Theor Appl Genet (1995) 90:755–761

N. Diwan · M. S. McIntosh · G. R. Bauchan

Methods of developing a core collection of annual *Medicago* species

Received: 25 February 1994 / Accepted: 30 May 1994

Abstract A core collection is a subset of a large germplasm collection that contains accessions chosen to represent the genetic variability of the germplasm collection. The purpose of the core collection is to improve management and use of a germplasm collection. Core collections are usually assembled by grouping accessions and selecting from within these groups. The objective of this study was to compare 11 methods of assembling a core collection of the U.S. National collection of annual *Medicago* species. These methods differed in their use of passport and evaluation data as well as their selection strategy. Another objective was to compare core collections with sample sizes of 5%, 10% and 17% of the germplasm collection. Core collections assembled with evaluation data and cluster analysis better represented the germplasm collection than core collections assembled based solely on passport data and random selection of accessions. The Relative Diversity and the logarithm methods generated better core collections than the proportional method. The 5% and 10% sample size core collection were judged insufficient to represent the germplasm collection.

Key words Core collection · Germplasm collection · Annual *Medicago* species · Relative Diversity method

Introduction

Germplasm collections exist to conserve the genetic diversity of crop species and their wild relatives (Williams 1991). Nevertheless, the size of many large germplasm collections may be an obstacle to their evaluation and utilization (Holden 1984). The management and use of germplasm collections could be enhanced if a limited number of genetically diverse accessions within the collection were selected as the core collection (Frankel 1984) and given priority in evaluation and hybridization (Brown 1989a).

The core collection generally contains 5–10% of the germplasm collection and ideally conserves at least 70% of the alleles in the whole collection (Brown 1989a). A good core collection should have no redundant entries, represent the whole collection in regard to species, subspecies, and geographical regions and be small enough to manage easily (Brown 1989b). The rest of the collection should be maintained as the "reserve collection".

Recently, there is a growing interest in the development of core collections (Hodgkin 1990). Core collections have been developed for several germplasm collections [okra (*Abelmoschus esculentus* Moneh.) (Hamon and van Stolen 1989), perennial *Glycine* spp. (Brown et al. 1987), winter wheat (*Triticum aestivum* L.) (Mackay 1986, 1989), peanut (*Arachis hypogae* L.) (Holbrook et al., 1993) and annual *Medicago* spp. (Diwan et al. 1994)] using various criteria and sampling techniques.

Brown (1989b) suggested assembling the core collection by determing groups within the germplasm collection (species, subspecies, geographical regions, maturity groups etc.) and selecting entries from each group. The number of entries chosen from a group depends on the core size and can be determined by using the constant, proportional or logarithm methods (Brown 1989b).

The constant, proportional and logrithm methods utilize diversity among but not within groups. Germplasm collections should represent the genetic diversity of a species, and be assembled from the range of geographical and ecological zones of the crop gene pool without bias (Williams 1991). However, many germplasm collections have not been assembled systematically. As a result, these collections may over- or underrepresent certain geographical and ecological zones, and a small group of accessions in the germplasm collection

Communicated by P. M. A. Tigerstedt

N. Diwan (✉) · G. R. Bauchan
USDA/ARS, Soybean and Alfalfa Research Laboratory, Beltsville Agriculture Research Center-West, Beltsville, MD 20705, USA

M. S. McIntosh
Department of Agronomy, University of Maryland, College Park, MD 20742, USA

may be quite diverse while a large group may be uniform. Therefore, we suggested the Relative Diversity method as an alternative method to determine the number of accessions to be selected for a core collection from each group. The Relative Diversity method is based on the variability within a group in the germ plasm collection, and it selects the number of accessions within a group based on the group's relative phenotypic or genetic diversity. The Relative Diversity method was used to assemble the U.S. annual *Medicago* core collection (Diwan et al. 1994).

The objectives of the study presented here were (1) to compare methods of assembling core collections that differ in their use of evaluation and passport data, as well as the selection strategy (proportional vs logarithm vs Relative Diversity, and direct selection versus random vs selection of accessions), and (2) to determine the effect of sampling proportion (5%, 10% or 17%) on the core collection. The 11 methods compared used evaluation data from the US National collection of annual *Medicago* species.

Materials and methods

Plant material

The US annual *Medicago* species collection has been described by Diwan et al. (1994). This germplasm collection consists of 36 species and 3159 accessions. The number of accessions per species in the germplasm collection ranges from 1 to 651 (Table 1).

An initial subset of 40% (1240 accessions) was selected for field evaluation (Table 1). The subset included accessions from all annual *Medicago* species in the germplasm collection chosen to represent proportionally the countries of origin for that species (Diwan et al. 1994). Specific accessions within the country of origin were randomly selected.

Field evaluation

The initial subset of accessions was evaluated in the field at Beltsville, Maryland, in 1990. The field preparation and conditions and the traits evaluated and analyzed have been described by Diwan et al. (1994). Traits evaluated were: days to flower, days to full pod production, growth habit, biomass within species, variability within accession, pod production, pod spines, plant height, plant maximal spread, length of middle leaflet, width of middle leaflet, number of flowers per

Table 1 The number of accessions of each annual *Medicago* species in the germplasm collection, in the initial subset, and in the core collections chosen using the Relative Diversity, proportional, and logarithm methods

Medicago species	Germplasm collection	Initial subset	Method		
			Relative Diversity	Proportional	Logarithm
arabica (L.)	71	35	2	6	7
blancheana	18	18	8	3	6
ciliaris (L.)	73	31	6	5	7
constricta	48	30	3	5	7
coronata (L.)	23	3	2	1	2
disciformis DC.	50	30	4	5	7
doliata Carmign	127	40	3	7	8
granadesis	14	13	4	2	5
heyniana Greuter	2	2	1	1	1
intertexta (L.)	22	19	6	3	6
italica (Miller)	83	32	9	6	7
laciniata (L.)	130	52	10	9	8
lanigera Winkl.	1	1	1	1	1
lesinsii E.	5	2	2	1	1
littoralis Rohde	120	50	6	9	8
lupulina L.	170	63	14	11	8
minima (L.)	274	101	4	17	9
murex Willd.	78	36	6	6	7
muricoleptis	7	7	1	1	4
noeana Boiss.	19	14	3	3	5
orbicularis (L.)	251	86	8	15	9
platycarpa (L.)	6	5	1	1	3
polymorpha L.	651	217	36	38	11
praecox DC.	21	20	2	3	6
radiata L.	12	11	4	2	5
rigidula (L.)	329	104	6	18	10
rotata Boiss.	21	20	8	3	6
rugosa Desr.	43	28	11	5	7
sauvagei Negre	5	5	2	1	3
scutellata (L.)	60	37	18	6	7
secundiflora	2	2	1	1	1
shepardii Post	4	4	1	1	3
soleirolli Duby	10	10	3	2	5
tenoreana Ser.	6	5	1	1	3
truncatula	325	71	8	12	9
turbinata (L.)	83	38	6	7	7
Total	3159	1240	211	218	209

raceme, internode length and seed size. The traits evaluated were chosen from the alfalfa (*M. sativa* L.) descriptor list developed by the Alfalfa Crop Advisory Committee (vs Deparatment of Agriculture 1989).

Selection of core collections

Eleven methods were used to assemble the core collections (Fig. 1). Each species was considered a group, and accessions were selected within every species for all core collections. Cores 1–7 were assembled on the basis of passport (species, variety, place of origin) and field evaluation data; cores 8–11 were assembled using passport data only.

For cores 1–6, accessions within species were grouped into clusters based on the 14 measured phenotypic traits that were standardized with a mean of 0 and a variance of 1. A Statistical Analysis System (SAS) macro (Jacobs 1990) calculated Euclidean distances between traits for each of the *Medicago* species. The distance matrices were entered into another SAS macro (Jacobs 1990) to conduct cluster analysis using an unweighted pair group method with arithmetic averages. The number of accessions selected per species for cores 1–6 was determined by the proportional or logarithm methods as described by Brown (1989b) or by the Relative Diversity method (Diwan et al. 1994) (Fig. 1). For the Relative Diversity method, accessions were selected within species for the core collections based on the diversity of the traits measured among accessions within species. Diversity within each species was determined by the number of clusters in each species. An Euclidian distance of 3.0 was used to determine clusters because it generated the desired core collection size 200–250 accessions. Species with more clusters were considered more diverse. One accession per cluster was selected for the core collection. Accessions within clusters were chosen either at random (Cores 2, 4 and 6), or to maximize the representation of geographical regions for the species in the germplasm collection (Cores 1, 3 and 5) (Fig. 1). In order to achieve the greatest representation of geographical regions in cores 1, 3 and 5, countries of origin as recorded in the passport data were grouped into geographical regions (Diwan et al. 1994). Within a species, accessions were chosen from different geographical regions. Core 1 was designated as the US annual *Medicago* species core collection (Diwan et al. 1994).

Core 7 was determined by clustering geographical regions within species. The species means of the measured traits for each geographical region were standardized with a mean of 0 and a variance of 1, and entered into the SAS macro (Jacobs 1990) to calculate a distance matrix. A cluster analysis was executed as described for cores 1–6. The number of accessions per species selected for core 7 was based on the Relative Diversity method. Diversity within each species was determined by counting the number of clusters created in each dendrogram at an Euclidian distance of 2.0.

Cores 8–11 were selected based only on the accessions' passport data information. For cores 8 and 9, a stratified random sampling method within species was used. Accessions were first grouped into geographical regions based on their places of origin (Table 2) and then randomly selected within these regions within species. The number of accessions selected from each geographical region was determined by the proportional (Core 8) or the logarithm (Core 9) methods. For cores 10 and 11, accessions were randomly selected from each species ignoring geographical information. The number of accessions slected from each species was determined by the proportional (Core 10) or the logarithm (Core 11) methods.

In order to investigate the influence of sample size on the core collection, core collections containing 5% or 10% of the initial subset were assembled using the selection strategies of cores 1, 3 and 5. Euclidean distances of 4.5 and 3.6 were used to cluster the two Relative Diversity cores, since these generated the desired number of accessions for the 5% and 10% core collections, respectively.

Wilcoxon rank-sum non-parametric tests were conducted to determine whether the 11 core collections represented the initial subset. These tests compared the number of means for each of the species and the number of ranges averaged over species that were significantly different between each of the core collections and the initial subset (Table 3). Wilcoxon rank-sum tests were performed using the SAS NPAR1WAY procedure Wilcoxon option (SAS Institute 1989). Additionally, a "range ratio" was calculated to determine the proportion of the range retained by each core collection. The "range ratio" is the average ratio of the range of the core collection to the range of the initial subset. Data from 18 species that had at least 28 accessions in the initial subset were used (Table 1).

$$\text{Range ratio} = \Sigma[(\Sigma_j RC_{ij}/RG_{ij})/n]/s$$

where RC = the range for the ith trait of the jth species in the core collection, RG = the range for the ith trait of the jth species in the

Fig. 1 Methods used for the assemblage of the annual *Medicago* species core collection

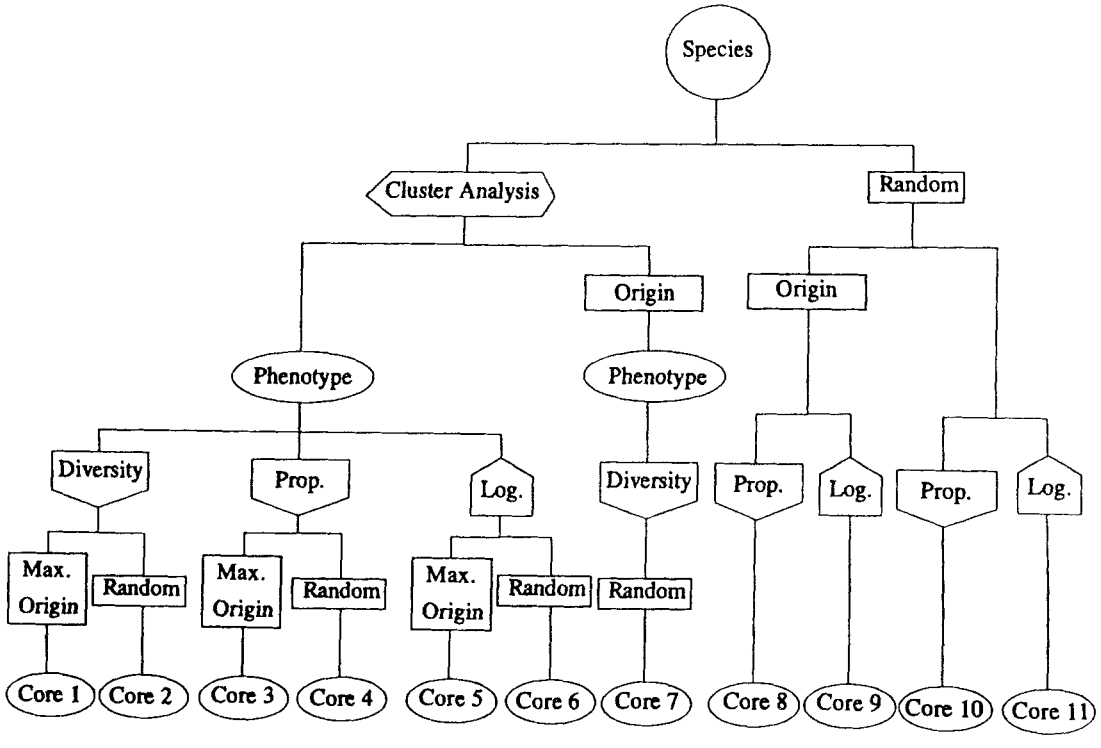

Table 2 Percentage of means and ranges significantly different ($\alpha = 0.05$) between the core collection and the germplasm collection, based on a Wilcoxon rank-sum non-parametric test and range ratio

Statistic	Core number										
	1	2	3	4	5	6	7	8	9	10	11
Mean[a]	3	5	5	3	3	3	<1	<1	<1	<1	<1
Range	14	7	7	29	0	0	86	71	71	79	86
Range ratio[b]	74 ± 14	74 ± 14	78 ± 14	81 ± 12	79 ± 9	79 ± 8	62 ± 16	59 ± 14	60 ± 15	56 ± 12	54 ± 14

[a] Number of comparisons per core collection: means = 504 and ranges = 14
[b] Mean range ratio ± SD based on 18 species with a sample size of at least 28 accessions

Table 3 Percentage of means significantly different ($\alpha = 0.05$) between each of the 5% and 10% sample size core collections and the initial subset, based on a Wilcoxon rank-sum non-parametric test and percentage of range ratios smaller than 0.70 (*prop* proportional method, *log* logarithm method, *div* Relative Diversity method)

Statistic	Sample size					
	5%			10%		
	Core method					
	Div	Prop	Log	Div	Prop	Log
Means[a]	2	3	4	3	4	4
Ranges[b]	0	36	14	0	0	7
Range ratio[c]	33 ± 22	43 ± 23	34 ± 17	55 ± 21	61 ± 17	62 ± 11

[a] Number of means compared = 504
[b] Percentage of ranges with R = (Range of the core-collection)/(Range of the initial subset) ratio smaller than 0.70, number of ranges compared = 14
[c] Mean range ratio ± SD based on 18 species with a sample size of at least 28 accessions

initial subset, n = number of ranges with non-missing values for the species and s = number of species.

Core collections were considered to be representative of the initial subset and therefore acceptable, if (1) 30% or fewer of the means and ranges of the core collection were significantly different ($P = 0.05$) from the initial subset; and (2) the percentage of the range retained by the core collection (range ratio) was at least 70% of the range of the initial subset.

Wilcoxon rank-sum non-parametric tests were used to compare the means of the 5% and 10% sample size core collections to the initial subset. Because of the small size of the 5% and 10% core collection, Wilcoxon rank-sum non-parameteric tests were not used to compare ranges of the 5% and 10% core collections to the initial subset. Instead, a ratio (R) of the range of core collection to the range of the initial subset overall species was calculated. The 5% and 10% sample size core collections were judged acceptable if they met the criteria listed above for cores 1–9 and if at least 70% of the ranges of the core collection had $R \geq 0.7$.

Results and discussion

Although the sizes of the 11 core collections were similar, the number of accessions per species in the different core collections varied greatly depending on selection method (Table 1). The number of accessions per species fluctuated the least among groups for the logarithm method (Table 1). In contrast, the proportional method tended to emphasize large groups. Therefore, species such as *M. polymorpha* L., *M. minima* (L.) Bart and *M. rigidula* (L.) All. are represented by more accessions in the proportional method core collections than in the other core collections (Table 1). Species such as *M. blancheana* Boiss., *M. italica* (Miller) Fiori, *M. lupulina* L., *M. rugosa* Desr. and *M. scutellata* (L.) Miller that expressed a wide range of variability for the evaluated traits were represented by more accessions in the Relative Diversity core collections than in other core collections. *M. arabica* (L.) Huds., *M. doliata* Carmign., *M. minima* and *M. rigidula* were represented in the Relative Diversity core collections by fewer accessions than other core collections due to their narrow range of variability (Table 1).

The differences between the three methods are illustrated by *M. scutellata* (Fig. 2), and *M. murex* Willd. (Fig. 3). The two species were represented by the same number of accessions with the proportional and logarithm methods. However, because *M. scutellata* was much more variable than *M. murex* it had 3 times more accessions in the Relative Diversity core collections (Cores 1 and 2) than *M. murex* (Table 1).

The trait means of the core collections were generally the same as the means in the initial subset regardless of the assembling method (Table 2). More than 70% of the ranges of cores 7–11 were significantly different from those of the initial subset. The range ratio for these core collections showed over a 30% average reduction in the range. Therefore, core collections assembled by the random selection of accessions (Cores 8–11) or by the selection of accessions from a cluster of geographical regions within species (Core 7) were not considered representative of the initial subset. However, stratified random sampling within species (Cores 8 and 9) resulted in slightly better core collections than a completely random selection (Cores 10 and 11).

The logarithm method based on evaluation data produced 2 acceptable core collections (Cores 5 and 6). Brown (1989b) argued that the logarithm method is efficient when the level of variation among and within groups in the germplasm collection is unknown. The logarithm method was found here to be efficient only when the level of variation among groups was known. When the level of variation in the collection was unknown, the logarithm method generated unacceptable

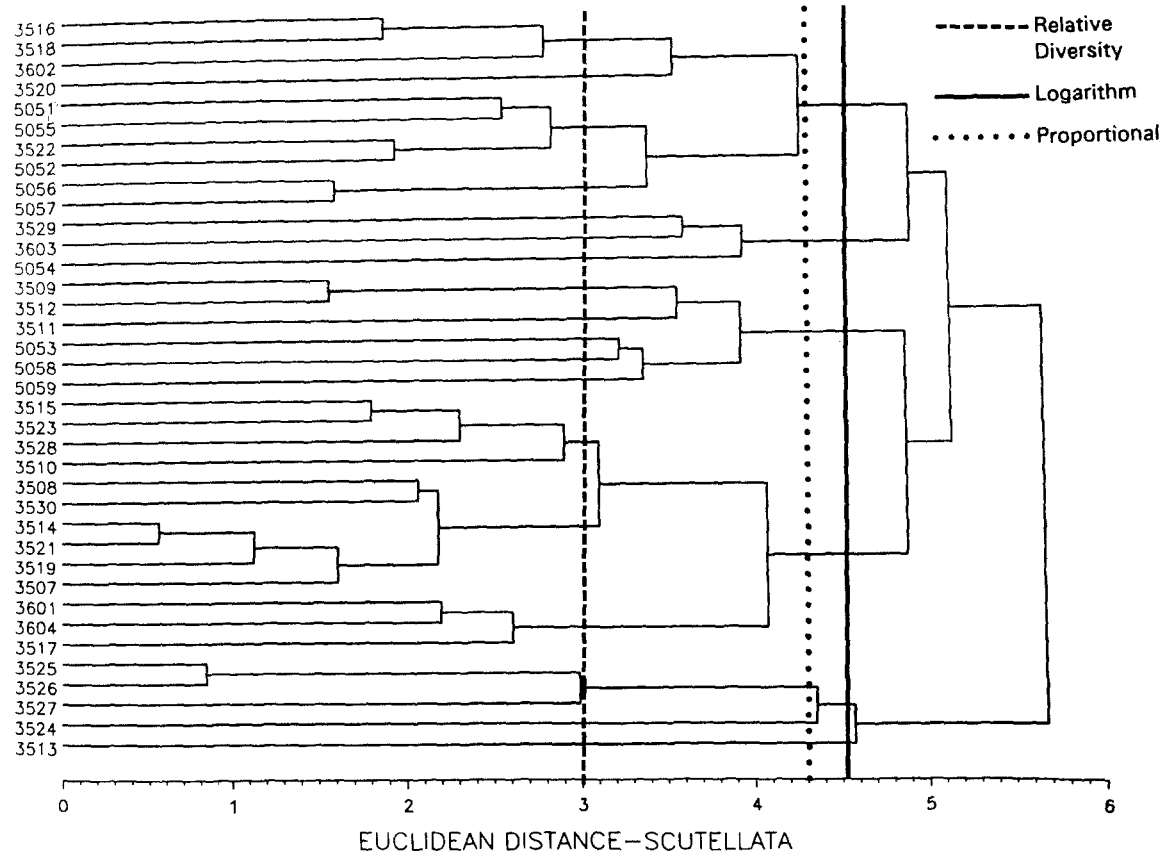

Fig. 2 Dendrogram of clusters of 37 *M. scutellata* accessions based on 14 morphological traits. The number of clusters in the dendrogram was determined by either the Relative Diversity, proportional or logarithm methods as indicated by the *lines*

core collections (Cores 9 and 11) (Table 2). Among the acceptable core collections (Cores 1–6), the proportional method produced the largest differences in ranges (29%) between the core collection (Core 4) and the initial subset. The proportional method generally biases the core collection in favor of the large groups in which the level of redundancy is likely to be high (Brown 1989b). In this study, the proportional method probably underrepresented the variability of small groups such as *M. balancheana*, *M. granadensis* Willd., *M. intertexta* (L.) Miller etc. These *Medicago* species were quite variable, and a relatively large percentage of their accessions was selected for the Relative Diversity core collections (Table 1). These species with a small number of accessions were also represented by a relatively large number of accessions using the loga-rithm method, which tends to balance the number of accessions selected within groups.

The Relative Diversity method generated good but not the best core collections (Table 2), probably due to a random sampling effect with small sample size. The annual *Medicago* core collections contained about 210 accessions, which is 17% of the initial subset but only 7% of the entire germplasm collection. Brown (1989b) showed that as the core collection sample size decreases to 10% the percentage of variability retained by the core collection decreases slowly. After 10%, the proportion of the original variability retained by the core collection decreases rapidly. Therefore, differences among sampling methods for large core collections should be smaller than those for small cores. For the annual *Medicago* species, when the core collection size was 10% (Table 3) or larger (Table 2), there were few differences among the three selection methods. However, for the 5% sample size, 36% of the ranges for the proportional method were reduced more than 30% while only 14% of the logarithm methods and none of the Relative Diversity method ranges were reduced, indicating better retention of the ranges with the Relative Diversity and logarithm methods than with the proportional method. However, all 5% and 10% core collections could not be acceptable, since the 5% and 10% core collections retained only 62% or less of the initial subset ranges, as indicated by the range ratio (Table 3). Core collections with a small sample size were not representative of the initial subset, probably because the annual *Medicago* species germplasm collection contains many species with very few accessions.

Core collections only contain the diversity present in the germplasm collection, and the material held in these collections is often unrepresentative of the total diversity of a species. For many crop species a large number of similar accessions are held in genebanks, and many of these accessions may be related (Hodgkin 1990). Recently, Schoen and Brown (1993) studied the conservation

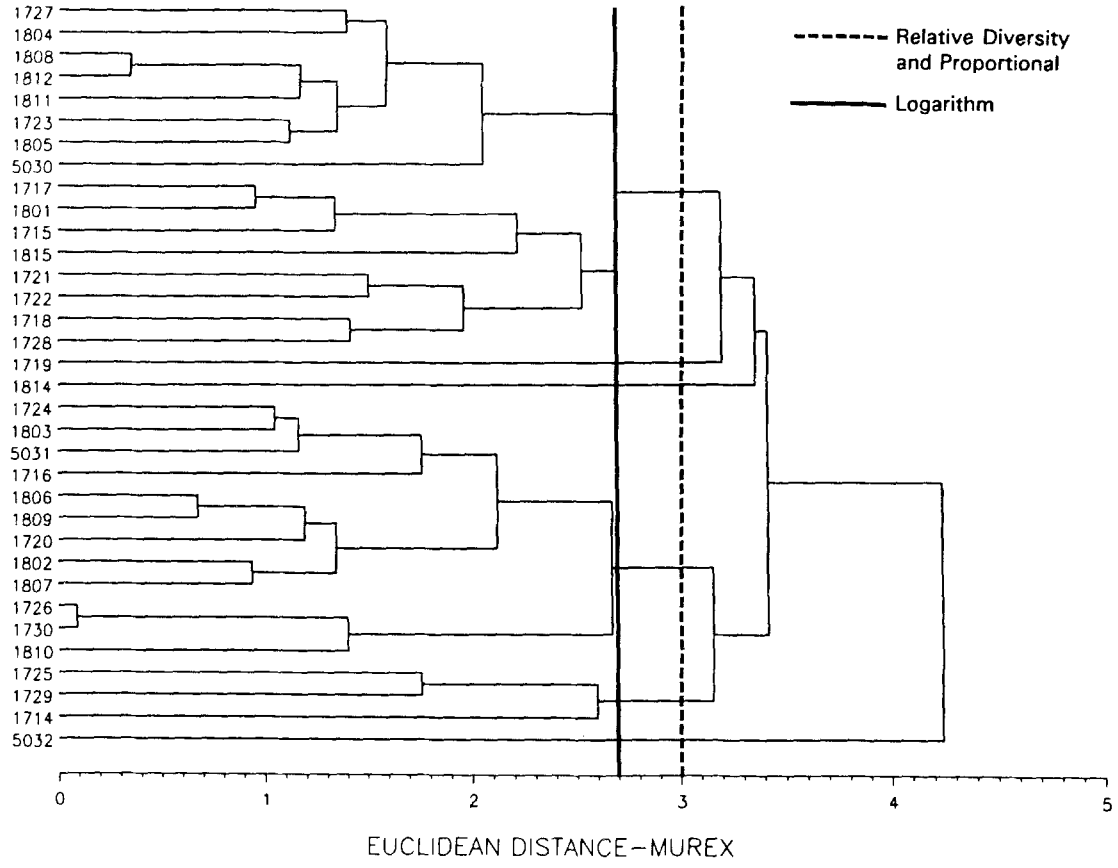

Fig. 3 Dendrogram of clusters of 34 *M. murex* accessions based on 14 morphological traits. The number of clusters in the dendrogram was determined by either the Relative Diversity, proportional or logarithm methods as indicated by the *lines*

of allelic richness in wild crop relatives and indicated that maximizing allelic richness at marker loci in inbreeders (such as the annual *Medicago* species) can lead to increased allelic richness at other loci. Therefore, when information about the range of genetic or phenotypic diversity of groups in the collection is available, the Relative Diversity method should be the method of choice for the assemblage of core collections. However, if evaluation data is not available and the selection of core collections is based solely on passport data, the logarithm method, which selects relatively few accessions from large groups, should be preferred over the proportional method, which may over-represent large uniform groups and under-represent small diverse groups.

Assemblage of core collections by random sampling using either all or part of the available passport data is relatively simple and rapid, while the utilization of germplasm collections evaluation data requires field work and complex statistical analysis. Nevertheless, this study has shown that for the annual *Medicago* species germ plasm collection, core collections assembled on the basis of evaluation data and cluster analysis of accessions are more representative core collections than those assembled on the basis of passport data and through the random selection of accessions, or by cluster analysis of geographical regions. Thus, the use of all available information (evaluation and passport data) was found to be very valuable for the assemblage of the annual *Medicago* species core collection.

References

Brown AHD (1989a) The case for core collections. In: Brown AHD, Frankel OH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 136–156

Brown AHD (1989b) Core collection: a practical approach to genetic resources management. Genome 31:818–824

Brown AHD, Grace JP, Speer SS (1987) Designation of a core collection of perennial *Glycine*. Soybean Genet Newsl 14: 59–17

Diwan N, Bauchan GR, McIntosh MS (1994) A core collection for the United States annual *Medicago* germplasm collection. Crop Sci 34:279–285

Frankel OH (1984) Genetic perspective of germplasm conservation. In: Arber WK, Llimensee K, Peacock WJ, Stalinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170

Hamon S, van Sloten DH (1989) Characterization and evaluation of okra. In: Brown AD, Frankel OH, Marshall DR, Williams J (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 173–196

Hodgkin T (1990) The core collection concept. In: van Hintum TJL, Frese L, Perret PM (eds) Crop networks: searching for new concepts for colaborative genetic resources management. International Board for Plant Genetic Resources, Rome, Italy, pp 43–48

Holbrook CC, Anderson WF, Pittman RN (1993) Selection of a core collection from the U.S. germplasm collection of peanuts. Crop Sci 33:859–861

Holden JHW (1984) The second ten years. In: Holden JHW, Williams J (eds) Crop genetic resources: conservation and evaluation. George Allen and Unwin, London, UK, pp 277–285

Jacobs D (1990) SAS/Graph software and numerical taxonomy. In: Proc of 15th Annu SAS Users Group. International Conference, SAS Institute, Cary, N.C. pp 1413–1418

Mackay MC (1986) Ultilizing wheat genetic resources in Australia. In: McLean R (ed) Proc 5th Assembly Wheat Breed Soc Australia. Perth/Merredin, Australia, pp 56–61

Mackay MC (1989) Strategic planning for effective evaluation of plant germplasm. In: Strivastava JP, Damania AB (eds) Wheat Genetic Resources: Meeting Diverse Needs. John Wiley & Sons Ltd. Chichester, England. pp 21–25

SAS Institute (1989) SAS/STAT user's guide version 6, 4th edn. SAS Institute Inc, Cary, N.C.

Schoen DJ, Brown AHD (1994) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 90:10623–10627

US Department of Agriculture (1989) Germplasm resources information network user's manual. Beltsville Agriculture Research Center, Beltsville, M.D.

Williams TJ (1991) Plant genetic resources: some new directions. Advan Agron 45:61–91