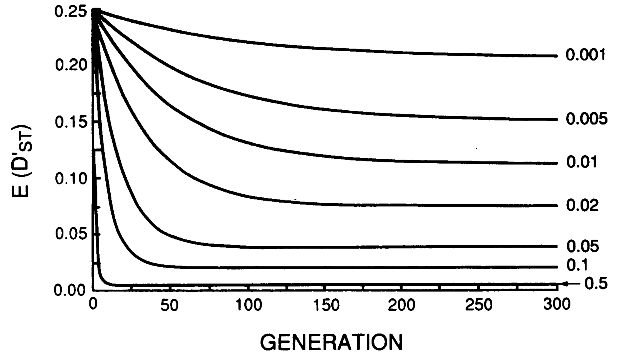

FIG. 1. $E(D'_{st})$ as a function of time for $r = 0.001, 0.005, 0.01, 0.02, 0.05, 0.1$, and 0.5. There are 10 subpopulations, each with $N = 50$.

Virtually all of D'_{st} is in the between-population component after a few generations. The effect of recombination on the disequilibrium maintained is substantial, while the rate of decay in the presence of mutation is very slow. The implication and importance of these results will be discussed further in the next section.

Although the main purpose of this paper is to consider the expected value of disequilibrium, the variance $\mathrm{VAR}(D'_{st})$ in both cases of mutation and migration was also studied by Ohta (9, 16). In the simplest case, with no mutation or migration, the asymptotic value of $E(D'_{st})$ can be derived by the following simple argument.

In each colony the population will be fixed eventually for one of the four gametes. The probability of being fixed for gamete k is

$$g_k = p_k(0) + \eta_k \cdot \frac{\bar{D}(0)}{2rN + 1}.$$

When each colony has achieved fixation, the process of sampling from the n subpopulations is analogous to taking a sample of size n from a multinomial distribution with parameters $\{g_1, \ldots, g_4\}$. From the moments of a multinomial distribution, it is easy to show that

$$E\{D'_{st}(\infty)\} = \frac{n-1}{n} \cdot \frac{\bar{D}(0)}{2rN + 1},$$

$$\mathrm{VAR}\{D'_{st}(\infty)\} = \frac{n-1}{n}\left\{\frac{1}{n}\left[p_\infty q_\infty(1 - p_\infty)(1 - q_\infty) + D_\infty(1 - 2p_\infty)(1 - 2q_\infty) - D^2\right] + \frac{1}{n^2}\left[2D_\infty^2 - D_\infty(1 - 2p_\infty)(1 - 2q_\infty)\right]\right\}. \quad [7]$$

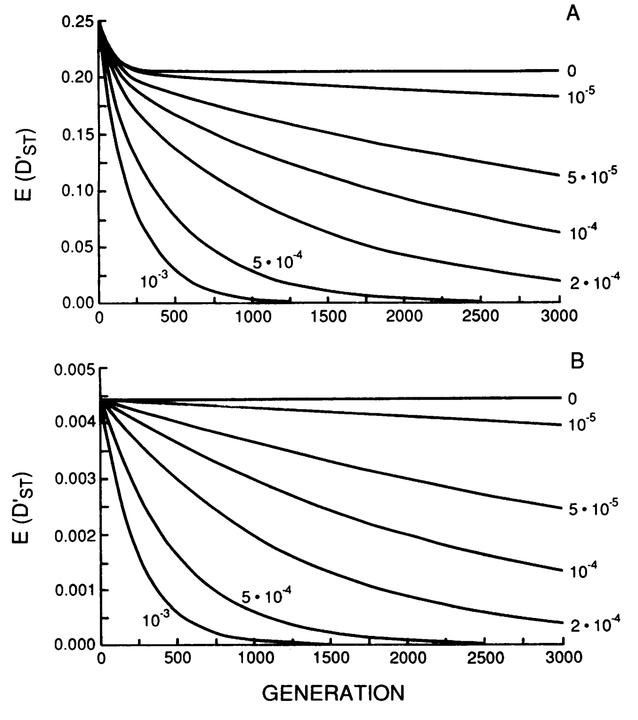

FIG. 2. $E(D'_{st})$ as a function of time for $U = 0, 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}$. (*A*) Nuclear system ($r = 0.001$). (*B*) Cytonuclear system. There are 10 subpopulations, each with $N = 50$.

When $n = 1$, both of the expressions in Eq. 7 are zero, as they should be. In Eq. 7, the variance in D decreases as the number of subpopulations increases, and we have:

$$p_\infty = g_1 + g_2 = p(0),$$

$$q_\infty = g_1 + g_3 = q(0),$$

$$D_\infty = g_1g_4 - g_2g_3 = 2rN\bar{D}(0)/(2rN + 1).$$

The expression $E\{D'_{st}(\infty)\}$ from this argument is exactly the same as in Eq. 5. The variances also agree well with our simulations. Examples in which $E\{D'_{st}\}$ and $\mathrm{VAR}(D'_{st})$ are calculated when subpopulations are fixed, are given in Table 3. The analogy to sampling from a multinomial distribution not only allows us to correct the biased estimate of D'_{st} but also provides the variance of the estimate. A normalized measure $D^*_{st} = [(n - 1)/n]\, D_{st}$ should be used in practice. The variance in disequilibrium increases over time until fixation of all subpopulations is achieved. Similar results to Table 3 were obtained with 1000 replicates.

Discussion

While it is well to know that genetic drift generates variance in linkage disequilibrium between two or more genetic loci (or RFLPs) in a subdivided population (9, 16) and that genetic drift with mutation eventually eliminates any memory of the disequilibrium in the ancestral stock, the exact dynamics of the allelic disequilibrium have not been considered. Here we have established that genetic drift alone is sufficient to generate permanent gametic-phase disequilibrium in a population subdivided by reproductive barriers. This result is analogous to the Wahlund Principle in the one locus case (17). If there is an initial disequilibrium $\bar{D}(0)$ between unlinked nuclear RFLPs or RFLPs in a cytonuclear system and if each subdivision is of constant size N, this allelic disequilibrium is quickly converted into a between-population component and permanently maintained at a level $[(n - 1)/n]\bar{D}(0)/(N + 1)$. If the initial disequilibrium is large and the deme size N is small, this final steady-state disequilibrium can be substantial. If the RFLPs are linked (Fig. 1), a much larger fraction $[(n - 1)/n]\bar{D}(0)/(2rN + 1)$ of the initial disequilibrium will be maintained. Thus, when the subpopulations are reproductively isolated, by chance we obtain both concordant and discordant segregation of RFLPs across reproductive barriers.

As shown here and by Ohta (9), mutation will eventually remove the memory of the historical events leading to these associations. What is remarkable, however, is that it takes thousands of generations (Fig. 2) for mutation to halve the initial associations created by drift. For example, founder events or range expansion of a subpopulation by means of small peripheral populations are two ways by which drift could establish the initial associations. In these cases, the

Table 3. Comparison of simulation and theoretical results (in parentheses) on asymptotic $E(D'_{\mathrm{st}})$ and $\mathrm{VAR}(D'_{\mathrm{st}})$ with initial parameters $p_0 = q_0 = 0.5$ and $D_0 = 0.25$, its maximum value

		Number of subpopulations							
		2		5		10		20	
N	r	$E(D'_{\mathrm{st}})$	$\mathrm{VAR}(D'_{\mathrm{st}})$	$E(D'_{\mathrm{st}})$	$\mathrm{VAR}(D'_{\mathrm{st}})$	$E(D'_{\mathrm{st}})$	$\mathrm{VAR}(D'_{\mathrm{st}})$	$E(D'_{\mathrm{st}})$	$\mathrm{VAR}(D'_{\mathrm{st}})$
50	0.01	0.0550	0.0151	0.1070	0.0084	0.1167	0.0042	0.1225	0.00270*
		(0.0625)	(0.0156)	(0.1000)	(0.0085)	(0.1125)	(0.0045)	(0.1188)	(0.00230)†
50	0.50	0.0050	0.0137	0.0016	0.0107	0.0042	0.0057	0.0049	0.00315*
		(0.0025)	(0.0156)	(0.0039)	(0.0100)	(0.0044)	(0.0056)	(0.0047)	(0.00297)†

*Results from simulation (200 replicates).
†Calculated from Eq. 7.

steady-state solution for the mean in disequilibrium may never be achieved and thus serves as a poor predictor of the association of molecular markers across a hybrid zone or recently derived subspecies separated by reproductive barriers. One extreme example of this is discussed by Gyllensten and Wilson (18).

There are two circumstances under which these models could be useful in interpreting RFLP data on hybrid zones. When random clones from a clonal library are used to identify RFLPs and hence the recombination fraction r is sizeable ($r > 0.001$), then the $\ln[E(D'_{\mathrm{st}}(t)]$ in Eq. **6** could be approximated by:

$$\ln\left[\frac{n-1}{n}\cdot\overline{\mathrm{cov}}(p,q)\right] = t\ln(1-U) + \ln\left[\frac{n-1}{n}\cdot\frac{\bar{D}(0)}{2rN+1}\right] + \ln\left[1-\left(1-\frac{1}{2N}-r\right)^{t}\right], \qquad [8]$$

where the last term is neglected for $r > 0.001$.

The left-hand side could be estimated from samples on either side of a hybrid zone, and the right-hand side could be estimated by a nonlinear regression on r utilizing the map distances along an RFLP map. The result would be information about the concordance of two markers measured by $\overline{\mathrm{cov}}$ (p, q) relative to other pairs in the regression, the history of the zone $[t, \bar{D}(0)]$, and population structure (N). The validity of Eqs. **4** and **6** is predicated on the markers being neutral and the site variation being generated by a substitution process. These models could also be used in interpreting site variation identified by other means, such as the amplification of DNA fragments by the polymerase chain reaction (19) utilizing random primers.

The second circumstance arises when a series of haplotypes are scored for the presence (absence) of restriction sites by using clones covering a particular gene region ($r < 0.001$). In this situation the within-subpopulation component $\bar{D}$ and between-population component $(n - 1/n)\cdot\overline{\mathrm{cov}}(p, q)$ would be estimated separately. Expression **8** would be used in conjunction with a similar expression for the within-population component:

$$\ln \bar{D}(t) = t\ln\left[(1-U)\left(1-\frac{1}{2N}-r\right)\right] + \ln \bar{D}(0). \qquad [9]$$

In both Eqs. **8** and **9**, the appropriate component of $E[D'_{\mathrm{st}}(t)]$ would be regressed on the physical distance d, where $r = ad$. The unknown constant of proportionality could be estimated in the regression or measured directly by pulsed field electrophoresis (21).

To make Eqs. **8** and **9** operational, we will need to consider the variance in allelic disequilibrium due to genetic sampling as well as that introduced by "statistical sampling" (22). The utility of Eqs. **7** and **8** will depend, for example, on the magnitude of $\mathrm{VAR}(D'_{\mathrm{st}})$. The size of $\mathrm{VAR}(D'_{\mathrm{st}})$ relative to $E(D'_{\mathrm{st}})$ is promising for $r = 0.01$ and $n = 20$ in Table 3, but not so for $r = 0.50$ and $n = 20$. Hill and Weir (23) warn that the utility of $E(D^2)$ as a function of r is low.

Hybrid zones among many species (1) serve as barriers to gene exchange. The history of how the subpopulations are established will create chance associations between genes on either side of the zone, and these associations can be expected to persist in the absence of mutation (18).

The hybrid zone itself may also be a mosaic of habitats (28). So long as individuals remain reproductively isolated in different patches, genetic drift and subdivision will conspire to preserve a fraction of the novel associations between genes among subpopulations generated by drift.

We thank A. J. Cuticchia, M. Cellino, E. McCarthy, and two reviewers for comments. This work was supported by National Science Foundation Grant BSR-8716804.

1. Hewitt, G. M. (1989) *Speciation and Its Consequences* (Sinauer, Sunderland, MA), pp. 85–110.
2. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
3. Wilson, A., Cann, R. L., Carr, S. M., Georges, M., Gyllensten, U. B., Helm-Bychaski, K. M., Higuchi, R. G., Palumbi, S. R., Prager, E. M., Sage, R. D. & Stoneking, M. (1985) *Biol. J. Linn. Soc.* **26**, 375–400.
4. Avise, J. C., Arnold, J., Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E., Reeb, C. A. & Saunders, N. C. (1987) *Annu. Rev. Ecol. Syst.* **18**, 489–522.
5. Cracraft, J. (1989) *Speciation and Its Consequences* (Sinauer, Sunderland, MA), pp. 28–59.
6. Avise, J. C. & Ball, M. (1990) *Oxford Surveys in Evolutionary Biology* (Oxford Univ. Press, Oxford), Vol. 7, pp. 45–67.
7. Ball, M., Neigel, J. E. & Avise, J. C. (1990) *Evolution* **44**, 360–370.
8. Arnold, J., Asmussen, M. A. & Avise, J. C. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 1893–1896.
9. Ohta, T. (1982) *Genetics* **101**, 139–155.
10. Nei, M. & Li, W. H. (1973) *Genetics* **75**, 213–219.
11. Asmussen, M. A. & Arnold, J. (1991) *Theor. Popul. Biol.*, in press.
12. Kimura, M. (1963) *Biometrics* **19**, 1–17.
13. Watterson, G. A. (1970) *Theor. Popul. Biol.* **1**, 72–87.
14. Hill, W. G. & Robertson, A. (1968) *Theor. Appl. Genet.* **38**, 226–231.
15. Karlin, S. & McGregor, J. (1968) *Genetics* **58**, 141–159.
16. Ohta, T. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1940–1944.
17. Feldman, M. W. & Christiansen, F. G. (1975) *Genet. Res.* **24**, 151–162.
18. Gyllensten, U. & Wilson, A. C. (1987) *Genet. Res.* **49**, 25–29.
19. Williams, J. G. K., Kubelik, A. R., Livsk, K. J., Rafalski, J. & Tingey, S. V. (1990) *Nucleic Acids Res.* **18**, 6531–6535.
20. Weir, B. S. (1990) *Genetic Data Analysis* (Sinauer, Sunderland, MA).
21. Meagher, R. B., McLean, M. D. & Arnold, J. (1988) *Genetics* **120**, 809–818.
22. Tachida, H. & Cockerham, C. C. (1986) *Theor. Popul. Biol.* **29**, 161–197.
23. Hill, W. G. & Weir, B. S. (1988) *Theor. Popul. Biol.* **33**, 54–78.
24. Arnold, M. L., Shaw, D. D. & Contreras, N. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 3946–3950.
25. Marchant, A. D., Arnold, M. L. & Wilkinson, P. (1988) *Heredity* **61**, 321–328.
26. DeSalle, R. & Giddings, L. V. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 6902–6906.
27. Carson, H. L. (1987) *Annu. Rev. Genet.* **21**, 405–423.
28. Harrison, R. & Rand, D. M. (1989) *Speciation and Its Consequences* (Sinauer, Sunderland, MA), pp. 111–113.