# Sampling Theory for Cytonuclear Disequilibria

**Marjorie A. Asmussen and Christopher J. Basten**[1]

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

Manuscript received March 9, 1994
Accepted for publication August 24, 1994

ABSTRACT

We examine the statistical properties of cytonuclear disequilibria within a system including one diploid nuclear locus and one haploid cytoplasmic locus, each with two alleles. The results provide practical guidelines for the design and interpretation of cytonuclear surveys seeking to utilize the novel evolutionary information recorded in the observed pattern of cytonuclear associations. Important applications include population studies of nuclear allozymes in conjunction with genes from mitochondria, chloroplasts, or cytoplasmically inherited microorganisms. Our attention focuses on the allelic and genotypic disequilibria, which respectively measure the nonrandom associations between the cytotypes and the nuclear alleles and genotypes. We first derive the maximum likelihood estimators and their approximate large sample variances for each disequilibrium measure. These are each in turn used to set up an asymptotic test of the null hypothesis of no disequilibrium. We then calculate the minimum sample sizes required to detect the disequilibria under specified alternate hypotheses. The work also incorporates the deviation from Hardy-Weinberg equilibrium at the nuclear locus, which can significantly affect the results. The practical utility of this new sampling theory is illustrated through applications to two nuclear-mitochondrial data sets.

BECAUSE of[1] the contrasting modes of inheritance of biparentally inherited nuclear loci and uniparentally inherited cytoplasmic loci, joint nuclear-cytoplasmic data can provide important and qualitatively new insights into the evolutionary forces acting on natural populations. Much of this novel information is encoded by the nonrandom associations that are increasingly observed between nuclear and cytoplasmic markers (SAGHAI-MAROOF *et al.* 1992; AVISE *et al.* 1990; LAMB and AVISE 1986; SPOLSKY and UZZELL 1984; AVISE *et al.* 1984; FERRIS *et al.* 1983). There is now a substantial theoretical framework from which to analyze cytonuclear data and use it to make inferences about a variety of important evolutionary processes. Initial applications to hybrid zones have been particularly fruitful, yielding formal statistical estimates of the rates of gene flow and assortative mating, which appear to be more sensitive than, and may be unobtainable from, nuclear or cytoplasmic systems alone (ARNOLD *et al.* 1988; ASMUSSEN *et al.* 1989; AVISE *et al.* 1990). The theoretical foundation has also been laid for using cytonuclear data as markers of admixture, population subdivision, and genetic drift (ASMUSSEN and ARNOLD 1991; FU and ARNOLD 1991, 1992a), and in plant populations, to decompose gene flow into diploid (seed) and haploid (pollen) components (ASMUSSEN and SCHNABEL 1991; SCHNABEL and ASMUSSEN 1989, 1992).

The present study provides a crucial link for these and other applications by formally developing the statistical properties of the cytonuclear disequilibrium statistics introduced by ASMUSSEN *et al.* (1987). Our approach is based on that summarized by WEIR (1979, 1990) for two locus nuclear systems. We begin by reviewing the cytonuclear parameterization and the disequilibrium measures which account for the nonrandom associations between cytoplasmic alleles and nuclear alleles or genotypes. We then show how to estimate these disequilibria and their sampling variances, together with how to use them to construct tests of the null hypothesis, namely that a given disequilibrium is zero. Finally, we show how to calculate the sample size required to detect a level of either an allelic or a genotypic disequilibrium specified in an alternative hypothesis. These procedures are illustrated through applications to two recent nuclear-mitochondrial data sets.

## GENERAL DEVELOPMENT

**Basic cytonuclear system:** We are concerned with estimating the nonrandom associations (disequilibria) in a diploid population with two alleles ($A$, $a$) at an autosomal nuclear locus and two alleles ($M$, $m$) at a haploid cytoplasmic locus. The populational frequencies of the six possible cytonuclear genotypes are denoted as in Table 1, together with the marginal genotypic frequencies at the individual loci. Note that we have adopted a more informative notation in which the $P$ symbol denotes a frequency, with nuclear genes superscripted and cytoplasmic alleles (cytotypes) subscripted. $P_M^{AA}$, for instance, replaces $u_1$ as the frequency of individuals who are homozygous for the A allele and have the $M$ cytotype, while $P^{AA}$ replaces $u$ as the frequency of $AA$

[1] Present address: Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203.

**TABLE 1**

**Genotypic frequencies and equivalent measures from Asmussen *et al.* (1987)**

| Cytotype | Nuclear genotypes *AA* | *Aa* | *aa* | Total |
|---|---|---|---|---|
| *M* | $P_M^{AA} \equiv u_1$ | $P_M^{Aa} \equiv v_1$ | $P_M^{aa} \equiv w_1$ | $P_M \equiv x$ |
| *m* | $P_m^{AA} \equiv u_2$ | $P_m^{Aa} \equiv v_2$ | $P_m^{aa} \equiv w_2$ | $P_m \equiv y$ |
| Total | $P^{AA} \equiv u$ | $P^{Aa} \equiv v$ | $P^{aa} \equiv w$ | 1.0 |

homozygotes, $P_M$ replaces $x$ as the frequency of the $M$ cytotype, and $P^A = P^{AA} + \frac{1}{2}P^{Aa}$ replaces $p$ as the frequency of allele $A$. The full correspondence between this and our earlier genotypic frequency notation is indicated in Table 1.

We will also make use of the frequency measure

$$P_M^A = P_M^{AA} + \tfrac{1}{2}P_M^{Aa}$$

that intuitively corresponds to the frequency of gametes carrying the $A$ nuclear allele and the $M$ cytotype. Formally, $P_M^A$ is the probability that when you sample an individual from the population, it has the $M$ cytotype, and a randomly sampled nuclear allele from that individual is $A$. In previous work this has been denoted by $e_1$ (Clark 1984; Asmussen *et al.* 1987) or $P_1$ (Asmussen and Schnabel 1991).

Analogous notation will be used with the symbols $n$ and $\tilde{P}$ in place of $P$ to denote the sample counts and estimators, respectively, associated with the various frequency variables.

**Disequilibrium measures:** We focus on two types of cytonuclear associations. The first is the allelic disequilibrium

$$D_M^A = P_M^A - P^A P_M \tag{1}$$

which is analogous to the standard gametic disequilibrium for a two locus nuclear system and measures nonrandom associations between the nuclear alleles and the cytotypes. We also consider the three genotypic disequilibria

$$D_M^{AA} = P_M^{AA} - P^{AA}P_M \tag{2}$$

$$D_M^{Aa} = P_M^{Aa} - P^{Aa}P_M \tag{3}$$

$$D_M^{aa} = P_M^{aa} - P^{aa}P_M \tag{4}$$

which similarly measure nonrandom associations between each of the nuclear genotypes and the two cytotypes. From these we obtain the basic cytonuclear parameterization shown in Table 2. Note that $D_M^A$, $D_M^{AA}$, $D_M^{Aa}$ and $D_M^{aa}$ are identical to $D$, $D_1$, $D_2$ and $D_3$, respectively, of Asmussen *et al.* (1987), and reduce to two independent disequilibrium measures as a result of the interrelationships

$$D_M^A = D_M^{AA} + \tfrac{1}{2}D_M^{Aa} \tag{5}$$

and

$$D_M^{AA} + D_M^{Aa} + D_M^{aa} = 0. \tag{6}$$

For completeness, we will also specify the nuclear Hardy-Weinberg disequilibrium following Weir (1990) as $D^A = P^{AA} - P^{A^2}$. The three nuclear genotypic frequencies can be decomposed in terms of $D^A$ and $P^A$ as

$$P^{AA} = P^{A^2} + D^A \tag{7}$$

$$P^{Aa} = 2P^A(1 - P^A) - 2D^A \tag{8}$$

$$P^{aa} = (1 - P^A)^2 + D^A. \tag{9}$$

The six cytonuclear frequencies can then be parameterized by the nuclear and cytoplasmic allele frequencies, the Hardy-Weinberg disequilibrium ($D^A$), and two of the cytonuclear disequilibria ($D_M^A$, $D_M^{AA}$, $D_M^{Aa}$ and $D_M^{aa}$). We choose to emphasize the parameterization based on $D_M^A$ and $D_M^{AA}$ (Table 3) although we will discuss $D_M^{Aa}$ since it is important for making inferences about migration and mating patterns. (Note that through symmetry, arguments for $D_M^{AA}$ are the same for $D_M^{aa}$.) Under a null hypothesis of no disequilibria, the nuclear allele and the cytotype frequencies define the joint cytonuclear genotypic frequencies in the population. The disequilibria are constrained by the marginal frequencies, as shown in Table 4. (The derivation of these bounds will appear elsewhere.)

**Estimators of frequencies and disequilibria:** We assume that when we sample individuals from a large population, obtaining an individual of a specific type does not alter the probabilities of selecting individuals of any type in the future. The distribution of classes in a sample thus follows the multinomial. (In small populations this would not be true, in which case sampling would be based on the hypergeometric distribution.) If we have a sample of $n$ individuals, we write the counts of the different cytonuclear genotypes as the vector, $(n_M^{AA}, n_M^{Aa}, n_M^{aa}, n_m^{AA}, n_m^{Aa}, n_m^{aa})$. Clearly, any one of these counts could be written as the sample size minus the sum of the remaining counts, so that there are five independent classes. From this data set we would like estimates of the five independent variables $P^A$, $P_M$, $D^A$, $D_M^A$, $D_M^{AA}$. Since we have five parameters and five independent classes of data, we can use Bailey's (1951) method for calculating maximum likelihood estimators for the parameters. We merely need to set the observed counts equal to the expected values from the sample and solve the ensuing five equations (see pp. 53–55 in Weir 1990). This yields the following maximum likelihood estimators for the three disequilibria

$$\tilde{D}^A = \tilde{P}^{AA} - (\tilde{P}^A)^2 \tag{10}$$

$$\tilde{D}_M^A = \tilde{P}_M^A - \tilde{P}^A\tilde{P}_M \tag{11}$$

$$\tilde{D}_M^{AA} = \tilde{P}_M^{AA} - \tilde{P}^{AA}\tilde{P}_M \tag{12}$$

**TABLE 2**

**Basic cytonuclear parameterization**

| | Nuclear genotypes | | | |
|---|---|---|---|---|
| Cytotype | $AA$ | $Aa$ | $aa$ | Total |
| $M$ | $P_M^{AA} = P^{AA} P_M + D_M^{AA}$ | $P_M^{Aa} = P^{Aa} P_M + D_M^{Aa}$ | $P_M^{aa} = P^{aa} P_M + D_M^{aa}$ | $P_M$ |
| $m$ | $P_m^{AA} = P^{AA}(1 - P_M) - D_M^{AA}$ | $P_m^{Aa} = P^{Aa}(1 - P_M) - D_M^{Aa}$ | $P_m^{aa} = P^{aa}(1 - P_M) - D_M^{aa}$ | $1 - P_M$ |
| Total | $P^{AA}$ | $P^{Aa}$ | $P^{aa}$ | 1 |

**TABLE 3**

**Cytonuclear parameterization in terms of allele frequencies and disequilibria**

| | Nuclear genotypes | | |
|---|---|---|---|
| Cytotype | $AA$ | $Aa$ | $aa$ |
| $M$ | $P_M^{AA} = P^{AA} P_M + D_M^{AA}$ | $P_M^{Aa} = P^{Aa} P_M + 2(D_M^A - D_M^{AA})$ | $P_M^{aa} = P^{aa} P_M + D_M^{AA} - 2D_M^A$ |
| $m$ | $P_m^{AA} = P^{AA}(1 - P_M) - D_M^{AA}$ | $P_m^{Aa} = P^{Aa}(1 - P_M) - 2(D_M^A - D_M^{AA})$ | $P_m^{aa} = P^{aa}(1 - P_M) + 2D_M^A - D_M^{AA}$ |
| Total | $P^{AA} = P^{A^2} + D^A$ | $P^{Aa} = 2P^A(1 - P^A) - 2D^A$ | $P^{aa} = (1 - P^A)^2 + D^A$ |

where

$$\tilde{P}^{AA} = \frac{1}{n}(n_M^{AA} + n_m^{AA}) \tag{13}$$

$$\tilde{P}^A = \frac{1}{2n}[2(n_M^{AA} + n_m^{AA}) + n_M^{Aa} + n_m^{Aa}] \tag{14}$$

$$\tilde{P}_M^A = \frac{1}{n}\left(n_M^{AA} + \frac{1}{2} n_M^{Aa}\right) \tag{15}$$

$$\tilde{P}_M = \frac{1}{n}(n_M^{AA} + n_M^{Aa} + n_M^{aa}) \tag{16}$$

$$\tilde{P}_M^{AA} = \frac{1}{n} n_M^{AA} \tag{17}$$

are the maximum likelihood estimators for the associated frequencies. The maximum likelihood estimators for $D_M^{Aa}$ and $D_M^{aa}$ are equivalent to that for $D_M^{AA}$ with $Aa$ (or $aa$) substituted for $AA$ in (12–13) and (17).

**Sampling properties of the disequilibrium estimators:** The expected value of each cytonuclear disequilibrium estimator ($\tilde{D}_M^A$, $\tilde{D}_M^{AA}$, $\tilde{D}_M^{Aa}$ and $\tilde{D}_M^{aa}$) has the form

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

indicating a slight bias in the estimators. (Details of the derivation can be found in APPENDIX A.) The expected value for $\tilde{D}^A$ is

$$\mathcal{E}\tilde{D}^A = D^A - \frac{1}{2n}[P^A(1 - P^A) + D^A]$$

(WEIR 1990).

To determine the statistical significance of observed disequilibria, we may use the variances of our estimators. There are two ways to approach this. The first utilizes the Delta method, which is based on a first order Taylor's expansion of the function whose variance is to be calculated (WEIR 1990). Ths approach allows us to calculate the approximate sampling variances which are of immediate use for developing test statistics to study natural populations, and is what we will focus on here.

Applying the Delta method to each of (10–12) yields the approximate expected values of the sampling variances for the three basic disequilibrium estimators

$$\mathrm{Var}(\tilde{D}^A) \approx \frac{1}{n}[P^{A^2}(1 - P^A)^2 + D^A(1 - 2P^A)^2 - D^{A^2}] \tag{18}$$

$$\mathrm{Var}(\tilde{D}_M^A) \approx \frac{1}{2n}[P^A(1 - P^A)P_M(1 - P_M) + D^A P_M(1 - P_M) + D_M^{AA}(1 - 2P_M) + D_M^A(1 - 4P^A)(1 - 2P_M) - 2D_M^{A^2}] \tag{19}$$

$$\mathrm{Var}(\tilde{D}_M^{AA}) \approx \frac{1}{n}[P^{AA}(1 - P^{AA})P_M(1 - P_M) + D_M^{AA}(1 - 2P^{AA})(1 - 2P_M) - D_M^{AA^2}]. \tag{20}$$

The approximations for $\mathrm{Var}(\tilde{D}_M^{Aa})$ and $\mathrm{Var}(\tilde{D}_M^{aa})$ are equivalent to that in (20) for $\mathrm{Var}(\tilde{D}_M^{AA})$ with $AA$ replaced by $Aa$ or $aa$. The variances given in (18–20) can be used to obtain sample variances by inserting the estimators for the measures on the right hand sides of the equations.

The second way to calculate variances is to derive the total variances using our indicator variables as has been done for the two locus nuclear case (BASTEN and WEIR 1992). These will help us clarify the effects of evolutionary and sampling forces on our statistics, for they take into account the genetic sampling which gives rise to variation between replicate populations. Preliminary work on this problem indicates that the large sample variances above are an $O(1/n)$ approximation of the total variances in infinite populations. Further

**TABLE 4**

**Bounds on the disequilibria from the marginal frequencies**

| $D$ | Lower bound (min $D$) | Upper bound (max $D$) |
|---|---|---|
| $D^A$ | $-\min[P^{A2}, (1-P^A)^2]$ | $P^A(1-P^A)$ |
| $D_M^A$ | $-\min[P^A P_M, (1-P^A)(1-P_M), \frac{1}{2}P^{AA} P_M + \frac{1}{2}P^{aa}(1-P_M)]$ | $\min[P^A(1-P_M), (1-P^A)P_M, \frac{1}{2}P^{AA}(1-P_M) + \frac{1}{2}P^{aa} P_M]$ |
| $D_M^{AA}$ | $-\min[P^{AA} P_M, (1-P^{AA})(1-P_M)]$ | $\min[P^{AA}(1-P_M), (1-P^{AA}) P_M]$ |
| $D_M^{Aa}$ | $-\min[P^{Aa} P_M, (1-P^{Aa})(1-P_M)]$ | $\min[P^{Aa}(1-P_M), (1-P^{Aa}) P_M]$ |
| $D_M^{aa}$ | $-\min[P^{aa} P_M, (1-P^{aa})(1-P_M)]$ | $\min[P^{aa}(1-P_M), (1-P^{aa}) P_M]$ |

**TABLE 5**

**Values of $\delta_0^2$ and $\delta_1^2$**

| $D$ | $\delta_0^2 = n\,\mathrm{Var}(\tilde{D} \mid D = 0)$ | $\delta_1^2 = n\,\mathrm{Var}(\tilde{D} \mid D \neq 0)$ |
|---|---|---|
| $D^A$ | $P^{A2}(1-P^A)^2$ | $\delta_0^2 + D^A(1-2P^A)^2 - D^{A2}$ |
| $D_M^A$ | $[P^A(1-P^A)P_M(1-P_M) + R]/2$ | $\delta_0^2 + [D_M^A(1-4P^A)(1-2P_M) - 2D_M^{A2}]/2$ |
| $D_M^{AA}$ | $P^{AA}(1-P^{AA})P_M(1-P_M)$ | $\delta_0^2 + D_M^{AA}(1-2P^{AA})(1-2P_M) - D_M^{AA2}$ |
| $D_M^{Aa}$ | $P^{Aa}(1-P^{Aa})P_M(1-P_M)$ | $\delta_0^2 + D_M^{Aa}(1-2P^{Aa})(1-2P_M) - D_M^{Aa2}$ |
| $D_M^{aa}$ | $P^{aa}(1-P^{aa})P_M(1-P_M)$ | $\delta_0^2 + D_M^{aa}(1-2P^{aa})(1-2P_M) - D_M^{aa2}$ |

Note: $R = D^A P_M(1-P_M) + D_M^{AA}(1-2P_M)$.

development and analysis of the total variance will be left for a future report.

**Testing hypotheses:** The distributions of our disequilibrium estimators are approximately normal with means and variances calculated in the previous section. Suppose we are interested in testing the null hypothesis defined by $H_0$: $D = 0$, where $D$ is one of ($D^A$, $D_M^A$, $D_M^{AA}$, $D_M^{Aa}$, $D_M^{aa}$). Under this null hypothesis, the estimator for $D$ ($\tilde{D}$) has a normal distribution with mean zero and variance $V_0 = \delta_0^2/n$, where for each $D$, $\delta_0^2$ is $n$ times the variance expressions of (18–20) assuming $D = 0$. The values of $\delta_0$ are given in Table 5. The statistic $n\tilde{r}^2$, where $\tilde{r} = \tilde{D}/\delta_0$, provides a useful test statistic because $n\tilde{r}^2 = \tilde{D}^2/V_0$ has an approximately $\chi^2$ (1) distribution (since the square of a normal random variable divided by its variance has a $\chi^2(1)$ distribution). This statistic is traditionally used in measuring nuclear gametic disequilibrium ($D$), where using our present allele frequency notation,

$$\tilde{r} = \frac{\tilde{D}}{\sqrt{\tilde{P}^A(1-\tilde{P}^A)\tilde{P}_M(1-\tilde{P}_M)}} \tag{21}$$

is often known as the correlation of genes. In the case of the cytonuclear genotypic disequilibria, say $D_M^{AA}$,

$$\tilde{r} = \frac{\tilde{D}_M^{AA}}{\sqrt{\tilde{P}^{AA}(1-\tilde{P}^{AA})\tilde{P}_M(1-\tilde{P}_M)}} \tag{22}$$

takes a similar form and could be termed the correlation of genotypes, while for the cytonuclear allelic disequilibrium the statistic takes a more complicated form

$$\tilde{r} = \frac{\sqrt{2}\tilde{D}_M^A}{\sqrt{\tilde{P}^A(1-\tilde{P}^A)\tilde{P}_M(1-\tilde{P}_M) + R}} \tag{23}$$

where

$$R = \tilde{D}^A\tilde{P}_M(1-\tilde{P}_M) + \tilde{D}_M^{AA}(1-2\tilde{P}_M)$$

includes additional terms involving the estimators for the Hardy-Weinberg disequilibrium, $\tilde{D}^A$, and the genotypic disequilibrium, $\tilde{D}_M^{AA}$.

For a test at the 0.05 significance level, we reject the null hypothesis that $D = 0$ if $n\tilde{r}^2 > 3.84$. The order of testing the various disequilibria is naturally suggested by the dependencies among the variances needed to calculate their test statistics (Table 5). In general, we start by testing the Hardy-Weinberg disequilibrium $D^A$ (whose test statistic is independent of the other disequilibria) and then test the genotypic disequilibria, $D_M^{AA}$, $D_M^{Aa}$ and $D_M^{aa}$ (whose test statistics depend on $D^A$ through the decompositions in (7–9)). Finally we test the allelic disequilibrium, $D_M^A$ (whose test statistic depends on both $D^A$ and $D_M^{AA}$). In order to implement this procedure, it is necessary to decide whether or not the estimates for $D^A$ and $D_M^{AA}$ should be included in calculating the test statistics of the higher order measures if we fail to reject the null hypothesis that $D = 0$ for either of these lower order test statistics. We performed simulation studies following the methodology of Boos and Browne (1989) as applied by Muse and Weir (1992) to resolve this issue, based on which form of the cytonuclear test statistics better fit a $\chi^2(1)$ distribution. The results indicate that the estimate for $D^A$ (and the observed nuclear genotypic frequencies) should be used in calculating the test statistics based on (22–23) for the other disequilibria even if we fail to reject the null hypothesis $H_0$: $D^A = 0$. The estimate for $D_M^{AA}$, on the other hand, should be included in calculating the test statistic based on (23) for $D_M^A$ only

if the null hypothesis $H_0$: $D_M^{AA} = 0$ has been rejected; in cases where $D_M^{AA}$ is not significantly different from 0, it should be set to 0 in the test for $D_M^A$.

We also used these simulations to determine over what range of sample sizes and allele frequencies the asymptotic $\chi^2(1)$ distribution can be invoked for the test statistics. The results indicate that the test statistics, $n\tilde{r}^2$, behave as desired for reasonable sample sizes and intermediate allele frequencies. Allele (or genotypic) frequencies that are extreme tend to skew the distribution of the test statistics, although larger sample sizes can overcome this. More specifically, the test criteria above based on the $\chi^2(1)$ approximation can be satisfactorily used for samples of size 100 or more for which the estimates for $P^A$ and $P_M$ are in the range [0.2,0.8]. The larger the sample size, the more extreme the allele frequencies can be, while for much smaller sample sizes such as 20–30, the two allele frequencies must both be in [0.4, 0.6] to use the $\chi^2(1)$ approximation. In general, the test statistic for $D_M^{AA}$ is more sensitive to lower values of $P^A$, while that for $D_M^{aa}$ is more sensitive to higher values since the frequency of the corresponding nuclear homozygote is then lower. With a sample size of 100, for instance, allele frequencies can be in [0.2,0.8] for testing $D_M^A$ and $D_M^{Aa}$, whereas the allele frequencies really should be in [0.3, 0.8] for $D_M^{AA}$ and in [0.2, 0.7] for $D_M^{aa}$.

A computer program implementing our testing procedure for cytonuclear disequilibria is available upon request. In cases with small sample sizes or extreme allele frequencies for which the asymptotic $\chi^2(1)$ distribution cannot be used, this program simulates the distribution of the statistic(s) to get a distribution for testing purposes. Such a simulation is only an intermediate step; ultimately we hope to incorporate exact tests for the analysis of such data sets.

**Sample sizes to detect disequilibria:** The method here is a generalization of BROWN's (1975) analysis of nuclear linkage disequilibrium. In testing the null hypothesis $H_0$: $D = 0$ against the alternate hypothesis $H_1$: $D = D_1 \neq 0$, we would like to calculate the sample size required to detect the disequilibrium assuming $H_1$. [Here, $D_1$ is the value of $D$ under $H_1$, not the genotypic disequilibrium defined in ASMUSSEN *et al.* (1987).] Recall that $\tilde{D}$ has an approximately normal distribution with mean zero and variance $\delta_0^2/n$ under $H_0$ and mean $D_1$ and variance $\delta_1^2/n$ under $H_1$ (where $\delta_1$ is calculated from Table 5 assuming that $D = D_1$ and we ignore the slight bias in $\tilde{D}$). In standard practice one defines the size of a test, $\alpha$, to be the probability of a type I error (rejecting a true hypothesis). The power of a test, $1 - \beta$, is the ability to detect disequilibria if it exists, where $\beta$ is the probability of a type II error (failing to reject a false hypothesis). We accept (or fail to reject) the null hypothesis if and only if the value of $\tilde{D}$ lies in the interval $[-(z_{\alpha/2})\delta_0/\sqrt{n}, (z_{\alpha/2})\delta_0/\sqrt{n}]$, where $z_x$ is the standard normal deviate defined by the relations, $x = P(Z \leq -z_x) = P(Z \geq z_x)$ for $Z \sim N(0, 1)$.

The approximate sample size to detect disequilibrium with power $1 - \beta$ is obtained by setting to $\beta$ the probability that the sample estimate $\tilde{D}$ is in the acceptance region for $H_0$ when $H_1$ holds. If $\tilde{D} > 0$, we approximate this by

$$\beta \approx P\left[\tilde{D} \leq \frac{z_{\alpha/2}\delta_0}{\sqrt{n}} \,\middle|\, H_1 \text{ true}\right] \tag{24}$$

while if $\tilde{D} < 0$ we use the approximation

$$\beta \approx P\left[\frac{-z_{\alpha/2}\delta_0}{\sqrt{n}} \leq \tilde{D} \,\middle|\, H_1 \text{ true}\right]. \tag{25}$$

In either case, the sample size required to detect disequilibrium with probability $1 - \beta$ (or fail to detect the disequilibria with probability $\beta$) when the true value is $D = D_1$ is given by

$$n \approx \left(\frac{z_\beta\delta_1 + z_{\alpha/2}\delta_0}{D_1}\right)^2 \tag{26}$$

for a hypothesis test of size $\alpha$.

## RESULTS AND DISCUSSION

The power to detect nonrandom cytonuclear associations is illustrated in Figures 1–5, which plot the $\log_{10}$ of the minimum sample sizes calculated from (26) to detect a given normalized level of disequilibrium $\dot{D}$. The latter corresponds to LEWONTIN's (1964) $D'$, which is the actual disequilibrium $D$ divided by the maximum possible magnitude for a disequilibrium of that sign in a population with the observed marginal frequencies. Formally,

$$\dot{D} = \begin{cases} D/|\min D| & \text{if } D < 0 \\ D/\max D & \text{if } D \geq 0 \end{cases} \tag{27}$$

where min $D$ and max$D$ are the lower and upper bounds on the disequilibrium $D$ (Table 4). Positive values indicate a proportion of the maximum disequilibrium while negative values indicate a proportion of the minimum. We use the notation $\dot{D}$ here to avoid clashes with our superscripts denoting the nuclear alleles and genotypes. The sample sizes in Figure 1–5 were calculated with $\alpha = 0.05$ and $\beta = 0.1$, so that $z_{\alpha/2} = 1.96$ and $z_\beta = 1.28$. They therefore represent the minimum sample sizes in order to have a 90% probability of detecting the specified level of disequilibrium when detection is based on the estimator $\tilde{D}$ falling outside the 95% confidence interval under the null hypothesis that $D = 0$. For each disequilibrium there is a symmetry with respect to the cytotype frequency in that the sample size to detect a specified level of cytonuclear disequilibrium $D$ in a population with cytotype frequency $P_M$ also applies to the detection of the disequilibrium $-D$ in a population with cytotype frequency $1 - P_M$. In the case of the heterozygote disequilibrium there is an additional symmetry: the sample sizes are the same for $D_M^{Aa}$ whether the nuclear allele frequency is $P^A$ or $1 - P^A$.

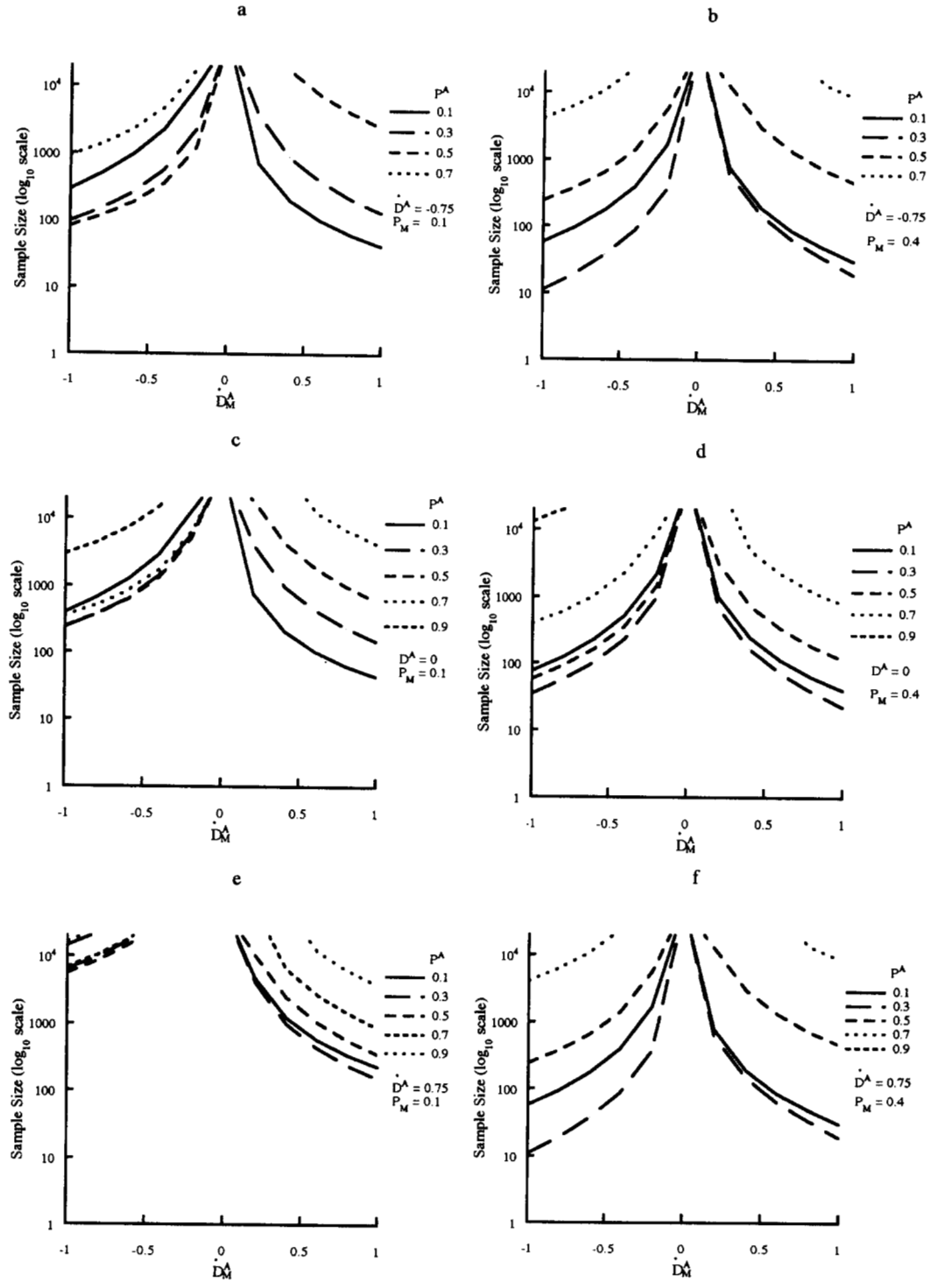


FIGURE 1.—The minimum sample sizes required to detect the specified levels of normalized disequilibrium $\dot{D}^A_M$ for a series of marginal frequency combinations when $\alpha = 0.05$ and $\beta = 0.1$. In all cases shown, we set $D^{AA}_M = 0.0$. The actual (nonnormalized) values of $D^A$ and the minimum and maximum of $D^A_M$ for each curve are given in APPENDIX B. In the graphs with no line shown for $P^A = 0.9$, the minimum sample sizes all exceeded $10^4$.

The examples in Figures 1–3 illustrate how, as in the case of nuclear linkage disequilibrium (BROWN 1975), the power to detect nonrandom associations varies widely, depending on the associated marginal frequencies and magnitude of the disequilibrium. The number of individuals needed to be sampled to have a 90% chance of detecting a given nonrandom association ranges from on the order of 10 to well over $10^4$. The power for detection is naturally greatest if the associated marginal frequencies are intermediate and the disequilibrium in question is near its minimum negative or maximum positive possible value.

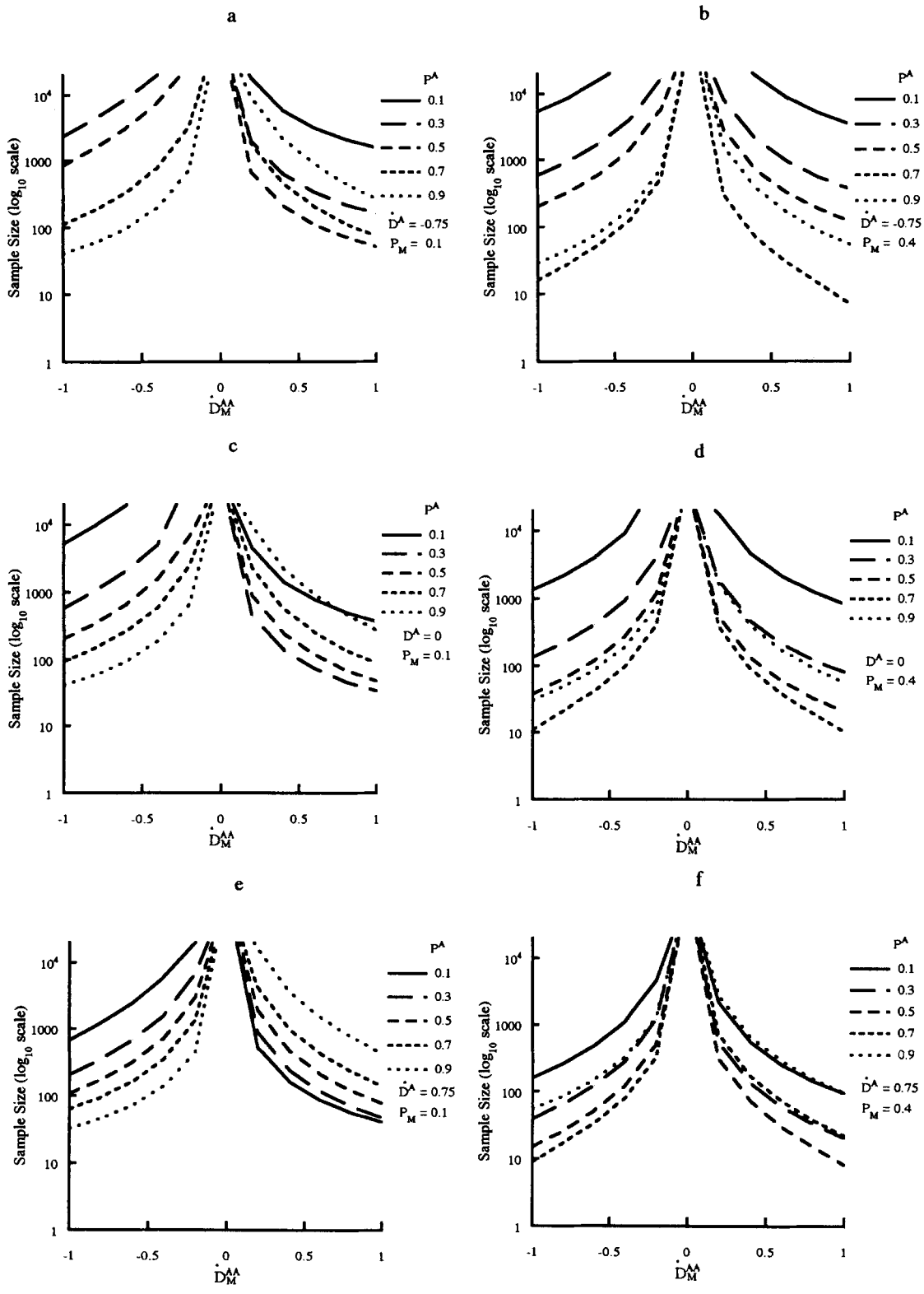


FIGURE 2.—The minimum sample sizes required to detect the specified levels of normalized disequilibrium $\dot{D}_M^{AA}$ for a series of marginal frequency combinations when $\alpha = 0.05$ and $\beta = 0.1$. The actual (nonnormalized) values of $D^A$ and the minimum and maximum of $D_M^{AA}$ for each curve are given in APPENDIX B.

It is clear from Figures 4 and 5 that the level of Hardy-Weinberg disequilibrium at the nuclear locus can greatly affect the sample sizes needed to detect specified levels of genotypic disequilibria for a given nuclear allele frequency. This is because Hardy-Weinberg disequilibrium determines the frequency of the nuclear genotypes, which in turn affects the maximum and minimum possible values of the cytonuclear disequilibria and thus

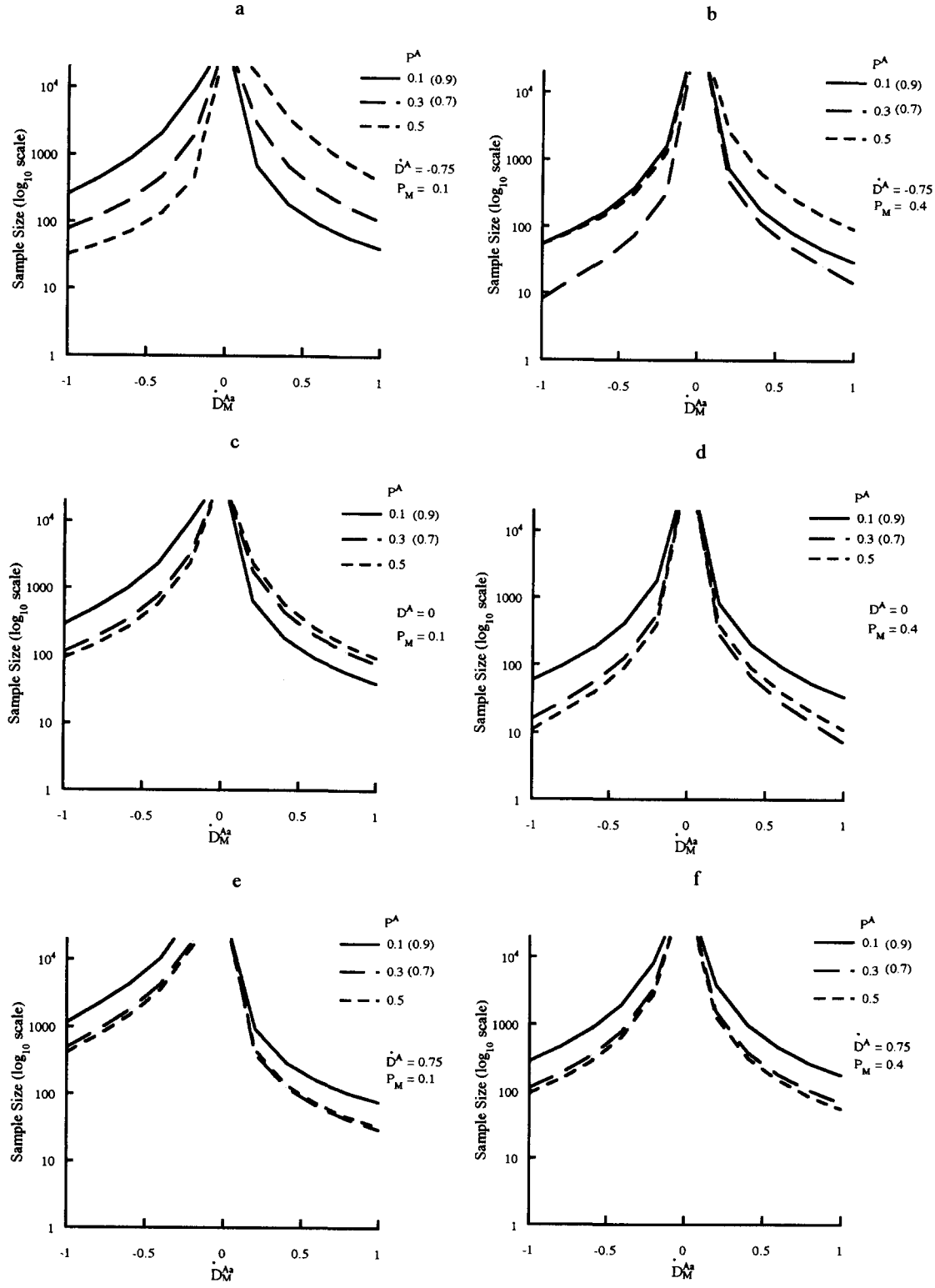


FIGURE 3.—The minimum sample sizes required to detect the specified levels of normalized disequilibrium $\dot{D}_M^{Aa}$ for a series of marginal frequency combinations when $\alpha = 0.05$ and $\beta = 0.1$. The actual (nonnormalized) values of $D^A$ and the minimum and maximum of $D_M^{Aa}$ for each curve are given in APPENDIX B.

the minimum sample sizes required for their detection. This phenomenon is illustrated for $D_M^{AA}$ in Figure 4 which plots the $\log_{10}$ of the minimum sample sizes calculated from (26) to detect a given normalized level of disequilibrium as a function of $\dot{D}^A$ for $P^A = 0.3$ and $P^A = 0.7$. When $P^A < 0.5$ (Figure 4a), the minimum sample sizes strongly depend on the Hardy-Weinberg disequilibrium, monotonically increasing to infinity in a loglinear

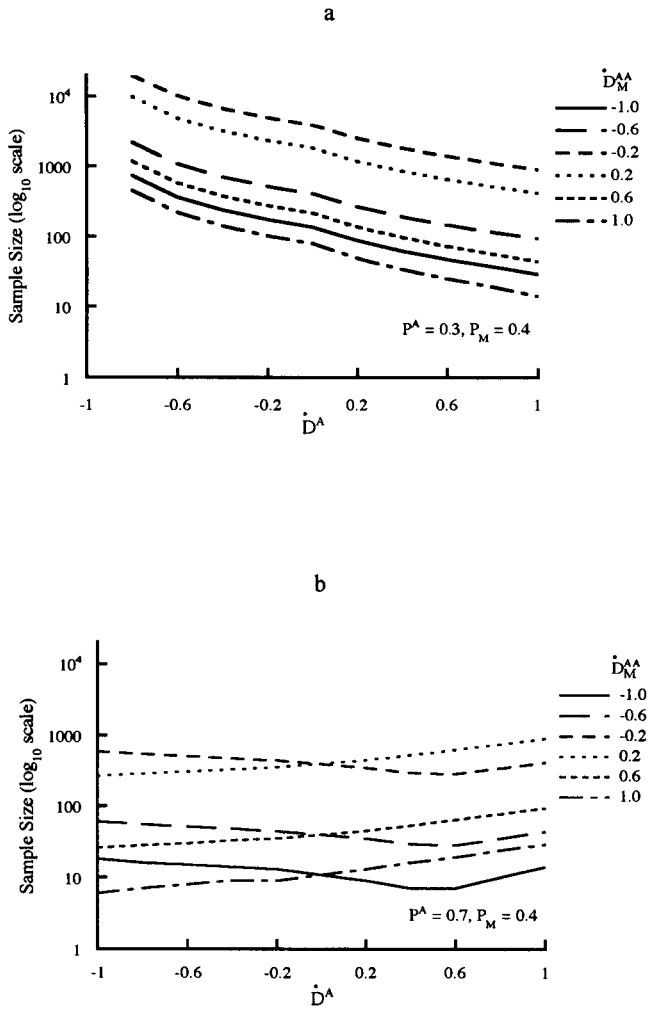


FIGURE 4.—Sample sizes required to detect specified levels of the normalized genotypic cytonuclear disequilibrium $\dot{D}_M^{AA}$ over the range of Hardy-Weinberg disequilibrium, $\dot{D}^A$, for $\alpha = 0.05$ and $\beta = 0.1$. a: $P^A = 0.3$, $P_M = 0.4$; b: $P^A = 0.7$, $P_M = 0.4$.

fashion as $D^A$ decreases to its minimum value. This is because in this case the minimum of $D^A$ corresponds to the absence of $AA$ homozygotes. Consequently, as $D^A$ decreases from its maximum to its minimum value, the frequency of $AA$ homozygotes and the range of admissible $D_M^{AA}$ values shrinks to zero. The sample sizes required to detect the ever smaller $D_M^{AA}$ disequilibrium accordingly increase without bound. In contrast, when $P^A > 0.5$ (Figure 4b), there is generally very little effect of Hardy-Weinberg disequilibrium upon the detection of $D_M^{AA}$ because the extreme values of $D^A$ then correspond to the absence of $aa$ or $Aa$ individuals. Although not shown, the curves for $D_M^{aa}$ are identical to those of $D_M^{AA}$ with $P^A$ replaced by $1 - P^A$. This suggests that if one has a choice between these two disequilibria, one should choose $D_M^{AA}$ when $P^A > 0.5$ and $D_M^{aa}$ when $P^A < 0.5$ in order to minimize the required sample sizes and the effect of Hardy-Weinberg disequilibrium. The situation is different for the heterozygote disequilibrium, $D_M^{Aa}$. As Hardy-

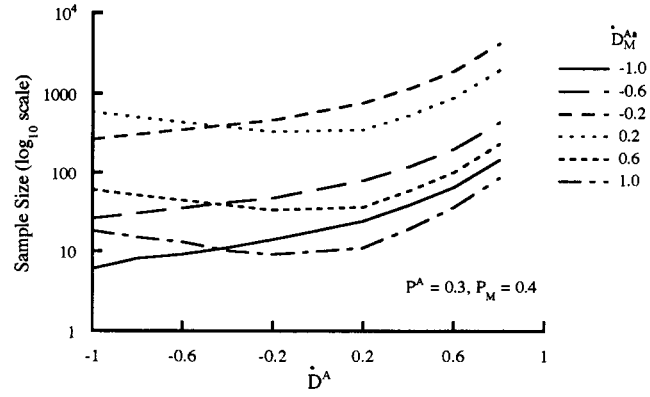


FIGURE 5.—Sample sizes required to detect specified levels of the normalized genotypic cytonuclear disequilibrium $\dot{D}_M^{Aa}$ over the range of Hardy-Weinberg disequilibrium, $\dot{D}^A$, when $P^A = 0.3$, $P_M = 0.4$, $\alpha = 0.05$ and $\beta = 0.1$.

Weinberg disequilibrium increases, the heterozygote class shrinks as does the range of admissible values for $D_M^{Aa}$. In Figure 5 we see how this increases the required sample sizes, especially when the normalized Hardy-Weinberg disequilibrium is above 0.2.

It should be emphasized that these minimum sample sizes are approximations based on the large sample variances of the disequilibrium estimators. FU and ARNOLD (1992b) calculated the sample sizes for Fisher's exact test of independence in 2 × 2 tables, and their work can be used as a rough benchmark by which to gauge the accuracy of our approximations to the sample sizes required to detect $D_M^A$ for a test of size $\alpha$ and power $1 - \beta$. We have used their program to calculate the exact sample sizes to detect $\dot{D}_M^A = 0.75$ for various values of $P^A$. Results are shown in Table 6 for $P_M = 0.4$ and $D^A = 0$, with $\dot{D}_M^{AA}$ set to $-0.75$, 0 and 0.75 in the top, middle and bottom thirds of the table, respectively. Since the sample size is discrete, the achieved values of $\alpha$ and $1 - \beta$ will not be exactly 0.05 and 0.9, so they are presented in Table 6 as well. The approximate sample sizes are not too far off from those calculated by FU and ARNOLD's exact test, and have the advantage of providing very fast results. There is some question about the appropriateness of this exact test for testing the allelic disequilibrium in cytonuclear systems, however, because it requires converting the 2 × 3 table of joint genotypic counts to a 2 × 2 table of joint allelic counts. This doubles the count of cytoplasmic alleles in the sample to $2n$, where $n$ is the number of individuals, and effectively treats each individual as a diploid homozygote at the cytoplasmic locus. The unique features of the cytonuclear system are thereby compromised. The close agreement between the two methods is nonetheless encouraging and consistent with preliminary results suggesting that the large sample variances on which the minimum sample size calculations are based can be reasonably good approximations to the exact values.

**TABLE 6**

**Comparison of approximation with the exact sample sizes for 2 × 2 tables in numbers of gametes ($n_g$) to detect $\dot{D}_M^A = 0.75$ for $P_M = 0.4$ and $D^A = 0.00$**

| | | Exact | | | Approximate | | |
|---|---|---|---|---|---|---|---|
| $P^A$ | $D_M^A$† | $n_g$ | $\alpha$ | $1 - \beta$ | $n_g^*$ | $\alpha$ | $1 - \beta$ |
| $\dot{D}_M^{AA} = -0.75$ | | | | | | | |
| 0.1 | 0.0383 | 166 | 0.0361 | 0.9017 | 155 | 0.0352 | 0.8755 |
| 0.2 | 0.0630 | 109 | 0.0374 | 0.9022 | 89 | 0.0360 | 0.8263 |
| 0.3 | 0.0634 | 140 | 0.0401 | 0.9008 | 106 | 0.0382 | 0.7954 |
| 0.4 | 0.0360 | 480 | 0.0455 | 0.9005 | 366 | 0.0441 | 0.8030 |
| $\dot{D}_M^{AA} = 0.0$ | | | | | | | |
| 0.1 | 0.0405 | 148 | 0.0349 | 0.9015 | 142 | 0.0346 | 0.8866 |
| 0.2 | 0.0720 | 83 | 0.0356 | 0.9008 | 71 | 0.0344 | 0.8385 |
| 0.3 | 0.0735 | 105 | 0.0382 | 0.9030 | 86 | 0.0373 | 0.8318 |
| 0.4 | 0.0540 | 216 | 0.0426 | 0.9012 | 189 | 0.0418 | 0.8540 |
| $\dot{D}_M^{AA} = 0.75$ | | | | | | | |
| 0.1 | 0.4388 | 124 | 0.0343 | 0.9010 | 125 | 0.0343 | 0.9010 |
| 0.2 | 0.0855 | 57 | 0.0330 | 0.9001 | 53 | 0.0319 | 0.8705 |
| 0.3 | 0.0887 | 71 | 0.0369 | 0.9000 | 67 | 0.0357 | 0.8771 |
| 0.4 | 0.0810 | 97 | 0.0377 | 0.9028 | 99 | 0.0384 | 0.9070 |

* Twice the minimum number of individuals from (26)

† Based on equation 27, with max $D_M^A$ further constrained by the value of $D_M^{AA}$.

**TABLE 7**

**Application of the statistics to the Albumin-mtDNA data from a hybrid population of Hyla treefrogs (LAMB and AVISE 1986)**

| Disequilibrium | $D^A$ | $D_M^{AA}$ | $D_M^{Aa}$ | $D_M^{aa}$ | $D_M^A$ |
|---|---|---|---|---|---|
| Estimator ($\tilde{D}$) | 0.1362 | 0.1902 | −0.06316 | −0.1271 | 0.1587 |
| Normalized ($\tilde{D}$) | 0.561 | 0.7839 | −0.6365 | −0.8858 | 0.8217 |
| Standard error ($H_0$) | 0.0139 | 0.01427 | 0.0117 | 0.01319 | 0.01327 |
| Standard error ($H_1$) | 0.01206 | 0.009313 | 0.01075 | 0.01049 | 0.008335 |
| Test Statistic ($n\tilde{r}^2$) | 95.99 | 177.8 | 29.16 | 92.87 | 143 |
| MSS[a] ($\beta = 0.1$) | 30 | 13 | 103 | 29 | 16 |
| MSS[a] ($\beta = 0.5$) | 12 | 7 | 40 | 13 | 8 |

[a] Minimum sample size when $\alpha = 0.05$.

**TABLE 8**

**Application of the statistics to the *Es-3* by mtDNA data from a hybrid swarm of bluegill (AVISE *et al.* 1984)**

| Disequilibrium | $D^A$ | $D_M^{AA}$ | $D_M^{Aa}$ | $D_M^{aa}$ | $D_M^A$ |
|---|---|---|---|---|---|
| Estimator ($\tilde{D}$) | −0.0287 | −0.0258 | 0.0497 | −0.0239 | −0.00097 |
| Normalized ($\tilde{D}$) | −0.1262 | −0.245 | 0.2112 | −0.1838 | −0.008753 |
| Standard error ($H_0$) | 0.0203 | 0.0162 | 0.0202 | 0.0175 | 0.0135[b] |
| Standard error ($H_1$) | 0.0202 | 0.0163 | 0.0198 | 0.0175 | 0.0135[b] |
| Test statistic ($n\tilde{r}^2$) | 1.996 | 2.532 | 6.053 | 1.865 | 0.005 |
| Probability | 0.158 | 0.112 | 0.014 | 0.172 | 0.943 |
| MSS[a] ($\beta = 0.1$) | 790 | 628 | 258 | 851 | 309961[b] |
| MSS[a] ($\beta = 0.5$) | 291 | 229 | 96 | 311 | 113475[b] |

[a] Minimum sample size when $\alpha = 0.05$.

[b] Calculated with $D_M^{AA} = 0$ having failed to reject $H_0 : D_M^{AA} = 0$.

As an illustration of the usage of our procedure, we apply it to two data sets. The first is the data of LAMB and AVISE (1986) from a hybrid zone of Hyla treefrogs. This data includes a joint survey of the mitochondrial types and isozyme genotypes at the *Alb* (Albumin) locus for 305 individuals. The cytonuclear genotypic counts for this data set are $(n_M^{AA}, n_M^{Aa}, n_M^{aa}, n_m^{AA}, n_m^{Aa}, n_m^{aa}) = (126, 11, 5, 20, 54, 89)$, where the $AA/M$ and $aa/m$ genotypes are characteristic of the two parental species, *Hyla cinerea* and *Hyla gratiosa*, respectively. Table 7 presents the estimators for the disequilibria, their normalized values, their standard errors, the test statistic, $n\tilde{r}^2$, and the sample sizes required to detect the observed levels of disequilibrium for $1 - \beta = 0.9$ and 0.5, and $\alpha = 0.05$. The probabilities of obtaining the observed values of $n\tilde{r}^2$ are all less than $10^{-3}$, and thus we reject the null hypothesis of zero disequilibrium for all the disequilibria. The estimators and their standard errors are similar to those of Table 9 in ASMUSSEN *et al.* (1987) which were obtained via a complex hierarchical computer

algorithm. Because the observed disequilibria are fairly near their maximum negative or positive values, extremely small sample sizes would be sufficient to ensure detection of the observed levels of cytonuclear associations with 90% probability.

The second data set comes from a nuclear-mtDNA survey of 151 individuals from a hybrid zone between two subspecies of bluegill sunfish (*Lepomis macrochirus*) involving the *Es-3* allozyme locus. The joint genotypic counts are (12, 52, 16, 18, 32, 21), where *AA/M* is diagnostic for *L. m. macrochirus* and *aa/m* is diagnostic for *L. m. purpurescens*. An application of our sampling theory yields the results in Table 8. As noted by ASMUSSEN *et al.* (1987), only $D_M^{Aa}$ is significantly different from zero. The actual sample size is 50% larger than that necessary for detection of this disequilibrium with 50% probability, but only about 60% of that necessary for detection with 90% probability. This emphasizes that the minimum sample sizes calculated here are neither necessary nor sufficient for detection. They simply ensure rejection of the null hypothesis of no disequilibrium with the specified probability. Note that roughly 800 individuals would be necessary to detect the levels of disequilibrium observed for the other nuclear genotypes with 90% probability, whereas $\tilde{D}_M^A$ is so small as to be virtually undetectable with any sample size, even with 50% probability.

This analysis of the formal statistical properties of cytonuclear disequilibria fills an important missing link in understanding the proper experimental design of cytonuclear surveys and the subsequent data analysis. One final practical point is that although our treatment explicitly applies to codominant nuclear loci, it also has applications to systems with complete dominance. In particular, if *a* is recessive to *A*, our results still hold for $D_M^{aa}$, although not for the other cytonuclear disequilibria. Thus, we can accommodate cytonuclear data where the nuclear component is generated from randomly amplified polymorphic DNAs (RAPDs). The amount of information from such systems is much reduced, however, since only one disequilibrium measure can be calculated. Further work is needed to calculate the exact cytonuclear variances, which will allow a true test of the accuracy of the theory here which is based on the large-sample variances from the standard Fisher-approximation. Ultimately, in order to properly design and interpret joint nuclear-mitochondrial-chloroplast surveys in plant populations, which should prove uniquely informative, it will be desirable to extend this statistical framework to the various three locus associations possible in nuclear-dicytoplasmic systems.

We thank JONATHAN ARNOLD, JOHN AVISE, BRIAN GOLDING and BRUCE WEIR for many valuable suggestions on the manuscript. JONATHAN ARNOLD also kindly provided the program to calculate the exact sample sizes for $2 \times 2$ tables. 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

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 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.

BAILEY, N. T. J., 1951 On estimating the size of mobile populations from recapture data. Biometrika **38:** 293–306.

BASTEN, C. J., and B. S. WEIR, 1992 Effect of gene conversion on measures of digenic disequilibrium, pp. 345–362 in *Population Paleo-Genetics: Proceedings of the Seventeenth Taniguchi International Symposium on Biophysics*, edited by N. TAKAHATA, Japan Scientific Societies Press, Tokyo.

BOOS, D. D., and C. BROWNIE, 1989 Bootstrap methods for testing homogeneity of variances. Technometrics **31:** 69–82.

BROWN, A. H. D., 1975 Sample sizes required to detect linkage disequilibrium between two or three loci. Theor. Popul. Biol. **8:** 184–201.

CLARK, A. G., 1984 Natural selection with nuclear and cytoplasmic transmission. I. A deterministic model. Genetics **107:** 679–701.

FERRIS, S. D., R. D. SAGE, C.-M. HUANG, J. T. NIELSEN, U. RITTE *et al.*, 1983 Flow of mitochondrial DNA across a species boundary. Proc. Natl. Acad. Sci. USA **80:** 2290–2294.

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 $2 \times 2$ tables. Biometrics **48:** 1103–1112.

LAMB, T., and J. C. AVISE, 1986 Directional introgression of mitochondrial DNA in a hybrid population of treefrogs: the influence of mating behavior. Proc. Natl. Acad. Sci. USA **83:** 2525–2530.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

MUSE, S. V., and B. S. WEIR, 1992 Testing for equality of evolutionary rates. Genetics **132:** 269–276.

SAGHAI-MAROOF, M. A., Q. ZHANG, D. B. NEALE and R. W. ALLARD, 1992 Associations between nuclear loci and chloroplast DNA genotypes in wild barley. Genetics **131:** 225–231.

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.

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

WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235–254.

WEIR, B. S., 1990 *Genetic Data Analysis.* Sinauer Associates, Sunderland, Mass.

Communicating editor: G. B. Golding

## APPENDIX A: EXPECTED VALUES OF THE DISEQUILIBRIUM ESTIMATORS

To study the statistical properties of the MLEs (11–12), we will make use of indicator variables in the same fashion as WEIR (1990). To this end, we index individuals in the sample by $i = 1, \ldots, n$. Furthermore, we arbitrarily index the nuclear alleles within an individual by $k = 1, 2$. With this convention, we define the following indicator variables for the nuclear locus

$$x_{ik} = \begin{cases} 1 & \text{if gene } k \text{ in individual } i \text{ is A} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, n$ and $k = 1,2$. Over all possible samples of size $n$ from the population, the $x_{ik}$ have the identical expectation

$$\mathcal{E}x_{ik} = 1 \cdot \Pr[x_{ik} = 1] + 0 \cdot \Pr[x_{ik} = 0] = P^A \qquad i = 1, \ldots, n;\ k = 1, 2.$$

The frequencies of the nuclear homozygotes can then be expressed as expectations of products of these variables. For $AA$ homozygotes, for instance, we have

$$\mathcal{E}x_{i1}x_{i2} = P^{AA} \qquad i = 1, \ldots, n.$$

We define a similar set of indicator variables for the cytoplasmic locus,

$$y_i = \begin{cases} 1 & \text{if individual } i \text{ has cytotype } M \\ 0 & \text{otherwise} \end{cases} \qquad \text{where} \qquad \mathcal{E}y_i = P_M \qquad i = 1, \ldots, n.$$

The expectations of various products of the nuclear and cytotype indicator functions yield the joint cytonuclear frequencies in the population. For example,

$$\mathcal{E}x_{ik}y_i = 1 \cdot \left[P_M^{AA} + \frac{1}{2}P_M^{Aa}\right] = P_M^A \qquad i = 1, \ldots, n; \quad k = 1, 2 \qquad \text{and} \qquad \mathcal{E}x_{i1}x_{i2}y_i = P_M^{AA} \qquad i = 1, \ldots, n.$$

By expanding (11)–(12) in terms of the indicator variables, we easily obtain the expected values of the cytonuclear disequilibrium estimators. For example, the allelic disequilibrium measure defined in (1) and estimated by (11) becomes

$$\tilde{D}_M^A = \tilde{P}_M^A - \tilde{P}^A\tilde{P}_M = \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{2} x_{ik}y_i - \left(\frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{2} x_{ik}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

$$= \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{2} x_{ik}y_i - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\sum_{k=1}^{2} x_{ik}y_{i'} = \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{2} x_{ik}y_i - \frac{1}{2n^2}\left[\sum_{i=1}^{n}\sum_{k=1}^{2} x_{ik}y_i + \sum_{i=1}^{n}\sum_{i'\neq i}\sum_{k=1}^{2} x_{ik}y_{i'}\right].$$

Under our assumption that different individuals are sampled independently, we find that

$$\mathcal{E}\tilde{D}_M^A = 2n\frac{1}{2n}P_M^A - \frac{1}{2n^2}[2nP_M^A + 2n(n-1)P^AP_M] = \left(1 - \frac{1}{n}\right)(P_M^A - P^AP_M) = \left(1 - \frac{1}{n}\right)D_M^A \tag{A1}$$

indicating a slight bias in the estimator. In a similar fashion, we can develop estimators for the genotypic disequilibria. The maximum likelihood estimator for $D_M^{AA}$ is

$$\tilde{D}_M^{AA} = \tilde{P}_M^{AA} - \tilde{P}^{AA}\tilde{P}_M = \frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2}y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2}\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2}y_i - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n} x_{i1}x_{i2}y_{i'}$$

where

$$\mathcal{E}\tilde{D}_M^{AA} = \left(1 - \frac{1}{n}\right)(P_M^{AA} - P^{AA}P_M) = \left(1 - \frac{1}{n}\right)D_M^{AA}.$$

The expected values of $\tilde{D}_M^{Aa}$ and $\tilde{D}_M^{aa}$ are equivalent to that for $\tilde{D}_M^{AA}$, with $Aa$ or $aa$ substituted for $AA$.

## APPENDIX B

Actual values of the bounds for the disequilibria examined in Figures 1–3 are given in Table 9.

**TABLE 9**

**Actual values of the bounds for the disequilibria examined in Figures 1–3**

| | | | $D_M^A$† | | $D_M^{AA}$ | | $D_M^{Aa}$ | |
|---|---|---|---|---|---|---|---|---|
| $P^A$ | $P_M$ | $D^A$* | Min | Max | Min | Max | Min | Max |
| 0.1 | 0.1 | −0.0075 | −0.00975 | 0.04013 | −0.00025 | 0.00225 | −0.0195 | 0.0805 |
| | | 0.0 | −0.009 | 0.0405 | −0.001 | 0.009 | −0.018 | 0.082 |
| | | 0.0675 | −0.0023 | 0.02025 | −0.00775 | 0.06975 | −0.0045 | 0.0405 |
| | 0.4 | −0.0075 | −0.039 | 0.0585 | −0.001 | 0.0015 | −0.078 | 0.117 |
| | | 0.0 | −0.036 | 0.054 | −0.004 | 0.006 | −0.072 | 0.108 |
| | | 0.0675 | −0.009 | 0.0135 | −0.031 | 0.0465 | −0.018 | 0.027 |
| 0.3 | 0.1 | −0.0675 | −0.0278 | 0.02113 | −0.00225 | 0.02025 | −0.0555 | 0.0445 |
| | | 0.0 | −0.021 | 0.0245 | −0.009 | 0.081 | −0.042 | 0.058 |
| | | 0.1575 | −0.0053 | 0.03238 | −0.02475 | 0.07525 | −0.0105 | 0.0895 |
| | 0.4 | −0.0675 | −0.111 | 0.0845 | −0.009 | 0.0135 | −0.222 | 0.178 |
| | | 0.0 | −0.084 | 0.098 | −0.036 | 0.054 | −0.168 | 0.232 |
| | | 0.1575 | −0.021 | 0.0315 | −0.099 | 0.1485 | −0.042 | 0.063 |
| 0.5 | 0.1 | −0.1875 | −0.02813 | 0.00313 | −0.00625 | 0.05625 | −0.0875 | 0.0125 |
| | | 0.0 | −0.025 | 0.0125 | −0.025 | 0.075 | −0.05 | 0.05 |
| | | 0.1875 | −0.00625 | 0.02188 | −0.04375 | 0.05625 | −0.0125 | 0.0875 |
| | 0.4 | −0.1875 | −0.01875 | 0.0125 | −0.025 | 0.0375 | −0.075 | 0.05 |
| | | 0.0 | −0.075 | 0.05 | −0.1 | 0.15 | −0.2 | 0.2 |
| | | 0.1875 | −0.025 | 0.0375 | −0.175 | 0.225 | −0.05 | 0.075 |
| 0.7 | 0.1 | −0.0675 | −0.01013 | 0.00113 | −0.04225 | 0.05775 | −0.0555 | 0.0445 |
| | | 0.0 | −0.021 | 0.0045 | −0.049 | 0.051 | −0.042 | 0.058 |
| | | 0.1575 | −0.00525 | 0.01238 | −0.06475 | 0.03525 | −0.0105 | 0.0895 |
| | 0.4 | −0.0675 | −0.00675 | 0.0045 | −0.169 | 0.231 | −0.222 | 0.178 |
| | | 0.0 | −0.027 | 0.018 | −0.196 | 0.204 | −0.168 | 0.232 |
| | | 0.1575 | −0.021 | 0.0315 | −0.2115 | 0.141 | −0.042 | 0.063 |
| 0.9 | 0.1 | −0.0075 | −0.00113 | 0.00013 | −0.08025 | 0.01975 | −0.0195 | 0.0805 |
| | | 0.0 | −0.0045 | 0.0005 | −0.081 | 0.019 | −0.018 | 0.082 |
| | | 0.0675 | −0.00225 | 0.00388 | −0.08775 | 0.01225 | −0.0045 | 0.0405 |
| | 0.4 | −0.0075 | −0.00075 | 0.0005 | −0.1185 | 0.079 | −0.078 | 0.117 |
| | | 0.0 | −0.003 | 0.002 | −0.114 | 0.076 | −0.072 | 0.108 |
| | | 0.0675 | −0.009 | 0.0135 | −0.0735 | 0.049 | −0.018 | 0.027 |

* The first in each trio for a given $P_M$ value sets $\dot{D}^A$ to −0.75, the third to 0.75.
† $D_M^A$ is calculated from (27), with max $D_M^A$ further constrained from assuming that $D_M^{AA} = 0.0$.