Statistics in Molecular Biology
IMS Lecture Notes – Monograph Series (1999) Volume 33

SOME STATISTICAL ASPECTS OF CYTONUCLEAR DISEQUILIBRIA

BY SUSMITA DATTA

Georgia State University

The purpose of this paper is to review the statistical properties of cytonuclear disequilibria, which measure the association of cytoplasmic genes with nuclear genes and genotypes within a hybrid zone, under different evolutionary models. We report the exact dynamics of the expected cytonuclear genotypic disequilibria for both the homozygotes and heterozygotes in a finite population, with or without having reproductively isolated subdivisions, under random drift alone and random drift along with mutation. The dynamics for the variance is studied using Monte Carlo simulation for a subdivided population, whereas its exact formula for a single undivided population is available. The asymptotic formulas for both the expectation and variance are obtained which are compared between populations with and without reproductive barriers. Construction of a of goodness of fit type statistical test using the dynamics of the cytonuclear disequilibria is discussed. An existing test in an undivided population is reviewed and a new test for a subdivided population is outlined.

1. Introduction. Scientists have noticed dramatic nonrandom associations of cytoplasmic markers with nuclear markers in a variety of hybrid populations such as mice (Ferris *et. al.*, 1983), waterfrogs (Spolsky and Uzzel, 1984), treefrogs (Lamb and Avise, 1986) etc. One needs to correctly define these associations and check whether these cytonuclear associations can be explained without invoking natural selection. These nonrandom associations of cytoplasmic genes with nuclear genes and genotypes, i.e., cytonuclear disequilibria (Lamb and Avise, 1986; Asmussen *et al.*, 1987; Arnold, 1993) can be used to infer the natural history of a particular species. For example, in hybrid zones, cytonuclear disequilibria provide important information about the directionality of the mating events between hybridizing taxa, levels of assortative mating by conspecifics and the kinds of selection on hybrids (Arnold 1993).

Rand and Harrison (1989) have suggested that some hybrid zones may be fundamentally different in character from the more clinal picture as in Mallet *et al.* (1990). For example, in the case of an extensive hybrid zone between two types of cricket along the Appalachians (*Gryllus pennsylvanicas* and *Gryllus firmus*), individuals may sort themselves according to the soil type. As a consequence, the hybrid zone becomes a mosaic of populations of one species or the

Research partially supported by NIH grant AI07442.
AMS 1991 *subject classifications.* Primary 92D25, 92D15; secondary 62P06, 92D10.
Key words and phrases. Genetic drift, neutrality hypothesis, goodness of fit test.

other reflecting the patchwork character of soil types. This kind of subdivided population structure is exactly what Wright (1968) hypothesized as necessary for the shifting balance theory. Under this kind of structure a population has the opportunity to explore repeatedly new gene combinations created by hybridization, and natural selection can retain those complexes that are adaptive.

Until recently, only single locus models of cytoplasmic diversity have been considered by Birky *et al.* (1989). Fu and Arnold (1991) have discussed the behavior of the allelic disequilibria in a subdivided population or a mosaic hybrid zone under a cytonuclear system. Earlier, Ohta (1982) considered this problem in a nuclear system.

In a recent article, Datta and Arnold (1998) studied the dynamics of two other genotypic disequilibria measures (to be defined in the next section) in a subdivided cytonuclear system. These results are contrasted with those in the case of a single (or undivided) cytonuclear population (Fu and Arnold, 1991; Datta *et al.*, 1996a).

In the next section, we define the various cytonuclear disequilibria in a subdivided population. In Section 3, we include the dynamics (in terms of number of generations) of the overall cytonuclear disequilibria under random drift alone and random drift in combination with mutation, respectively, for a subdivided population and compare the results with those in an undivided population. In Section 4, we present the asymptotic results for the same disequilibria under the random drift model. Construction of statistical tests using the expected trajectory of the cytonuclear disequilibria is discussed in Section 5. An overall discussion of the results of Sections 3 and 4 including their biological significance is given in Section 6.

2. Cytonuclear disequilibria. We consider a population consisting of n isolated subpopulations each having the same discrete, nonoverlapping generations. These subpopulations are isolated reproductively either by intrinsic or extrinsic barriers to gene exchange. Hence they evolve independently. We observe the whole population at two restriction sites, one at a nuclear site with allelic types A and a and the other at a cytoplasmic site with alleles M and m. As a result, there are six possible cytonuclear genotypes with (relative) frequencies denoted by p_k, $k = 1, \ldots, 6$ (Table 1). The corresponding frequencies within the subpopulations will be denoted by double suffixes. For example, p_{1i} is the frequency of the AA/M cytonuclear genotype in the ith subpopulation. The measures of the cytonuclear disequilibria for the homozygous case AA/M and for the heterozygous case Aa/M, within subpopulation i are defined by

$$D_{1i} = p_{1i} - u_i q_i, i = 1, \ldots, n, \quad D_{2i} = p_{2i} - v_i q_i, i = 1, \ldots, n,$$

TABLE 1
Frequencies of cytonuclear genotypes

	Nuclear genotype			
Cytoplasm	AA	Aa	aa	Total
M	p_1	p_2	p_3	q
m	p_4	p_5	p_6	$1-q$
Total	u	v	w	1

respectively. Here n is the total number of subpopulations, and for each i,

$$u_i = p_{1i} + p_{4i},$$

$$q_i = p_{1i} + p_{2i} + p_{3i}, \quad v_i = p_{2i} + p_{5i}.$$

The overall frequencies of these genotypes in the entire population are denoted by $P_1, \ldots, P_6$, where each P_k is defined as

$$P_k = (\sum_{i=1}^{n} n_i p_{ki})/(\sum_{i=1}^{n} n_i);\ k = 1, \ldots, 6,$$

with n_i being the size of the ith subpopulation, $1 \leq i \leq n$. Thus, the **overall cytonuclear disequilibria** for the entire subdivided population corresponding to the homozygous case AA/M and the heterozygous case Aa/M, are given by

$$D_{1,ST} = P_1 - UQ \quad \text{and} \quad D_{2,ST} = P_1 - VQ,$$

respectively, where

$$U = P_1 + P_4,\ V = P_2 + P_5,\ Q = P_1 + P_2 + P_3.$$

An overall gametic disequilibria can be defined along the same line as

$$D_{ST} = e_{1,ST} - PQ,$$

where P is given by

$$P = e_{1,ST} + e_{2,ST}.$$

$e_{1,ST}$ is the frequency of the gametic type A/M and $e_{2,ST}$ is the frequency of the gametic type A/m in the entire population. Results concerning the behavior of $E(D_{ST})$ and $Var(D_{ST})$ over time can be found in Fu and Arnold (1992) for the case $n = 1$, and Fu and Arnold (1991) for $n \geq 1$, respectively. In the next

section, we will discuss the dynamics of the expected disequilibria $E(D_{1,ST})$ and $E(D_{2,ST})$ over time (generation). Datta and Arnold (1998) show that for the special case when all the subpopulations have the same size, $E(D_{1,ST})$ can be written as

$$E(D_{1,ST}) = \overline{D}_1 + \frac{n-1}{n}\overline{cov}(u,q).$$

Here the quantity $\overline{D}_1 = \sum_{i=1}^{n} E(p_{1i} - u_i q_i)/n$ measures the average homozygous cytonuclear disequilibria within subpopulations and $\overline{cov}(u,q) = \frac{1}{n}\sum_{i=1}^{n}[E(u_i q_i) - E(u_i)E(q_i)]$, is the covariance (cov) in site frequencies u and q averaged across subpopulations. Similarly, $E(D_{2,ST})$ can be represented as

$$E(D_{2,ST}) = \overline{D}_2 + \frac{n-1}{n}\overline{cov}(v,q).$$

Since under the scenario of a random drift model, all the individual subpopulations will reach equilibrium eventually (Datta *et al.*, 1996a), the average cytonuclear disequilibria converges to zero with time. Note that this is in contrast with the eventual behavior for an undivided population, i.e., $n = 1$, where there is no covariance term. Hence any nonzero value of the overall cytonuclear disequilibria after a long period of time will have to come from the between population covariance term. However the behavior of the cytonuclear disequilibria for a relatively small generation number is much more complicated as the results in the next section show.

3. Dynamics of the cytonuclear disequilibria. The behavior of the expectation and variance curves over time (generation number) for the cytonuclear disequilibria $D_{1,ST}$ and $D_{2,ST}$ for $n = 1$ were studied by Datta *et al.* (1996a) and were later generalized to the case $n \geq 1$ in Datta and Arnold (1998). Here, we present the more general results for an arbitrary n first and then note its various ramification for an undivided population by specializing to the case $n = 1$. Unlike Fu and Arnold (1991), Datta and Arnold (1998) considered the most general case, where the subpopulation sizes are allowed to be different, and they are subject to change over generations. We denote the size of the ith subpopulation at time t by $n_i(t)$, $(i = 1, \ldots, n,\ t \geq 1)$ where there are n subpopulations in the entire population.

The formulas reported in this paper are somewhat more complicated than the ones in Fu and Arnold (1991). This is mostly due to the fact that we are dealing with genotypic disequilibria rather than allelic disequilibria and also partly due to the fact that we made no special assumption on the subpopulation sizes. However, they are not too difficult to calculate in a computer program.

Sometimes to save space and to keep the notation simple we may not show this dependence on t, but it is to be understood.

3.1. *Random drift alone.* Recall that the overall cytonuclear disequilibria for the entire population due to the homozygote AA at the nuclear locus and M at the mtDNA locus is

$$D_{1,ST} = P_1 - UQ = P_1 - (P_1^2 + P_1P_2 + P_1P_3 + P_1P_4 + P_2P_4 + P_3P_4),$$

where $P_k(t) = \sum_{i=1}^{n} n_i(t)p_{ki}(t)/N(t)$; $k = 1, \ldots, 6$, and $N(t) = \sum_{i=1}^{n} n_i(t)$ is the size of the entire population at time t. The six cytonuclear genotypic frequencies in the ith subpopulations are random variables with joint probability mass function (p.m.f) g_i. These random variables have the Markov property, i.e., this joint density of the current generation can be calculated by conditioning on the previous generation. See, e.g. Datta *et al.* (1996a) for the exact description of this conditional distribution. We assumed that the subpopulations are independent and hence the cytonuclear genotypic frequencies of different subpopulations are independent. Therefore all the expectations are calculated with respect to the product p.m.f $g_1, \ldots, g_n$. Thus the expected value of $D_{1,ST}$ is given by

$$\begin{aligned}
E(D_{1,ST}) &= \frac{1}{N}E(\sum_{i=1}^{n} n_ip_{1i}) - \frac{1}{N^2}E\{\sum_{i=1}^{n}\sum_{j=1}^{n} n_in_j(p_{1i}p_{1j} + p_{1i}p_{2j} \\
&\quad + p_{1i}p_{3j} + p_{1i}p_{4j} + p_{2i}p_{4j} + p_{3i}p_{4j})\}, \\
&= \frac{1}{N}E(\sum_{i=1}^{n} n_ip_{1i}) - \frac{1}{N^2}[(\sum_{i=1}^{n} n_iEp_{1i})^2 + (\sum_{i=1}^{n} n_iEp_{1i})(\sum_{j=1}^{n} n_jEp_{2j}) \\
&\quad + (\sum_{i=1}^{n} n_iEp_{1i})(\sum_{j=1}^{n} n_jEp_{3j}) + (\sum_{i=1}^{n} n_iEp_{1i})(\sum_{j=1}^{n} n_jEp_{4j}) \\
&\quad + (\sum_{i=1}^{n} n_iEp_{2i})(\sum_{j=1}^{n} n_jEp_{4j}) + (\sum_{i=1}^{n} n_iEp_{3i})(\sum_{j=1}^{n} n_jEp_{4j})] \\
&\quad + \frac{1}{N^2}\sum_{i=1}^{n} n_i^2ED_{1i} - \frac{1}{N^2}\sum_{i=1}^{n} n_i^2Ep_{1i} \\
&\quad + \frac{1}{N^2}[\sum_{i=1}^{n} n_i^2\{Ep_{1i}^2 + (Ep_{1i})(Ep_{2i}) + (Ep_{1i})(Ep_{3i}) \\
&\quad + (Ep_{1i})(Ep_{4i}) + (Ep_{2i})(Ep_{4i}) + (Ep_{3i})(Ep_{4i})]. \qquad (3.1)
\end{aligned}$$

Note that all the quantities in the above expression are evaluated at a given time t. The expectation ED_{1i} is the value of the cytonuclear disequilibrium within the ith subpopulation. To derive the above term one uses the fact that the subpopulations are independent of each other and hence $E(p_{ki}p_{kj}) = E(p_{ki})E(p_{kj})$ for $i \neq j$. To find the expectation of p_{ki}'s one assumes the RUZ (Random Union of Zygotes) model for the mating within each cytonuclear subpopulation (Fu and Arnold, 1992). For each subpopulation under the random drift model the conditional moment generating function can be written as

$$\begin{aligned} M(\theta_1, \ldots, \theta_6) &= E[exp\ \{\sum_k \theta_k p_{ki}(t)\} | p_{ki}(t-1)] \\ &= \sum_{f,m} Pr(Y_{fmi}(t)) exp\ (\ \sum_k \theta_k\ p_{ki}(t)\ ), \\ &= (\sum_{f,m} e_{fi}(t-1)\ e_{mi}(t-1) exp(\sum_k \theta_k \alpha_{fmk}/n_i(t)))^{n_i(t)} \end{aligned} \tag{3.2}$$

Here, i stands for the ith subpopulation. f, m stand for the father and the mother, respectively, and α_{fmk} are known constants (Datta *et al.*, 1996a). The count Y_{fmi} is the number of individuals in the ith subpopulation receiving gametes of type f from the father and type m from the mother. Each f and m can be one of four gametes A/M, A/m, a/M or a/m, and e_{fi}, e_{mi} are the gametic frequencies in fathers and mothers respectively, for the ith subpopulation. For details, we refer to Datta *et al.* (1996a).

From the above moment generating function one can find the expected values of the genotypic frequencies p_{ki} and also D_{1i} within a subpopulation. To that end, Datta and Arnold (1998) defined the following variables:

$$\begin{aligned} &x_1 = D,\ x_2 = Dp,\ x_3 = pq,\ x_4 = p^2q, \\ &x_5 = p^2,\ x_6 = q,\ x_7 = p. \end{aligned}$$

One can then verify from the moment generating function (3.2) that

$$\underline{X}(t) = \mathbf{A}(t)\underline{X}(t-1), \tag{3.3}$$

where the $\underline{X} = X_1, \ldots, X_7)$ is a vector containing the expectations of the x's mentioned above and $\mathbf{A}(t)$ is a 7×7 matrix whose nonzero elements are polynomials of the subpopulation sizes at time t, $n_i(t)$. It is easy to solve the linear recursion (3.3) and obtain the X's at a given time t knowing their initial values (at $t = 0$). Moreover, Datta and Arnold (1998) noted that the expectations of all the p_{ki} and D_{1i} can be written in terms of X's. Consequently, from (3.1),

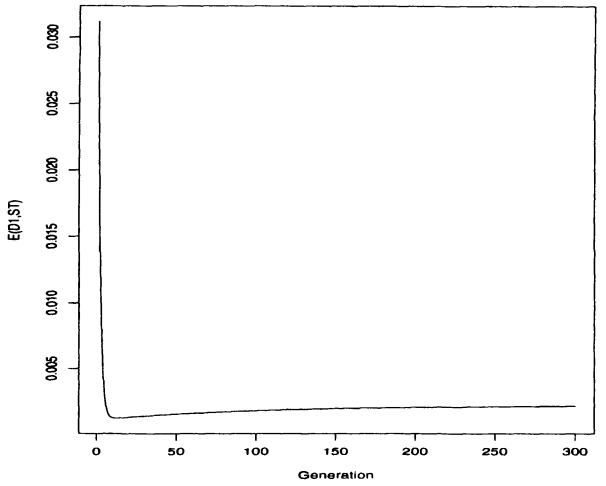

FIG. 1. *Trajectory of the expectation of overall cytonuclear disequilibria, $E(D_{1,ST})$, over time for a population consisting of 10 subpopulations. The initial values are p(0) = q(0) = 0.25, D(0) = 0.125.*

the expectations of the overall cytonuclear disequilibria $D_{1,ST}$ and $D_{2,ST}$ can be expressed in terms of the X's.

If we assume that the sizes of the subpopulations remain constant over all generations and they are the same for all the subpopulations, then the above expressions become somewhat simpler. Graphs of $E(D_{1,ST})$ and $E(D_{2,ST})$ for this special case are given in Figures 1 and 2, respectively. We choose the initial values $p(0) = 0.25$, $q(0) = 0.25$, $D(0) = 0.125$ which are taken to be the same for all the subpopulations. In both the figures the number of subpopulations $n = 10$ and we consider that all the subpopulations are of the same size $s = 50$, which remains constant over the generations. In Figure 1, we see that the expectation of $D_{1,ST}$ drops rapidly in the first few generations and afterwards it steadily approaches its asymptotic limit given by (4.2) in Section 4. The nonzero asymptotic value is 0.0022058, for $n = 10$. In Figure 2, we see that $E(D_{2,ST})$ approaches zero (its asymptotic value, (see (4.3)) quite fast even in a subdivided population.

If the population is not subdivided i.e., $n = 1$ then the recursions for the expectations of the cytonuclear disequilibria $E(D_1)$ and $E(D_2)$ over the generations are given in Datta *et al.* (1996a). It is shown that, unlike in the presence of the subdivision in the population, the undivided population reaches equilibrium after only a few generations and all the expected disequilibria measures approach zero after just a few generations in the presence of just random drift.

For the subdivided population, the exact formulas for the variance function

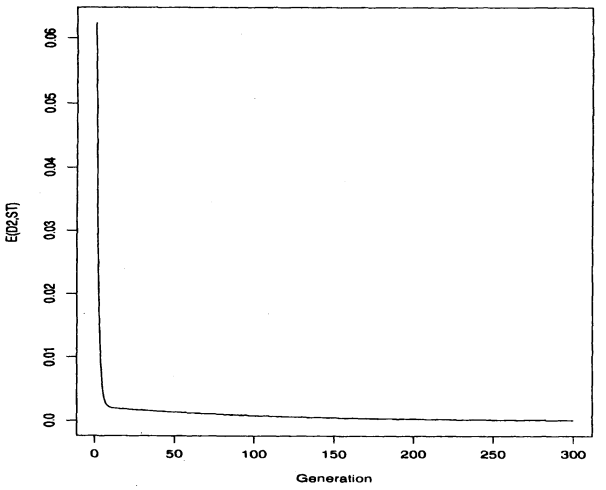

FIG. 2. *Trajectory of the expectation of overall cytonuclear disequilibria, $E(D_{2,ST})$, over time for a population consisting of 10 subpopulations. The initial values are p(0) = q(0) = 0.25, D(0) = 0.125.*

over time (generation) of the overall disequilibria are likely to be extremely complicated even for the case when all the subpopulations are of the same size and have the same initial condition. Therefore, we only report the results of a Monte Carlo simulation to describe the approximate trajectory of the overall variances of the disequilibria $D_{1,ST}$ and $D_{2,ST}$. We have used a smoothed version of the simulated variance curve to partially remove the simulation error. The graphs of the trajectories are given in Figures 3 and 4, respectively. For both the trajectories the number of subpopulations $n = 10$ and we assume that all the subpopulations are of the same size $s = 50$, which remains constant over generations. From Figure 5, we can see that $Var(D_{1,ST})$ decreases for just a few initial generations and after that it increases and eventually it converges to its predicted asymptotic value given in equation (4.2) in the next section. The pattern remains the same for different number of subpopulations (result not shown). It can also be shown that the magnitude of $Var(D_{1,ST})$ remains higher for smaller number of subdivisions consistently for all the generations (Datta and Arnold, 1998). Consequently, the asymptotic values are also higher for smaller number of subdivisions (Datta and Arnold, 1998). In Figure 6, we observe that unlike $Var(D_{1,ST})$, $Var(D_{2,ST})$ increases for a few initial generations and after that decreases and converges to zero. The pattern remains the same for different numbers of subdivisions (Datta and Arnold, 1998).

When the population is undivided, i.e., $n = 1$, the exact calculation of the variance under the random drift model is shown in Datta *et al.* (1996a). It is

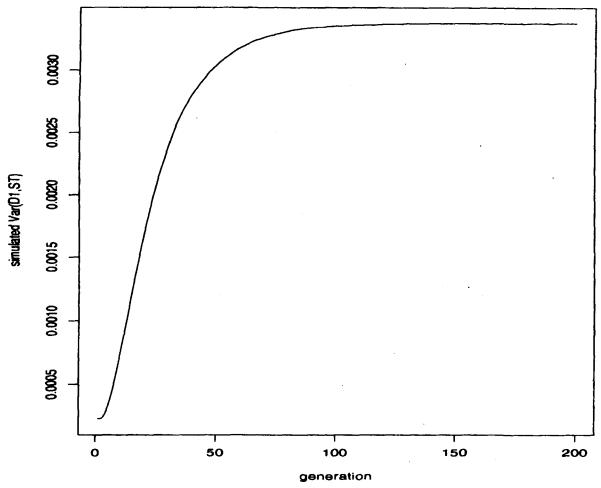

FIG. 3. *Trajectories of the simulated variance of overall cytonuclear disequilibria* $Var(D_{1,ST})$ *over 200 generations for a population consisting of 10 subpopulations. The initial values are* $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

clear from the graphs of the variances that both the variances go to zero after a few generations (Datta *et al.*, 1996a).

3.2. *Random drift with mutation.* In this section we will consider methods to calculate the expected value of the overall cytonuclear disequilibrium $E(D_{1,ST})$ under the random drift model in the presence of mutation.

Suppose the mutation rate at the nuclear locus from A to a is μ_1, a to A is ν_1, and at the mtDNA locus is μ_2 for M to m and ν_2 for m to M, respectively. Following the same argument as in Ohta and Kimura (1969), we have

$$p_m(t) = (1 - \mu_1 - \nu_1)p(t) + \nu_1, \quad q_m(t) = (1 - \mu_2 - \nu_2)q(t) + \nu_2,$$

$$D_m(t) = (1 - \mu)D(t),$$

where

$$\mu = \mu_1 + \nu_1 + \mu_2 + \nu_2,$$

and the higher order terms are ignored. The recursions of the expectations $E(D_{1,ST})$ and $E(D_{2,ST})$ can be found by solving the recursive relationship given below.

$$X_m^\star(t) = (\prod_{j=1}^{t} H^\star A^\star(t - j + 1))X_m^\star(0)$$

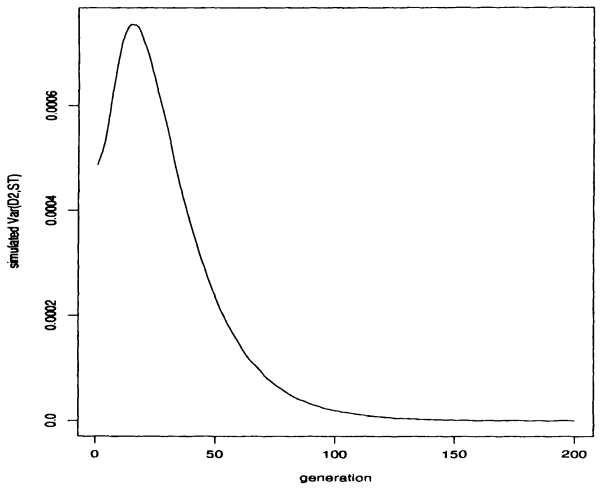

FIG. 4. *Trajectories of the simulated variance of overall cytonuclear disequilibria* $Var(D_{2,ST})$ *over 200 generations for a population consisting of 10 subpopulations. The initial values are* $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

where $X_m^\star$ are the values of the X vector discussed in Section 2 in the presence of mutation. $H^\star$ and $A^\star$ are 8×8 matrices. For the details, see Datta and Arnold (1998).

We draw the trajectories of the expected values of expected values of $D_{1,ST}$ and $D_{2,ST}$ in Figures 5 and 6 respectively. From Figure 5, we notice that in the presence of mutation, the expectation of the overall D_1 decays down to zero eventually, although it takes a long time for the mutation to remove the non zero asymptotic value of the expectation under random drift alone. Hence the rate of decay under the mutation could be extremely slow. In Figure 5, we find that when the mutation rate is larger, the rate of decay is faster, as to be expected. In the case of $E(D_{2,ST})$ however, even in the presence of mutation the value converges to zero much faster than $E(D_{1,ST})$ irrespective of the magnitude of the mutation rates (Figure 6).

Datta *et al.* (1996a) obtained the asymptotic result for the expected cytonuclear disequilibria in the presence of mutation if the population is not subdivided. It can be shown that all the steady state expectations are zero in this case.

4. Asymptotic results. In this section, we discuss the asymptotic results for the expected values and the variances of the cytonuclear disequilibria under the random drift model in the case of a subdivided population. Assume that all

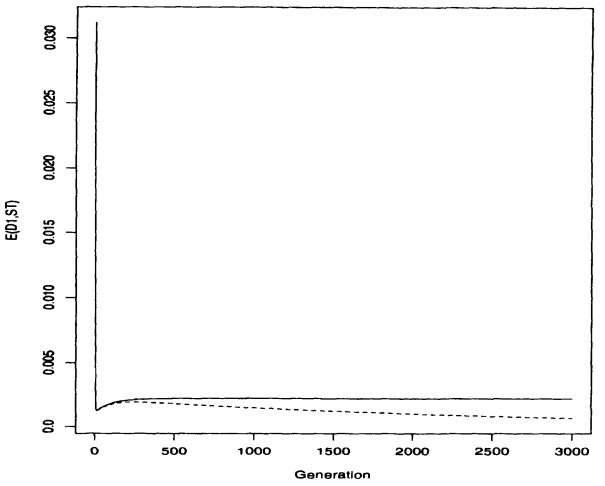

FIG. 5. *Trajectories of the expectation of overall cytonuclear disequilibria* $E(D_{1,ST})$ *over 3000 generations for 10 subpopulations with two different mutation rates. The solid line represents mutation rates of* $\mu_1 = \nu_1 = 10^{-6}$, $\mu_2 = \nu_2 = 10^{-6}$. *The dashed line represents mutation rates of* $\mu_1 = \nu_1 = 10^{-4}$, $\mu_2 = \nu_2 = 2 \times 10^{-4}$. *The initial values in all cases are* $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

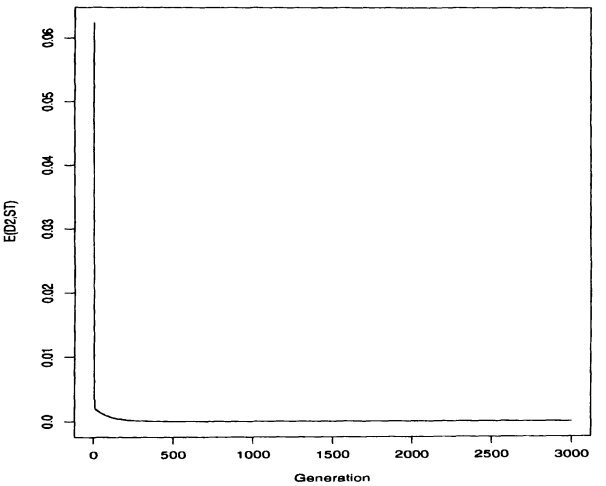

FIG. 6. *Trajectories of the expectation of overall cytonuclear disequilibria* $E(D_{2,ST})$ *over 3000 generations for 10 subpopulations for two different mutation rates. Sets of mutation rates are* $\mu_1 = \nu_1 = 10^{-4}$, $\mu_2 = \nu_2 = 2 \times 10^{-4}$, *and* $\mu_1 = \nu_1 = 10^{-6}$, $\mu_2 = \nu_2 = 10^{-6}$. *Trajectories are almost the same for both sets of rates. The initial values in all cases are* $p(0) = q(0) = 0.25$, $D(0) = 0.125$.

the subpopulations are of the same constant (in time) size s and they all have the same initial conditions. Note that, eventually, each of the subpopulations will be fixed for one of the four possible genotypes p_1, p_4, p_3 and p_6. Let the corresponding probabilities be g_1, g_4, g_3, g_6, respectively. The values of g's can be determined from the initial values $D(0)$, $p(0)$ and $q(0)$ which are assumed to be the same for all the subpopulations. Thus, when all the subpopulations have reached fixation, sampling from n subpopulations is equivalent to taking a sample of size n from a multinomial distribution with parameters g_1, g_4, g_3, g_6. From the moment generating function of the multinomial distribution we can find

$$E(D_{1,ST}(\infty)) = (\frac{n-1}{n})D^*(0), \tag{4.1}$$

$$\begin{aligned} Var(D_{1,ST}(\infty)) = (\frac{n-1}{n^3})\{&(n-1)D^*(0) - 2D^*(0)^2(n-1) \\ &- 2D^*(0)p(0)(n-1) - 2D^*(0)q(0)(n-1) \\ &- 4D^*(0)p(0)q(0)(n-1) + np(0) - np^2(0)q(0) \\ &- np(0)q^2(0) + np^2(0)q^2(0)\}, \end{aligned} \tag{4.2}$$

where

$$D^*(0) = (\frac{1}{s+1})D(0).$$

Note that when $n = 1$, i.e, when the population is not subdivided then both the above quantities are zero.

Since at time $t = \infty$, each of the sub-populations will be fixed at one of the four genotypes mentioned above, $P_2 = P_5 = 0$. Therefore, asymptotically $D_{2,ST} = 0$ and $D_{1,ST} = D_{ST}$, with probability one. Indeed it can be checked that the above asymptotic expressions for $D_{1,ST}$ agrees with those for D_{ST} as given in Fu and Arnold (1991). Furthermore,

$$E(D_{2,ST}(\infty)) = 0;\ Var(D_{2,ST}(\infty)) = 0. \tag{4.3}$$

5. Statistical tests based on the dynamics of cytonuclear disequilibria. One can use the formulas for the expectation and the variance for D_1 and D_2 over time to construct a goodness of fit type statistical test which assesses departure from a given model. In particular, using the moment formulas under random drift, this approach yields a test of the neutrality hypothesis using the dynamics of cytonuclear disequilibria. A number of tests following this idea can be constructed depending on the sampling scheme used to obtain the necessary data.

5.1. *Test for random drift in undivided populations.* Consider a single undivided population. In this case, the formulas for the expectations and variances under the random drift model can be found by specializing to $n = 1$ in the general formulas for a subdivided population or directly from the results in Datta *et al.* (1996a). The above paper also shows how to calculate the covariance between D_1 and D_2 at a given time following the approach described in Section 3.1.

Suppose, data are available on a number of populations of the same species. We assume that the initial conditions are known for each and that the ith population is completely sacrificed at time i so that one can calculate $D_1(t)$ and $D_2(t)$ only for $t = i$ for this population. For example, this sampling scheme was used by Scribner and Avise (1994). In this scheme, the statistics for different generations are independent of each other. Note that under this sampling scheme the counts are based on a complete census of the population and there is no sampling variability (only the genetic variability). Letting $\underline{D}(t) = (D_1(t), D_2(t))$, $\underline{\mu}(t) = (ED_1(t), ED_2(t))$, and

$$\Sigma(t) = \begin{pmatrix} Var(D_1(t)), & Cov(D_1(t), D_2(t)) \\ Cov(D_1(t), D_2(t)), & Var(D_2(t)) \end{pmatrix}$$

where all the moments are calculated under the random drift model, one can construct a test statistic

$$T = \sum_{t=1}^{k} (\underline{D}(t) - \underline{\mu}(t))^T \Sigma^{-1}(t) (\underline{D}(t) - \underline{\mu}(t))$$

which measures the total distance across time between the observed and the expected disequilibria under random drift. The asymptotic null distribution of the above statistic was shown to be chi-square with $2k$ degrees of freedom by Datta and Arnold (1996). Therefore one would reject the random drift model if $T > \chi^2_\alpha(2k)$. Datta *et al.* (1996b) proposed a test along the same line under a different sampling scheme which results in both sampling as well as genetic variation.

5.2. *Tests of random drift in subdivided populations.* Consider a subdivided population with n components which had identical initial conditions. Then in the notation of Section 2, $ED_{1i}(t)$ and $ED_{2i}(t)$ are constant in i, up to terms that are $O(n_i^{-1})$. Denote the common value by $\mu_1(t)$ and $\mu_2(t)$ respectively. Suppose, the entire population is sacrificed at time t so that one can calculate $D_{1i}(t)$ and $D_{2i}(t)$ for $i = 1, \ldots, n$. Combining these disequilibria measures from

each subpopulation one can construct an overall estimate $\hat{\mu}_j(t), j = 1,2$ by

$$\hat{\mu}_j(t) = g_j(\sum_{i=1}^{n} \hat{v}_{ij}\lambda_i g_j^{-1}(D_{ij}(t))/\sum_{i=1}^{n} \hat{v}_{ij}\lambda_i)),$$

where g_j is a smooth function, $\lambda_i = n_i/N$, v_{ij} is the variance of D_{ij}. For example, g_j could be such that $g_j(u) = \sqrt{u}$ for $j = 1,2$. Note that unlike $D_{j,ST}(t)$, $\hat{\mu}_j(t)$ is both asymptotically (as n_i grows) unbiased as well as efficient. A test of random drift based on $\underline{\hat{\mu}}(t)$ can be constructed using the statistic

$$T = (\underline{\hat{\mu}}(t) - \underline{\mu}(t))^T \Sigma^{-1}(t)(\underline{\hat{\mu}}(t) - \underline{\mu}(t))$$

where

$$\Sigma(t) = \begin{pmatrix} Var(\hat{\mu}_1(t)), & Cov(\hat{\mu}_1(t), \hat{\mu}_2(t)) \\ Cov(\hat{\mu}_1(t), \hat{\mu}_2(t)), & Var(\hat{\mu}_2(t)) \end{pmatrix}.$$

Note that Σ can be calculated from the variance covariance formulas for $D_{1i}(t)$, $D_{2i}(t)$ via the delta method and the independence among subpopulations. Moreover it is possible to show that T has an approximate chi-square distribution with two degrees of freedom.

If one has data from a number of independent subdivided populations as in the previous subsection then one can obtain an overall test statistic by adding the T's obtained from each subdivided population.

6. Discussion. In this paper, we have reviewed some recent results on the exact dynamics of the expectation of two measures of overall cytonuclear disequilibrium, $D_{1,ST}$ and $D_{2,ST}$, in a subdivided population with n demes under genetic drift and genetic drift with mutation. The variance curves of these disequilibria measures are also studied and construction of statistical tests using the above results are discussed.

It is an established fact that genetic drift generates variation in linkage disequilibrium between two or more genetic loci in a subdivided population (Ohta, 1982a; Ohta, 1982b) and that mutation eventually eliminates that permanent disequilibria caused by genetic drift. In this paper, in the absence of mutation we see that the expectation of cytonuclear disequilibrium $E(D_{1,ST})$ eventually goes to a nonzero asymptotic value $[(n-1)/n]D^*(0)$, ($D^*(0)$ is defined in Section 4) which is the same as the asymptotic value of the allelic disequilibrium given in Fu and Arnold (1991). However, the exact dynamics of $E(D_{1,ST})$ are different from that of the expected allelic disequilibrium. Initially, both the within-subpopulation and between-subpopulation component of the disequilibria decrease very rapidly. Then the between-subpopulation component starts

increasing slowly and it finally settles down for the between-population component $[(n-1)/n]D(0)/(s+1)$. An increased subdivision (larger n, the number of subpopulations) increases the value of $E(D_{1,ST})$ (Datta and Arnold, 1998) and consequently yields a higher steady-state value. The between-population component of $E(D_{1,ST})$ is always a little larger for larger n. It is easy to see that increasing $D(0)$ and/or decreasing s will result in larger steady-state values. In the presence of mutation the permanent association is eventually removed and $E(D_{1,ST})$ decreases to zero, but it may take a very long time (Figure 2). Note that higher mutation rates imply a shorter time till the expectation of $D_{1,ST}$ goes to zero (Figure 2).

Unlike the expectation of $D_{1,ST}$, expectation of $D_{2,ST}$ goes down to zero (its asymptotic value) quite fast under the random drift model. It is interesting to see that the between-population component of $D_{2,ST}$ also goes down to zero for the heterozygote. Even with the presence of mutation it goes down to zero quite fast irrespective of the different mutation rates. Genetic drift weeds out the nuclear heterozygotes Aa/M and Aa/m, and no 'F_1 hybrids' remain to generate a nonzero D_2.

For an undivided single population both the disequilibria measures eventually go to zero under the random drift model and in the presence of mutation along with random drift. The variance of cytonuclear disequilibria decays asymptotically under the random drift model if there is no other source of variability like mutation or migration. In the presence of mutation it has non-zero asymptotic value. For the detailed discussion on this matter we refer to Datta *et al.* (1996a).

For a subdivided population, trajectory of the simulated variance of the disequilibrium for homozygotes $Var(D_{1,ST})$ under the random drift model is shown in Figure 5. In the initial generations, value of the variance is low but it increases quite rapidly under the random drift to reach the predicted non-zero steady-state value as expected. We have discussed in Section 4 that asymptotically $D_{1,ST}$ and D_{ST} are the same. Hence the steady-state variance is also the same. The asymptotic value of the variance $Var(D_{1,ST})$ increases as number of subpopulations decrease (Datta and Arnold, 1998). Variance of $D_{2,ST}$ (Figure 6) on the other hand goes down to zero within the first 150 generations approximately for all the different number of subpopulations and that is what one would expect, because under the random drift overall disequilibria due to the heterozygotes goes to zero. The nonzero values of the variance $Var(D_{2,ST})$ are higher for smaller number of subdivisions (Datta and Arnold, 1998).

Clearly the results in this paper offer more insight into the behavior of a subdivided system under a neutral model than those using just the allelic disequilibrium. This will be reflected in the additional statistical power if one

constructs a test comparing the observed dynamics of the pair $(D_{1,ST}, D_{2,ST})$ with the expected dynamics under a neutral model. Such neutrality tests in the case of an undivided population have been proposed by Datta and Arnold (1996) and Datta *et al.* (1996b). An extension of the Datta and Arnold (1996) test to a subdivided population has been briefly described in Section 5.2 of this paper. This test is applicable if the data are collected in an ideal experimental setting as described in Section 5.2. In the case of an undivided population, data using such a sampling scheme have been collected by Scribner and Avise (1994). An extension of the test to handle a more practical setup incorporating genetic as well as statistical sampling can also be done.

Acknowledgments. I would like to thank Professor P. K. Sen for some useful discussions. Thanks are also due to two anonymous referees for their critical reviews which have improved the manuscript to a great extent.

REFERENCES

ARNOLD, J. (1993). Cytonuclear Disequilibria in hybrid zones. *Annu. Rev. Ecol. Syst.* **24** 521–554.

ASMUSSEN, M. A., ARNOLD, J. and ARNOLD, J. (1987). Definition and properties of disequilibrium statistics for associations between nuclear and cytoplasmic genotypes. *Genetics* **115** 1351–1363.

BIRKEY, C. W., FUERST, P. and MARUYAMA, T. (1989). Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. *Genetics* **121** 613–627

DATTA, S. and ARNOLD, J. (1996). Diagnostics and a Statistical test of neutrality hypotheses using the dynamics of cytonuclear disequilibria. *Biometrics* **52** 1042–1054.

DATTA, S. and ARNOLD, J. (1998). Dynamics of cytonuclear disequilibria in subdivided populations. *Journal of Theoretical Biology*, **192** 99–111.

DATTA, S., FU, Y. X. and ARNOLD, J. (1996a). Dynamics and equilibrium behavior of cytonuclear disequilibria under genetic drift. *Theor. Pop. Biol.* **50** 298–324.

DATTA, S., KIPARSKY, M., RAND, D. M. and ARNOLD, J. (1996b). A statistical test of a neutral model using the dynamics of cytonuclear disequilibria. *Genetics* **144** 1985–1992.

FERRIS, S. D., SAGE, R. D., HUANG, C. M., NIELSEN, J. T., RITTE, U. and WILSON, A. C. (1983). Flow of mitochondrial DNA across a species boundary. *Proc. Natl. Acad. Sci. USA* **80** 2290–2294.

FU, Y. X. and ARNOLD, J. (1991). On the association of fragment length polymorphisms across species boundaries. *Proc. Natl. Acad. Sci. USA* **88** 3967–3971.

FU, Y. X. and ARNOLD, J. (1992). Dynamics of cytonuclear disequilibria in finite populations and a comparison with a two-locus nuclear system. *Theor. Popul. Biol.* **41**, 1–25.

LAMB, T. and AVISE, J. C. (1986). Directional introgression of mitochondrial DNA in a hybrid population of tree frogs: the influence of mating behavior. *Proc. Natl. Acad. Sci. USA* **83** 2526–2530.

MALLET, J., BARTON, N., LAMAS, G., SANTISTEBAN, J., MUEDAS, M. and EELEY, H. (1990). Estimates of cline width and linkage disequilibrium in *Helicious* hybrid zones. *Genetics* **124** 921–936.

OHTA, T. and KIMURA, M. (1969). Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63** 229–238.

OHTA, T. (1982). Linkage disequilibrium with the island model. *Genetics* **101** 139–155.

OHTA, T. (1982). Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci. USA* **79** 1940–1944.

RAND, D. M. and HARRISON, R. G. (1989). Ecological genetics of a mosaic hybrid zone: mitochondrial, nuclear, and reproductive differentiation by soil type. *Evolution* **43** 432–449.

SCRIBNER, K. T. and AVISE, J. C. (1994). Population cage experiments with a vertebrate: genetics of hybridizaion in *Gambusia* fishes. *Evolution* **48** 155–171.

SPOLSKY, C. and UZZEL, T. (1984). Natural interspecies transfer of mitochondrial DNA in amphibians. *Proc. Natl. Acad. Sci. USA*. **81** 5802–5805.

WRIGHT, S. (1968). *Evolutions and Genetics of Populations, Vol II. The Theory of Gene Frequencies.* University Press of Chicago.

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE
GEORGIA STATE UNIVERSITY
UNIVERSITY PLAZA
ATLANTA GA 30303
SDATTA@CS.GSU.EDU