

May 6, 2002

Measuring Intraspecies Genetic Diversity Using RAPD-PCR Profiles

Thaddeus Tarpey

Department of Mathematics and Statistics, Wright State University
Dayton, Ohio 45435
email: ttarpey@mail.wright.edu
phone: (212) 543-5812
fax: (212) 543-5599

and

Dan Krane

Department of Biological Sciences, Wright State University
Dayton, Ohio 45435

Abstract. Determining the extent to which environmental insults have effected contaminated sites is an important step in rational and efficient remediation efforts. Among other differences, comparisons between pristine and polluted sites often reveals diminished levels of genetic diversity within naturally occurring populations of organisms. An intraspecies diversity index is defined and applied to multivariate binomial data generated from Randomly Amplified Polymorphic DNA Polymerase Chain Reaction (RAPD-PCR) profiles. The diversity index is compared to well known interspecies diversity index from the ecological literature. The sampling distribution for the intraspecies diversity index is investigated and a permutation test is used to test for a decrease in genetic diversity at impacted sites compared to reference sites.

Key words: Analysis of quadratic entropy; binomial distribution; bootstrap; entropy; multinomial distribution; permutation Test; Simpson's index;

1. INTRODUCTION

Pielou (1988) calls a diversity index a measure of qualitative dispersion of a population of individuals belonging to different categories. Diversity indices have received considerable attention in the literature because the ecological health of an area is generally related to its species diversity (e.g. Alam and Williams 1993; Patil and Taille 1979, 1982; Morales, Taneja, and Pardo 1993; Fritsch and Hsu 1999; a comprehensive bibliography of diversity related papers is given by Dennis et al 1979). Areas subject to environmental stresses generally exhibit both diminished numbers and different kinds of species. A variety of statistical problems arise such as how to “penalize” a site for the presence of invading undesirable species, which diversity index should be used, how to estimate the diversity and what are its sampling properties. Even more important is the question of how well a single number statistic can capture the diversity of a complex community of organisms.

Diversity indices are typically defined as a function of multinomial frequencies that correspond to the frequency of different species found within a given area. In this paper, we define and investigate an alternative diversity index that focusses instead upon measuring *intraspecies genetic diversity*, i.e., genetic diversity within single sentinel species. In order to measure intraspecies diversity, data obtained from Randomly Amplified Polymorphic DNA Polymerase Chain Reaction (RAPD-PCR) profiles is used. Krane et al (1999) use RAPD-PCR profiles from crayfish (*Orconectes rusticus*) to compare the genetic diversity of the crayfish populations to other measures of environmental integrity at a variety of reference and impacted sites. RAPD-PCR as well as other means of assaying genetic

diversity are also reviewed by Theodorakis and Wirgin (2002). In the present paper, RAPD-PCR profiles are discussed in Section 2. In Section 3 we develop a diversity index to be used to measure intraspecies genetic diversity. The connection of diversity and evenness for intraspecies experiments is discussed in Section 4. In Section 5 we propose an estimator of the diversity and investigate the sampling distribution of the estimated intraspecies diversity index. RAPD-PCR data from different sites is used in permutation tests in Section 6 to test for differences in genetic diversity at impacted sites and reference sites. The genetic diversity measure is compared to a genetic similarity measure in Section 7 and we conclude the paper in Section 8.

2. GENETIC DIVERSITY and RAPD-PCR PROFILES

RAPD-PCR profiles have been generated from snails (*Physella gyrina*) collected from five watersheds and a total of seventeen sites in Ohio as part of an on-going ecological study (Krane and Sternberg, in preparation). We will illustrate the estimation and testing of hypotheses of genetic diversity using RAPD-PCR data profiles from snails collected at two of those sites (Elk Creek and Dick's Creek). One of the sites (Dick's Creek) has been chronically exposed to high levels of a variety of pollutants including polycyclic biphenyls (PCBs) while Elk Creek, a stream of similar order and drainage area, has consistently been determined to be a suitable uncontaminated reference site on the basis of a variety of chemical and biotic measures (Rowland and Burton 1996). We also shall look at RAPD-PCR data on snails obtained at six locations along the Black River in northern Ohio.

RAPD-PCR profiles were obtained for each of the $n = 24$ randomly selected snails at each site. A primer (B01: 5'-CAGGCCCTTC-3') was found to be capable of generating a total of 18 different scorable bands within the snail populations at Elk and Dick's Creek. For each snail and a given band, either the band is amplified (denoted by a "1" in the data) or is not amplified (denoted by a "0" in the data) in the RAPD-PCR process. The data for the profiles of snails at each of the test sites are given in Tables 1 and 2. From the tables we see that the data for each snail consists of a multivariate Bernoulli vector of size $k = 18$ for the 18 different bands.

The RAPD-PCR approach has several features that makes it particularly well suited for the determination of intraspecific genetic diversity levels. Like conventional PCR, it is capable of amplifying visually detectable quantities of DNA from extremely small amounts of starting material. The RAPD-PCR differs however in that it utilizes short (typically 10-nucleotide long) primers to amplify multiple anonymous and often polymorphic loci (Tinker et al., 1993). Unlike conventional PCR, no prior knowledge of the test organism's DNA sequence is required and the multiple bands that are generated from each primer are generally independent of those generated by other primers. The markers that result from RAPD-PCR amplification can be readily resolved with agarose gel electrophoresis because of differences in their sizes and can be reproducibly amplified (Penner et al., 1993) to distinguish between even closely related individuals of the same species (del Tufo and Tingey, 1994; Grosberg, Levitan, and Cameron 1995; Skepner and Krane 1995).

The Hardy-Weinberg model of population genetics theory predicts that a given genetic

marker's frequency of occurrence (and therefore the diversity at that locus) within a population will not change so long as nine assumptions are met (Hartl and Clark 1997). The choice of snails as a test organism assures that six of those assumptions are reasonably met for this data set (the organisms are diploid, sexually reproducing, allele frequencies are identical in males and females, possess non-overlapping generations, have minimal migration and mate randomly). Of the remaining three assumptions, departures from two have a tendency to reduce diversity (sampling error through the action of selective forces on or near the locus and failure to maintain large population sizes) and one (mutation) to increase it but only on the scale of thousands of generations. Consequently, anthropogenic stressors that impose strong selective pressure and/or induce population bottlenecks such as pollution and channelization of waterways are generally expected to diminish the genetic resources a population can bring to bear in response to changing environmental conditions and thereby weaken the stability of the species (Thorpe and Koonce, 1981; Allendorf and Leary, 1986).

3. A DIVERSITY INDEX

Let \mathbf{X} denote the random vector response for a given snail, where $\mathbf{X} = (X_1, \dots, X_k)'$ and k equals the dimension of the data (i.e. the number of potentially detectable bands in the RAPD-PCR profile). For a given component X_j of \mathbf{X} , let $p_j = P(X_j = 1)$, $j = 1, \dots, k$. Thus, \mathbf{X} follows a multivariate Bernoulli distribution with an associated vector of success probabilities $\mathbf{p} = (p_1, \dots, p_k)'$. For the RAPD-PCR application, the dependence structure between the X_j is unknown and would have to be estimated. If \mathbf{X}_i is the

response for the i th snail, then $\mathbf{X} = \sum \mathbf{X}_i$ follows a multivariate binomial distribution, $\mathbf{X} \sim \text{MVB}_k(\mathbf{p}, n, \mathbf{D})$ (Patil et al 1984, p 121). The matrix \mathbf{D} specifies the dependence structure.

The vector $\mathbf{p} = (p_1, \dots, p_k)'$ is of primary importance. Our goal is to define a reasonable measure of intraspecies diversity. For the j th band on the RAPD-PCR profile with corresponding component X_j , a low diversity occurs when $p_j \approx 1$ or $p_j \approx 0$ in which case the profiles of most snails will either be amplified or not be amplified at j th band. The diversity is maximized when $p_j = 1/2$ since this maximizes the variance of the j th component X_j . A simple measure of diversity then is to consider the squared Euclidean distance of \mathbf{p} from the vector $\mathbf{J}/2$ given by $\|\mathbf{p} - \mathbf{J}/2\|^2$, where \mathbf{J} is a vector of ones. We can rewrite this expression as:

$$\begin{aligned} \|\mathbf{p} - \mathbf{J}/2\|^2 &= (\mathbf{p} - \mathbf{J}/2)'(\mathbf{p} - \mathbf{J}/2) \\ &= \mathbf{p}'\mathbf{p} - \mathbf{p}'\mathbf{J} + k/4 \\ &= \sum_{j=1}^k p_j^2 - \sum_{j=1}^k p_j + k/4 \\ &= k/4 - \sum_{j=1}^k p_j(1 - p_j). \end{aligned}$$

Therefore, \mathbf{p} is closer to $\mathbf{J}/2$ in terms of squared Euclidean distance when the sum of variances for each of the components of \mathbf{X} is large. Thus, we shall denote as our diversity index Δ ,

$$\Delta = \sum_{j=1}^k p_j(1 - p_j)/k. \quad (1)$$

The index Δ can be motivated by the following consideration. Let d denote a measure

of separation between two independent random profiles \mathbf{X}_1 and \mathbf{X}_2 so that $d(\mathbf{X}_1, \mathbf{X}_2) = d(\mathbf{X}_2, \mathbf{X}_1) \geq 0$. One can define the diversity as the average separation (based on d) between two randomly selected individuals from the population: $\mathcal{E}[d(\mathbf{X}_1, \mathbf{X}_2)]$ (see Patil and Taille 1979, p 23). If we take $d(\mathbf{X}_1, \mathbf{X}_2) = \|\mathbf{X}_1 - \mathbf{X}_2\|^2$, then it follows that

$$\mathcal{E}[d(\mathbf{X}_1, \mathbf{X}_2)] = E\|\mathbf{X}_1 - \mathbf{X}_2\|^2 = 2 \sum_{j=1}^k p_j(1 - p_j) = 2k\Delta. \quad (2)$$

Rao (1982, p 72) notes that the univariate version of d is a conditionally negative definite function from which it follows that the multivariate version is also conditionally negative definite. If the function d is conditionally negative definite, then $-\mathcal{E}[d(\mathbf{X}_1, \mathbf{X}_2)]$ is completely convex – Rao (1982) calls such a diversity measure a *quadratic entropy*. Therefore, the diversity Δ is a quadratic entropy which satisfies certain natural requirements. In particular, the diversity of a mixture of two populations will not be less than the average of the diversity within the individual populations (Rao 1982, p 69).

The index Δ given above is closely related to Simpson’s interspecies diversity index (Simpson 1949). For an infinite population consisting of s distinct species in proportions p_1, \dots, p_k , a concentration index C is the probability that two individuals drawn at random will be of the same species. Clearly $C = \sum p_j^2$ (Pielou 1977, p 309). A measure of diversity should increase as concentration decreases, so one can consider the expression $1 - \sum p_j^2$. Since selecting organisms at random from a collection of s species yields a multinomial distribution, we have $\sum p_j = 1$ giving $1 - \sum p_j^2 = \sum p_j(1 - p_j)$. Taking the logarithm of this function admits it to the general family of entropy’s of order α for an s symbol code. Taking $\alpha = 2$ gives Simpson’s interspecies diversity index.

Simpson's index is just one of many diversity indices which have been proposed in the literature for measuring interspecies diversity. Patil and Talle (1982) suggest a family of indices given by

$$\Delta_\beta = \sum_{i=1}^k p_i(1 - p_i^\beta)/\beta$$

for $\beta > -1$. Simpson's index corresponds to $\beta = 1$. In the interspecies multinomial setup, the values of $\beta = -1$ and 0 correspond to the species count and Shannon's diversity index respectively. However, in the multivariate binomial intraspecies setting, a value of $\beta = 1$ seems a natural choice since it gives a measure of the total variance.

4. CONNECTION WITH THE MULTINOMIAL DISTRIBUTION AND EVENNESS

Interspecies diversity measures attempt to quantify a community's *richness* which corresponds to the number of species and the relative abundance of the species [their *evenness* (e.g. see Solomon 1979)].

Since evenness is an attribute of a community comprised of different types of individuals, how do we measure evenness in a genetic diversity study within a population of a single species? One can think of all 2^k possible RAPD-PCR profiles as a collection of different types of individuals. For the snail data it is easy to confirm that no two snails share the same profile at all possible band positions.

In the interspecies multinomial setup, the evenness is maximized when all the multinomial probabilities are equal, i.e. equal proportions of each species. Simpson's interspecies index $1 - \sum p_j^2$ in the multinomial setting is thus maximized when one has maximal evenness which occurs when the p_j 's are equal to each other. We shall now connect the

multinomial maximal evenness with the multivariate binomial setup.

To illustrate matters, consider a trivariate binomial random vector $(X_1, X_2, X_3)'$ corresponding to $k = 3$ possible bands on the RAPD-PCR profile. Then for each organism there are eight possible outcomes

$$(1, 1, 1), (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 0, 0), (0, 1, 0), (0, 0, 1), (0, 0, 0)$$

with associated probabilities

$$p_{111}, p_{110}, p_{101}, p_{011}, p_{100}, p_{010}, p_{001}, p_{000}.$$

From a sample of n organisms, let

$$Y_{111}, Y_{110}, Y_{101}, Y_{011}, Y_{100}, Y_{010}, Y_{001}, Y_{000}$$

correspond to the multinomial counts for each of the 8 possible outcomes. The connection between the multivariate binomial and the multinomial can be seen by noting that that X_1 , the number of successes with variable 1 is just $X_1 = Y_{111} + Y_{110} + Y_{101} + Y_{100}$ (see Patil et al 1984, p 122). A similar relation holds for X_2 and X_3 .

In the current setting, maximal evenness occurs when all multinomial probabilities are equal to 2^{-k} . However, when the multinomial probabilities are all equal, then the components of the multivariate binomial distributions, the X_j , will have success probabilities $\sum p_{ijl}$ summing over all possible values of i and l which gives a sum of $2^{k-1}/2^k = 1/2$. In other words, maximal evenness in the multinomial setup leads to maximal variance of the binomial components giving success probabilities equal to $1/2$. In the example of RAPD-PCR profiles of organisms, a higher measure of evenness corresponds to a larger

variety of profiles for a particular group. Therefore, the connection of the multivariate binomial in terms of maximal multinomial evenness is another justification for the use of Δ as an index of intraspecies diversity.

5. ESTIMATING THE DIVERSITY Δ

One of the goals of the current study is to determine a benchmark value of diversity which can be used to compare sample diversities from other stressed areas. When investigators collect RAPD-PCR profiles from organisms of a possibly contaminated site and compute the diversity index $\hat{\Delta}$, the sampling distribution and variability of the estimated index needs to be determined.

In order to estimate the diversity, which is an average of component variances, one may consider replacing the p_j by the sample proportion \hat{p}_j to get $\sum_{j=1}^k \hat{p}_j(1 - \hat{p}_j)$. However, it is easy to show that $\mathcal{E}[\hat{p}_j(1 - \hat{p}_j)] = \frac{n-1}{n}p_j(1 - p_j)$. To correct for this bias, we shall define the sample diversity index $\hat{\Delta}$ as:

$$\hat{\Delta} = \frac{n}{n-1} \sum_{j=1}^k \hat{p}_j(1 - \hat{p}_j)/k. \quad (3)$$

From equation (2) it is easy to verify that

$$2k\hat{\Delta} = \sum_{i=1}^N \sum_{j=1}^N \|\mathbf{X}_i - \mathbf{X}_j\|^2 / (N(N-1)). \quad (4)$$

Elk Creek, the reference site, has a sample diversity index of 0.179 which is higher than the sample diversity index of 0.143 for Dick's Creek. In Section 6 we look to see if the diversity at Dick's Creek is significantly lower than the diversity at Elk Creek.

By the multivariate central limit theorem, $\sqrt{n}(\hat{\mathbf{p}} - \mathbf{p}) \Rightarrow N_k(\mathbf{0}, \Psi)$ as $n \rightarrow \infty$ where

Ψ is the covariance matrix of $\hat{\mathbf{p}}$ and is given by $p_j(1 - p_j)/n$ on the main diagonals and $(p_{ij} - p_i p_j)/n$ on the off-diagonals where $p_{ij} = P(X_i = 1, X_j = 1)$. Defining the function $f(\hat{\mathbf{p}}) = \sum \hat{p}_j(1 - \hat{p}_j)$, it follows that $\sqrt{n}(\hat{\Delta} - \Delta) \rightarrow N(0, \mathbf{F}'\Psi\mathbf{F})$ where \mathbf{F} is the vector of partial derivatives of f evaluated at the p_j (e.g. see Seber 1984, p 532; see also Liu and Rao 1995). Differentiating, we get $\mathbf{F} = (1 - 2p_1, \dots, 1 - 2p_k)'$ and thus, $\hat{\Delta}$ is asymptotically normal with mean Δ and variance

$$\sigma_{\hat{\Delta}}^2 = \sum_{i=1}^k \sum_{j=1}^k (1 - 2p_i)(1 - 2p_j)(p_{ij} - p_i p_j)/(k^2 n)$$

(p_{ii} is taken to equal p_i).

From Tables 1 and 2, we see that some of the estimated \hat{p}_j 's are generally close to 1 which indicates that the distribution of certain components of $\hat{\mathbf{p}}$ will be skewed left and the normal approximation may be poor for a sample size of $n = 24$. The exact distribution of $\hat{\Delta}$ for finite sample sizes is not known since the dependence structure \mathbf{D} is unknown and would have to be estimated. Therefore, a bootstrap analysis was undertaken to determine the sampling distribution of $\hat{\Delta}$. Letting \hat{F} denote the empirical distribution of the sample and let \hat{p}^* denote an estimate of a proportion from a bootstrap sample, then it is not difficult to show that $\mathcal{E}_{\hat{F}}[\hat{p}^*(1 - \hat{p}^*)] = (n - 1)\hat{p}(1 - \hat{p})/n$ (see Efron and Tibshirani 1993, p 125). If B bootstrap samples are taken, then in order to correct for the bootstrap bias, and the bias associated with $\hat{p}(1 - \hat{p})$, we shall rescale our bootstrap estimates of the diversity by a factor of $n^2/(n - 1)^2$.

Figure 1 below shows a relative frequency polygon based on $B = 10,000$ bootstrap samples for the intraspecies diversity index of snails at Dick's Creek and Elk Creek. The

relative frequency polygons indicate the sampling distributions for the sample diversities for the two sites have roughly the same shape, but different means. Also, the sampling distribution of the diversity at each site is skewed to the left somewhat as expected. Figure 1 also shows the normal approximation based on the asymptotic result above – the normal density curve for both sites are indicated by the dotted curves. The bootstrap distribution and the normal distribution approximations coincide considerably, with the major difference occurring in the left tail which one may expect since several of the sample proportions are larger than one half.

insert Figure 1 near here

6. A PERMUTATION TEST

In order to compare the genetic diversity of an organism at an impacted site with a reference site, we shall utilize a permutation test. Let Δ_1 and Δ_2 denote the diversity indices at sites 1 and 2 respectively. Our hypotheses are

$$H_0 : \Delta_1 = \Delta_2 \text{ versus}$$

$$H_1 : \Delta_1 > \Delta_2$$

when site 1 refers to the reference site. The test statistic will be defined to be the difference $t = \hat{\Delta}_1 - \hat{\Delta}_2$. A natural way to test the null hypothesis is in terms of a permutation test. First, the raw data is used to compute the test statistic t . Next, the data for the two groups are combined and shuffled. We randomly pick n_1 vectors without replacement for group 1

and n_2 vectors for group 2 and re-compute the test statistic. From the $\binom{n_1 + n_2}{n_1}$ possible permutations, a large number of permutations are randomly selected and the test statistic t_p is re-computed for each permutation. Note that in the permutation procedure, we are re-sampling the entire vectors which insures that the population dependence structure is preserved. Westfall and Young (1989) also consider such a permutation procedure for multivariate binomial data where interest was focused on comparisons of individual components of the vector for inference.

First we consider snail RAPD-PCR data at Elk Creek and Dick's Creek. Currently there are biologically based indices used to measure the impact of anthropogenic stressors at aquatic sites. One such index is the *Index of Biotic Integrity* (IBI) (Karr 1993). The IBI values for Elk and Dick's Creek are 28 and 14 respectively. Higher values of IBI indicate greater ecological health. Thus, according to the IBI values, Elk Creek is less impacted than Dick's Creek. For the permutation test, $P = 10,000$ random permutations were obtained using the software GAUSS. The permutation test statistics t_p were ordered from smallest to largest. Recall that we reject the null hypothesis H_0 for large values of t (which would indicate a higher genetic diversity at Elk Creek than at Dick's Creek). The estimated p -value of the test is taken as the percentage of t_p which exceed the observed t from the raw data. For the snail data, we have $t = 0.0360$ with $\hat{p} = 0.0022$ giving very strong evidence of a higher genetic diversity at the reference site. Figure 2 below shows the permutation distribution of test statistics t_p . The vertical line corresponds to the observed value of the test statistic t which occurs in the extreme right tail of the

permutation distribution.

insert Figure 2 near here

RAPD–PCR data was also collected at six sites along the Black River. At each site, a sample of $n = 24$ snails was obtained. The RAPD–PCR profiles indicated a total of $k = 23$ scorable bands. Table 3 shows the estimated diversity and the associated IBI values for the six sites on the black river.

insert Table 3 near here

Figure 3 shows the bootstrap frequency polygons for the sampling distribution of diversity indices at the six Black River sites (based on 10,000 bootstrap samples). Also shown are the associated IBI values at each site. Clearly the fourth site (with IBI = 18) appears to have the lowest diversity as well as the lowest IBI value.

insert figure 3 near here

In order to test for a difference in diversities at the six sites, we can perform an Analysis of Quadratic Entropy (ANOQE) similar to a single factor analysis of variance (ANOVA) (Liu and Rao 1995). The null hypothesis of interest is

$$H_0 : \Delta_1 = \Delta_2 = \cdots = \Delta_6$$

versus

$$H_a : \text{not all } \Delta_i \text{ are equal.}$$

Let $\hat{\Delta}_i$ denote the estimated diversity at the i th site with corresponding sample size n_i .

The within population diversity is defined to be

$$W = \frac{1}{N} \sum_{i=1}^m n_i \hat{\Delta}_i$$

where m = the number of sites and $N = n_1 + \dots + n_m$. The total entropy is found by simply pooling all the data together and computing (3). Liu and Rao (1995) suggest using $T - W$ as a test statistic and they derive its asymptotic distribution. However, due to the small sample sizes, a permutation testing approach will be used here. The test statistic $T - W$ is computed for the original data yielding a value of $T - W = .04650$. Next, 10,000 random permutations of the data were obtained. For each permutation, the test statistic is re-computed. Not one of the 10,000 permutations yielded a test statistic value greater than the observed value of .0465. Thus, the estimated p -value is $\hat{p} = 0.0000$ which provides very strong evidence that the genetic diversities at the six sites are not all equal. Permutation tests comparing pairs of diversities at individual sites were also computed. The only significant differences found were comparisons involving site 4 which consistently had a significantly lower diversity.

Note that for two sites ($m = 2$), the test statistic $T - W$ is equivalent to $(\hat{\Delta}_1 - \hat{\Delta}_2)^2$.

7. GENETIC SIMILARITY AND DIVERSITY

Instead of measuring the intraspecies genetic diversity, one could also consider a measure of genetic similarity using RAPD-PCR profiles. Clark and Lanigan (1993) considered one such similarity measure in order to study the genetic relatedness of individual organisms. Let \mathbf{X}_1 and \mathbf{X}_2 denote the zero-one Bernoulli profiles for two independent organisms obtained from their respective RAPD-PCR profiles. A measure of genetic similarity be-

tween the two individuals is

$$f = \frac{2\mathbf{X}_1'\mathbf{X}_2}{\|\mathbf{X}_1\|^2 + \|\mathbf{X}_2\|^2}.$$

Thus, f can be regarded as a proportion of shared amplified bands between the two organisms. Clearly $0 \leq f \leq 1$ and $f = 1$ for two individuals with identical profiles (if both \mathbf{X}_1 and \mathbf{X}_2 are identically zero, then define $f = 1$). The *mean genetic similarity* (MGS) is $E[f]$.

Krane et al (1999) estimate MGS from RAPD-PCR data for crayfish (*Orconectes rusticus*) as follows: compute f for all pairs of organisms involving the first organism in the sample and average these values obtaining \hat{f}_1 . Repeat this for each of the other organisms in the sample. Finally, an estimate of MGS is obtained by averaging $\hat{f}_1, \dots, \hat{f}_n$.

In order to test for a difference in genetic diversity at different sites in crayfish, Krane et al (1999) base their inference on MGS using Tukey's procedure. However, the MGS is an average of the non-independent values $\hat{f}_1, \dots, \hat{f}_n$ which may be problematic. Also, $E[f]$ is a complicated expression in terms of multinomial parameters discussed in section 4.

Recall from equation (2) that $2k\Delta = \mathcal{E}\|\mathbf{X}_1 - \mathbf{X}_2\|^2$. Now $\|\mathbf{X}_1 - \mathbf{X}_2\|^2 = \|\mathbf{X}_1\|^2 + \|\mathbf{X}_2\|^2 - 2\mathbf{X}_1'\mathbf{X}_2$. Therefore, $\|\mathbf{X}_1 - \mathbf{X}_2\|^2 = (\|\mathbf{X}_1\|^2 + \|\mathbf{X}_2\|^2)(1 - f)$. This relationship between f and $\|\mathbf{X}_1 - \mathbf{X}_2\|^2$ indicates that in practice one might expect to see a negative correlation between the estimated diversity and estimated mean genetic similarity. Figure 4 below shows a scatterplot of the estimated intraspecies diversity $\hat{\Delta}$ versus estimated mean genetic diversity for snails gathered from 22 different sites in southwestern Ohio.

There is clearly a strong negative association between MGS and genetic diversity. The sample correlation from these 22 sites is -0.773 . However, one of the weaknesses of using MGS is that it is possible that two very similar organisms may have a genetic similarity measure f close to zero. To illustrate, suppose $k = 6$ and $X_1 = (0, 0, 0, 0, 0, 0)'$ and $X_2 = (0, 0, 0, 0, 0, 1)'$: then the genetic similarity f is zero although the two vectors are almost identical. If a primer is used which amplifies very few bands, then it is possible that similar organisms may receive low values of f . In particular, the three sites on the far left in figure 4 correspond to snails from one particular watershed (Little Scioto River) where there were considerably fewer amplifications of bands than for snails at all the other sites. These three sites stick out considerably in the plot of figure 4. If they are removed, then the correlation between diversity and MGS jumps to $-.991$.

insert Figure 4 near here

The problem with MGS is that f is a measure of the proportion of shared *amplified* bands. Why not simply compute the proportion of all bands (amplified or unamplified) which are shared between two organisms? If there are k possible bands, then define g to be the proportion of bands which are shared between two independent organisms. That is,

$$\begin{aligned}
 g &= \frac{\text{number of matches}}{k} \\
 &= \frac{\mathbf{X}_1' \mathbf{X}_2 + (\mathbf{J} - \mathbf{X}_1)' (\mathbf{J} - \mathbf{X}_2)}{k} \\
 &= \frac{k + 2\mathbf{X}_1' \mathbf{X}_2 - \|\mathbf{X}_1\|^2 - \|\mathbf{X}_2\|^2}{k} \\
 &= 1 - \|\mathbf{X}_1 - \mathbf{X}_2\|^2 / k.
 \end{aligned}$$

Therefore, the estimated genetic diversity measure will be perfectly (negatively) correlated with the estimated mean genetic similarity measure g .

8. DISCUSSION

As we have seen, the intraspecies diversity index for RAPD-PCR profiles is very useful for comparing the genetic diversity of snails at different sites. Studies are currently being conducted using RAPD-PCR profiles using a wide variety of other organisms as well.

The intraspecies diversity index Δ discussed in this paper is simply the total variance for a multivariate binomial distribution. The permutation test discussed in Section 7 using $\hat{\Delta}$ is a natural and effective means of testing for a decrease in genetic diversity at two sites. It would be interesting to see if a multivariate test statistic similar to Hotelling's T^2 which directly takes into consideration the correlation structure would be useful. In such a case, one sided multivariate tests would be of interest. We leave these possibilities for future research.

Acknowledgement. The authors would like to thank David Sternberg who helped collect and organize the data for this study.

REFERENCES

- Allendorf F. W., Leary R.F. (1986), "Heterozygosity and Fitness in Natural Populations of Animals," In M.E. Soule, ed., *Conservation Biology: The Science of Scarcity and Diversity*, Sinauer Associates, Inc. Sunderland, MA, p 57-76.

- Alam, K. and Williams, C. (1993), "Relative Difference in Diversity Between Populations," *Annals of the Institute of Statistical Mathematics*, **45**, 383–399.
- Clark, A. G. and Lanigan, C. M. S. (1993), "Prospects for Estimating Nucleotide Divergence Times with RAPDs," *Molecular Biology and Evolution*, **10**, 1096–1111.
- Dennis, B. Patil, G. P. Rossi, O. and Taille, C. (1979). "A Bibliography of Literature on Ecological Diversity and Related Methodology," *Ecological Diversity in Theory and Practice*, 319–354, International Co-operative Publishing House, Fairland, Maryland.
- del Tufo J.P., Tingey S.V. (1994), "RAPD assay: A Novel Technique for Genetic Diagnostics," *Methods in Molecular Biology*, **28**, 237-241.
- Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- Fritsch, K. and Hsu, J. (1999), "Multiple Comparison of Entropies with Application to Dinosaur Biodiversity," *Biometrics*, **55**, 1300–1305.
- Grosberg R.K., Levitan D.R., Cameron B.B. (1995), "Characterization of Genetic Structure and Genealogies Using RAPD-PCR Markers: A Random Primer for the Novice and Nervous," In J. D. Ferraris and S.R. Palumbi, eds., *Molecular Zoology: Advances, Strategies and Protocols*, John Wiley & Sons, New York, New York, USA, p 67-100.

- Hartl, D. L. and Clark, A. G. (1997), *Principals of Population Genetics*, Sinauer Associates, Inc., Sunderland, MA.
- Karr J. R. (1993), "Defining and Assessing Ecological Integrity: Beyond Water Quality," *Enviromental Toxicology Chem*, **12**, 1521–1531.
- Krane, D. Sternberg, D., Burton, G. (1999), "Randomly Amplified Polymorphic DNA Profile–Based Measures of Genetic Diversity in Crayfish Correlated with Environmental Impacts," *Enviromental Toxicology and Chemistry*, **18**, 504–508.
- Liu, Z. J., and Rao, C. R. (1995), "Asymptotic Distribution of Statistics Based on Quadratic Entropy and Bootstrapping," *Journal of Statistical Planning and Inference*, **43**, 1–18.
- Muirhead, R. (1982), *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New York.
- Penner G.A., Bush A., Wise R., Kim W., Domier L., Kasha K., Laroche A., Scoles G., Molinar S.J., Fedak R. (1993), "Reproducibility of Random Amplified Polymorphic DNA (RAPD) Analyses Among Laboratories," *PCR Methods and Applications*, **2**, p 341-345.
- Patil, G. P., and Taillie, C. (1979), "An Overview of Diversity," in *Ecological Diversity in Theory and Practice*, Grassle, J., Patil, G., Smith, W., and Taillie, C. (Eds.), International Co-operative Publishing House, Fairland, Maryland, p 3–27.

- Patil, G. P., and Taillie, C. (1982), "Diversity as a Concept and its Measurement,"
Journal of the American Statistical Association, **77**, 548–561.
- Patil, G. P., Boswell, M. T., Joshi, S. W., and Ratnaparkhi, M. V. (1984), *Dictionary
and Classified Bibliography of Statistical Distributions in Scientific Work, Volume 1
Discrete Models*, International Co-operative Publishing House, Fairland, Maryland.
- Pielou, E. C. (1977), *Mathematical Ecology*, John Wiley & Sons, New York.
- Pielou, E. C. (1988) "Diversity Indices," In *Encyclopedia of Statistical Sciences*, Volume
2, S. Kotz, N. Johnson, and C. Read, eds, Wiley, New York, 408–412.
- Rao, C. R. (1982), "Convexity Properties of Entropy Functions and Analysis of Diver-
sity," *Inequalities in Statistics and Probability*, IMS Lecture Notes – Monograph
Series, **5**, 68–77, Hayward, California.
- Rowland, C. and Burton, G. A. (1996), "Effect of Exposure Method on Benthic Organism
Responses," *Abstracts, 17th Annual Meeting, Society of Environmental Toxicology
and Chemistry*, Washington, D.C. November 17–21, p 284.
- Skepner, A. P. and Krane, D. E. (1999), "RAPD Reveals Genetic Similarity of *Acer
Saccharum* and *Acer Nigrum*," *Heredity*, **80**, 422–428.
- Seber, G. A. F. (1984), *Multivariate Observations*, Wiley, New York.
- Solomon, D. (1979), "A Comparative Approach to Species Diversity," in *Ecological Di-
versity in Theory and Practice*, Grassle, J., Patil, G., Smith, W., and Taillie, C.

(Eds.), International Co-operative Publishing House, Fairland, Maryland, p 29–35.

Theodorakis, C. W. and Wirgin, I. (2002), “Genetic Responses as Population-Level Biomarkers of Stress in Aquatic Organisms,” in *Bioindicators of Stress in Aquatic Ecosystems*, S. M. Adams (Ed.), Amer. Fish. Soc. (in press).

Thorpe J.E., Koonce J.F. (with Borgeson D., Henderson B., Lamsa A., Maitland P.S., Ross M.A., Simon RC, Walters C.), (1981), “Assessing and Managing Man’s Impact on Fish Genetic Resources,” *Canadian Journal of Fisheries and Aquatic Science*, **38**, p 1899-1907.

Tinker N.A., Fortin M.G., Mather D.E. (1993), “Random Amplified Polymorphic DNA and Pedigree Analysis in Spring Barley,” *Theoretical and Applied Genetics*, **85**, p 976-984.

Westfall, P. and Young, S. (1989), “*p* Value Adjustments for Multiple Tests in Multivariate Binomial Models,” *Journal of the American Statistical Association*, **84**, 780–786.

| Snail | Bands | | | | | | | | | | | | | | | | | |
|-------------|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 3 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 4 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 6 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 7 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 8 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |
| 9 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 10 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 11 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| 12 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 13 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 14 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 15 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 16 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 17 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 18 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 19 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 20 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 21 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
| 22 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 23 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| 24 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| \bar{p}_i | .375 | .958 | .583 | .708 | .458 | .750 | .417 | .875 | .833 | .625 | .833 | .625 | .917 | .375 | .625 | .583 | .958 | 1.00 |

Table 1 Data for snails (*Physella gyrinus*) using primer B01 at Elk Creek.

| Snail | Bands | | | | | | | | | | | | | | | | | |
|-------------|-------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 7 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 8 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 9 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 10 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 11 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| 12 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 13 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 14 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 15 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 16 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 17 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 18 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |
| 19 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 20 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 21 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 22 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| 23 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 24 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| \bar{p}_i | .750 | .917 | .875 | .875 | .833 | .875 | .917 | .750 | .833 | .625 | .875 | .750 | .958 | .458 | .667 | .500 | .958 | 1.00 |

Table 2 Data for snails (*Physella gyrinus*) using primer B01 at Dick's Creek.

| Site | 1 | 2 | 3 | 4 | 5 | 6 |
|-----------|------|------|------|------|------|------|
| Diversity | .139 | .128 | .125 | .076 | .137 | .136 |
| IBI | 32.6 | 28 | 25.3 | 18 | 40 | 31 |

Table 3: Estimated diversity and Index of Biotic Integrity (IBI) values for six sites along the Black River based on $n = 24$ RAPD-PCR profiles of snails at each site.

Figure Captions

Figure 1. Sampling distributions of the sample diversity indices for Elk Creek and Dick's Creek. The bootstrap distribution (based on 10,000 bootstrap samples) is denoted by the solid (Elk Creek) and dashed (Dick's Creek) curves. The dotted curves correspond to the estimated asymptotic normal distribution approximation. The IBI values for Dick's Creek (impacted) and Elk Creek (reference site) are 14 and 28 respectively.

Figure 2. The permutation distribution for comparing the diversity indices for Elk Creek and Dick's Creek based on 10,000 random permutations. The vertical line corresponds to the observed value of the test statistic. The p -value for the test is 0.0022.

Figure 3: Bootstrap distributions of estimated diversity at the six Black River sites. Also shown are the associated Index of Biotic Integrity (IBI) values at each site.

Figure 4. A scatterplot of intraspecies diversity versus mean genetic similarity (MGS) for snails at 22 sites in southwestern Ohio. The sample correlation is -0.891 .

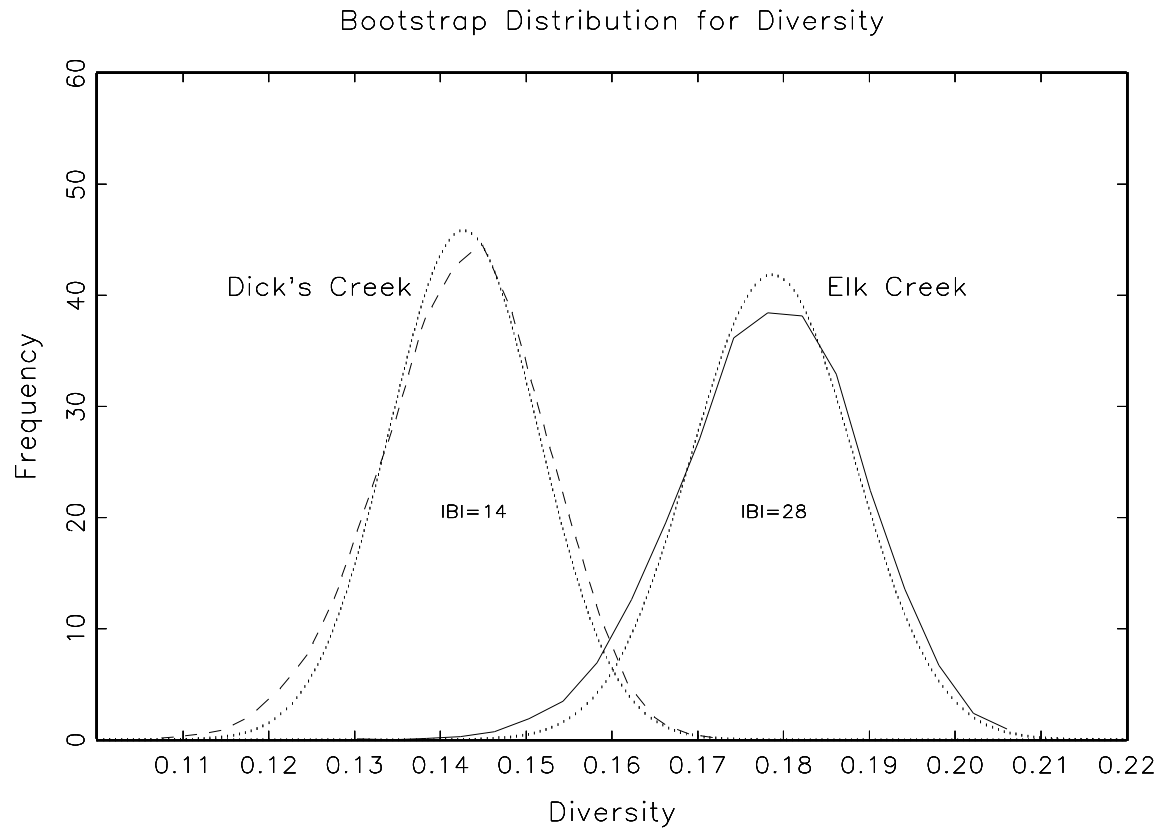


Figure 1

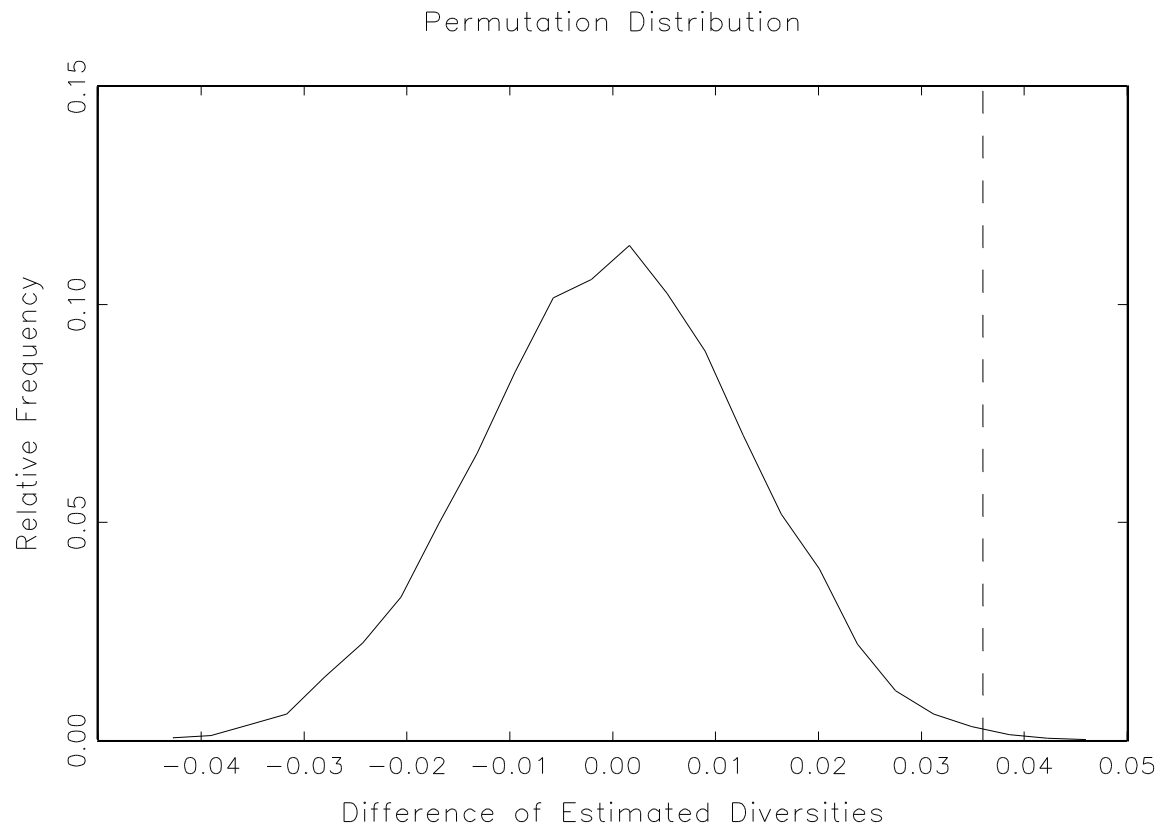


Figure 2

Bootstrap Distribution for Diversity at Black River Sites

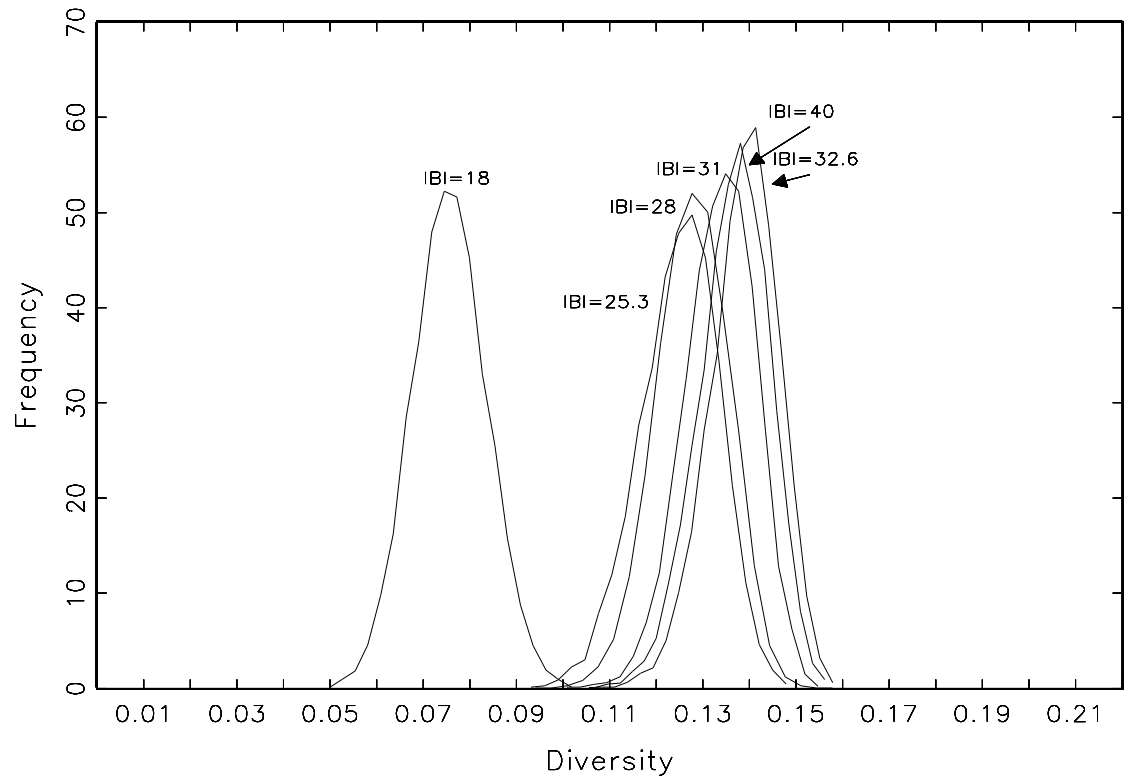


Figure 3

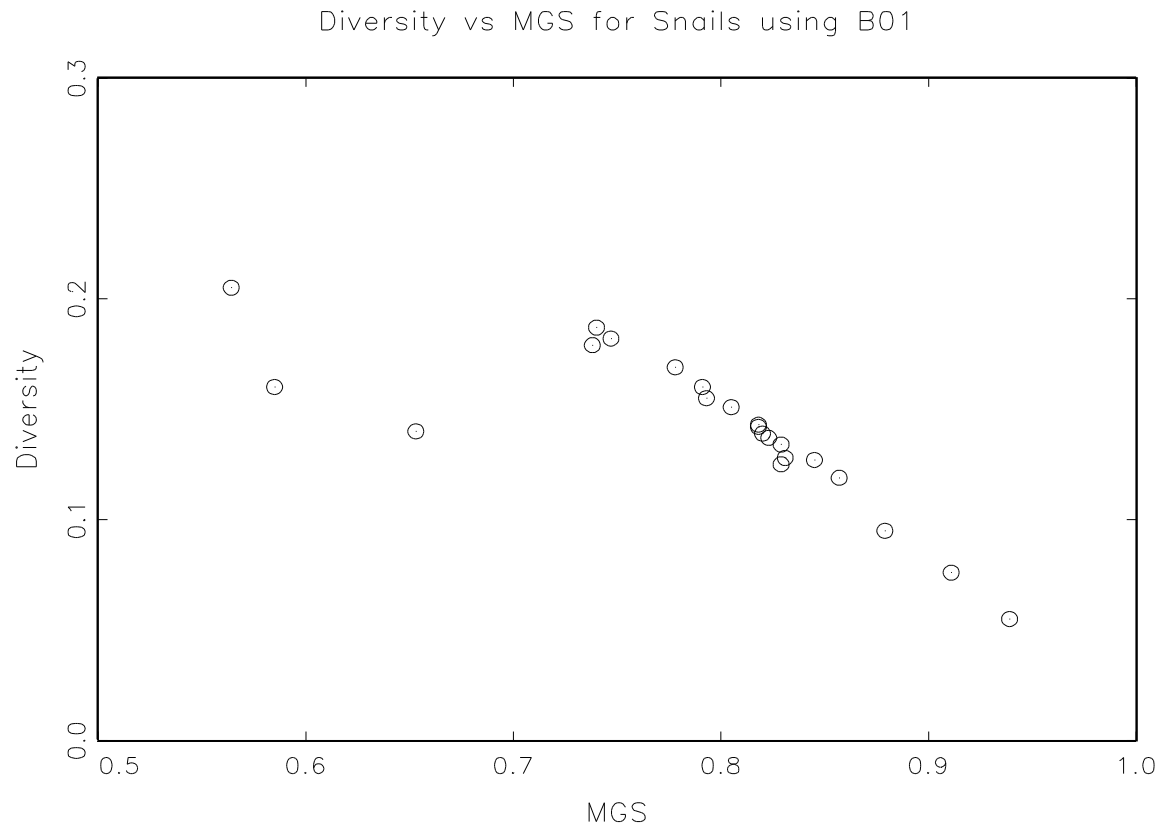


Figure 4