Y. B. Fu

Effectiveness of bulking procedures in measuring population-pairwise similarity with dominant and co-dominant genetic markers

Received: 20 October 1999 / Accepted: 11 November 1999

Abstract Two bulking procedures (bulking individuals before and after genotyping) are commonly applied in similarity based studies of genetic distance at the population or higher level, but their effectiveness is largely unknown. In this study, expected population-pairwise similarity for both bulking procedures is derived with dominant and co-dominant diallelic markers. Numerical examples for the derived formulae are given with up to ten individuals randomly selected from each population. The procedure of bulking individuals after genotyping with either marker system is generally more informative than the procedure of bulking individuals before genotyping, because the former incorporates the information from marker alleles of intermediate frequency. Both procedures are effective with 5–10 individuals selected randomly from either population, but the procedure of bulking before genotyping requires a genotyping effort several-fold less than the procedure of bulking after genotyping. For either bulking procedure, a co-dominant marker system is generally more informative than a dominant marker system.

Key words Population similarity · Genetic relationship · Co-dominant marker · Dominant marker · Bulk analysis

Introduction

In recent years, the similarity based (or DNA band-sharing) approach has been widely applied to describe genetic relationships among individuals, populations and species, particularly in evolutionary studies of natural populations (Avise 1994), in the development of breeding programs (Bohn et al. 1999), and in the management of genetic resources (Ayad et al 1997). This broad application largely contributes to the availability of numerous molecular genetic markers, such as RFLPs, RAPDs, AFLPs and SSRs, and to the simplicity of the similarity based approach. This approach requires no genetic models to understand the causes and mechanisms of the genetic variations observed, but rather focuses on the variation patterns measured by the proportion of shared DNA fragments (or marker genotypes). Such model-free measurements can reveal various patterns of genetic relationship over several levels through the use of many advanced multivariate analyses such as clustering and multidimensional scaling. With more markers identified to cover the whole genome, such an approach can be more informative in describing various patterns of genetic relationship that are environmentally independent and essentially reflect the true evolutionary history. It is expected that this approach will continue to play an important role in studies of genetic relationships over various levels.

While many similarity based studies are intended to describe genetic relationships below the species level, studies at the level of species, or higher, are also common. Molecular characterization of hundreds (or thousands) of germplasm accessions collected from various regions, and representing different related species or taxa, can be an example for measuring genetic distance at the population or higher level (Ayad et al. 1997). This characterization usually requires genotyping of a large number of individuals and consequently a large experimental effort. A commonly used procedure to reduce the genotyping effort is to bulk individuals from each population (or accession in gene bank systems) before genotyping and measure the genotypic similarity between bulked samples directly, assuming genetic variations within a population or accession are low as in most selfing species (e.g., Margale et al. 1995; De Bustos et al. 1999; Divaret et al. 1999; Gilbert et al. 1999). Another procedure, particularly used for outcrossing species, is to first genotype a few individuals from each population

Communicated by H.F. Linskens

Y.B. Fu
Plant Gene Resources of Canada, Saskatoon Research Centre,
Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon,
Saskatchewan S7 N 0X2, Canada
e-mail: fuy@em.agr.ca
Tel.: +1-306–956–7642, Fax: +1-306–956–7246

and then calculate an average of the similarity measurements at the individual level for all the individuals combined from two populations (e.g., Miller and Tanksley 1990). This procedure can also be seen in those studies of genetic relationship below the species level in which the patterns of genetic relationship are often needed in order to summarize at a higher level (e.g., Wang et al. 1992). Thus, it is most desirable to know how effective the similarity measure with these bulking procedures is at the population or higher level and how many individuals should be selected to represent either population. To my knowledge, no theoretical treatments have been made so far on the population-pairwise similarity.

The objective of the present study was to examine the effectiveness of the two bulking procedures in measuring genetic distances among populations with co-dominant and dominant genetic markers. Specifically, this was done by a derivation of the expected population-pairwise similarities for both bulking procedures with both marker systems and with a numerical illustration.

Materials and methods

Expected population similarity

For a pair of individuals within a population, the similarity measure (*S*) is commonly calculated as:

$$S = \frac{1}{n}\sum_{i=1}^{n} S_i, \qquad (1)$$

where S_i is the number of matches in genotype per locus *i* for two individuals, and *n* is the total number of marker loci scored. This measure, unlike the other similarity formulae (e.g., Nei and Li 1979; Lynch 1990), has a fixed denominator and takes into account the genotypes inferred from recessive phenotypes such as those with absent bands in dominant markers. Thus it shows some advantage, particularly when different marker systems are considered.

To derive a similarity for a pair of populations (here we call it population similarity for short), we must consider how individuals of either population are genotyped and analyzed. Both dominant and co-dominant markers are commonly applied, but they can show a difference in their effectiveness for inferring genetic relationships even for the same populations used (Fu, unpublished). This difference largely reflects the informativeness of the two marker systems. For a marker locus with two alleles (*M* and *m*), three genotypes (*MM*, *Mm*, and *mm*) can be visualized for individuals on a gel for RFLP and SSR systems. For RAPD and AFLP systems, a DNA band (or fragment) is either present or absent on the gel; its presence can be interpreted as the expression of a genotype of either *MM* or *Mm* (normally encoded as *M*-) when the *M* allele is dominant and its absence for the genotype *mm*.

There are two commonly used procedures in genotyping individuals from either population, as mentioned above. The first procedure is to bulk individuals randomly selected from each population, genotype each bulked sample, and then calculate the population similarity between two bulked samples based on marker genotypes (here we call it bulking before genotyping, for short). The second procedure is to select a few individuals from each population, genotype them individually, and then calculate the population similarity based on individual-pairwise similarity measures (here we call it bulking after genotyping, for short). In what follows, we first derive the expected population similarity for bulking before genotyping for each of the two marker systems and then for bulking after genotyping.

Bulking before genotyping

Let us consider two random mating populations from which *N* individuals are randomly selected each to form two bulks. Bulked DNAs are analyzed with dominant marker systems such as RAPDs and AFLPs for *n* diallelic loci. Let us denote the frequencies of the two alleles m_{1i} and M_{1i} for the locus $i(i=1..n)$ in the first population as q_{1i} and $1–q_{1i}$, and the frequencies of the two alleles m_{2i} and M_{2i} in the second population as q_{2i} and $1–q_{2i}$, respectively. For the first bulk, the probability of having the genotype $m_{1i}m_{1i}$ for the locus $i(P_{1i}^{mm})$ is $[q_{1i}^2]^N$ and the probability of having the other genotypes (P_{1i}^{M-}) is $1–[q_{1i}^2]^N$. Similarly for the second bulk, the probability of having the genotype $m_{2i}m_{2i}$ for the locus $i(P_{2i}^{mm})$ is $[q_{2i}^2]^N$ and the probability of having the other genotypes (P_{2i}^{M-} is $1–[q_{2i}^2]^N$. For the locus *i*, the expected similarity between the two bulks is calculated as:

$$S_i = P_{1i}^{mm}P_{2i}^{mm}(1) + P_{1i}^{mm}P_{2i}^{M-}(0) + P_{1i}^{M-}P_{2i}^{mm}(0) + P_{1i}^{M-}P_{2i}^{M-}(1). \qquad (2)$$

For *n* loci, the expected population similarity can be derived as:

$$S_N = \frac{1}{n}\sum_{i=1}^{n} S_i = \frac{1}{n}\sum_{i=1}^{n}[q_{1i}^{2N}q_{2i}^{2N} + (1-q_{1i}^{2N})(1-q_{2i}^{2N})]. \qquad (3)$$

For co-dominant markers, there are three possible genotypes (*MM*, *Mm*, and *mm*) for each locus. For each bulk, the probability of having each of the three genotypes m_im_i M_iM_i, and M_im_i for the locus *i* is $[q_i^2]^N$, $[1–q_i^2]^N$, and $1–[q_i^2]^N–[1–q_i^2]^N$, respectively. For the locus *i*, there are nine possible pairs of genotypes between the two bulks with their probability and similarity values as below:

	$m_{1i}\,m_{1i}$	$M_{1i}\,M_{1i}$	$M_{1i}\,m_{1i}$
	$[q_{1i}^2]^N$	$[1–q_{1i}^2]^N$	$1–[q_{1i}^2]^N–[1–q_{1i}^2]^N$
$m_{2i}\,m_{2i}[q_{2i}^2]^N$	1	0	0
$M_{2i}M_{2i}[1–q_{2i}^2]^N$	0	1	0
$M_{2i}m_{2i}1–[q_{2i}^2]^N–[1–q_{2i}^2]^N$	0	0	1

Summing up the nine products (probability × similarity value) for each locus and all the loci, one can derive the expected population similarity between the two bulks as:

$$S_N = \frac{1}{n}\sum_{i=1}^{n}\{q_{1i}^{2N}q_{2i}^{2N} + (1-q_{1i})^{2N}(1-q_{2i})^{2N}$$
$$+[1-q_{1i}^{2N}-(1-q_{1i})^{2N}][1-q_{2i}^{2N}-(1-q_{2i})^{2N}]\}. \qquad (4)$$

Bulking after genotyping

This procedure generates marker genotype data available for each of *N* individuals randomly selected from each population. Because of this, a population similarity can be calculated in different ways. One of the commonly practiced methods is to calculate the average individual-pairwise similarity for each population and then to take a mean of the two averages as the population similarity measure (e.g., Wang et al. 1992). This can be expressed as:

$$S_N = \frac{1}{2n}\sum_{i=1}^{n}\left[\frac{1}{N_p}\sum_{j=1}^{N_p}\left(S_{1ij}+S_{2ij}\right)\right], \quad \text{where} \quad N_p = \frac{N(N-1)}{2}. \qquad (5)$$

Note that *S* with subscripts represents the individual-pairwise similarity for either population for each marker locus. Clearly, such a similarity measure does not take into account the similarity information from the pairings of individuals between populations and thus can be biased, particularly when the information from such pairings is significant. One way to correct such bias is to calculate an average individual-pairwise similarity on all the possible pairs of 2*N* individuals pooled from both populations. This can be formulated as:

$$S_N = \frac{1}{n}\sum_{i=1}^{n}(N_{p1}S_{1i} + N_{p2}S_{2i} + N_{p12}S_{12i})/(N_{p1}+N_{p2}+N_{p12}), \qquad (6)$$

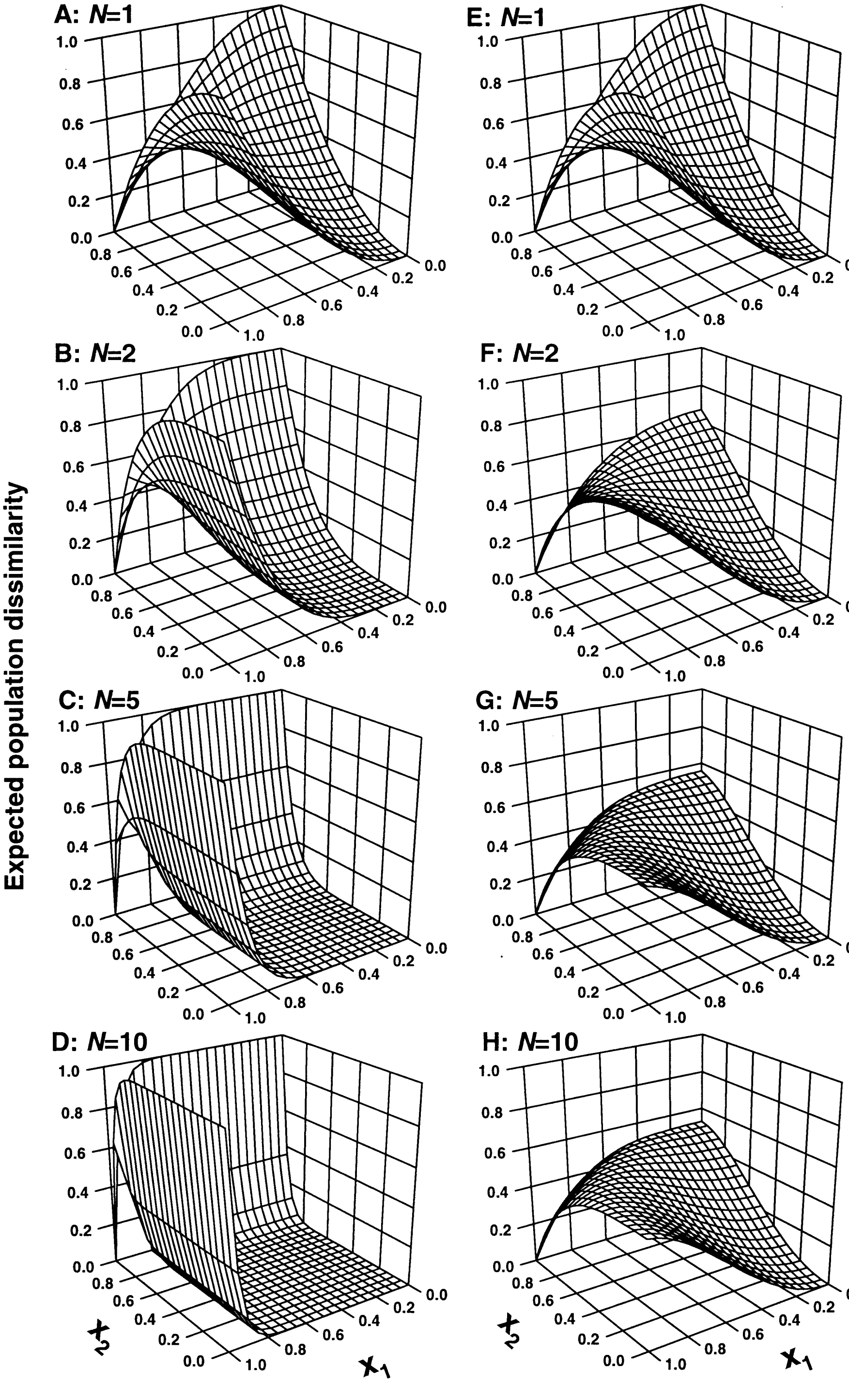


Fig. 1A–H Expected population dissimilarity for two bulking procedures with dominant diallelic markers, as a function of the expected marker allele frequencies in two populations (x_1 and x_2) and the number of the selected individuals (N). **A–D** represents the procedure of bulking before genotyping and **E–H** the procedure of bulking after genotyping

where $N_{p1} = N_{p2} = \frac{N(N-1)}{2}$ for all the possible pairs of N individuals from each population, N_{p12}=NN for all the possible individual pairs between the two populations, while S with subscripts represents the pairwise genotypic similarity for the three components of pairing (i.e., pairing individuals within each of the two populations and among the two populations). To derive S_N, we need to derive the similarity measures for the three components.

For a dominant diallelic marker locus, the similarity measures for the three components can be obtained following the same derivation procedure given above for S_N as:

$$S_{1i} = q_{1i}^4 + (1 - q_{1i}^2)^2,$$

$$S_{2i} = q_{2i}^4 + (1 - q_{2i}^2)^2, \quad \text{and}$$

$$S_{12i} = q_{1i}^2 q_{2i}^2 + (1 - q_{1i}^2)(1 - q_{2i}^2).$$

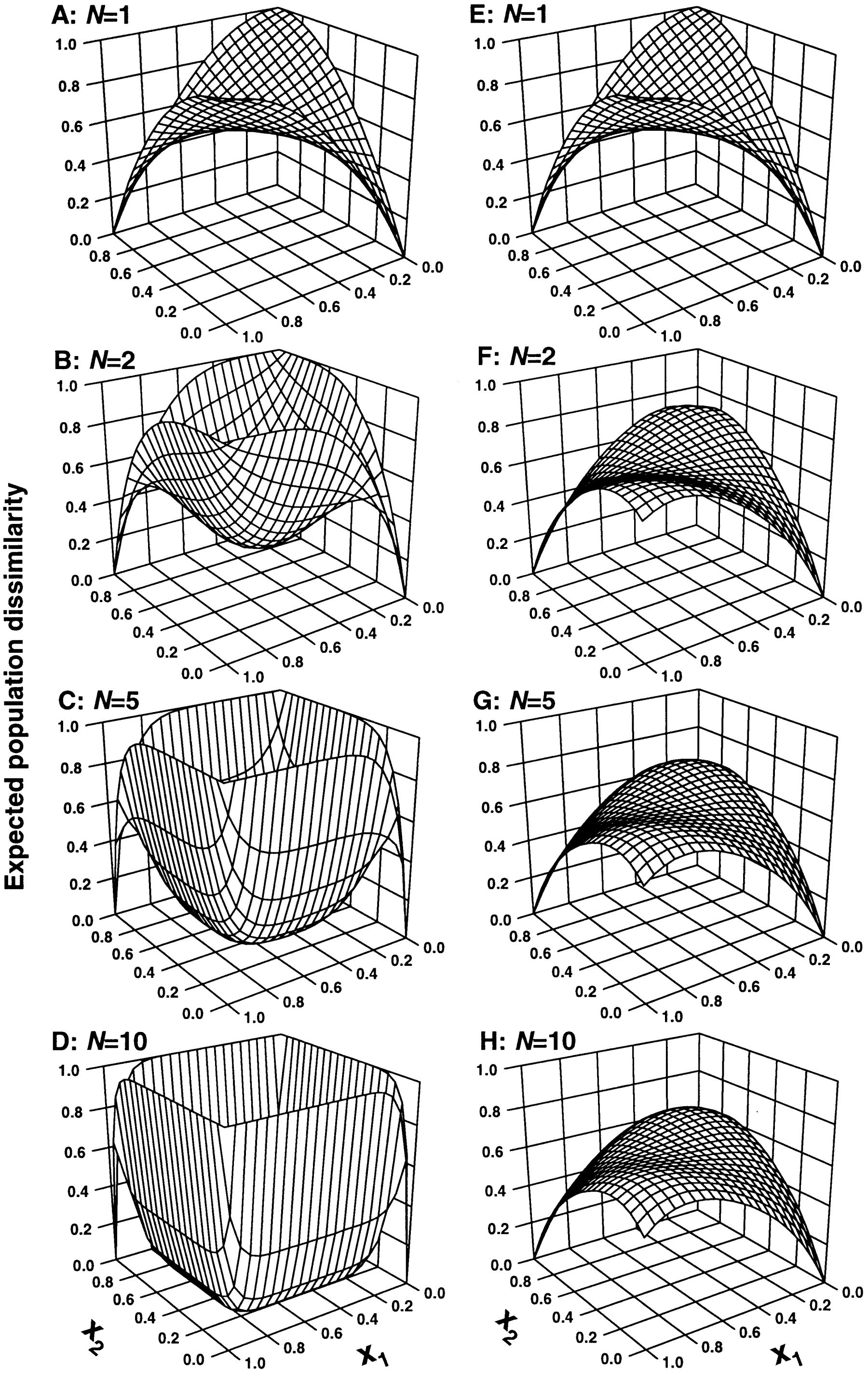
For a co-dominant diallelic marker locus, the similarity measures for the three components can be derived for S_N as:

$$S_{1i} = q_{1i}^4 + (1 - q_{1i})^4 + [1 - q_{1i}^2 - (1 - (q_{1i})^2]^2,$$

$$S_{2i} = q_{2i}^4 + (1 - q_{2i})^4 + [1 - q_{2i}^2 - (1 - q_{2i})^2]^2, \quad \text{and}$$

$$S_{12i} = q_{1i}^2 q_{2i}^2 + (1 - q_{1i})^2 (1 - q_{2i})^2 + [1 - q_{1i}^2 - (1 - q_{1i})^2][1 - q_{2i}^2 - (1 - q_{2i})^2].$$

Fig. 2A–H Expected population dissimilarity for two bulking procedures with co-dominant diallelic markers, as a function of the expected marker allele frequencies in two populations (x_1 and x_2) and the number of the selected individuals (*N*). **A–D** represents the procedure of bulking before genotyping and **E–H** the procedure of bulking after genotyping



Numerical illustration

The formulae derived above for the expected population similarity are lengthy and cumbersome, and thus difficult to evaluate. Here we present some numerical examples to illustrate the effectiveness of the two bulking procedures under the two marker systems to measure population similarity. This was done with a computer program written in SAS IML (1995) by specifying the expected marker allele frequency from 0 to 1 and with the number of individuals randomly selected up to 30 from either population. The expected marker allele frequency was used mainly because the true distributions of marker allele frequencies are usually unknown and may differ for different populations as well as for different marker systems. Since population dissimilarity (1-S_N) is commonly used for the measurement of genetic distances among populations, the expected population dissimilarities for the two populations under the four scenarios of bulking and marker systems were generated. The results for 1, 2, 5, and 10 individuals selected are plotted in Fig. 1 for dominant markers and in Fig. 2 for co-dominant markers.

Under either marker system, bulking after genotyping seems to be more informative in measuring population similarity than bulking before genotyping (Figs. 1 and 2). The latter is informative only for marker alleles of expected extreme frequencies (i.e., for fixed marker genotypes) in either population, while the former incorporates information from marker alleles of both expected extreme and intermediate frequencies. For both populations represented by five selected individuals, for example, the expected population dissimilarity is close to zero for bulking before genotyping (Fig. 1C) and around 0.4 for bulking after genotyping in the dominant marker system (Fig. 1G), when the expected marker allele frequencies are 0.4. When the expected marker allele frequencies are close to 1 in either population, the expected population dissimilarities are toward 1, larger than those for bulking after genotyping (Figs. 2C and 2G).

For either bulking procedure, the expected population dissimilarities do not seem to change much with 5–10 individuals selected, even for both marker systems (Figs. 1 and 2). This implies that applications of a few (5–10) individuals from each population are largely sufficient. The results are consistent with those reported from empirical studies on *Brassica* and Lupin collections (Divaret et al. 1999; Gilbert et al. 1999). In spite of this, however, bulking after genotyping can still be several-fold more than bulking before genotyping in terms of genotyping effort.

Comparing between the marker systems, one can find that bulking before genotyping with co-dominant markers is more informative than with dominant markers, as the procedure takes into account the fixation of either co-dominant marker allele (*m* or *M*) (Fig. 2A–D), but only the fixation of the recessive marker allele (Fig. 1A–D). Such patterns of difference also apply to bulking after genotyping, particularly for expected intermediate marker allele frequencies in which the expected population dissimilarities seem to be symmetric under co-dominant markers (Fig. 2E–H) and skewed toward recessive marker alleles of high frequencies (Fig. 1E–H) for dominant marker systems.

Discussion

In this study we derived the expected similarity for a pair of two random mating populations, considering the applications of the two commonly used bulking procedures (i.e., bulking individuals before and after genotyping) and both co-dominant and dominant diallelic markers. With numerical examples, we showed that the procedure of bulking individuals after genotyping with either dominant or co-dominant marker systems is generally more informative than the procedure of bulking individuals before genotyping in terms of measuring population similarity. Both procedures are effective with a few (5–10) individuals selected randomly from each population, whereas bulking before genotyping requires a genotyping effort several-fold less than bulking after genotyping. This presents a trade-off between the two bulking procedures, i.e., bulking after genotyping is more informative, but requires more effort in genotyping than bulking before genotyping. When the experimental effort in genotyping is relatively small, the procedure of bulking after genotyping is recommended. For the procedure of bulking before genotyping, applications of a co-dominant marker system are generally more informative than that of a dominant marker system. Such a difference in informativeness also applies to the procedure of bulking after genotyping for which the expected population dissimilarities are symmetric and skewed with co-dominant and dominant markers, respectively.

Our derivation considered only the diallelic marker loci and was based on only one commonly used formula for measuring genotypic similarity (i.e., equation 1), but the derivation procedures can be extended to multiallelic loci and other similarity formulae. This extension can be a subject of further study to assess the difference in patterns of expected population similarity among populations and species. Also, our derivation assumed for simplicity a random mating population. Violation of such an assumption may affect the sample distribution of marker allele frequencies (i.e., the variance) but not necessarily the expected marker allele frequency (i.e., the mean), and thus the observed patterns of expected population dissimilarity should largely hold.

Acknowledgments The author thanks Drs. K.W. Richards and A. Diederichsen for their encouragement during the study and Dr. Rob Baker for his helpful comments on an earlier version of the manuscript.

References

Avise JC (1994) Molecular markers, natural history and evolution. Chapman and Hall, New York

Ayad WG, Hodgkin T, Jaradat A, Rao VR (1997) Molecular genetic techniques for plant genetic resources. Report of an IPGRI workshop, 9–11 October 1995, Rome, Italy. International Plant Genetic Resources Institute, Rome, Italy pp 137

Bohn M, Utz HF, Melchinger AE (1999) Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. Crop Sci 39: 228–237

De Bustos A, Soler C, Jouve N (1999) Analysis by PCR-based markers using designed primers to study relationships between species of *Hordeum* (Poaceae). Genome 42: 129–138

Divaret I, Margale E, Thomas G (1999) RAPD markers on seed bulks efficiently assess the genetic diversity of a *Brassica oleracea* L. collection. Theor Appl Genet 98: 1029–1035

Gilbert JE, Lewis RV, Wilkinson MJ, Caligari PDS (1999) Developing an appropriate strategy to assess genetic variability in plant germplasm collections. Theor Appl Genet 98: 1125–1131

Lynch M (1990) The similarity index and DNA fingerprinting. Mol Biol Evol 7: 478–484

Margale E, Herve Y, Hu J, Quiros CF (1995) Determination of genetic variability by RAPD markers in cauliflower, cabbage and kale local cultivars from France. Genetic Resour Crop Evol 42: 281–289

Miller JC, Tanksley SD (1990) RFLP analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theor Appl Genet 80: 437–448

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76: 5269–5273

SAS Institute Inc. (1995) SAS user guide, version 6.12 edition. SAS Institute Incorporated, Cary, North Carolilna

Wang ZY, Second G, Tanksley SD (1992) Polymorphism and phylogenetic relationships among species in the genus *Oryza* as determined by analysis of nuclear RFLPs. Theor Appl Genet 83: 565–581