Theor Appl Genet (2007) 115:343–349
DOI 10.1007/s00122-007-0566-5

ORIGINAL PAPER

Data resolution: a jackknife procedure for determining the consistency of molecular marker datasets

Th. J. L. van Hintum

Received: 4 July 2006 / Accepted: 23 April 2007 / Published online: 15 May 2007

Abstract The results of genetic diversity studies using molecular markers not only depend on the biology of the studied objects but also on the quality of the marker data. Poor data quality may hamper the correct answering of biological questions. A new statistic is proposed to estimate the quality of a marker data set with regard to its ability to describe the structure of the biological material under study. This statistic is called data resolution (DR). It is calculated by splitting a marker data set at random into two sets each with half the number of markers. In each set, similarities between all pairs of objects are calculated. Subsequently, the similarities obtained for the two sets are correlated. This process is repeated a large number of times. The average of the correlation coefficients obtained in this way is the DR of the dataset. In the present paper, the DR statistic is applied to four studies involving amplified fragment length polymorphism as well as micro-satellite markers. In addition, some properties and possible applications of DR are discussed, including the prediction of the added value of scoring additional markers, and the determination of which similarity measure is, apart from genetical considerations, most appropriate for analyzing the data.

Introduction

Molecular markers such as micro-satellites (Morgante and Olivieri 1993) and amplified fragment length polymorphism (AFLP) (Vos et al. 1995) are popular tools in studying the genetic structure of a set of genotypes or populations. This type of studies are used to determine the genetic relationship amongst the accessions in a gene bank or breeding program, e.g. to identify duplicates and describe flux across time, to predict heterosis and compose heterotic groups for hybrid breeding, or to identify essentially derived varieties in plant variety protection (cf., Reif et al. 2005). Often genetic relationships are visualized by a principle component plot (e.g. Jolliffe 1986) or by a dendrogram obtained from cluster analysis, such as the unweighted pair group method with arithmetic average (Sneath and Sokal 1973). Using these methods, samples appearing close to each other are considered genetically similar, whereas samples appearing far apart are considered genetically different. Especially if the marker data reveal a lack of structure, the question arises whether this is due to a true lack of population structure or the result of poor data quality.

Data quality has many aspects. One of these aspects is the frequency of obvious errors in the data set, think of occurrence of score '2' where only '0', '1' and '*x*' have been defined, or rows with more scores than there are columns defined. Another aspect is the availability of metadata required to properly interpret the data, or repeat the experiment, think of the protocols used in the molecular characterization and proper references to the material that was analyzed. The purpose of this paper is to present a method based on calculating an intra-class correlation using a jackknife approach for estimating yet another aspect of quality: the ability of datasets to describe biological structure, by quantifying the internal consistency of the data. This biological structure may depend on the type of markers that are used. Neutral markers may describe the biological structure of the same biological material differently as compared to functional markers, simply because different

Communicated by A. Bervillé.

Th. J. L. van Hintum (✉)
Centre for Genetic Resources, The Netherlands (CGN),
Wageningen University and Research Centre,
P.O. Box 16, 6700 AA Wageningen, The Netherlands
e-mail: theo.vanhintum@wur.nl

processes have shaped it (random vs. selection processes). Also AFLPs might result in a different structure as compared to simple sequence repeats (SSRs) because of the different properties of these markers systems in terms of coverage of the genome, proximity to selective genes, etc.

The proposed quality estimate is called data resolution (DR), and will be illustrated by applying it to four datasets that were used in published molecular studies: an AFLP and a SSR dataset describing wild lettuce (*Lactuca serriola*), an AFLP dataset describing barley (*Hordeum vulgare*) and one with AFLP band frequencies describing white cabbage (*Brassica oleracea*).

Materials and methods

Definition of "data resolution"

The following assumptions are made: (1) The set of markers used is a random sample from an infinite universe of similar markers such as AFLP (Vos et al. 1995), SSR (Morgante and Olivieri 1993), single nucleotide polymorphisms (Jenkins and Gibson 2002), or diversity arrays technology (Wenzl et al. 2004). Different types of markers have different abilities to describe the structure of a population. Markers will to a varying degree be correlated with each other. (2) If an infinite number of markers is used, the structure of the population (as defined by the type of marker) will be perfectly described. This implies that a very large set of markers is expected to result in the same description of the structure as any other very large set of markers of the same type. As a consequence, the DR for large marker data sets will tend to unity. (3) The structure of a population can be described by the similarities of all pairs of individuals. Principle coordinate analysis and clustering analysis are merely used to visualize this matrix of similarities.

The data resolution is calculated as follows. Consider an analysis based on a dataset consisting of N markers and M objects, where for reasons of simplicity N is taken to be an even number. The data set is randomly split into two independent data sets D_1 and D_2, both consisting of $N/2$ markers $\times$ M objects each. For each of the data sets D_1 and D_2, the $M(M-1)/2$ similarities between all pairs of objects are calculated. Subsequently the correlation coefficient (Edwards 1976) between these two sets of values is calculated. This is repeated a large number of times using different random divisions of the dataset. The average correlation coefficient is calculated. This average correlation coefficient is called the DR of the dataset.

As concluded from assumption (2) the DR of a data set consisting of an infinite number of markers is equal to unity. If the quality of a dataset is low, which means that the data are not able to properly describe the genetic structure of the population, a division of the datasets will result in two halves which will each describe a different structure, and will thus have a low correlation and a low DR. In other words, a low DR indicates that the data set (before splitting) will give an unstable and therefore unreliable description of the structure of the objects. However, it should be realized that a strong structure will result in high values for DR (close to 1.0) whereas a weak structure will be hardly recovered, thus a low DR, even if the data are of high quality. In other words, a dataset mixing close and distant objects (small and high dissimilarities), high correlations are expected whatever the consistency. So DR is valid for a given set of objects but care should be taken when comparing between different sets of objects.

Calculation of "data resolution curve"

To get an indication of the effect of adding markers to an existing data set, a "data resolution curve" can be calculated. This curve consists of DR values plotted against number of markers, ranging from two to the total numbers of markers in the dataset. For example to calculate the point on the data resolution curve for ten markers, two sets of five markers are randomly sampled without replacement from the whole marker data set, and used to calculate all pair wise similarity estimates of the objects and their correlation (as described above). This is repeated many times. The average correlation will give an expectation of the DR of ten markers from the marker set. This curve will show the effect of the use of less or more (via extrapolation) markers on the DR.

The "data resolution curve" gives the average DR of a set of a given number of markers. Obviously it is also possible to calculate the DR of specific subsets of markers. For example, it is possible to calculate the DR of the "first two" SSR markers, calculate the effect of adding the third and so on. The resulting curve is called the "sequential data resolution curve".

Similarity measures

Since DR is calculated from the pair wise similarities between objects, it requires a measure for determining this similarity. To calculate similarities, five commonly used methods are applied and compared. The first three are solely for calculating similarities between single objects with presence/absence data: (1) the Jaccard similarity coefficient (Jaccard 1908; Sneath 1957), where similarity is defined as the fraction of band positions with common bands relative to the total number of positions with bands: n11/(n01 + n10 + n11), (2) the simple matching coefficient (Sokal and Michener 1958), which is the fraction of positions with a common state (present or absent) relative to the

total number of positions: (n11 + n00)/(n11 + n01 + n10 + n00) and (3) the Nei and Li similarity (Nei and Li 1979) also known as the Dice coefficient (Dice 1945), which is the fractions of bands shared by both individuals relative to the total number of bands, thus $2 \times n11/(2 \times n11 + n01 + n10)$. These three measures cannot be used for allele frequency data. The simple matching coefficient, in which the absence of an allele in both objects is considered an indication of similarity, has no interpretation in a diversity study using multi allelic loci such as SSRs.

The other two measures used in this study are quantitative measures that were used for calculating dissimilarity based on allele frequencies in populations: (4) the Euclidean and (5) Manhattan distance, the latter also known as the City Block distance. These distances can also be used for absence/presence data using allele frequencies of zero and one. If the loci under study are assumed to be bi-allelic, such as in the AFLP studies, the Euclidean distance is the square root of the Manhattan distance, and the Manhattan distance corresponds to the simple matching coefficient. Furthermore, in the bi-allelic case the Manhattan distance corresponds to the Rogers' distance and the Euclidean distance corresponds to the modified Rogers' distance (Rogers 1972; Sokal and Rohlf 1962). (For an excellent discussion of the different similarity measures see Reif et al. 2005). In the multi-allelic SSR case, the Mahattan distance was calculated as relative number of SSR markers that differed (and thus one minus the Nei-Li similarity), and the Euclidean distance as the square root of the Manhattan distance.

Data

The behavior of DR was tested using a number of datasets. Most calculations were performed with a dataset consisting of 100 *Lactuca serriola* plants characterized with three AFLP primer combinations, yielding 179 polymorphic bands. The fingerprint for each plant was coded as a sequence of 179 and 1's, 0 indicating the absence and 1 indicating the presence of a band. The 100 plants were randomly selected from all characterized plants with complete fingerprints. This work was done in the framework of a much larger EU project aimed at exploring the possible uses of molecular markers for genetic resources management (Hintum 2003). The second data set originated from the same project, and consisted of the scores of ten micro satellites (SSR) markers describing 100 *Lactuca serriola* plants. These scores were recorded in a binary matrix, in which each of the alleles observed corresponded to a column and each allele was scored per plant as present (1) or absent (0). The plants were also randomly selected from all characterized plants that showed one allele for each marker. Heterozygotes and plants with missing values were excluded to facilitate the analysis. The third dataset consisted of the data used for comparing the structure of a *Brassica oleracea* genebank collection with the effects of standard regeneration protocols (Hintum et al. 2007). This set consisted of 56 accessions described by the frequency of occurrence of 101 polymorphic AFLP bands in 50 plants per accession, recorded as a sequence of 101 values between 0 and 1, indicating the frequency of the corresponding band in the accession. One accession from the original dataset was removed since it contained missing values. The fourth data set consisted of 75 polymorphic AFLP bands on a set of 51 barley varieties that were genotyped to determine how molecular fingerprints could support decisions concerning acquisition for a gene bank collection (Treuren et al. 2006). The fingerprints were recorded in the same format as the first dataset. In this case one variety was removed from the data set since it contained missing values.

The removal of hetrozygotes and missing values was purely to increase the simplicity of the calculations. The alternative would be to adjust the similarity calculations accommodating hetrozygosity and missing values, this would not have any consequence for the subsequent analysis.

Data analysis

All data analyses were performed with tailor-made software programs written in visual basic for applications in a MS-Excel environment. The number of replications used to calculate the results presented in the present paper was 10,000.

Results

The calculation of the DR of datasets appeared relatively straightforward; no complicated calculations need to be made.

To obtain reliable estimates of the DR large number of runs (each run represents one case of splitting the data set) had to be performed. Figure 1 shows the data resolution curve for the *Lactuca serriola* data set. It also shows the standard deviation of the correlation coefficients in individual runs for a range of number of markers. It can be observed that especially for a small number of markers the standard deviation is large. For the cases presented in this paper, using 10,000 runs, this resulted in a standard error of around 0.001.

Figure 2 shows the data resolution curves for all three AFLP datasets. All curves show a similar shape, and, probably due to the similarity in the type of markers all being characterized in the same lab, also show remarkable similar values for a given number of markers. This is somewhat surprizing given the dependency of the DR on the structure of the studied material. Figure 3 shows the data resolution

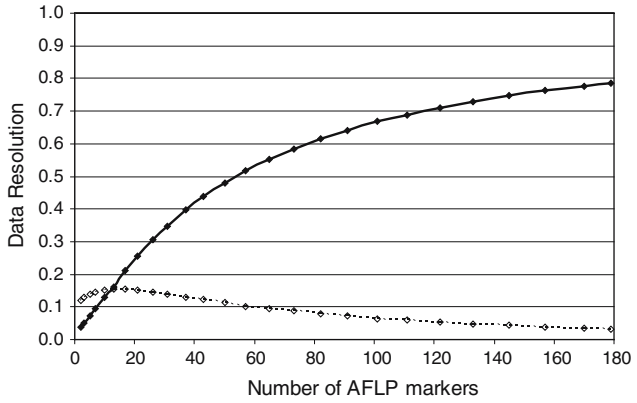

Fig. 1 Data resolution curve (Jaccard similarity) for the *Lactuca serriola* AFLP data set (*fat line*) and standard deviation of the correlation coefficients in 10,000 runs (*dotted line*)

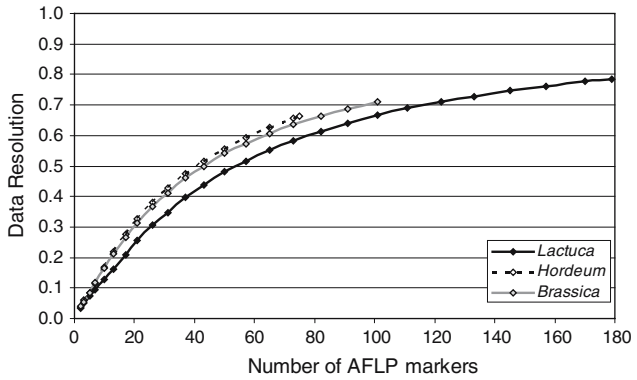

Fig. 2 Data resolution curves for the *Lactuca serriola*, *Hordeum vulgare* and *Brassica oleracea* AFLP data sets based on 10,000 runs (based on Jaccard similarities for the *Lactuca* and *Hordeum* data set, and the Euclidean for the *Brassica* dataset)

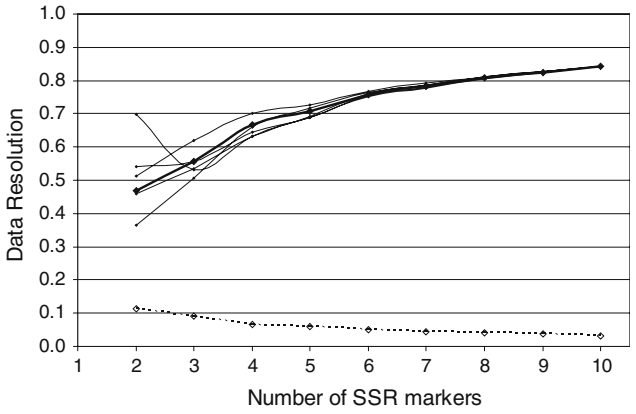

Fig. 3 Data resolution curves (Jaccard similarity) for the *Lactuca serriola* SSR data set (*fat line*), standard deviation of the correlation coefficients in 10,000 runs (*dotted line*) and five sequential data resolution curves of randomly ordered markers

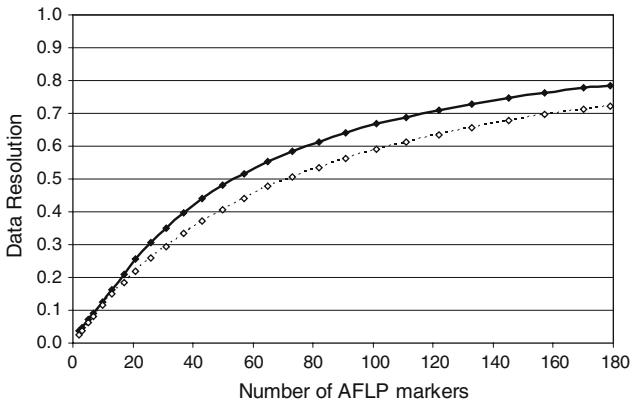

Fig. 4 Data resolution curves for the *Lactuca serriola* AFLP data set, using the Jaccard (*fat line*) and Simple matching (*dotted line*) similarity indexes

curve for the *Lactuca serriola* SSR data set and its standard deviation again showing the familiar shape. This figure also shows five sequential data resolution curves of random sequences of the ten SSR markers. It can be observed that especially with small number of markers the shape can be quite different from the average curve, as could be expected looking at the relatively large standard deviation.

The DR can be used to select in a family of biologically relevant similarity measures the best index from a consistency perspective (e.g. between Euclidean or Manhattan for frequencies). To illustrate this possible application,the DR for the data sets were calculated using all five (dis-)similarity measures, as far as they could be calculated given the dataset. The results are presented in Table 1. Here it can be observed that in the bi-allelic AFLP data sets (*Lactuca* and *Hordeum* data), the DR for simple matching is equal to that using the Manhattan distance, as expected. To show that these differences can have a large impact on the number of markers needed to reach a certain DR the data resolution curves of the best (Jaccard) and worst (simple matching) measures for the *Lactuca serriola* AFLP data set are presented in Fig. 4. It shows that if a DR of 0.7 is to be realized, using the simple matching similarity, 158 markers are required, whereas the same DR can be reached with 133 markers if the Jaccard similarity index distance is used. This implies a potential saving of 25 markers based on the choice of the Jaccard similarity measure in favor of simple matching. Similarly, in the *Lactuca serriola* SSR case, the resolution of the complete set of ten markers when using Nei-Li (0.816) is reached with less than nine markers if Jaccard is used (nine markers with Jaccard have a DR of 0.825).

Discussion

Calculation of DR

In order to calculate DR, large numbers of runs are required. In this study the DR of the complete datasets

Table 1 Data resolution of the four data sets using different (dis-)similarity measures

Data set	(Dis-)Similarity measure				
	Jaccard	Nei-Li	Simple matching	Euclidean	Manhattan
Lactuca 179 AFLP markers	0.784	0.762	0.722	0.752	0.721
Hordeum 75 AFLP markers	0.662	0.652	0.672	0.676	0.672
Brassica 101 AFLP frequencies	–	–	–	0.710	0.672
Lactuca 10 SSR markers	0.842	0.816	–	0.834	0.816

ranged from 0.652 to 0.842; the corresponding standard deviations were 0.060 and 0.033. This implies that when 10,000 runs are used the standard error of the DR estimates is around 0.001. For the calculation of the data resolution curves the standard deviation is larger (see Figs. 1, 3) since the sampling of the markers in each run introduces an additional source of variation.

This study used data sets of up to 100 objects. When the DR of larger data sets is to be determined alternative approaches have to be considered since the number of similarities that need to be calculated for each run becomes too large. Random sampling of a limited number of similarities, i.e., random combinations of cases, can be considered. The effects of this and other alternative methods for calculating DR on the variance of the estimates needs further study.

Another issue concerns the calculation of the DR in case of an odd number of markers. If an odd number of markers (M) is present, and an equal split of markers is not possible, the markers were divided into two sets of $(M + 1)/2$ and $(M - 1)/2$ markers respectively. Since the highest correlation can be expected between data sets of equal size, this inequality will result a slight reduction of the DR, as can be observed in Fig. 3; the connection between the values for 2 and 4 shows a small but observable decrease, between 4 and 6 this effect has nearly disappeared, and it can hardly be observed for larger values. Since the DR of such a small number of markers will rarely be relevant this effect has virtually no impact. To avoid this effect, it is also possible and in some cases preferable to only consider even number of markers in the construction of the data resolution curve, allowing a split in equal numbers of markers.

Data resolution as function of number of markers

When studying the properties of the data resolution curves, the influence of the structure of the objects should first of all be taken into account: much structure, this is a mixture of clusters of similar material, will result in high DR values. All four figures show the expected shape of the data resolution curve: a relatively steep start, approaching the value unity asymptotically. Despite the differences of the populations analyzed, all three AFLP datasets that were analyzed resulted in similar curves (Fig. 2). The starting point of the SSR curve (Fig. 3) was much higher: the starting point for the AFLP curves was at about 0.04, the SSR curve started at 0.47. This was expected since a multiallelic SSR marker contains much more information than a single AFLP marker displaying only the presence or absence of a band, which might even correspond to more than one DNA fragment due to homoplasy (Koopman and Gort 2004).

The individual sequential data resolution curves (Fig. 3) showed large variation. The five random marker sequences that were calculated (out of the 1,814,400 that are possible) show a large variation of DR values for small number of markers. This is not surprising. If the first two markers are relatively highly correlated they will result in similar similarities and thus a high DR. A third marker can destroy this congruence and decrease the correlation considerably, as can be seen in one of the curves drawn in Fig. 3, starting at a DR of 0.70 for two markers, dropping to 0.53 for three markers and recovering to 0.63 and 0.69 for four and five markers respectively. This effect also occurs in AFLP datasets for small number of markers, which is reflected in the relatively high standard deviations of the average DR for small numbers of AFLP markers (Fig. 1).

Data resolution as function of similarity measure

The calculation of the DR requires a similarity measure. As a result, the choice of this measure will influence the resolution of a dataset. This choice of a measure to calculate the genetic similarity based on molecular marker data is primarily based its mathematical properties in relation to the biological characteristics of the markers: qualitative versus quantitative, dominant versus codominant, homozygote versus heterozygote, bands versus alleles, etc. However, DR can be used to compare the ability of several relevant measures from a consistency perspective. The absolute differences between DRs based on different similarity measures is small (Table 1), however since the slope of the data resolution curve becomes low relatively soon (Fig. 4), these small absolute differences translate in large difference in number of markers required to reach a certain DR level, as was illustrated by the examples in the results section.

From a consistency perspective, and only based on this descriptive study, some observation about the performance of the similarity measures can be made. Of the "classic" similarity measures for individual marker scores, Jaccard

performed consistently better than Nei-Li, the performance of simple matching varied (Table 1). This occurred in all cases where these measures could be calculated, including the SSR case. Of the two quantitative measures the Euclidean distance performed consistently better than the Manhattan distance. Surprizingly the Euclidean distance could compete with the "molecular distances", in the *Hordeum* AFLP case both quantitative measures performed better than Jaccard and Nei-Li. Further study will be required to explain these observations.

Importance of data resolution

The data resolution as presented in this paper provides a way to quantify the consistency of a data set in terms of the ability of the dataset to describe the structure of the characterized material as defined by the markers used. Besides this widely applicable property, the data resolution also allows prediction of the value of additional markers and comparison of alternative ways to calculate the similarity between individuals. Until now such quantity was unavailable. In principal component analysis it is possible to indicate the "percentage of explained variance" for each principal component, which can be considered an indication for the degree to which the data describe the structure in the material. However, this percentage will obviously be largely dependent on the structure of the material and the number of markers used (the more markers, the lower the explained variance). Also in the case of hierarchical clustering algorithms it is possible to give an indication of the stability of the tree by calculating "bootstrap values" for each node (Felsenstein 1985). An other approach that can be used to validate the structure represented in a dendrogram is the cophenetic correlation coefficient (Rohlf 1972). It is a simple correlation between the cophenetic distances obtained from the tree (the height of the link between two objects in the dendrogram) and the original distances in the distance matrix that was used to construct the dendrogram. These analyses are obviously very dependent of the structure of the material itself and also of the clustering algorithm used. The data resolution is also to some extent determined by the structure of the material, as discussed above. A proper validation of the method, including a determination of the effect of the population structure should be determined in future studies probably using simulated data, where both structure and consistency of the data can be defined a priori.

Zhang et al. (2002) and You et al. (2004) used the correlation between the similarity matrix based on a subset of alleles or markers with that based on the complete set of markers as a measure of the quality of the subset. This approach however requires one to have a sufficiently large set of markers before the quality of a set can be determined, as obviously will not be the case in most situations.

In conclusion, the DR is an additional tool in gaining insight in the consistency of the dataset, and its ability to describe the structure of the objects under, and allows prediction of the effect of adding additional markers. However, as with any resampling method, the results are based on the dataset itself and not on any additional information. Any bias in the data will also be present in the subsamples, and can thus influence the value of the DR. But given this restriction, it provides a new method to quantify an important aspect of quality of the data before analysis by calculating an intra-class correlation using a jackknife approach. It is widely deployable and relative intuitive.

Acknowledgments The author would like to thank Rob van Treuren, Hans Jansen, Jean Christophe Glaszmann and Graham McLaren for suggestions and comments. The author would also like to thank the anonymous referees for their excellent feedback that greatly helped to improve the manuscript. This work is supported by the Generation Challenge Programme.

References

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:297–302

Edwards AL (1976) The correlation coefficient: an introduction to linear regression and correlation, Chap 4. W. H. Freeman, San Francisco

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

Hintum TJL van (2003) Molecular characterisation of a lettuce germplasm collection. Eucarpia leafy vegetables, In: Proceedings of the Eucarpia meeting on leafy vegetables genetics and breeding, Noordwijkerhout, The Netherlands, 19–21 March, 2003. Centre for Genetic Resources, Wageningen, pp 99–104

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bull Soc Vaud Sci Nat 44:223–270

Jenkins S, Gibson N (2002) High-throughput SNP genotyping. Comp Funct Genom 3:57–66

Jolliffe IT (1986) Principal component analysis. Springer, New York

Koopman WJM, Gort G (2004) Significance tests and weighted values for AFLP similarities, based on *Arabidopsis* in silico AFLP fragment length distributions. Genetics 167:1915–1928

Morgante M, Olivieri AM (1993) PCR-amplified microsatellites as markers in plant genetics. Plant J 3:175–182

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

Reif JC, Melchinger AE, Frisch M (2005) Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

Rogers JS (1972) Measures of genetic similarity and genetic distance. Studies in genetics VII. University of Texas Publication 7213, Austin, pp 145–153

Rohlf FJ (1972) An empirical comparison of three ordination techniques in numerical taxonomy. Syst Zool 21:271–280

Sneath PHA (1957) Some thoughts on bacterial classification. J Gen Microbiol 17:184–200

Sneath PHA, Sokal RR (1973) Numerical taxonomy. W.H. Freeman, San Francisco, pp 230–234

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–1438

Sokal RR, Rohlf FJ (1962) The comparison of dendrograms by objective methods. Taxon 11:33–40

van Hintum TJL, van Treuren R, van de Wiel CCM, Visser DL, Vosman B (2007) The distribution of AFLP variation in a *Brassica oleracea* genebank collection in comparison with the effects of regeneration on diversity. Theor Appl Genet 114:777–786

van Treuren R, Tchoudinova I, van Soest LJM, van Hintum TJL (2006) Marker-assisted acquisition and core collection formation of plant genetic resources: a case study in barley using AFLPs and pedigree data. Genet Resour Crop Evol 53:43–52

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. PNAS 10:9915–9920

You GX, Zhang XY, Wang LF (2004) An estimation of the minimum number of SSR loci needed to reveal genetic relationships in wheat varieties: information from 96 random accessions with maximized genetic diversity. Mol Breed 14:397–406

Zhang XY, Li CW, Wang LF, Wang HM, You GX, Dong YS (2002) An estimation of the minimum number of SSR alleles needed to reveal genetic relationships in wheat varieties. I. Information from large-scale planted varieties and cornerstone breeding parents in Chinese wheat improvement and production. Theor Appl Genet 106:112–117