The Exact Test for Cytonuclear Disequilibria

Christopher J. Basten* and Marjorie A. Asmussen†

* *Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203 and*
† *Department of Genetics, University of Georgia, Athens, Georgia 30602-7223*

Manuscript received October 11, 1996
Accepted for publication March 19, 1997

ABSTRACT

We extend the analysis of the statistical properties of cytonuclear disequilibria in two major ways. First, we develop the asymptotic sampling theory for the nonrandom associations between the alleles at a haploid cytoplasmic locus and the alleles and genotypes at a diploid nuclear locus, when there are an arbitrary number of alleles at each marker. This includes the derivation of the maximum likelihood estimators and their sampling variances for each disequilibrium measure, together with simple tests of the null hypothesis of no disequilibrium. In addition to these new asymptotic tests, we provide the first implementation of Fisher's exact test for the genotypic cytonuclear disequilibria and some approximations of the exact test. We also outline an exact test for allelic cytonuclear disequilibria in multiallelic systems. An exact test should be used for data sets when either the marginal frequencies are extreme or the sample size is small. The utility of this new sampling theory is illustrated through applications to recent nuclear-mtDNA and nuclear-cpDNA data sets. The results also apply to population surveys of nuclear loci in conjunction with markers in cytoplasmically inherited microorganisms.

JOINT nuclear-cytoplasmic data are a valuable tool for deducing the evolutionary history of natural populations. Their potential utility has been affirmed by the growing theoretical framework showing how such data can provide estimates of the rates of nonrandom mating and gene flow in hybrid zones (ARNOLD *et al.* 1988; ASMUSSEN *et al.* 1989) as well as a means to detect and estimate migration, admixture, population subdivision, selection, mutation and genetic drift in more general contexts (ASMUSSEN and ARNOLD 1991; FU and ARNOLD 1991, 1992a; BABCOCK and ASMUSSEN 1996; DATTA and ARNOLD 1996; DATTA *et al.* 1996a,b). The methodology is also in place for using cytonuclear data from plant populations to decompose gene flow into diploid (seed) and haploid (pollen) components (ASMUSSEN and SCHNABEL 1991; SCHNABEL and ASMUSSEN 1992; M. E. ORIVE and M. A. ASMUSSEN, unpublished results), another critical task that has proven difficult by traditional methods based on nuclear or cytoplasmic data alone. Even more importantly, the special utility of the new cytonuclear-based methods in these and other biological situations is now being exploited in many empirical applications to natural populations (*e.g.*, ASMUSSEN *et al.* 1989; AVISE *et al.* 1990; FORBES and ALLENDORF 1991; PAIGE *et al.* 1991; CRUZAN and ARNOLD 1993, 1994; SCRIBNER and AVISE 1993, 1994; ABERNETHY 1994; SITES *et al.* 1996; AVISE *et al.* 1997).

Much of the novel information from this new class of data is encoded within the observed pattern of nonrandom associations between the cytoplasmic types (cytotypes) and the alleles and genotypes at the nuclear marker. These associations can be quantified by formal measures of cytonuclear disequilibria in two-locus systems with either diallelic (ASMUSSEN *et al.* 1987) or multiallelic (ASMUSSEN and BASTEN 1996) markers, as well as in three-locus, nuclear-dicytoplasmic systems (SCHNABEL and ASMUSSEN 1989). In the basic, two-locus systems additional biological insight can be gained by calculating normalized disequilibrium measures, which take into account the constraints imposed on the cytonuclear associations by the marginal frequencies at the two markers (ASMUSSEN and BASTEN 1996).

The present paper further facilitates practical applications by extending the formal statistical guidelines for the proper experimental design and analysis of standard, two-locus cytonuclear surveys. As the first step to this end, ASMUSSEN and BASTEN (1994) developed maximum likelihood disequilibrium estimators and the asymptotic sampling theory for the original diallelic measures defined by ASMUSSEN *et al.* (1987). The latter includes analytic formulas for the approximate standard errors in the estimators, test statistics for the null hypothesis of random associations, and approximate minimum sample sizes for their detection (ASMUSSEN and BASTEN 1994). These asymptotic results provide satisfactory approximations as long as the sample size is reasonably large (*e.g.*, 100 or more), and the marginal genotypic and allelic frequencies are intermediate (*e.g.*, between 0.2 and 0.8). If the sample size is small or the marginal frequencies extreme, exact tests should instead be employed. FU and ARNOLD (1992b) pro-

Corresponding address: Christopher J. Basten, Department of Statistics, North Carolina State University, 220 Patterson Hall, Hillsborough St., Raleigh, NC 27695-8203.
E-mail: basten@statgen.ncsu.edu

vided such a test for allelic associations based on Fisher's exact test for 2 × 2 tables, and DEAN and ARNOLD (1996) have recently calculated the exact variances for the genotypic associations. All these previous efforts were restricted to diallelic markers. Here, we fill the remaining gaps in the statistical theory for two-locus cytonuclear disequilibria by deriving maximum likelihood estimators and their asymptotic properties for the cytonuclear disequilibria between multiallelic markers and by developing exact tests for genotypic cytonuclear associations. Finally, we extend these procedures to the allelic cytonuclear disequilibria in multiallelic systems.

GENERAL DEVELOPMENT

Basic cytonuclear system: We examine a system with one nuclear and one cytoplasmic locus, and allow for multiple alleles at both. Following ASMUSSEN and BASTEN (1996), we assume there are r alleles A_i, $i = 1, r$ at the nuclear locus and m cytotypes M_k, $k = 1, m$. The populational frequency of the A_i allele is written as P^i, while that of each nuclear genotype A_iA_j is P^{ij}. The frequency of individuals carrying the M_k cytotype will be denoted P_k. In general, nuclear genotypic and allelic frequencies will be indicated by superscripts while cytotypes are indicated by subscripts. The joint frequency of A_iA_j/M_k individuals is P_k^{ij}. Since we choose not to order the nuclear alleles, $P_k^{ij} = P_k^{ji}$ and $P^{ij} = P^{ji}$ for $j \neq i$. The genotypic and allelic frequencies are related to the joint cytonuclear frequencies as follows:

$$P^{ij} = \sum_{k=1}^{m} P_k^{ij}, \tag{1}$$

$$P^i = P^{ii} + \tfrac{1}{2} \sum_{j \neq i} P^{ij}, \tag{2}$$

$$P_k = \sum_{i=1}^{r} \sum_{j=i}^{r} P_k^{ij}, \tag{3}$$

$$P_k^i = P_k^{ii} + \tfrac{1}{2} \sum_{j \neq i} P_k^{ij}. \tag{4}$$

The final measure, P_k^i, intuitively corresponds to the frequency of A_iM_k gametes. Formally, it is the probability that a randomly sampled individual from the population has the M_k cytotype, and a randomly sampled nuclear allele from that individual is A_i.

Disequilibrium measures: We focus on the two types of cytonuclear associations defined by ASMUSSEN and BASTEN (1996). The first are the allelic disequilibria

$$D_k^i = P_k^i - P^iP_k, \tag{5}$$

which are analogous to the standard gametic disequilibria for a two-locus nuclear system and measure nonrandom associations between each nuclear allele A_i and each cytotype M_k. We also consider the genotypic disequilibria

$$D_k^{ij} = P_k^{ij} - P^{ij}P_k, \tag{6}$$

TABLE 1

Observed counts of multiallelic cytonuclear genotypes

	Nuclear genotype					
Cytotype	A_1A_1	⋯	A_iA_j	⋯	A_rA_r	Total
M_1	n_1^{11}	⋯	n_1^{ij}	⋯	n_1^{rr}	n_1
⋮		⋱				
M_k	n_k^{11}	⋯	n_k^{ij}	⋯	n_k^{rr}	n_k
⋮				⋱		
M_m	n_m^{11}	⋯	n_m^{ij}	⋯	n_m^{rr}	n_m
Total	n^{11}	⋯	n_{ij}	⋯	n_{rr}	n

which similarly measure nonrandom associations between each nuclear genotype A_iA_j and each cytotype M_k. D_k^i is identical to D and D_k^{ij} to D_1, D_2 and D_3 of ASMUSSEN *et al.* (1987) for the diallelic case where $A_1 = A$, $A_2 = a$, $M_1 = M$ and $M_2 = m$. The interrelationships among these disequilibria and the constraints placed upon them by the marginal frequencies have been given previously (ASMUSSEN and BASTEN 1996).

Maximum likelihood estimators: The data will consist of a set of joint cytonuclear genotypic counts

$$\{n_k^{ij}\} = [[[n_k^{ij}]_{k=1}^{m}]_{i=1}^{r}]_{j=i}^{r} \tag{7}$$

as shown in Table 1, where n_k^{ij} is the number of A_iA_j/M_k individuals in the sample of size n. The array of counts in (7) requires $j \geq i$, but for convenience in the formulas below, we will assume that $n_k^{ij} = n_k^{ji}$ for $j < i$. If the sample size is much smaller than the total population size, then these counts will be multinomially distributed. The maximum likelihood estimators for the cytonuclear disequilibria are then

$$\tilde{D}_k^{ij} = \tilde{P}_k^{ij} - \tilde{P}^{ij}\tilde{P}_k \tag{8}$$

$$\tilde{D}_k^i = \tilde{P}_k^i - \tilde{P}^i\tilde{P}_k, \tag{9}$$

where the frequency estimators are as follows:

$$\tilde{P}_k^{ij} = \frac{1}{n} n_k^{ij}, \tag{10}$$

$$\tilde{P}^{ij} = \frac{1}{n} \sum_{k=1}^{m} n_k^{ij}, \tag{11}$$

$$\tilde{P}_k = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=i}^{r} n_k^{ij}, \tag{12}$$

$$\tilde{P}_k^i = \frac{1}{2n} \left(2n_k^{ii} + \sum_{j \neq i} n_k^{ij}\right), \tag{13}$$

$$\tilde{P}^i = \frac{1}{2n} \sum_{k=1}^{m} \left(2n_k^{ii} + \sum_{j \neq i} n_k^{ij}\right). \tag{14}$$

Sampling properties of the disequilibrium estimators: The expected value of each of the cytonuclear disequilibrium estimators, $\tilde{D}_k^{ij}$ and $\tilde{D}_k^i$, has the form

$$\mathcal{E}\tilde{D} = \left(1 - \frac{1}{n}\right)D, \tag{15}$$

indicating a slight bias in the estimators. [The details of the derivation are similar to those for the diallelic case outlined in Appendix A of ASMUSSEN and BASTEN (1994).] The approximate, large sample variances, which are generally sufficient for the asymptotic tests in the next section, are readily found via the Delta method (WEIR 1996) to be

$$\text{Var}\,(\tilde{D}_k^{ij}) \approx \frac{1}{n}\,[P^{ij}(1 - P^{ij})P_k(1 - P_k) + D_k^{ij}(1 - 2P^{ij})(1 - 2P_k) - (D_k^{ij})^2] \tag{16}$$

$$\text{Var}\,(\tilde{D}_k^{i}) \approx \frac{1}{2n}\,[P^{i}(1 - P^{i})P_k(1 - P_k) + D^{ii}P_k(1 - P_k) + D_k^{ii}(1 - 2P_k) + D_k^{i}(1 - 4P^{i})(1 - 2P_k) - 2(D_k^{i})^2], \tag{17}$$

where $D^{ii} = P^{ii} - (P^i)^2$ is the Hardy-Weinberg disequilibrium for the nuclear genotype A_iA_i (WEIR 1996). For two alleles at a locus, these large sample variances reduce to the variances given in ASMUSSEN and BASTEN (1994). Note that these sampling properties apply to data from a single population rather than over replicate populations.

Asymptotic tests: The asymptotic tests are set up similarly to those for the diallelic measures in ASMUSSEN and BASTEN (1994). For example, to test $H_0 : D_k^{ij} = 0$ we have

$$\tilde{r} = \frac{\tilde{D}_k^{ij}}{\sqrt{\tilde{P}^{ij}(1 - \tilde{P}^{ij})\tilde{P}_k(1 - \tilde{P}_k)}}, \tag{18}$$

where to test $H_0 : D_k^i = 0$ we have

$$\tilde{r} = \frac{\sqrt{2}\tilde{D}_k^i}{\sqrt{\tilde{P}^i(1 - \tilde{P}^i)\tilde{P}_k(1 - \tilde{P}_k) + \tilde{D}^{ii}\tilde{P}_k(1 - \tilde{P}_k) + \tilde{D}_k^{ii}(1 - 2\tilde{P}_k)}}. \tag{19}$$

For a test at the 0.05 significance level, we reject the null hypothesis that $D = 0$ if $n\tilde{r}^2 > 3.84$, where $n\tilde{r}^2$ has an approximately $\chi^2(1)$ distribution. We first test for the Hardy-Weinberg disequilibria, and then proceed to the cytonuclear genotypic, and lastly, the allelic disequilibria, following the order and rules developed in the two-allele case (ASMUSSEN and BASTEN 1994).

Exact tests: The traditional exact test of Fisher is performed for genotypic associations by enumerating all samples of size n such that the marginal counts $\{n^{ij}\} = [[n^{ij}]_{i=1}^{r}]_{j=i}^{r}$ and $\{n_k\} = [n_k]_{k=1}^{m}$ are the same as those observed, and then ordering these samples based on their probabilities under the hypothesis of no genotypic disequilibrium and conditioned on the observed marginal counts. The sum across all samples with probabilities less than or equal to that for the observed sample is the "exact" probability of the sample. When sample sizes are small, or the data matrix is sparse, then the asymptotic tests can be inaccurate, but exact tests tend to perform well. Two types of exact tests can be performed for cytonuclear genotypic associations. The first is whether the data set as a whole is consistent with the independent association of nuclear genotypes and cytotypes. The second is to test for disequilibria between specific combinations of nuclear genotypes and cytotypes. Even when the overall test fails to reject the hypothesis of independent association of nuclear genotypes and cytotypes, individual tests may show some disequilibria.

TABLE 2

The observed joint cytonuclear allelic counts

	Nuclear allele					
Cytotype	A_1	⋯	A_i	⋯	A_r	Total
M_1	n_1^1	⋯	n_1^i	⋯	n_1^r	$2n_1$
⋮		⋱				
M_k	n_k^1	⋯	n_k^i	⋯	n_k^r	$2n_k$
⋮				⋱		
M_m	n_m^1	⋯	n_m^i	⋯	n_m^r	$2n_m$
Total	n^1	⋯	n^i	⋯	n^r	$2n$

For the overall test of cytonuclear genotypic disequilibria, we require the conditional probability

$$P(\{n_k^{ij}\}|\{n^{ij}\}, \{n_k\}) = \frac{P(\{n_k^{ij}\})}{P(\{n^{ij}\}, \{n_k\})} = \frac{\prod_i \prod_{j \geq i} n^{ij}! \prod_k n_k!}{n! \prod_i \prod_{j \geq i} \prod_k n_k^{ij}!} \tag{20}$$

based on the assumption of no genotypic disequilibria. The analogous procedure for allelic cytonuclear associations is based on the $m \times r$ allelic data matrix of joint allelic counts in Table 2. Here $n_k^i = 2n\tilde{P}_k^i$, where $\tilde{P}_k^i$ is calculated from (13) and the sum of all counts is $2n$. The exact test is then based on the allelic analogue of (20)

$$P(\{n_k^i\}|\{n^i\}, \{n_k\}) = \frac{\prod_i n^i! \prod_k (2n_k)!}{(2n)! \prod_i \prod_k n_k^i!} \tag{21}$$

for all joint allelic samples, assuming no allelic disequilibria and conditioned on the observed marginal allelic counts, where $\{n^i\} = [n^i]_{i=1}^{r}$.

Individual disequilibria can be examined by collapsing the data into 2×2 tables and performing exact tests. Specifically, we test D_k^{ij} using the counts in Table 3 (which are derived from the data matrix in Table 1). For a specific allelic disequilibrium, D_k^i, we reduce the allelic data matrix in Table 2 in a similar fashion. These 2×2 exact tests can be performed with an efficient algorithm (ZAYKIN and PUDOVKIN 1993) for all pairwise comparisons of each cytotype with each nuclear genotype (or allele).

TABLE 3

The observed genotypic counts collapsed into a 2 × 2 array for testing the genotypic disequilibrium D_k^{ij}

	Nuclear genotype		
Cytotype	A_iA_j	Non-(A_iA_j)	Total
M_k	n_k^{ij}	$n_k - n_k^{ij}$	n_k
Non-M_k	$n^{ij} - n_k^{ij}$	$n - n_k - n^{ij} + n_k^{ij}$	$n - n_k$
Total	n^{ij}	$n - n^{ij}$	n

Approximations for the exact test: The exact tests seek to calculate the probabilities for all samples subject to fixed values of the marginal genotypic or allelic counts. Whereas exact tests are computationally feasible in the 2 × 2 tables for the individual disequilibria, often the exact tests for the full data set are not. For this reason we need approximations to the exact test for overall departures from random associations between nuclear genotypes (or alleles) and cytotypes. GUO and THOMPSON (1992) have summarized two approximations for the exact test for Hardy Weinberg disequilibria for multiallelic loci, which we have adapted to the cytonuclear context. We will refer to these approximations as the Monte Carlo and Markov chain methods. The Monte Carlo approach proceeds by shuffling the nuclear genotypes (or alleles) and cytotypes in a sample and recalculating the probability in (20) for genotypes and in (21) for alleles. The proportion of shuffled samples with a probability lower than or equal to the observed sample yields an approximation of the exact probability of the sample. The alternative, Markov chain method is based on a random walk across the set of samples with the same marginal counts as the observed sample. Our implementation of the Markov chain method is the special case of GUO and THOMPSON (1992, page 365) in which $\Delta = 0$; otherwise the details are identical.

The standard error for both approximations is obtained via a batching method (GUO and THOMPSON 1992) that divides the entire set of repetitions R into B batches of C consecutive observations ($R = B \times C$). The sample variance of the exact probability estimate, $\tilde{p}$, is

$$s_{\tilde{p}}^2 = \frac{1}{B(B-1)} \sum_{i=1}^{B} (\tilde{p}_i - \tilde{p})^2 = \frac{1}{B(B-1)} \left[\sum_{i=1}^{B} \tilde{p}_i^2 - \left(\sum_{i=1}^{B} \tilde{p}_i \right)^2 \Big/ B \right], \quad (22)$$

where $\tilde{p}_i$ is the estimate of the exact probability from the ith batch and $\tilde{p} = \Sigma_{i=1}^{B} \tilde{p}_i / B$. The second version of (22) is computationally useful: the sum and sum of squares can be calculated as the Monte Carlo or Markov chain approximation is run rather than storing the individual $\tilde{p}_i$ values. The standard error for the exact probability estimate ($\tilde{p}$) is simply the square root of (22).

TABLE 4

Application of the statistics to data on hybrid swarms of bluegill for nuclear system Got-2

Disequilibrium	D_M^{AA}	D_M^{Aa}	D_M^{aa}
Estimate (D_k^{ij})	0.0077	0.0325	−0.0402
Normalized estimate ($D_k^{ij\prime}$)	0.0885	0.1447	−0.3184
Standard error (H_0)	0.0158	0.0201	0.0173
Standard error (T)	0.0157	0.0199	0.0172
Test statistic ($n\tilde{r}^2$)	0.2391	2.6216	5.3997
Asymptotic probability	0.6249	0.1054	0.0201
Exact probability	0.6787	0.1379	0.0228
MSS ($\beta = 0.1$)	6605	602	293
MSS ($\beta = 0.5$)	2426	221	107

From AVISE *et al.* (1984).

NUMERICAL RESULTS AND DISCUSSION

When testing for cytonuclear disequilibria where there are multiple alleles, the first step is to test the entire data matrix of Table 1 or 3 for overall agreement with the hypothesis of row and column independence. Here due to the computational demands, the Monte Carlo or Markov chain approximations should be used. While both test the overall departure from random cytonuclear associations, they fail to distinguish the contributions to nonrandom assortments from the different combinations of the nuclear component (genotypes or alleles) and cytotypes. These can be tested individually by collapsing the data into 2 × 2 tables and applying asymptotic or exact tests. Asymptotic results will be accurate when the marginal frequencies are not too extreme ($0.2 < \tilde{P}^{ij}, \tilde{P}_k < 0.8$) and the sample sizes are reasonably large ($n > 100$). Otherwise, the exact test results will tend to be more accurate.

Although well suited for genotypic associations, our traditional Fisherian approach to exact tests of the allelic cytonuclear disequilibria has the same potential shortcoming found with the original approach developed by FU and ARNOLD (1992b): the matrix of joint allelic counts in Table 2 necessarily counts each cytotype twice, thereby effectively treating the cytoplasmic marker as a (homozygous) diploid genome. This is primarily because the allelic disequilibria are simply not natural variables in the cytonuclear context, since they involve the joint allelic frequencies, which are defined as probabilities rather than true frequencies. As a result, the joint allelic counts in Table 2 obscure the distinctive features of the cytonuclear system (diploid *vs.* haploid inheritance). In theory, one remedy would be to base the joint allelic counts solely on the maternally inherited nuclear alleles, but at present this is unlikely to be feasible in practice. It remains to be seen whether another type of exact allelic test, more appropriate to cytonuclear systems, can or should be devised.

TABLE 5

Joint nuclear-cpDNA counts for jack and lodgepole pine involving the ACO nuclear locus data

Chloroplast type	Nuclear genotype										Margin
	A_1A_1	A_1A_2	A_1A_3	A_1A_4	A_2A_2	A_2A_3	A_2A_4	A_3A_3	A_3A_4	A_4A_4	
M_1	3	2	1	1	1	0	0	0	0	0	8
M_2	1	0	0	0	0	0	0	0	0	0	1
M_3	0	0	0	0	0	1	0	0	0	0	1
M_4	1	2	0	0	0	0	0	0	0	0	3
M_5	1	3	1	0	0	0	0	0	0	0	5
M_6	14	13	4	0	5	2	0	0	2	0	40
M_7	0	1	0	0	0	0	0	0	0	0	1
M_8	0	0	0	0	1	0	0	0	0	0	1
M_9	69	58	27	6	11	14	1	3	1	0	190
M_{10}	6	4	3	0	1	0	0	2	0	0	16
Margin	95	83	36	7	19	17	1	5	3	0	266

From LI (1995).

An illustration of our methodology for a diallelic system is provided by data from a hybrid swarm of bluegill for the nuclear system *Got-2* (AVISE *et al.* 1984). For clarity, the two nuclear alleles are denoted as A and a, and the two cytotypes as M and m. The raw counts of the six possible joint cytonuclear genotypes (n_M^{AA}, n_M^{Aa}, n_M^{aa}, n_m^{AA}, n_m^{Aa} n_m^{aa}) are (16, 51, 13, 12, 36, 23). A test for overall departures from random genotypic associations yields a Monte Carlo probability of 0.0697 that agrees well with the Markov chain probability of 0.0678. Both approximations were calculated based on 100 batches of 1000 observations and yielded sample standard deviations of 0.0009 and 0.00033, respectively. Neither are significant and we proceed to testing the individual disequilibria.

The disequilibrium estimates in Table 4 come from (8) and (9), their normalized values are calculated as in ASMUSSEN and BASTEN (1996), and the standard errors under $H_0: D_k^{ij} = 0$ are computed from the square root of (16) with D_k^{ij} set to 0 while the total standard errors (T) set D_k^{ij} to its estimated value. The test statistics ($n\hat{r}^2$) are calculated from (18) and their asymptotic probabilities are from the χ_1^2 distribution. The exact probabilities for the individual disequilibria are calculated using the algorithms of ZAYKIN and PUDOVKIN (1993) after collapsing the 2 × 3 table of joint cytonuclear genotype counts into a series of three 2 × 2 tables, one for each nuclear genotype. MSS gives the approximate minimum sample sizes for detecting the observed level of disequilibrium with probability $1 - \beta$ when detection is based on the estimator $\hat{D}_k^{ij}$ falling outside the 95% confidence interval under H_0. These are calculated as in ASMUSSEN and BASTEN (1994) using the asymptotic variances in (16). Note that the probabilities (P values) of the test statistics based on asymptotic tests are similar to those obtained by the exact test. Although the overall test failed to detect nonrandom associations between the nuclear genotypes and cytotypes, the individual tests show that D_M^{aa} is significantly different from 0 at the 5% level.

TABLE 6

Statistics for chloroplast type M_9 of Table 5 data

Nuclear genotype	Statistics for M_9				
	D_9^{ij}	$D_9^{ij'}$	s_9^{ij}	A.P.	E.P.
A_1A_1	0.0043	0.0421	0.0133	0.7461	0.7811
A_1A_2	−0.0048	−0.0246	0.0128	0.7064	0.7726
A_1A_3	0.0048	0.1250	0.0095	0.6100	0.6953
A_1A_4	0.0038	0.5000	0.0044	0.3965	0.6770
A_2A_2	−0.0097	−0.1895	0.0071	0.1754	0.1920
A_2A_3	0.0070	0.3824	0.0068	0.3028	0.4104
A_2A_4	0.0011	1.0000	0.0017	0.5263	1.0000
A_3A_3	−0.0021	−0.1600	0.0038	0.5679	0.6259
A_3A_4	−0.0043	−0.5333	0.0029	0.1419	0.1973

We apply our multiallelic methodology to the data in Table 5 summarized from LI (1995) for a population of lodgepole pine (*Pinus contorta*) near an area of active hybridization with jack pine (*P. banksiana*). There are 10 chloroplast types and four nuclear alleles at the ACO locus resulting in 100 possible joint cytonuclear genotypes. Rather than present the many individual disequilibrium statistics for this highly multiallelic system, for purposes of illustration we focus on the genotypic associations involving cytotype M_9 and present in Table 6 the estimates of disequilibrium (D_9^{ij}), normalized disequilibrium ($D_9^{ij'}$), standard deviation (s_9^{ij}), asymptotic probability (A.P.) and exact probability (E.P.). (Note that there were no A_4A_4 nuclear genotypes). None of the M_9 disequilibria were significant at the 5% level for either test. Furthermore, the actual P values based on the asymptotic statistics agree well with those based on the exact test, with large differences only occurring for rare cytonuclear counts. For example, there is only one individual of nuclear genotype A_2A_4 and for D_9^{24} the exact and asymptotic probabilities differ greatly (1.0 *vs.* 0.52).

TABLE 7

Comparison of Monte Carlo and Markov chain methods for the overall test of genotypic associations in the data in Table 5

Reps (=BC)	B	C	Monte Carlo		Markov Chain	
			$\tilde{p}$	$s_{\tilde{p}}$	$\tilde{p}$	$s_{\tilde{p}}$
10000	100	100	0.6080	0.00532	0.6361	0.04401
100000	100	1000	0.6071	0.00159	0.6232	0.03282
1000000	100	10000	0.6064	0.00046	0.5887	0.01531
10000000	100	100000	0.6068	0.00015	0.6096	0.00520

The Markov chain and Monte Carlo approximations for overall departures can also differ slightly. In Table 7 we compare the two methods for the entire data array in Table 5. The Markov chain was dememorized for 1000 repetitions. Standard errors $s_{\tilde{p}}$ were obtained through the batch method as the square root of (22). The Monte Carlo P value does not vary more than 1.5 standard deviations from the Markov chain result. Although the former is more precise at any number of repetitions, the Markov chain method is computationally much quicker. GUO and THOMPSON (1992) accordingly suggest using the Monte Carlo approach for data sets with small sample sizes and the Markov approach for large sample sizes. We direct the reader to their report for a more thorough discussion on the relative merits of the two approximations. As a practical matter, setting the number of batches (B) and observations per batch (C) to 100 for either approximation would provide a good first estimate of the P value. Greater precision can be achieved with a larger value of C, and would be desired if the significance of the P value estimate is in doubt, *i.e.*, when $\tilde{p} \pm 1.96 s_{\tilde{p}}$ spans the significance level of interest.

This work extends the guidelines of ASMUSSEN and BASTEN (1994) for the proper experimental design of cytonuclear surveys and the use of cytonuclear disequilibria for the testing of evolutionary hypotheses. We have now provided exact tests for the overall association, and asymptotic and exact tests for the individual associations between the nuclear and cytoplasmic components for data with an arbitrary number of alleles at each marker. These overall and individual tests are different types of tests: small, nonsignificant individual disequilibria may combine yielding a significant overall disequilibrium. Conversely, significant individual disequilibria could be of such a pattern that no overall disequilibrium is detected. Furthermore, when data exist for multiple loci, the actual values of the individual disequilibria allow insights into whether there are forces acting at the population level *vs.* forces acting on individual loci. Examples of the former include migration or assortative mating based on pure species status and would be characterized by disequilibrium values that are concordant across loci. In the latter case we would expect a lack of such concordance, with disequilibria for some of the markers and not others, or of varying signs and magnitudes across loci.

Programs for calculating the statistics presented herein may be obtained via anonymous ftp at brooks.statgen.ncsu.edu in the directory /pub/cnd or by contacting the first author.

We thank BRUCE WEIR for much helpful advice and two anonymous reviewers for some useful comments. In addition, we are very grateful to DMITRI ZAYKIN for generously providing his program and to T. LI and D. WAGNER for their unpublished data. This investigation was supported in part by National Science Foundation grant DEB 92-10895 to M.A.A. and National Institutes of Health grant GM-45344 to North Carolina State University.

LITERATURE CITED

ABERNETHY, K., 1994 The establishment of a hybrid zone between red and sika deer (genus *Cervus*). Mol. Ecol. **3:** 551–562.

ARNOLD, J., M. A. ASMUSSEN and J. C. AVISE, 1988 An epistatic mating system model can produce permanent cytonuclear disequilibria in a hybrid zone. Proc. Natl. Acad. Sci. USA **85:** 1893–1896.

ASMUSSEN, M. A., and J. ARNOLD, 1991 The effects of admixture and population subdivision on cytonuclear disequilibria. Theor. Popul. Biol. **39:** 273–300.

ASMUSSEN, M. A., and C. J. BASTEN, 1994 Sampling theory for cytonuclear disequilibria. Genetics **138:** 1351–1363.

ASMUSSEN, M. A., and C. J. BASTEN, 1996 Constraints and normalized measures for cytonuclear disequilibria. Heredity **76:** 207–214.

ASMUSSEN, M. A., and A. SCHNABEL, 1991 Comparative effects of pollen and seed migration on the cytonuclear structure of plant populations. I. Maternal cytoplasmic inheritance. Genetics **128:** 639–654.

ASMUSSEN, M. A., J. ARNOLD and J. C. AVISE, 1987 Definition and properties of disequilibrium statistics for associations between nuclear and cytoplasmic genotypes. Genetics **115:** 755–768.

ASMUSSEN, M. A., J. ARNOLD and J. C. AVISE, 1989 The effects of assortative mating and migration on cytonuclear associations in hybrid zones. Genetics **122:** 923–934.

AVISE, J. C., E. BERMINGHAM, L. G. KESSLER and N. SAUNDERS, 1984 Characterization of mitochondrial DNA variability in a hybrid swarm between subspecies of bluegill sunfish (*Lepomis macrochirus*). Evolution **38:** 931–941.

AVISE, J. C., W. S. NELSON, J. ARNOLD, R. K. KOEHN, G. C. WILLIAMS *et al.*, 1990 The evolutionary genetic status of Icelandic eels. Evolution **44:** 1254–1262.

AVISE, J. C., P. C. PIERCE, M. J. VAN DEN AVYLE, M. H. SMITH, W. S. NELSON *et al.*, 1997 Cytonuclear introgressive swamping and species turnover of bass after an introduction. J. Hered. **88:** 14–20.

BABCOCK, C. S., and M. A. ASMUSSEN, 1996 Effects of differential selection in the sexes on cytonuclear polymorphism and disequilibria. Genetics **144:** 839–853.

CRUZAN, M. B., and M. L. ARNOLD, 1993 Ecological and genetic associations in an *Iris* hybrid zone. Evolution **47:** 1432–1445.

CRUZAN, M. B., and M. L. ARNOLD, 1994 Assortative mating and natural selection in an *Iris* hybrid zone. Evolution **48:** 1946–1958.

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

DATTA, S., Y.-X. FU and J. ARNOLD, 1996a Dynamics and equilibrium behavior of cytonuclear disequilibria under genetic drift, mutation, and migration. Theor. Popul. Biol. **50:** 298–324.

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

DEAN, R., and J. ARNOLD, 1996 Small sample properties for estimators of cytonuclear disequilibria. Heredity **77:** 396–399.

FORBES, S. H., and F. W. ALLENDORF, 1991 Associations between mitochondrial and nuclear genotypes in cutthroat trout hybrid swarms. Evolution **45:** 1332–1349.

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

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

FU, Y. X., and J. ARNOLD, 1992b A table of exact sample sizes for the use with Fisher's exact test for 2 × 2 tables. Biometrics **48:** 1103–1112.

GUO, S. W., and E. A. THOMPSON, 1992 Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics **48:** 361–372.

LI, T., 1995 Cytonuclear population genetic structure of jack pine (*Pinus banksiana Lamb.*) and lodgepole pine (*Pinus contorta Dougl.*). Ph.D. dissertation, University of Kentucky, Lexington, KY.

PAIGE, K. N., W. C. CAPMAN and P. JENNETTEN, 1991 Mitochondrial inheritance patterns across a cottonwood hybrid zone: cytonuclear disequilibria and hybrid zone dynamics. Evolution **45:** 1360–1369.

SCHNABEL, A., and M. A. ASMUSSEN, 1989 Definition and properties of disequilibria within nuclear-mitochondrial-chloroplast and other nuclear-dicytoplasmic systems. Genetics **123:** 199–215.

SCHNABEL, A., and M. A. ASMUSSEN, 1992 Comparative effects of pollen and seed migration on the cytonuclear structure of plant populations. II. Paternal cytoplasmic inheritance. Genetics **132:** 253–267.

SCRIBNER, K. T., and J. C. AVISE, 1993 Cytonuclear genetic architecture in mosquitofish populations and the possible roles of introgressive hybridization. Mol. Ecol. **2:** 139–149.

SCRIBNER, K. T., and J. C. AVISE, 1994 Cytonuclear genetics of experimental fish hybrid zones inside Biosphere 2. Proc. Natl. Acad. Sci. USA **91:** 5066–5069.

SITES, J. W., C. J. BASTEN and M. A. ASMUSSEN, 1996 Cytonuclear genetic structure of a hybrid zone in lizards of the *Sceloporus grammicus* complex (Sauria, Phrynosomatidae). Mol. Ecol. **5:** 379–392.

WEIR, B. S., 1996 *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.

ZAYKIN, D. V., and A. I. PUDOVKIN, 1993 Two programs to estimate significance of χ^2 values using pseudo-probability tests. J. Hered. **84:** 152.

Communicating editor: M. KIRKPATRICK