CALL #:	**QH359 .J68x**
LOCATION:	**NTE :: Main Library :: Internet,t2,t2p,WSU Owen**
TYPE:	Article
USER JOURNAL TITLE:	Journal of evolutionary biology
OCLC JOURNAL TITLE:	Journal of evolutionary biology
NTE CATALOG TITLE:	JOURNAL OF EVOLUTIONARY BIOLOGY
ARTICLE TITLE:	Assessing population genetic structure and variability with RAPD data: Application to Vaccinium macrocarpon (American Cranberry)
ARTICLE AUTHOR:	Stewart
VOLUME:	9
ISSUE:	2
YEAR:	1996
PAGES:	RUSH 153-171
ISSN:	1010-061X
OCLC #:	
CROSS REFERENCE ID:	452745
VERIFIED:	

BORROWER:	**TKN :: TKN-Library**
PATRON:	**Holden,Diana**
PATRON ID:	
PATRON ADDRESS:	
PATRON PHONE:	
PATRON FAX:	
PATRON E-MAIL:	
PATRON DEPT:	Libraries
PATRON STATUS:	Faculty
PATRON NOTES:	

J. Evol. Biol. 9: 153-171 (1996)

1010 061X/96/020153-19 $ 1.50 + 0.20/0

Assessing population genetic structure and variability with RAPD data: Application to *Vaccinium macrocarpon* (American Cranberry)

C. N. Stewart, Jr.[1,*] and L. Excoffier[2]

[1]*Department of Biology, University of North Carolina, Greensboro,*
NC 27412, USA
[2]*Department of Anthropology and Ecology, 12, rue G. Revilliod,*
University of Geneva, CH-1227 Carouge, Switzerland

Key words: Heteroscedasticity; random amplified polymorphic DNA (RAPD); analysis of molecular variance (AMOVA); population genetics; genetic variation.

Abstract

A method for estimating and comparing population genetic variation using random amplified polymorphic DNA (RAPD) profiling is presented. An analysis of molecular variance (AMOVA) is extended to accomodate phenotypic molecular data in diploid populations in Hardy-Weinberg equilibrium or with an assumed degree of selfing. We present a two step strategy: 1) Estimate RAPD site frequencies without preliminary assumptions on the unknown population structure, then perform significance testing for population substructuring. 2) If population structure is evident from the first step, use this data to calculate better estimates for RAPD site frequencies and sub-population variance components. A nonparametric test for the homogeneity of molecular variance (HOMOVA) is also presented. This test was designed to statistically test for differences in intrapopulational molecular variances (heteroscedasticity among populations). These theoretical developments are applied to a RAPD data set in *Vaccinium macrocarpon* (American cranberry) using small sample sizes, where a gradient of molecular diversity is found between central and marginal populations. The AMOVA and HOMOVA methods provide flexible population analysis tools when using data from RAPD or other DNA methods that provide many polymorphic markers with or without direct allelic data.

[*] Author for correspondence.

Introduction

RAPD profiling, a molecular genetic method that utilizes the polymerase chain reaction (PCR), is gaining wide usage in genetic, taxonomic, ecological, and behavioral research in bacteria, plants and animals (Hadrys et al., 1992). RAPD profiling uses single short (5–15 base) oligonucleotide primers and *Taq* DNA polymerase (or other thermostable polymerases) to amplify DNA segments between priming sites. Amplified DNA fragments may then be visualized on horizontal or vertical gels, and bands scored as presence/absence character states. A composite RAPD profile is therefore the product of all DNA bands amplified using primers that reveal useful polymorphisms. These bands have been shown to be inherited in Mendelian fashion, and therefore useful as molecular markers for qualitative and quantitative traits (Williams et al., 1990; Hadrys et al., 1992).

Molecular genetic methods such as RAPD profiling are being increasingly used in population surveys because of the ease of methodology and the numerous polymorphisms distinguishable. The advantages of RAPD profiling over allozyme analysis include:

1) Increased number of polymorphic markers available (indeed RAPD profiling has revealed polymorphisms in plant species that are monomorphic using allozymes (e.g., this study; Brauner et al., 1992).
2) DNA-level variation is revealed, which is a less biased estimator of genetic variation than gene product-level variation (allozymes).

The main disadvantage of RAPD profiling, when compared with allozyme analysis, lies in the fact that RAPD products are dominant, not codominant like allozymes (Tinker et al., 1993; Williams et al., 1993). With dominant markers heterozygotes are indistinguishable from homozygotes, and therefore allelic information is not directly available from RAPDs. Population genetics can be assessed using traditional analyses based upon allele frequencies if Hardy-Weinberg assumptions are made. However, in plants especially, this is often not appropriate because of small sample sizes and non-random mating. Therefore, researchers employing DNA typing methods often address populational quantities such as genetic heterogeneity or population structure in a qualitative and nonstatistical manner (e.g., Brauner et al., 1992), although attempts have been made to use these polymorphisms in a more rigorous population genetics framework (Huff et al., 1993; Lynch and Milligan, 1994).

Recently introduced was a general methodology based on an analysis of molecular variance (AMOVA) for the estimation and the testing of population genetic structure using both haplotype frequencies and molecular information (Excoffier et al., 1992). Its application to RAPD profile data is not straightforward because of the dominance problem resulting in the lack of information on the exact genotypes of diploid individuals. In this paper we show how the AMOVA framework can be extended to accommodate RAPD profile data, despite these obstacles, in populations with assumed amounts of self-fertilization. Thus, we use RAPD phenotypic data to estimate genotypic information. In several research areas (e.g. resource management, conservation, ecology), the comparison of the amount of genetic variability within populations is as important as the population genetic structure. We therefore introduce a nonparametric test for the homogeneity of molecular variance (HOMOVA) based on Bartlett's statistic (Bartlett, 1937). We illustrate our theoretical developments with *Vaccinium macrocarpon* (American cranberry) RAPD population data with small sample sizes. AMOVA will describe how RAPD variance is partitioned within and among populations, and test for significance against the null hypothesis of no population structure. In addition, the HOMOVA will test whether populations have different amounts of RAPD variation. The strength of the statistical tests presented here lies in their nonparametric nature. Since genetic data are not normally distributed this permutational approach is more appropriate than tests based upon Gaussian theory (e.g., G-test).

Materials and methods

Samples

We sampled 3 marginal populations (Dare County, North Carolina (NC) Johnson County, Tennessee (TN), and Pocahontas County, West Virginia (WV), and 3 central populations (Oswego County, New York (NYF), Schenectady County, New York (nye), and Houghton County, Michigan (MI)) (Fig. 1). Cranberry clonal spatial arrangements within sites are patchy and individual genets cannot be casually observed. The populations were small (< 1 ha) and discrete. The marginal populations (especially TN and NC) were geographically isolated. DNA was extracted from 4 leaf samples taken 20 m apart along a transect using a

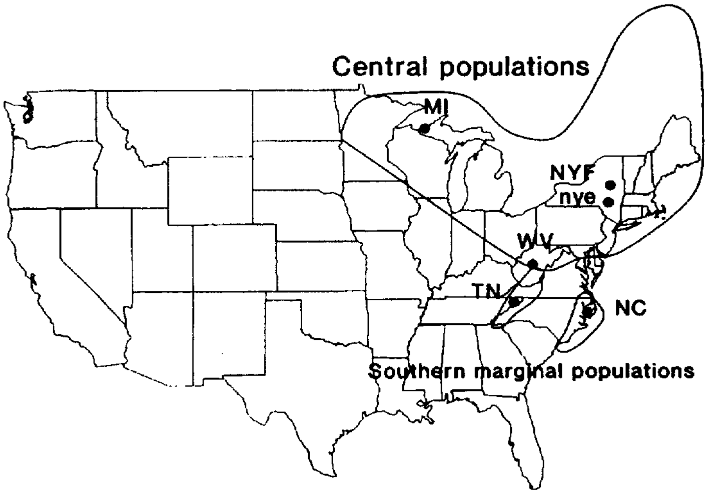

Fig. 1. Geographic location of cranberry populations sampled.

modified Doyle and Doyle (1987) procedure (Stewart and Via, 1993). RAPD reactions were performed according to methods outlined in Stewart and Via (1993) using 5 ng of template per sample. Forty primers were screened for polymorphisms (A and B RAPD primer kits from Operon Technologies, Alameda, California). Of these primers, seven revealed consistent profiles. Reproducible band states (90) were scored and data coded +/− for the 7 primers utilized (Tab. 2): OPA4 (5′ AATCGGGCTG), OPA7 (5′ GGTCCCTGAC), OPA9 (5′ GGGTAACGCC), OPA13 (5′ CAGCACCCAC), OPA18 (5′ AGGTGACCGT), OPB4 (5′ GGACTGGAGT), OPB18 (5′ CCACAGCAGT).

Estimating population genetic structure with RAPD data

The estimation of population genetic structure in an analysis of variance framework (AMOVA) has been fully described for codominant molecular markers (Excoffier et al., 1992). The principle of AMOVA is to extract variance components and analogs of F-statistics (called Φ-statistics) from a matrix of squared euclidian molecular distances between chromosomes collected into populations, themselves arranged into groups of populations. The main difference between codominant molecular markers, such as RFLP's or DNA sequences, and dominant RAPD markers, lies in the unavailability of genotypic information for the assayed individuals: one has to deal with phenotypic information only. Therefore, as the nature of the chromosomes present in one particular individual cannot be assessed, interchromosomal distances are not directly available. We are therefore led to try to estimate these unknown distances by making several assumptions. These additional assumptions allow the use of the highly polymorphic RAPD markers, but at the cost of a decreased precision of the analysis compared to that of codominant markers, as already pointed out by Lynch and Milligan (1994).

Throughout this study, we will assume the following properties of RAPD markers in populations of diploid individuals:

1) The banding pattern on gels is interpreted without ambiguity.
2) A band on a gel identifies a strictly two-allele locus. A (+) will denote the presence of a band, a (−) its absence and (+) is completely dominant over (−).
3) Different band positions can be considered as different loci.
4) Loci are independent.
5) A proportion S_i of individuals are considered as selfers in the population i.
6) At each locus, the genotypic proportions are at Hardy-Weinberg equilibrium (if $S_i = 0$) or at inbreeding equilibrium (if $S_i \neq 0$).

The consequence of possible departures from these assumptions will be addressed in the results and discussion.

We will use here the same hierarchical model of population genetic structure as described in Excoffier et al. (1992) and others (Cockerham, 1969, 1973; Weir and Cockerham, 1984; Long, 1986), with genotypes collected into populations, and populations nested into groups, leading to the analysis of variance format shown in

Table 1. General design for the analysis of molecular variance (AMOVA) of *Vaccinium macrocarpon* populations.

Source of variation	d.f.	MSD	Expected MSD
among populations	5	MSD(a)	$\sigma_w^2 + 8\sigma_a^2$
among individuals	42	MSD(w)	σ_w^2

Table 1. The model assumes that the j-th haplotype frequency vector from the i-th population in the g-th group is a linear equation of the form

$$\mathbf{x}_{ji} = \mathbf{x} + \mathbf{a}_i + \mathbf{w}_{ji}. \tag{1}$$

The vector $\mathbf{x}_{ji}$ takes the form of a boolean vector of dimension m, equal to the total number of RAPD loci is surveyed, where the presence of a site is coded as a one and its absence as a zero. The associated Φ_{ST} conventionally obtained as $\Phi_{ST} = \alpha_a^2/\sigma^2$.

A sum of squared deviations may be expressed as a function of squared distances (Excoffier et al., 1992) as (for the total sum of square deviations),

$$SSD(T) = \frac{1}{4N} \sum_{j=1}^{2N} \sum_{k=1}^{2N} \delta_{jk}^2 , \tag{2}$$

where N is the total number of diploid individuals surveyed in all populations. If the RAPD loci are assumed independent and providing the same amount of information, the genetic distance between chromosomes j and k (δ_{jk}^2) is of the form

$$\delta_{jk}^2 = (\mathbf{x}_j + \mathbf{x}_k)' \quad (\mathbf{x}_j + \mathbf{x}_\mathrm{k}) \tag{3a}$$

This Euclidian distance is equivalent to the number of different RAPD markers between two haploid genomes and may be rewritten as

$$\delta_{jk}^2 = \sum_{s=1}^{m} (x_{sj} - x_{sk})^2 , \tag{3b}$$

where the subscript 's' indexes the m RAPD loci.

As RAPD data consist of mostly dominant markers, we can only compare individual phenotypes. An AMOVA can however be performed if we can translate distances between multilocus vectors in terms of phenotypic distances between the RAPD profiles of two individuals, thus transforming equation (2) into

$$SSD(T) = \frac{1}{4N} \sum_{j=1}^{N} \sum_{k=1}^{N} \Delta_{jk}^2 , \tag{4}$$

where Δ_{jk}^2 is a squared distance between two individuals. In the diploid case, the inter-individual distance (Δ_{jk}^2) is simply the sum of four inter-haplotypic distances,

$$\Delta_{jk}^2 = \sum_{i=j_1}^{j_2} \sum_{l=k_1}^{k_2} \delta_{il}^2 , \tag{5}$$

the only problem being that we ignore the individual's genotype. We can however

Table 2. RAPD haplotypes found in six population samples.

Populations	RAPD halpotypes
NYF	1001110100010010100001000001001101101011111111111010000111111101010110111111011001111110101
	100111000001001010000100000100110110101111111111101000011111111010001101111101100111110101
	1001110000010010100001000011100101101011111111111010000111111110100011010111101100011110101
	10111100000100101000010110110001011010111111111111010000111111110101011010111101110111111 0101
NYE	1001110010010100001000010011011010111111111110100001111111001010110101111101100011110101
	10011100000110101000010000010011011010111101111101000011111100101011010011101011110101
	100111000001001010000100000100110110111111111111101000011111110010101100000111011000111 10101
	1001110000010010100001000001010111101011111111111111010000111111100101011010111101101010111 10111
MI	1001110010010110110101000001001111101011101111111101000011100100011100111101010111110101
	1101110010010010100101011101001111101011101111111101000010110100100011000111101010011110101
	1001110010010010110001000001001111101011101111111100000010111100100111100111101010011110111
	1001110010010010100101011101001111101011101111111101000010110100100011000111101010011110011
WV	1001110010010010101101000001001111111011101011111101000010110010001100011110101011011000111
	10011100100100101011010000010011111110111010111111010000101101001000110001111010101011100100 1
	1001110010010010101111000001001111111011101011111101000010110100100011010111101010110110001 11
	1001110011010011011010000010011111101110101111111110100101101001000110001111011111101101 0111
TN	1001110000010010100001000001001111111011101111111101000010110100101011000111101010011110101
	1001110000010010100001000001001111111011101111111101000010110100101011000111101010011110101
	1001110000010010100001000001001111111011101111111101000010110100101011000111101010011110101
	1001110000010010100001000001001111111011101111111101000010110100101011000111101010011110101
NC	1001010010110010100101000001001111111011101111111101100010101100110011000111101011011010101
	1001010010110010100101000001001111111011101111111101100010101100110011000111101011011010101
	1001010010110010100101000001001111111011101111111101100010101100110011000111101011011010101
	1001010010110010100101000001001111111011101111111101100010101100110011000111101011011010101

replace Δ^2_{jk}, by its expectation, which is the expected number of RAPD loci differences among four haploid genomes.

$$E(\Delta^2_{jk}) = \sum_{s=1}^{m} E\left(\sum_{i=j_1}^{j_2} \sum_{l=k_1}^{k_2} (x_{si} - x_{sk}) = \sum_{s=1}^{m} E(\Delta^2_{sjk} \right) \tag{6}$$

Hardy-Weinberg equilibrium

Let us first consider a single RAPD site (the s-th) and compare two diploid individuals drawn from arbitrary populations A and B, where the frequencies of the presence of the s-th site are p_A and p_B, respectively, the frequencies of its absence being q_A and q_B. We will turn later on to the problem of estimating site frequencies. The absence of a band for a particular individual implies a site homozygosity $-/-$, but its presence may be both due to homozygotes $+/+$ or heterozygotes $+/-$ arising in different proportions, depending on unknown population site frequencies. Four cases can happen when we compare two individuals j and k: both individuals are $[+]$; both individuals are $[-]$; individual j is $[-]$ and individual k is $[+]$; individual j is $[+]$ and individual k is $[-]$. If Hardy-Weinberg equilibrium can be assumed, the conditional expectations of Δ^2_{sjk} may be found in the following way:

$$\begin{aligned} E(\Delta^2_{sjk}\,|j=[+], k=[-]) &= \frac{4p_A^2 q_B^2}{p_A^2 q_B^2 + 2p_A q_A q_B^2} + \frac{4p_A q_A q_B^2}{p_A^2 q_B^2 + 2p_A q_A q_B^2} \\ &= \frac{4}{1+q_A}, \end{aligned} \tag{7a}$$

where the first term of the second member of the equation stands for the case where both individuals are homozygotes ($+/+$ and $-/-$), and the second term represents the case where indivual j is a heterozygote $+/-$. This estimator can vary in the range [2–4], depending on the proportion of heterozygotes $+/-$ in population A. Similarly we have

$$E(\Delta^2_{sjk}\,|j=[-], k=[+]) = \frac{4}{1+q_B}\,. \tag{7b}$$

when the s-th band is absent in both individuals, we have

$$E(\Delta^2_{sjk}\,|j=[-], k=[-]) = 0\,. \tag{7c}$$

In the case where both individuals present a band at the s-th site, we find

$$E(\Delta^2_{sjk}\,|j=[+], k=[+] = \frac{4(q_A+q_B)}{1+q_A q_B + (q_A+q_B)}\,. \tag{7d}$$

A particular case occurs however when an individual RAPD profile is compared to itself. When $j=k$, we have

$$E(\Delta^2_{sjj}\,|j=[-], j=[-]) = 0\,, \tag{8a}$$

and

$$E(\Delta_{sjj}^2 \mid j = [+], j = [+]) = \frac{8q_A^2}{p_A^2 + 4q_A^2}. \tag{8b}$$

Partial selfing

When Hardy-Weinberg equilibrium does not hold because of inbreeding, similar estimates can be found if one assumes that genotype inbreeding proportions have attained equilibrium in populations A and B with known amount of self-fertilization S_A and S_B such that

$$\begin{aligned} &P(\textit{individual } j \textit{ is } +/+) = p^2 + \frac{Spq}{2-S}, \\ &P(\textit{individual } j \textit{ is } +/-) = \frac{4pq(1-S)}{2-S}, \\ &P(\textit{individual } j \textit{ is } -/-) = q^2 + \frac{Spq}{2-S}. \end{aligned} \tag{9}$$

Using these equilibrium proportions, we find

$$E(\Delta_{sjk}^2 \mid j = [+], k = [-]) = \frac{4(2-S_A)}{2 + 2q_A(1-S_A) - S_A}, \tag{10a}$$

$$E(\Delta_{sjk}^2 \mid j = [-], k = [+]) = \frac{4(2-S_B)}{2 + 2q_B(1-S_B) - S_B}, \tag{10b}$$

$$E(\Delta_{sjk}^2 \mid j = [-], k = [-]) = 0, \tag{10c}$$

$$\begin{aligned} &E(\Delta_{sjk}^2 \mid j = [+], k = [+]) \\ &\quad = \frac{\begin{array}{c} 8[q_A(1-S_A)(2-S_B) + q_B(1-S_B)(2-S_A) \\ + q_A q_B S_A S_B (1-S_A)(1-S_B)] \end{array}}{\begin{array}{c} 4(1-S_A)(1-S_B) + 4q_A(1-S_A) + 4q_B(1-S_B) \\ + 4q_A q_B (1-S_A)(1-S_B) + S_A(2-S_B) + S_B(2-S_A) \end{array}}. \end{aligned} \tag{10d}$$

One can check that equation (10) is equivalent to equation (7) when S_A and S_B are both equal to zero. For complete selfing populations made up of homozygous individuals, the expectations of the inter-individual distances at one site are zero when both individuals share the presence or the absence of a band and four when a band shows up for one individual only. As before, when $j = k$, the analogue of equation (8b) is

$$E(\Delta_{sjj}^2 \,|j=[+],j=[+])$$
$$=\frac{32q_A^2(1-S_A)^2}{4p_A^2(1-S_A)^2+16q_A^2(1-S_A)^2+4S_A(1-S_A)p_A+S_A^2}. \quad (11)$$

Using equation (6b), the distance between two individual RAPD profiles is then simply obtained as the sum of the distances for each site, under the assumption that RAPD loci are independent. Thus, providing adequate transformations, an analysis of molecular variance can be done for RAPD data on the basis of a matrix of inter-individual distances instead of inter-chromosomal distances as in the case of the original AMOVA.

Testing population genetic structure

If site frequencies (p's) are not known, they must be estimated from the data. A problem occurs because site frequency estimates depend on the unknown population structure and the partition of the individuals into discrete populations. The simultaneous estimation of both site frequencies and population genetic structure does not appear possible, and we will adopt a two-step strategy to obtain estimates of population genetic structure indices and their significance. This detour is a consequence of the dominance pattern of RAPD data. Our first step will be to estimate RAPD site frequencies without preliminary assumptions on the unknown population structure. Under the null hypothesis of no population structure, all the individuals can be assumed to be drawn from a single panmictic population. If the proportion of individuals not showing any band for the s-th locus is Q_s, an estimator of q_s can be obtained under Hardy-Weinberg equilibrium as

$$\hat{q}_s=\sqrt{Q_s}\,, \quad (12a)$$

which is biased for low values of q. In case of partial selfing, an estimator can be found under inbreeding equilibrium as

$$\hat{q}_s=\frac{\sqrt{S^2(1+8Q_s)-24SQ_s+16Q_s-S'}}{4-4S'}\,, \quad S\neq 1, \quad (12b)$$

$$\hat{q}_s=Q_s\,, \quad S=1, \quad (12c)$$

where S is a weighted average of the selfing proportions over all sub-populations. Note that (12b) reduces to (12a) when $S=0$, thus having a similar bias. Estimators given by (12) can be used in equations (7)–(11) to define an appropriate inter-individual distance matrix between individuals (Δ_0) and performing a preliminary AMOVA. Such a distance matrix is insensitive to permutation of individuals across populations, as the total sums of squares remain unchanged through permutation and is very conservative for the null hypothesis of no population structure. Thanks to the former property, we can obtain variance components and Φ-statistics null

Table 3. Phenotypic and genotypic distances among individuals. Below diagonal: Number of RAPD site differences between pairs of phenotypes. On diagonal and above: Expected number of RAPD site differences among genotypes, taking into account population selfing proportions (0.9) and population site frequencies.

NYF				NYE				MI				WV				TN				NC			
2.14	14.36	23.13	30.34	19.87	23.56	23.56	22.81	56.99	63.63	56.92	60.48	62.67	62.67	62.71	81.88	43.58	43.58	43.58	43.85	70.29	70.29	70.29	70.29
4	0.48	0.62	23.43	20.47	24.16	24.16	22.87	50.93	57.58	50.87	54.42	56.62	56.62	56.66	75.83	44.24	44.24	44.24	44.24	64.24	64.24	64.24	64.24
7	3	1.18	20.98	29.64	33.33	33.33	26.03	60.74	67.38	60.67	64.23	66.47	66.47	66.51	85.69	54.11	54.11	54.11	54.11	74.11	74.11	74.11	74.11
9	7	6	3.64	35.81	39.50	39.50	32.20	67.27	63.30	67.21	60.14	72.94	72.94	72.98	92.15	53.84	53.84	53.84	53.84	80.56	80.56	80.56	80.56
5	5	8	10	2.09	17.95	17.95	18.85	47.86	49.20	47.80	46.04	52.61	52.61	53.37	72.54	39.37	39.37	39.37	39.37	61.00	61.00	61.00	61.00
6	6	9	11	5	0.91	14.29	20.80	57.83	64.48	57.77	61.33	55.54	55.54	56.29	82.17	43.17	43.17	43.17	43.17	71.17	71.17	71.17	71.17
6	6	9	11	5	4	0.91	20.80	50.67	57.32	50.61	54.16	55.54	55.54	63.00	69.54	35.97	35.97	35.97	35.97	63.97	63.97	63.97	63.97
6	6	7	9	5	6	6	1.98	51.41	58.06	51.35	54.90	56.19	56.19	56.94	76.12	36.59	36.59	36.59	36.59	64.59	64.59	64.59	64.59
15	13	16	18	12	15	13	13	2.04	27.94	16.81	24.78	26.40	26.40	33.86	60.45	29.23	29.23	29.23	29.23	35.31	35.31	35.31	35.31
17	15	18	16	12	17	15	15	8	2.04	33.63	7.96	26.40	26.40	33.86	60.45	29.23	29.23	29.23	29.23	42.02	42.02	42.02	42.02
15	13	16	18	12	12	13	13	4	10	1.82	30.48	33.56	33.56	41.02	67.61	29.20	29.20	29.20	29.20	42.49	42.49	42.49	42.49
16	14	17	15	11	16	14	14	7	1	9	1.34	23.25	23.25	30.70	57.30	26.07	26.07	26.07	26.07	38.87	38.87	38.87	38.87
16	14	17	19	13	14	14	14	7	7	9	6	0.01	0.61	8.06	41.52	20.30	20.30	20.30	20.30	32.30	32.30	32.30	32.30
16	14	17	19	13	14	14	14	7	7	9	6	0	0.01	0.06	41.52	20.30	20.30	20.30	20.30	32.30	32.30	32.30	32.30
16	14	17	19	13	14	16	14	9	9	11	8	2	2	0.11	48.97	27.76	27.76	27.76	27.76	39.76	39.76	39.76	39.76
21	19	22	24	18	21	17	19	16	18	18	15	11	11	13	0.38	54.36	54.36	54.36	54.36	66.36	66.36	66.36	66.36
11	11	14	14	10	11	9	9	8	8	8	7	5	5	7	14	0.00	0.00	0.00	0.00	36.00	36.00	36.00	36.00
11	11	14	14	10	11	9	9	8	8	8	7	5	5	7	14	0	0.00	0.00	0.00	0.00	36.00	36.00	36.00
11	11	14	14	10	11	9	9	8	8	8	7	5	5	7	14	0	0	0.00	0.00	36.00	36.00	36.00	36.00
11	11	14	14	10	11	9	9	8	8	8	7	5	5	7	14	0	0	0	0.00	36.00	36.00	36.00	36.00
18	16	19	21	15	18	16	16	9	11	11	10	8	8	10	17	9	9	9	9	0.00	0.00	0.00	0.00
18	16	19	21	15	18	16	16	9	11	11	10	8	8	10	17	9	9	9	9	0	0.00	0.00	0.00
18	16	19	21	15	18	16	16	9	11	11	10	8	8	10	17	9	9	9	9	0	0	0.00	0.00
18	16	19	21	15	18	16	16	9	11	11	10	8	8	10	17	9	9	9	9	0	0	0	0.00

distributions by randomly permuting individuals across populations, as was done for codominant markers (Excoffier et al., 1992): a random collection of samples is generated by allocating individuals to random populations, keeping sample sizes constant; population statistics (sums of squared deviations, molecular variances, Φ_{ST}) are then computed for these new samples; the null distribution is obtained by repeating this procedure many times. In case of significant population structure, meaning that populations are reproductively isolated, the second step of our strategy can be performed. The second step consists of computing site frequencies for each sub-population using equation (12), which implies new distance matrices among phenotypes (see the upper diagonal matrix in Table 3). This assumes Hardy-Weinberg equilibrium or, as in this case, assumes a selfing frequency estimated by biological means. These refined input distance matrices can be used to obtain more accurate estimates of variance components and related Φ-statistics. Note that the significance level of those estimates cannot be computed with a permutational approach as the sums of square deviations would change under permutation. However, these more accurate site frequencies can only increase population difference measures, and these estimates should be considered to have the same significance level than those computed in the first step. This implies that AMOVA is less powerful with dominant than with codominant markers.

Testing molecular variance homogeneity

Conventional in analyses of variance is the assumption of variance homogeneities among populations. With normality of the variates, these assumptions are crucial for the proper parametric testing of variance ratios, but they are not required when using nonparametric methods for testing variance-component significance (Excoffier et al., 1992). In many ecological or biodiversity studies, testing for within-population genetic diversity is important in itself (Antonovics, 1984, Ford-Lloyd and Jackson, 1986, Altukhov, 1990, Millar and Libby, 1991). Other quantitative approaches such as the utilization of the Shannon diversity index have been attempted to estimate RAPD diversity within populations (e.g., Russell et al., 1993) but significance testing was not possible.

Conventional testing procedures, such as Bartlett's test of homogeneity of variance (Bartlett, 1937) are parametric and assume normality of the variates for proper testing procedure. However, Bartlett's statistic (B) is meaningful as it expresses a deviation of population variances from the mean total variance. It can be formulated in terms of sum of squared deviations as

$$B = \frac{(N-P)\ln\left(\dfrac{SSD(T)}{N-P}\right) - \sum_{i=1}^{P} (N_i - 1)\ln\left(\dfrac{SSD(WP)_i}{(N_i-1)}\right)}{1 + \dfrac{1}{3(P-1)}\left(\sum_{i=1}^{P} \dfrac{1}{N_i - 1} - \dfrac{1}{N-P}\right)}, \qquad (13)$$

where P is the number of populations, $SSD(T)$ is the sum of squared deviations pooled over all samples, and $SSD(WP)_i$ is the sum of squared deviations within the

i-th population given by equation (2) with N replaced by N_i. B should follow a Chi-square distribution with $P - 1$ degrees of freedom if profile frequencies were normally distributed. This assumption certainly does not hold and several methods less sensitive to nonnormality have been proposed (reviewed in Martin and Games, 1977). Therefore, we propose to test if the observed B value is significant without assuming normality of the variates by computing Bartlett's statistics null distribution using the same permutational approach used for variance components. Note that this statistic can be computed provided that all population molecular variances differ from zero. However, heteroscedasticity can be safely assumed when some populations are homogeneous and others are not, thus avoiding the need of performing a formal test.

Results and discussion

Samples

The RAPD site states for the 24 sampled individuals are shown in Table 2, and typical results from one set of reactions visualized by ethidium bromide-stained RAPD products electrophoresed in an agarose-based gel are shown in Fig. 2. Two populations sampled (NC & TN) appeared to be monomorphic. In another study (Stewart and Nilsen, 1995b) where extensive ramet sampling was done in TN, only 4 of 22 samples had different RAPD profiles compared to the predominant clone, which, by chance, was only sampled here. So, TN is, in fact, nearly monomorphic. Intensive sampling of NC shows that it truly is monomorphic (Stewart, unpublished data). Although larger sample sizes could have been used, there were resource constraints that prevented this. Sample sizes can affect the analysis in two ways: 1) Estimates accuracy: the site frequencies will be better estimated with larger sample sizes, and therefore variance components and the F statistics will also be more accurate with larger sample sizes; 2) Power to detect significance; the resampling strategy used here will be more powerful with larger sample sizes, i.e., the same ϕ_{ST} value may not be significant with small sample sizes, but significant with larger sample sizes. For cranberry, larger sample sizes did not seem necessary for our

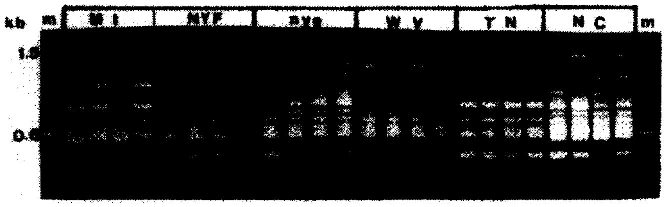

Fig. 2. Gel photograph of an ethidium bromide stained RAPD gel (0.8% agarose, 0.8% synergel). The fragments are a product of RAPD-PCR using primer OPA13 (5′ CAGCACCCAC). See text for population abbreviations and location. $m = 100$ base pair ladder (Gibco-BRL, Bethesda, Maryland).

purposes because of obvious population differentiation, which was detected with sufficient power by the AMOVA.

Selfing

Vaccinium macrocarpon is self-fertile and flower morphology and pollination biology suggest that autogamy is the reproductive norm (Reader, 1977). The species itself is highly homozygous, as miniscule levels of allozyme diversity have been revealed (Hugan et al., 1993). Stewart and Nilsen (1995b) mapped clones in the TN and WV population using RAPDs and found little clonal heterogeneity within populations. In homogeneous populations greater amounts of geitonogamy (pollination within a clone) would increase selfing frequencies to approach 100% (Handel, 1983). However, cross-pollination has been reported in cranberry (Bain, 1933) and fruit yield is reported to be greater in heterogeneous populations (Marucci and Filmer (1964). The lack of genetic diversity within populations and the factors mentioned above reinforce the paradigm of autogamy in *V. macrocarpon*. Because of this, a conservative estimate of selfing in all populations (of which there are none in the literature) was arbitrarily set at 90%. As we will see in the next section, the significance level for Φ_{ST} under various selfing rate assumptions did not change.

Population structure

Three types of analyses were performed according to different assumptions on the RAPD data, leading to different input distance matrices. We first carried out an AMOVA from a phenotypic distance matrix (**DP**), shown in the lower diagonal matrix in Table 3. This analysis is similar to that done in Huff et al. (1993) and is essentially an analysis of phenotypic variance. The results are reported in Table 4 and suggest a very strong phenotypic structure of the populations with a Φ_{ST} of 0.669.

Table 4. Population statistics estimated according to different assumptions on the data. **DP**: Matrix of phenotypic distances between individuals; **DG1 DG2 DG1′ DG2′**: Matrices of genotypic distances between individuals. See text for their definition. Bartlett statistics cannot be computed when some population molecular variances are equal to zero.

Population	Input square distance matrices				
Statistics	**DP**	**DG1**	**DG2**	**DG1′**	**DG2′**
σ_a^2	4.065***	3.537***	0.456***	4.317	4.184
σ_w^2	2.013***	2.237***	4.862***	1.622	1.588
Φ_{ST}	0.669***	0.613***	0.086***	0.727	0.725
B	N.A.	8.379**	0.745**	N.A.	N.A.

** $0.01 > p > 0.001$;
*** $p < 0.001$;
N.A. not available.

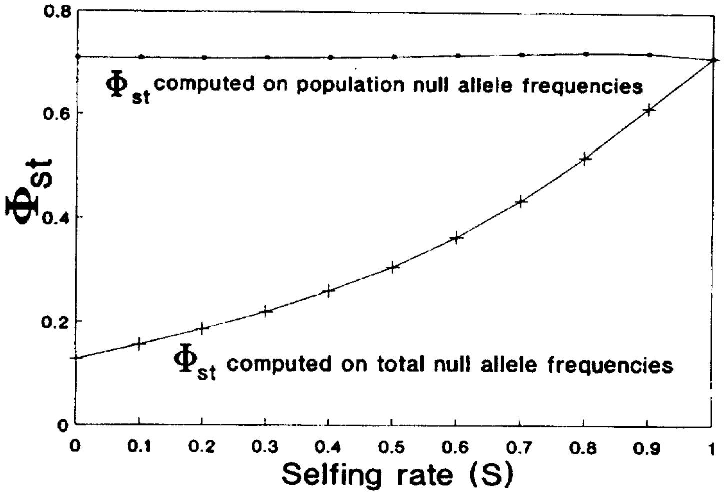

Fig. 3. The population structure of *V. macrocarpon* based upon different selfing rates and input allele frequencies.

Our two-step strategy was then applied to the data under two assumptions: (a) the selfing proportion is 0.9, and (b) an absence of selfing ($S = 0$) in all populations, giving rise to two different genotypic distance matrices **DG1** and **DG2**, respectively. The resulting AMOVA analyses gave very contrasting results, with more than 61% of the total genotypic variance due to differences among populations for **DG1**, and more than 91% of the total genotypic variance due to differences within populations for **DG2**. This resulting difference may be due to the fact that, when selfing is assumed to be absent, a large proportion of individuals are considered as heterozygotes by the analysis when a RAPD band is present, whereas, in fact, they are mostly homozygotes. Thus, assuming an absence of selfing and using the same raw profile data, interindividual differences within populations would be computed to be much larger than when no selfing is assumed and will thus account for a larger proportion of the total variance (Fig. 3). Thus, convergence occurs in Fig. 3 because, when the selfing rate is 1, all individuals are assumed homozygous and allele frequencies are inconsequential. However, it is obvious that no matter what selfing frequency is assumed in this data set, that Φ_{ST} although variable, is significant in the continuum from $S = 0$ to $S = 1$. In addition, and more importantly, the ϕ statistics derived from the second step of the analyses, for $S = 0.9$ and $S = 0$ (from **DG1′** and **DG2′**, respectfully as explained below) are nearly the same.

Although qualitatively very dissimilar for matrices **DG1** and **DG2**, δ_a^2 and Φ_{ST} are significantly different from zero in both cases. These significant results allow us to compute RAPD site frequencies separately for each population and generate two new genotypic distance matrices, **DG1′** and **DG2′**, where the prime indicates the second step of our strategy. For instance, **DG1′** is shown as the upper diagonal matrix in Table 3. It differs from the phenotypic matrix (the lower diagonal matrix in Table 3) by two main features: first, its diagonal is non-zero because shared RAPD sites at the individual level do not imply identical genotypes; second, off-diagonal elements differ by approximately a factor four, as genotypic distances correspond to four inter-chromosomal distances. **DG1**′ is our best estimate of the inter-chromosomal differences. The population structure indices computed from these two matrices are very similar, with Φ_{ST} values of 0.727 and 0.725 for **DG1**′ and **DG2**′, respectively. These figures attest a stronger genetic structure than for the phenotypic case inferred from **DP**, suggesting that population differences are slightly underestimated when the genotypic level is not considered. Nonetheless, the Φ_{ST} estimate derived by the original AMOVA (0.669) is not drastically different than the Φ_{ST} estimate using AMOVA for RAPDs in cranberry. The close agreement between Φ_{ST} estimates obtained in the 2nd step for cases $S=0$ and $S=0.9$ is quite striking. It suggests that the inferred genetic structure among populations is not very sensitive to the degree of population in breeding, a topic of additional inquiry. Although significance testing cannot be done at this stage, the population statistics are necessarily significant because the significant results of the former steps were obtained under less discriminant conditions. In the second-step, where site frequencies are estimated for each sub-population using equation (12), the use of correct selfing proportions is less crucial than in the first step to correctly infer genotypic distances, because the low site polymorphism observed at the population level results in the apparent prevalence of homozygous individuals.

Population heterogeneity

The computation of Bartlett's statistic reveals that molecular variances are significantly heterogeneous among populations in both cases. Note that a conventional parametric test, assuming a Chi-square distribution of B, would not have revealed significance for **DG2**. Because apparent monomorphism in two populations the entire data set cannot be tested for local differences. If we omit the NC data, which is truly monomorphic, and use the variance computed for intensive sampling of TN (22 samples) we get the following values for σ^2 revealing significant differences at the 0.01 level: NC (0), TN $(0.12) <$ WV $(1.0) <$ NYF $(1.74) \approx$ nye $(1.69) \approx$ MI (1.80). We assume homogeneous populations are different than heterogeneous populations. Thus we see a gradient of genetic heterogeneity from marginal to central populations.

Assumptions and RAPD reproducibility

There may be deviations from our assumptions that must be considered when interpreting the results. First, all bands may not represent dominant polymorphisms. Tinker et al. (1993) showed that approximately 1 in 30 RAPD polymorphisms in barley are codominant. Second, different band positions may not represent different loci because of the possibility of nested inverted repeats of primer sequences within a locus. This phenomenon may be revealed when performing duplicate RAPD reactions. In comparing duplicate reactions visualized on gels, one may observe few higher molecular weight bands replaced by several low molecular weight bands (Penner et al., 1993, Williams et al., 1993). Thus, analyzing duplicate reactions for reproducibility and also scoring RAPD products only in the middle molecular weight range (Stewart and Porter, 1995) will minimize the possibility of scoring products resulting from nested inverted repeat motifs. For example, primers OPA11 and OPA17 showed the inconsistent amplification as described above. Thus, they were not used in the analysis. Third, bands may co-migrate and be non-homologous. This occurrence is minimized by choosing primers that generate relatively few RAPD products per lane (McClelland and Welsh, 1994). Thus, by heeding the precautions mentioned above much of the irreproducibility reported by Penner et al. (1993) may be avoided (McClelland and Welsh, 1994), and more assumptions for the analysis will be validated. We had four criteria in the selection of primers and band scoring. The reactions had to: 1) reveal polymorphisms, 2) consistently produce strong (brightly staining) amplification products, 3) produce uniform reproducible markers between replicate PCRs, 4) be insensitive to DNA template concentrations varying from 1 ng/uL to 100 ng/uL (McClelland and Welsh, 1994). Furthermore, we only scored reproducible fragments (shared fragments between replicate RAPD reactions) that were in the mid-molecular weight range (see Penner et al. 1993; Stewart and Porter, 1995).

Conclusion

The methodology presented here allows the estimation of population genetic structure at the genotypic level with RAPD profile data. It extends the treatment presented in Huff et al. (1993) in explicitly dealing with the dominance pattern of RAPD data and in giving estimates of fixation indices that are similar in nature to those obtained with codominant-diploid or haploid data (Excoffier et al., 1992). It therefore has the advantage that population structure estimates can be directly compared to those obtained for allozyme data by using Weir and Cockerham's (1984) method and co-dominant molecular data such as RFLPs, DNA sequences, or microsatellites by using AMOVA (Excoffier et al., 1992). Like Lynch and Milligan (1994), the AMOVA approach computes an average F-statistic over all sites strictly equal to the weighted average defined by Weir and Cockerham (1984). In contrast to the methodology presented by Lynch and Milligan (1994), we chose not to assume any population structure before computing site frequencies in our first step. Paradoxically, for an analysis of variance, this leads us to consider that all populations have similar site frequencies. It is the uneveness in banding pattern among samples that will be actually used to test for sample differences. Therefore, it appears that a phenotypic analysis of variance as done by Huff et al. (1993) could be an alternative to our first step to assess the degree of population divergence. It should also be clear that the amount of genetic structure inferred from the first step is not meant to lead to correct estimates, as the purpose of this step is to only furnish significance levels. The second step must then be performed to obtain usable estimates of Φ statistics.

Vaccinium macrocarpon is a vegetatively spreading clonal plant that likely reproduces sexually primarily by selfing. Other studies have shown that clonal spread is positively associated with marginality (Stewart and Nilsen, 1995a,b). That is, in marginal sites there are few widespread clones compared to many smaller clones in central sites. Genetic diversity has not been successfully measured by allozyme analysis in cranberry presumably because of high homozygosity, although very low levels of allozyme polymorphisms have been detected in *V. macrocarpon* recently (Hugan et al. 1993). RAPD profiling, however, reveals many polymorphisms within and among populations. HOMOVA estimates differences within population genetic heterogeneity that are obvious in cranberry. However, if differences are obscure, then HOMOVA may resolve these using larger sample sizes. It is commonly thought that genetically homogeneous populations are less stable and flexible than genetically heterogeneous populations (Millar and Libby, 1991). Quantifying population genetic homogeneity is therefore important in conservation biology and ecology and the HOMOVA provides a direct estimation of this parameter. This generalized method may be used with other molecular markers, with few required assumptions. Thus, DNA typing methods such as RAPD profiling that do not provide allele frequencies, but many polymorphic markers may be used in a comprehensive populational analysis using the AMOVA and HOMOVA in tandem.

Acknowledgements

We appreciate Peter Smouse for his ideas and discussion. Thanks to B. J. Turner for suggesting RAPDs as a population-level markers, and thanks to Terry Sharik for collecting the Michigan cranberry samples. We appreciate the helpful comments of three anonymous reviewers on a prior draft. Laboratory work was performed in the lab of E. T. Nilsen, and supported by grants to C.N.S. from Ocean Spray Cranberries, Inc. and Sigma Xi, the Scientific Research Society, and USDA-BRARGP, and Swiss FNRS grant No 32-37821.93 to L.E. A computer program written for Windows 3.1 to carry out the computations done in this paper is available from L.E. upon request.

References

Altukhov, Yu. P. 1990. Population Genetics Diversity and Stability. Harwood Academic Publishers, London.

Antonovics, J. 1984. Genetic variation within populations. pp. 229–241. *In* R. Dirzo and J. Sarukhan (eds.), Pespectives on Plant Population Ecology Sinauer. Sunderland, Massachusetts.

Bain, H. F. 1933. Cross pollinating the cranberry. Proceedings of the Wisconsin State Cranberry Grower's Association 47th Annual Summer Meeting 7–11.

Bartlett, M. S. 1937. Some examples of statistical methods of research on agriculture and applied biology. Journal of the Royal Statistical Society Supplement 4: 137 170.

Brauner, S., D. J. Crawford and T. F. Stuessy. 1992. Ribosomal DNA and RAPD variation in the rare plant family Lactoridaceae. American Journal of Botany 79: 436–1439.

Cockerham, C. C. 1969. Variance of gene frequencies. Evolution 23: 72–83.

Cockerham, C. C. 1973. Analysis of gene frequencies. Genetics 74: 679–700.

Doyle, J. J. and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19: 11–15.

Excoffier, L., P. E. Smouse and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–491.

Ford-Lloyd, B. and M. Jackson. 1986. Plant Genetic Resources: an Introduction to their Conservation and Use. Edward Arnold, London.

Hadrys, H., M. Balick and B. Schierwater. 1992. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. Molecular Ecology 1: 55–63.

Handel, S. N. 1983. Pollination ecology, plant population structure, and gene flow. pp. 163–211. *In* L. Real (ed.), Pollination Biology. Academic Press, New York.

Huff, D. R., R. Peakall and P. E. Smouse. 1993. RAPD variation within and among natural populations of outcrossing buffalograss (Buchloe dactyloides (Nutt.) Engelm.). Theoretical and Applied Genetics 86: 927–934.

Hugan, M. S., L. P. Breuderle, N. Vorsa, 1993. Genetic variation in diploid and polyploid cranberry populations (*Vaccinium* section *Oxycoccus*). American Journal of Botany (abstract) 80: 153.

Long, J. C. 1984. The allelic correlation structure of Gainj and Kalam speaking people I. The estimation and interpretation of Wright's F-statistics. Genetics 112: 629–647.

Lynch, M. and B. G. Milligan. 1994. Analysis of population genetic structure with RAPD markers. Molecular Ecology 3: 91–99.

McClelland, M., J. Welsh. 1994. DNA fingerprinting by arbitarily primed PCR. PCR Methods and Applications 4: S59–S65.

Marucci, P. E. and R. S. Filmer. 1964. Preliminary cross pollination tests on cranberries. Proceedings of the 94th Annual Meeting of the American Cranberry Grower's Association 48–51.

Martin, C. G. and P. A. Games. 1977. Anova tests for homogeneity of variance: nonnormality and unequal samples. Journal of Educational Statistics 2: 187–206.

Michalakis, Y. and L. Excoffier. A generic estimation of population subdivision using distances between alleles with special reference to microsatellite data. Genetics (to appear).

Millar, C. I. and W. J. Libby. 1991. Strategies for conserving clinal, ecotypic, and disjunct population diversity in widespread species. pp. 149–170. *In* D. A. Falk and K. E. Holsinger (eds.). Genetics and Conservation of Rare Plants. Oxford University Press, New York.

Penner, G. A., A. Bush, R. Wise, W. Kim, L. Domier, K. Kasha, A. Laroche, G. Scoles, S. J. Molnar and G. Fedak. 1993. Reproducibility of random amplified polymorphic DNA (RAPD) analysis among laboratories. PCR Methods and Applications 2: 341–345.

Reader, R. J. 1977. Bog ericad flowers: self-compatibility and the relative attractiveness to bees. Canadian Journal of Botany 55: 2279–2287.

Russell, J. R., F. Hosein, E. Johnson, R. Waugh and W. Powell. 1993. Genetic differentiation of cocoa (*Theobroma cacao* L.) populations revealed by RAPD analysis. Molecular Ecology 2: 89–97.

Stewart Jr., C. N. and E. T. Nilsen. 1995a. Phenotypic plasticity and genetic variation of *Vaccinium macrocarpon*, the American cranberry. I. Reaction norms of clones from central and marginal populations in a common garden. International Journal of Plant Sciences 156: 687–697.

Stewart Jr., C. N. and E. T. Nilsen. 1995b. Phenotypic plasticity and genetic variation of *Vaccinium macrocarpon*, the American Cranberry. II Reaction norms and spatial clonal patterns in two marginal populations. International Journal of Plant Sciences 156: 698–708.

Stewart Jr., C. N. and D. M. Porter. 1995. RAPD profiling in biological conservation: an application to estimating clonal variation in rare and endangered *Iliamna* in Virginia. Biological Conservation 74: 135–142.

Stewart Jr., C. N. and L. E. Via. 1993. A rapid CTAB DNA isolation technique was for RAPD fingerprinting and other PCR applications. BioTechniques 14: 748–751.

Tinker, N. A., M. G. Fortin and M. E. Mather. 1993. Random amplified polymorphic DNA and pedigree relationships in spring barley. Theoretical and Applied Genetics 85: 976–984.

Weir, B. S., and C. C. Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38: 1358 1370.

Williams, J. G. K., A. R. Kubelik, K. J. Livak, J. A. Rafalski and S. V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research 18: 6531–6535.

Williams, J. G. K., M. K. Hanafey, J. A. Rafalski and S. V. Tingey. 1993. Genetic analysis using random amplified polymorphic DNA markers. Methods in Enzymology 218: 704–740.

Zar, Z. H. 1984. Biostatistical Analysis. Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Received 6 February 1995;
accepted 9 August 1995.
Corresponding Editor: T. R. Meagher