Theor Appl Genet (1994) 89:259–264

J. G. Tivang · J. Nienhuis · O. S. Smith

Estimation of sampling variance of molecular marker data using the bootstrap procedure

Received: 29 March 1993 / Accepted: 25 November 1993

Abstract Knowledge of genetic relationships among genotypes is useful in a plant breeding program because it permits the organization of germplasm and provides for more efficient sampling. The genetic distance (*GD*) among genotypes can be estimated using random restriction fragment length polymorphisms (RFLPs) as molecular markers. Knowledge of the sampling variance associated with RFLP markers is needed to determine how many markers are required for a given level of precision in the estimate of *GD*. The sampling variance for *GD* among all pairs of 37 maize (*Z. mays* L.) inbred lines was estimated from 1 202 RFLPs. The 1 202 polymorphisms were generated from 251 enzyme-probe combinations (EPC). The sampling variance was used to determine how large a sample of RFLPs was required to provide a given level of precision. The coefficient of variation (*CV*) associated with *GD* has a nearly linear relationship between its expected standard deviation and mean. The magnitude of the decrease in the mean *CV* for *GD* with increasing numbers of bands was dependent upon the sampling unit; e.g., individual polymorphic bands vs EPC, and the degree of relatedness among the inbreds compared. The rate of reduction in mean *CV* with increasing sample size was the same regardless of the restriction enzyme used, *Bam*HI, *Eco*RI or *Hin*dIII, when the bootstrap sampling units were individual polymorphic bands. In constrast, although the rate of reduction (slopes) was the same, the intercepts of the mean *CV*s were different when EPCs were used as the bootstrap sampling unit. This difference was due to the higher number of bands per EPC in *Bam*HI (4.94) compared with *Eco*RI (4.83) and *Hin*dIII (4.63).

Communicated by A. R. Hallauer

J. G. Tivang (✉) · J. Nienhuis
Deparment of Horticulture, University of Wisconsin, Madison, WI 53706, USA

O. S. Smith
Department of Data Management, Pioneer Hi-bred International, Johnston, IA 50131, USA

Key words RFLP · Bootstrap · Sampling variance · Zea mays

Introduction

Knowledge of genetic relationships among genotypes is useful in a plant breeding program because it permits the organization of germplasm and provides for more efficient sampling. Molecular markers, including restriction fragment length polymorphisms (RFLPs) and random amplified polymorphic DNAs (RAPDs), have been used to estimate the genetic relationships among genotypes in numerous crop species, including maize (Lee et al. 1989; Godshalk et al. 1990; Smith et al. 1990; Messmer et al. 1992), *Brassica* species (Song et al. 1988; Slocum et al. 1990; Hu and Quiros 1991; Nienhuis et al. 1993), *Cucumis melo* (Neuhausen 1992), *Solanum* species (Debener et al. 1990), and *Cucurbita* (Wilson et al. 1992). The number of polymorphic molecular marker bands used to estimate genetic relationships varies widely from 61 (Nienhuis et al. 1993) to 1 205 (Smith et al. 1990).

Sampling variance in the estimation of genetic relationships occurs when discrepancies are detected between a random subset of molecular marker bands and all possible bands. Larger numbers of random polymorphic bands will provide an increasingly more precise estimate of genetic relationships and will reduce the variance caused by over- or under-sampling certain regions of the genome. Moreover, if marker bands could be chosen to be uniformly distributed over linkage groups, the sampling variance due to over- or under-sampling could be reduced. Nevertheless, to obtain larger numbers of random polymorphic bands is now expensive in terms of time and resources, and linkage information among polymorphic bands is not always available. Thus, it would be desirable to estimate genetic relationships using the smallest set of polymorphic bands which minimize sampling variance (Smith et al. 1990).

The bootstrap is a computer-intensive sampling method designed to empirically estimate variance where theoretical computations may be problematic (Felsenstein 1985; Efron and Tibshirani 1986, 1991; Krajewski and Dickerman 1990). The bootstrap procedure has been used by Felsenstein (1985) to estimate the sampling variance associated with phylogenetic analysis. Molecular markers usually generate large amounts of data, providing an excellent opportunity for bootstrap sampling both of whole data sets and within smaller partitions of the data set. Partitions could be based on the degree of relatedness among the inbreds or on the restriction enzymes used in RFLP analysis.

The objectives of the present study were: (1) to determine the sampling variance using different size bootstrap samples from a large (1 202 bands) RFLP data set; (2) to estimate the sampling variance for different sampling units (e.g., individual polymorphic bands or enzyme probe combinations); and (3) to compare the bootstrap sampling variance associated with the use of different restriction enzymes.

Materials and methods

Germplasm and RFLP data

Thirty-seven elite *Zea mays* inbred lines (Pionere Hi-bred International, Johnston, Iowa) were used in this study. A detailed description of the inbreds is provided in previous publications (Smith and Smith 1989; Smith et al. 1990). The methods for DNA isolation, restriction endonuclease digestion, electrophoresis, Southern blotting, hybridization and auto-radiography have been previously described (Smith et al. 1990). The restriction fragments were scored from the autoradiograms resulting in a matrix of 37 inbred lines and 1 202 bands. The 1 202 bands were generated from 251 mapped enzyme-probe combinations which were evenly distributed across the genome (Smith et al. 1990).

Terminology and definitions

The term enzyme-probe combination (EPC) is equivalent to "restriction enzyme probe combinations" (Smith et al. 1990), "clone-enzyme combinations" (Lee et al. 1989; Messmer et al. 1992), and "probe/enzyme combinations" (Debener et al. 1990), while band is equivalent to polymorphic/monomorphic band. The presence of a restriction fragment was scored as "1", while absence of a fragment was scored as "0" (Lansman et al. 1981). The 1s and 0s are referred to as scores. The term "fragment frequency p(1)" refers to frequency of a type "1" score.

Statistical analyses of data

A computer program, written in Think C (Symantic, Cupertino Calif.), was developed to read in the data file, and to calculate the genetic distance (*GD*). Estimates of *GD* were calculated for all 666 pairs of inbreds according to the following equation which is the complement of the simple matching coefficient (Gower 1985);

$$GD(i, j) = \sum^{N}(i \neq j)/[\sum^{N}(i \neq j) + \sum^{N}(i = j)],$$

where *GD* is the measure of genetic distance between inbreds *i* and *j*, while $\sum^{N}(i \neq j)$ and $\sum^{N}(i = j)$ are the total number of scores discordant and concordant between inbreds *i* and *j*, respectively, over all *N* bands considered. A *GD* value of 0.0 and 1.0 indicates, respectively, no and maximal RFLP difference between two inbreds.

Bootstrap

To empirically estimate the sampling variance, a computer program was written to execute a bootstrap sampling procedure which is an adaptation of the procedures of Efron and Tibshirani (1986) and Felsenstein (1985). The program was designed to execute bootstrap sampling, using either individual bands or EPCs. Two hundred subsets, for a given number of *N* polymorphic bands, were sampled. The *N* bands were selected at random from the whole data set of 1 202 polymorphic bands. Sampling with replacement generated a probability of 1/1202 for any one band to be selected. A second bootstrap sampling method was also used, where the sampling unit was the the EPC. In this case 200 subsets of *M* arbitrary EPCs were sampled. Random selection from the whole data set with replacement generated a probability of 1/251 for any one EPC to be selected at any one time. The *GD* was calculated between all (666) inbred pairs. The variability among 200 bootstrap samples for each pair of inbreds was standardized using the coefficient of variation (*CV*). Normalization of the variance to the *CV* was possible since the relationship between the variance and the mean for the *GD* estimator in the targeted range is nearly linear (Tivang 1992).

The mean *CV* of the 666 inbred pairs for a given sample size was plotted against the sample size. When the EPC was used as the sampling unit, the mean and standard deviation for the number of bands in each sample was calculated. The mean *CV* was plotted against the mean number of bands included in each sample of *M* EPCs. Linear regression calculations were obtained by natural logarithmic transformation of mean *CV* and bootstrap sample size.

Bernoulli random variable

The *GD* statistic is based on two outcomes for each comparison; either the scores agree (00, 11) or disagree (01, 10) for any two inbreds. This dual outcome can be modelled as a Bernoulli random variable. The probability of observing an agreement or disagreement is then a function of the fragment frequency.

Results

A total of 251 enzyme probe combinations (EPC) were scored for each of the three restriction enzymes *Bam*HI, *Eco*RI, and *Hin*dIII, with a mean of 4.79 ± 2.06 (polymorphic) bands per EPC (Table 1). Among the 37 inbreds and the 1 202 polymorphic bands the probability of finding a fragment present (1) and absent (0) was 0.249 and 0.751, respectively.

Comparison of sampling units

Under the assumption that RFLP data can be modelled as a Bernoulli random variable, the following statistics

Table 1 Number of enzyme probe combinations (EPCs) and the number of polymorphic bands among the three restriction enzymes

Restriction enzyme	Number of EPC (probes)	Number of bands	Mean number of bands per EPC
*Bam*HI	78	385	4.94 ± 2.21
*Eco*RI	82	396	4.83 ± 2.21
*Hin*dIII	91	421	4.63 ± 1.81
Total	251	1202	4.79 ± 2.06

were computed:

$p(\text{disagree}) = (0.249)(0.751)(2) = 0.3738$

$q(\text{agree}) = (0.249)^2 + (0.751)^2 = 0.6262,$

where p and q indicate the expected fractions of concordant and discordant observations in the data set, respectively. The expected mean discordant observation in a sample of n polymorphic bands thus becomes

$\mu = (n)p$

with the associated expected variance of

$s^2 = (n)pq.$

It can be shown that p is equivalent to the mean GD, using the definition of GD described in Materials and methods, since $p + q = 1$. The coefficient of variation (CV) can be estimated using the following formula:

$$CV = (100)(\sqrt{npq})/(np) = (100/\sqrt{n})^*\sqrt{(q/p)}.$$

Treating each band as an independent sampling unit, the mean CVs were calculated and plotted against the bootstrap sample size (Fig. 1). The CVs for the bootstrap samples were larger than those obtained by the Bernoulli random variable.

Under laboratory conditions bands are not obtained as independent units, but rather several bands usually occur as a block associated with a specific enzyme probe complex (EPC). The bootstrap estimates of CV, sampling among the EPCs, was plotted against bootstrap sample size (Fig. 1). The grouping of bands by EPCs resulted in increased CV values compared to the Bernoulli random variable and bootstrap sampling of bands as independent units. The standard deviation (indicated by the horizontal bar through the diamond) of the number of bands within each EPC increased as the number of EPCs in the bootstrap sample increased. The upper and lower quartile (indicated by the vertical bar through the diamond) of the distribution of all 666 CVs was reduced as bootstrap sample size increased.

Natural log transformations of the scale used in Fig. 1 resulted in a linear relationship between CVs and bootstrap sample sizes (data not shown). Significant differences in the y-intercepts were detected, while the rate of reduction in mean CV (slope) was the same for all three sampling methods. The slope, representing precision, is a function of the sample size, while the y-intercept determines the level of precision. The number of bands for a 10% CV for each method was 167, 265, and 388, for Bernoulli random variable, random band, and EPC, as independent sampling units, respectively.

Partitioning by degree of relatedness

The skewed distribution of CV values at each bootstrap sampling level (illustrated by the unequal quartile range in Fig. 1) is reflected in the skewed frequency distribution of GD for all 666 comparisons among the 37 inbred lines (Fig. 2). From this distribution three groups of five inbred pairs representing closely-, intermediately- and distantly-related inbred pairs were sampled, with mean GDs of 0.09, 0.19, and 0.37, respectively. Using EPCs as the bootstrap sampling unit, the mean CVs for each group were plotted against bootstrap sample size

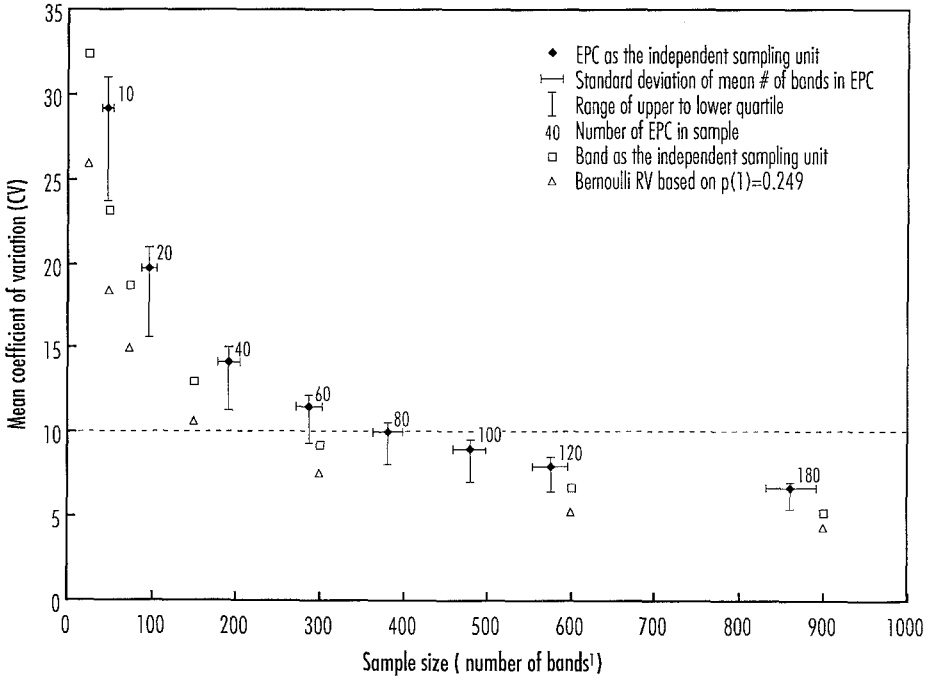

Fig. 1 Plot of the relationships between the mean CV and the sample size for the different sample units. The *dotted line* indicates a 10% CV'. The mean number of bands were plotted for EPC

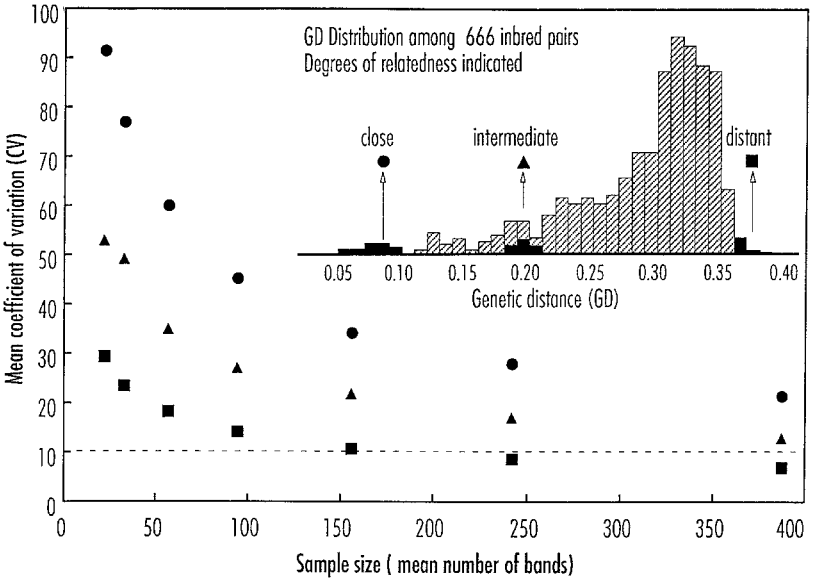

Fig. 2 Plot of the relationships between the mean *CV* and the sample size for closely-, intermediately- and distantly-related lines. The *dotted line* indicates a 10% *CV*

(Fig. 2). The scatter plot shows the most distantly-related group with the lowest *CV* values compared to closely- and intermediately-related inbreds. Natural log transformation of the scales used in Fig. 2 resulted in a linear relationship between the mean *CV*s and bootstrap sample sizes (data not shown). The rate of reduction was constant among the three groups whereas the intercepts were significantly different. The number of bands required for a *CV* of 10% was 388, 150, and 38, for closely-, intermediately- and distantly-related inbreds, respectively.

Restriction enzyme partitions

Three different restriction enzymes, *Bam*HI, *Eco*RI, and *Hind*III, were used in different EPCs. Each restriction enzyme subset was evaluated independently using random polymorphic bands and EPCs as bootstrap sampling units. Natural log transformations revealed a linear relationship between the mean *CV*s for each restriction enzyme partition sample size when the bootstrap sampling unit was bands (Fig. 3 upper panel). In this case no differences were observed among the three restriction enzymes in either the y-intercept or the slope. In contrast, when EPCs were used as the bootstrap sampling unit, the slopes were the same but the y-intercepts were different among the three restriction enzyme partitions (Fig. 3 lower panel). This lack of intercept homogeneity can be attributed to the mean number of bands per probe generated by each restriction enzyme (Table 1). The higher mean *CV*s appear to be a function of the number of bands per EPC. The number of bands required (271) for a 10% *CV* was determined by applying the common regression equation (Fig. 3 upper panel). The average polymorphic band provides an equivalent amount of information regardless of the restriction enzyme used. In contrast, the EPCs regression equations

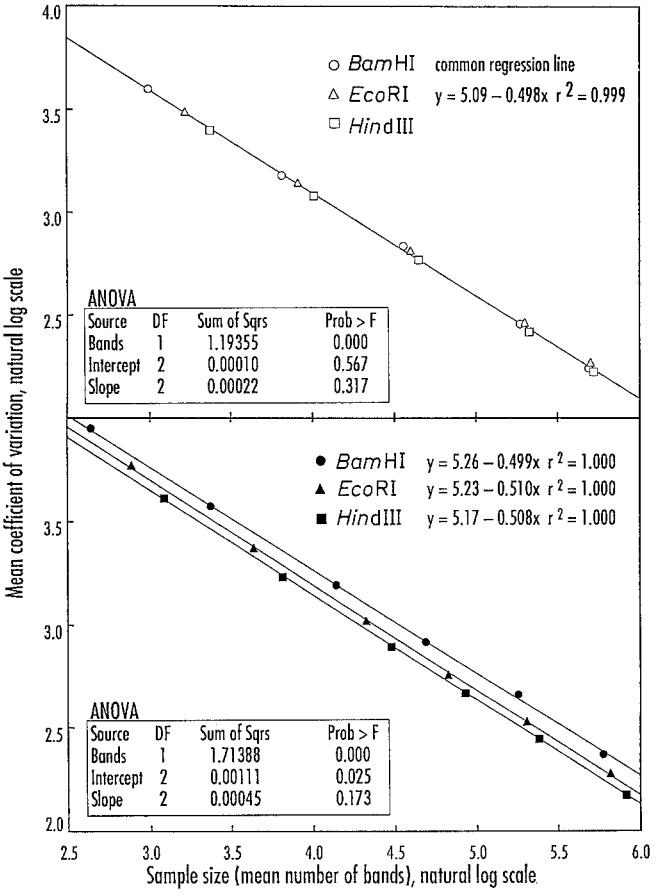

Fig. 3 Plot of relationships between the mean *CV* and the sample size for RFLP bands produced using three different restriction enzymes, *Bam*HI, *Eco*RI, and *Hind*III

(Fig. 3 lower panel) solving for a 10% *CV* require 377, 313, and 284 bands, respectively, for *Bam*HI, *Eco*RI, and *Hind*III. Converting these bands into EPCs, for each respective restriction enzyme, results in a mean number

of 77, 65, and 62 EPCs for *Bam*HI, *Eco*RI, and *Hin*dIII, respectively, when used on the same 37 inbreds.

Discussion

The bootstrap procedure was effective in determining the sampling variance of a molecular marker data set. The results indicate that 80 EPCs (about 375 bands) were required to achieve a *CV* of 10% in the estimation of genetic distance among the 37 inbreds used in this study. The Bernoulli random variable consistently underestimated the *CV*, because linkage among fragments resulted in redundancy which was not consistent with the assumption of independence of the data points.

Sampling of molecular data was affected by the sampling unit, whether bands are sampled at random or in (EPC) blocks. This discrepancy can be attributed to redundancy of information in the bands belonging to an EPC. The ideal EPC consists of independent polymorphic bands that discriminate among all inbreds, where redundancy is absent. Deviation from this ideal results in redundancy that reduces the efficiency of the EPC. The optimal performance of an EPC is equivalent to the discrimination ablity of an equal number of bands obtained at random, disregarding the EPC structure. In such a case, the greater the number of bands per EPC, the better the performance. In this study, random independent bands consistently had lower *CV*s than when EPCs were used as the bootstrap sampling unit. This difference suggests that, within an EPC, polymorphic bands carry some redundant information. Thus the greater the number of bands present in an EPC, the less discriminatory power will be conveyed by the average band within that block. In this case the EPC with the lower number of bands performed better, since the redundant information was minimized, thus improving sampling efficiency.

Closely-related inbreds require more data for discrimination than do more-distantly-related lines. This observation suggests that a modified approach in the data collection could provide greater precision levels with a reduced effort by reassigning resources. To achieve this objective the data collection process should be terminated the moment a genotype can be classified as distant in relation to all others. This classification will naturally be dependent on an arbitrary precision level. The resources originally assigned to the terminated genotype should then be reassigned to the remaining more-closely-related individuals.

That the restriction enzyme partitions performed equivalently on a band per band basis was not unexpected, since all restriction enzymes used required six-base sequences. This fact suggests that each enzyme has the same restriction frequency. A surprising aspect was the discrepancy among the y-intercepts observed when sampling the restriction enzymes in terms of blocks determined by the probes. The polymorphic band distributions among the three restriction enzyme were not drastically different, and they all contain roughly the same number of total polymorphic bands (about 400). The number of bands per probe were similar: 4.94, 4.83, and 4.63 for *Bam*HI, *Eco*RI, and *Hin*dIII, respectively. When the bootstrap sampling unit was the EPC, minor differences in the mean number of bands per EPC reflected significant differences in *CV*. Comparing the random band sample, regardless of enzyme source, with the best of the restriction enzyme partitions (*Hin*dIII) suggests that the bands obtained at random provided more information, i.e., had lower *CV*s, than the *Hin*dIII partition. However, in terms of the number bands required, the differences diminished (271, 284 for random bands and *Hin*dIII, respectively).

The selection of *GD* as an estimator of genetic relationship is related to the use of *CV*s as a measure of precision. In this study the complement to the simple matching coefficient (Gower 1985) was used and termed genetic distance (*GD*). We realize that this is not the most common estimator of genetic relationships. Other estimators were evaluated such as Nei's I (Nei 1987) which appeared similar to Jaccards coefficient (Gower 1985; Debener et al. 1990), and MRD (Rogers 1972; Lee et al. 1989). Although empirical estimates of variance could be obtained by the bootstrap method, the display of the variance proved problematic since the relationship between the variance and the mean was not linear (Tivang 1992).

References

Debener T, Salamini F, Gebhardt C (1990) Phylogeny of wild and cultivated *Solanum* species based on nuclear restriction fragment length polymorphisms (RFLPs). Theor Appl Genet 79:360–368

Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1:54–77

Efron B, Tibshirani R (1991) Statistical analysis in the computer age. Science 253:390–395

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Gower JC (1985) Measures of similarity, dissimilarity, and distance. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 5. Wiley, New York, pp 397–405

Godshalk EB, Lee M, Lampkey KR (1990) Relationship of restriction fragment length polymorphisms to single-cross hybrid performance in maize. Theor Appl Genet 80:273–280

Hu J, Quiros CF (1991) Identification of broccoli and cauliflower cultivars with RAPD markers. Plant Cell Rep 10:505–511

Krajewski C, Dickerman AW (1990) Bootstrap analysis of phylogenetic trees derived from DNA hybridization distances. Systematic Zool 39:383–390

Lansman RA, Shade RO, Shapira JF, Avise JC (1981) The use of restriction endonucleases to measure mitochondrial DNA sequence relatedness in natural populations. J Mol Evol 17:214–226

Lee M, Godshalk EB, Lamkey KR, Woodman WW (1989) Association of restriction fragment length polymorphisms among maize inbreds with agronomic performance of their crosses. Crop Sci 29:1067–1071

Messmer MM, Melchinger AE, Boppenmaier J, Hermann RG, Brunklaus-Jung E (1992) RFLP analysis of early-maturing European maize germ plasm. Theor Appl Genet 83:1003–1012

Nei M (1987) Molecular evolutionary genetics. Colombia University Press, New York

Nienhuis J, Slocum MK, DeVos DA, Muren R (1993) Genetic similarity among *Brassica oleracea* L. genotypes as measured by restriction fragment length polymorphisms. J Am Soc Hort Sci 118:298–303

Neuhausen S (1992) Evaluation of restriction fragment length polymorphism in *Cucumis melo*. Theor Appl Genet 83:379–384

Rogers JS (1972) Measures of genetic similarity and genetic distance. Studies in genetics VII. Univ Tex Publ 7213:145–153

Slocum MK, Figdore SS, Kennard WE, Suzuki JY, Osborn TC (1990) Linkage arrangement of restriction fragment length polymorphism loci in *B. oleracea*. Theor Appl Genet 80:57–64

Smith JSC, Smith OS (1989) The description and assessment of distances between inbred lines of maize. II. The utility of morphological, biochemical and genetic descriptors and a scheme for the testing of distinctiveness between inbred lines. Maydica 34:141–50

Smith OS, Smith JSC, Bowen SL, Tenborg RA, Wall SJ (1990) Similarities among a group of elite maize inbreds a measured by predigree, F_1 grain yield, grain yield heterosis and RFLPs. Theor Appl Genet 80:833–840

Song KM, Osborne TC, Williams PH (1988) *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). II. Preliminary analysis of subspecies within *R. rapa* (syn. *campestris*) and *B. oleracea*. Theor Appl Genet 76:593–600

Tivang JG (1992) Sampling variance of molecular marker data using the bootstrap procedure. Masters thesis, University of Wisconsin, Madison

Wilson HD, Doebley J, Duvall M (1992) Chloroplast diversity among wild and cultivated members of *Cucubita*. Theor Apppl Genet 84:859–865