CBCS
http://agrobiol.sggw.waw.pl/cbcs

Communications in Biometry and Crop Science
Vol. 5, No. 2, 2010, pp. 127–137

International Journal of the Faculty of Agriculture and Biology,
Warsaw University of Life Sciences, Poland

REGULAR ARTICLE

The efficiency and effectiveness of sampling strategies used to develop a core collection for the Polish spring triticale (×*Triticosecale* Wittm.) germplasm resources

Marcin Studnicki[1]*, Wiesław Mądry[1], Wanda Kociuba[2]

[1]Department of Experimental Design and Bioinformatics, Warsaw University of Life Sciences – SGGW, Poland.
[2]Institute of Plant Genetics, Breeding and Biotechnology, University of Life Sciences, Lublin, Poland.
* Corresponding author: Marcin Studnicki E-mail: marcin_studnicki@sggw.pl

CITATION: Studnicki, M., Mądry, W., Kociuba, W. (2010). The efficiency and effectiveness of sampling strategies used to develop a core collection for the Polish spring triticale (×*Triticosecale* Wittm.) germplasm resources. *Communications in Biometry and Crop Science* 5 (2), 127–137.

Received: 8 August 2010, Accepted: 25 November 2010, Published online: 30 December 2010

ABSTRACT

Triticale (×*Triticosecale* Wittm.) is a hybrid of wheat and rye, which is bred using conventional plant breeding methods. A core collection is defined as a representative sample of the entire plant genetic resources collection that reflects the diversity in the entire collection. A core collection may simplify management and improve the utilisation of the considered germplasm resources. This paper describes and evaluates the efficiency (in sample representativeness sense) of 50 sampling strategies used to establish core collections of Polish spring triticale germplasm resources. Five fractions of core collections (10%, 15%, 20%, 25% and 30% of the entire collection), two clustering methods (Ward's and UPGMA) and five sample allocation methods based on agro-morphological quantitative traits were compared for their effectiveness using two indices. The first index refers to the average of absolute differences between means across all of the traits in the core and entire collections relative to the means in the entire collection, MD%. The other index is the average of the absolute differences between variances across all of the traits in the core and entire collections relative to the variances in the entire collection, VD%. The results showed that for the studied spring triticale germplasm collection 1) the core collection including at least 20% of the entire collection should be sufficient to provide good representativeness, 2) two of the five sample allocation methods (the proportional and the D_2) were characterized by highest level of sample strategies effectiveness, 3) Ward's method of cluster analysis enabled us to stratify the entire collection in a way that draws more representative core collections than those created using the UPGMA method.

Key Words: *core collection; genetic resources; phenotypic diversity; sampling strategies; spring triticale.*

INTRODUCTION

Triticale (×*Triticosecale* Wittm.) is a grain crop developed from a hybrid of wheat and rye, which is produced using conventional plant breeding methods. The first grain of this hybrid was bred in 1875. Triticale has the grain quality characteristics of wheat and is able to withstand difficult soils, tolerate drought, withstand cold and resist disease, and has low-input requirements that are common to rye (Mergoum and Gómez-Macpherson, 2004). Triticale is an important animal feed in Central and Eastern Europe that is commonly used for feeding pigs, but it can be, and is, fed to poultry and ruminants. In 2007, according to the Food and Agriculture Organisation, just over 3.6 million ha were grown in 32 countries across the world. Poland is the world's largest producer of triticale, with 1.26 million ha of harvest area producing 4.1 million metric tons of grain yields. The breeding of triticale in Poland began in the 1960s and the first cultivars were registered in 1982. Triticale has also been considered a potential crop for human nutrition and as an energy crop for bioethanol production.

Plant genetic resources will be the main contributing factor to future progress in developing new cultivars (Upadhyaya et al., 2007). Collections of plant genetic resources in gene banks are often so large that their size interferes with achieving the main goals for which the collections have been established, namely, the conservation and utilisation of the genetic diversity of a crop species and its relatives through accessions. These collections can include wild populations, traditional landraces, modern cultivars, elite (modern) forms that have contributed to the more recent progress in selective breeding and the registration of improved cultivars, genetic stocks or other research materials. Genetic collections are currently facing problems caused by the large size of collections, and the resultant costs of their maintenance (Franco et al., 2006; Jansen and van Hintum, 2007). The size of the germplasm collections often has hindered their evaluation and utilisation for specific breeding purposes. To solve these problems, Frankel (1984) proposed the establishment of a core collection that could be created from the existing collection of crop species resources in a gene bank.

A core collection (called also a "core subset") derived from an existing entire collection (a gene bank) within a crop species comprises a chosen set of accessions that represents the genetic and phenotypic variation available in the collection with minimal duplication (Frankel and Brown, 1984; Brown, 1989, 1995; van Hintum et al., 2000; Franco et al., 2006). Then, the core collection consists of a limited number of the accessions from the existing collection that represent the diversity (or spectrum) in the entire collection. Representativeness is the most important property for a core collection. This term is defined as similarity of the genotypic or/and phenotypic diversity in a core collection with the respective diversity in the entire collection.

A core collection provides a convenient way to study and utilize germplasm resources, and this method has been receiving extensive attention all over the world. The purpose of forming core collections is generally to make easier and more effective evaluations of the phenotypic and genetic diversity from the total genetic resources and for maintaining, managing and utilization of these resources in crop breeding programs. Core collections also allow for more effective examinations of the allelic variation in genes that are of interest and for the assessment of genotype-phenotype associations.

Several statistical methods, referred to as sampling strategies or sampling methods, have been introduced for the selection of accessions from an existing genetic resources collection to form core collections that are as representative as possible (Marita et al., 2000; van Hintum et al., 2000; Upadhyaya et al., 2007; Wang et al., 2007). These methods include simple random sampling (Brown, 1989; Charmet and Balfourier, 1995) and stratified random sampling (Spagnoletti Zeuli and Qualset, 1993; Franco et al., 2005, 2006; Xu et al., 2006; Wang et al. , 2007). In stratified random sampling, the following main steps to establish a core collection are: (1) determine the size of the core; (2) stratify or cluster the entire collection in distinct groups using miscellaneous criteria; (3) calculate a fraction of selected entries in each group using a sample allocation method and (4) select at random or not-random (stepwise clustering) entries in each group to form the core (Diwan et al., 1995; van Hintum et al., 2000; Wang et al., 2007). There is limited information on thorough evaluation and comparison of the efficiency among many sampling strategies established as combinations of the above mentioned steps 1-4. Also, newly sample allocation methods, suggested by Franco et al. (2005, 2006) and being a modification of proportional and logarithmic strategies as including the mean squared Euclidean distance in any stratified group, have not been yet thoroughly tested.

The objective of this research was to evaluate the efficiency of 50 sampling strategies used for forming core collections from the Polish spring triticale genetic resources based on agro-morphological quantitative traits. The sampling strategies are combined as schemes (combinations) for five sample fractions of core collections (sampling intensity), two clustering methods and five sample allocation methods, and they are subsequently analyzed for their ability to create reliable sub-samples. Selection of entries in each group to form the core was done at random.

MATERIALS AND METHODS

PLANT MATERIAL AND DATA

The entire collection of spring triticale genetic resources consisted of 133 accessions (cultivars and advanced lines). The samples were derived and are held at the Institute of Plant Genetics, Breeding and Biotechnology at Lublin University of Agriculture, Poland. The accessions were evaluated in a field experiment at the Institute of Plant Genetics, Breeding and Biotechnology over a 6-year period (1996–2001). Within each year, the accessions were observed on a single-replicate plot. The 10 agro-morphological traits were recorded on 1 (low) to 9 (high resistance) point scale (lodging, susceptibility to leaf diseases, susceptibility to spikes Septoria), or quantitative (days to anthesis, plant height (cm), spike length, numbers of spikelet per spike, spikelet fertility, the number of grains per spike, 1000-grain weight, the weight of the grains per spike and the grain protein content). The data were arranged into incomplete two-way accession x year classifications for each trait.

STATISTICAL MODEL AND ANALYSIS OF VARIATION IN THE ENTIRE COLLECTION

To assess the phenotypic variability of the accessions in the entire collection, both univariate and multivariate methods were used. First, the observed values of each trait were expressed through the following mixed model (Hartung and Piepho, 2005; Piepho and Mohring, 2005; Upadhyaya et al., 2007):

$$y_{ij} = m + g_i + r_j + e_{ij},$$

where y_{ij} is the value of i-th accession in the j-th year, m is the population mean, g_i is the genotypic effect of the i-th accession, r_j is the effect of the j-th year, and e_{ij} is the residual effect including both the GE (accession × year) interaction effect and the experimental error. We assumed the year as a fixed factor and accessions as a random factor.

The estimates of the genotypic effects were obtained using the Best Linear Unbiased Predictor (BLUP) with the Residual Maximum Likelihood (REML) method. Estimates of the genotypic means were obtained using the formula:

$$\hat{m}_i = \hat{m} + \hat{g}_i^{BLUP},$$

where $\hat{m}_i$ is the estimate of genotypic means for the i-th accession, $\hat{m}$ is the estimate of population mean, and $\hat{g}_i^{BLUP}$ is the BLUP estimate of genotypic effect of the i-th accession.

Estimates of genotypic means for the studied traits were used in the calculation of the distances between the accessions in the cluster analysis and calculations of other statistical measures.

SAMPLING STRATEGIES USED FOR CONSTRUCTING THE CORE COLLECTIONS

Two categories of statistical methods, including the clustering methods and sample allocation methods, were used to construct the differently sized core collections. The fractions of the core collections were 10%, 15%, 20%, 25% and 30% of the entire collection. Two cluster analysis methods, e.g., UPGMA (unweighted pair group method with arithmetic mean) and Ward's method, were included in these studies (Williams, 1976; Jahufer et al., 1997). For each method of cluster analysis, the 10 traits were standardized, and a matrix of squared Euclidean distances was included. We used the R^2 (measures the heterogeneity of the cluster solution formed at a given step) criterion of 0.70 for defining the number of groups required in the analysis (Upadhyaya et al., 2003). Five sample allocation methods were considered, e.g., proportional (Pro), logarithmic (Log) – (Brown 1989) and three strategies based on the mean squared Euclidean distance (D_1, D_2, and D_3) – (Franco et al. 2005, 2006).

Brown (1989) proposed two sample allocation methods based on the group size, which are usually known as the proportional (Pro) and logarithmic (Log) strategies. The proportional strategy allocates n_t accessions from the t-th cluster (group) in proportion to the number of accessions in the cluster, N_t, which are calculated using the formula

$$n_t^{\mathrm{Pro}} = n \times \frac{N_t}{\sum_{t=1}^{g} N_t}$$

where n is the size (accession number) of core collections and g is the number of clusters obtained in cluster analysis. The logarithmic sampling strategy uses the proportion of the logarithm of the number of accessions in the cluster. The number of accessions allocated from the t-th cluster, n_t, is represented by

$$n_t^{Log} = n \times \frac{\log(N_t)}{\sum_{t=1}^{g} \log(N_t)}$$

Franco et al. (2005) proposed allocation methods for determining the number of accessions taken from a cluster based on the mean squared Euclidean distance between the accessions within the cluster. Groups that are more diverse will have a larger mean distance and, therefore, will have larger samples drawn from them. The D_1 sampling strategy used in this study determines that the size of the sample to be drawn from each cluster should be proportional to the mean squared Euclidean distance between the accessions within that cluster. The number of accessions, n_t, to be drawn from the t-th cluster is

$$n_t^{D_1} = n \times \frac{d_t}{\sum_{t=1}^{g} d_t}$$

where d_t is the mean squared Euclidean distance between the accessions within the t-th cluster. In the D_2 sampling strategy the number of accessions, n_t, to be drawn from the t-th cluster is calculated using a formula including the size of the t-th cluster, N_t, as weighted by the diversity measured as the squared Euclidean distance, d_t, obtaining

$$n_t^{D_2} = n \times \frac{N_t \times d_t}{\sum_{t=1}^{g} N_t \times d_t}$$

The D_3 sampling procedure allocates the number of entries per cluster into the logarithm of the number of accessions in the t-th group (N_t) and is weighted by the diversity measured as the mean squared Euclidean distance (d_t)

$$n_t^{D_3} = n \times \frac{\log(N_t) \times d_t}{\sum_{t=1}^{g} \log(N_t) \times d_t}$$

Combinations of the three statistical approaches described earlier established the 50 sampling procedures (5 core fractions × 2 clustering methods × 5 sample allocation methods) that were assessed in this study. Each sampling procedure was used five times, establishing five core collections assumed to be the 'replicates' in the experiment.

THE INDICES FOR EVALUATING REPRESENTATIVENESS OF CORE COLLECTION

For the five developed (replicated) core collections within each of 50 sampling strategies, two indices of validities (goodness or quality in a sense of representativeness) were used (Kim et al., 2007). The first index refers to the average of absolute differences between means across all of the traits in the core and entire collections relative to the means in the entire collection, MD%. The other index is the average of the absolute differences between variances across all of the traits in the core and entire collections relative to the variances in the entire collection, VD%. The goodness indices were calculated according to the formulas (Kim et al., 2007):

$$\mathrm{MD\%} = \frac{\sum_{\tau=1}^{p} \frac{|\bar{x}_{C\tau} - \bar{x}_{E\tau}|}{\bar{x}_{E\tau}}}{p} \times 100\%$$

$$\mathrm{VD\%} = \frac{\sum_{\tau=1}^{p} \frac{|\hat{\sigma}^2_{C\tau} - \hat{\sigma}^2_{E\tau}|}{\hat{\sigma}^2_{E\tau}}}{p} \times 100\%$$

where $\bar{x}_{C\tau}$ is the mean of the τ-th trait ($\tau = 1,2,\ldots,p$) for a core collection, $\bar{x}_{E\tau}$ is the mean of the τ-th trait for the entire collection, $\hat{\sigma}^2_{C\tau}$ is the variance of the τ-th trait for a core collection, $\hat{\sigma}^2_{E\tau}$ is the variance of the τ-th trait for the entire collection.

Smaller values of MD% and VD% for the sampling strategy indicate a more effective strategy, e.g., smaller values show a better ability of the sampling strategy to establish a representative core collection. The calculated values of MD% and VD% for all of the 50 procedures and the 'replicate' generated data were used in a three-way cross-classification with 5 'replicates'. The data were analyzed by three-factor ANOVA to check significance of the effects of the three statistical approaches on both the MD% and the VD% indices. Also Tukey's multiple mean comparison procedure was used to test significance of means differences. In this way it is possible to find optimal size of the core collection, the clustering method and the sampling strategy. There are also opportunities to compare all of the 50 sampling procedures with regard to both indices and to recommend some of the procedures that showed the best efficiency (effectiveness and validity), that is the procedures having the highest relative ability to establish a representative core collection.

STATISTICAL ANALYSIS AND COMPUTATIONS

The statistical analyses were carried out using SAS (SAS Institute 2004). The MIXED procedure was used to estimate the Best Linear Unbiased Predictor (BLUP) for random effects. The squared Euclidean distance between the accessions was calculated using the DISTANCE procedure. The CLUSTER procedure was used to apply the two cluster analysis methods.

RESULTS

The 133 accessions were grouped into 6 clusters using the two cluster analysis methods. The number of accessions per cluster under Ward's method varied from 44 accessions in cluster II to 7 accessions in cluster VI. The number of accessions in the individual clusters under the UPGMA method ranged from 1 in cluster V and VI to 102 in cluster I.

The mean values (across 5 'replicates') of the MD% in a set of 50 sampling procedures ranged from 0.55 to 3.07%, while the respective means for the VD% ranged from 12.44 to 45.69 % (Table 1). These results suggest that discrepancies between trait means in the entire collection and the core collections that developed using the suggested procedures were generally much smaller compared to the respective discrepancies between trait variances.

The most effective sampling strategies were those for which both the MD% and the VD% indices were relatively small. These procedures proved to be the ones that used Ward's clustering method and the proportional or the D2 sample allocation method (Ward's and Pro, and Ward's and D_2) as well the UPGMA clustering method and the proportional allocation method (UPGMA and Pro). These strategies enabled us to draw core collections of the spring triticale germplasm collection belonging that were the most representative at each studied sample fraction, i.e., reflecting and maintaining the majority of the phenotypic diversity existing in the entire collection.

Table 1. The mean values of the MD% and the VD% across the 5 'replicates' for the 50 sampling strategies used to develop a core collection of the Polish spring triticale germplasm resources.

Cluster method	Sample allocation method	Sample fraction (%)									
		10		15		20		25		30	
		MD%	VD%	MD%	VD%	MD%	VD%	MD%	VD%	MD%	VD%
Ward	Pro	0.94	26.44	0.71	21.51	0.75	15.66	0.67	15.65	0.64	12.68
	Log	1.19	30.94	1.11	27.77	1.13	26.32	1.18	27.26	0.91	22.43
	D_1	1.88	45.69	1.93	33.37	1.65	30.75	2.00	36.12	1.85	34.58
	D_2	0.90	26.48	0.71	21.51	0.74	17.72	0.73	15.98	0.66	12.44
	D_3	1.88	45.69	1.11	27.77	1.25	26.03	1.24	24.34	0.99	22.41
UPGMA	Pro	1.20	27.69	0.92	20.54	0.78	16.85	0.84	15.55	0.55	12.85
	Log	2.73	36.50	2.54	34.52	2.63	30.95	2.60	30.30	2.48	28.27
	D_1	2.73	36.50	2.86	41.73	3.07	34.65	2.99	33.65	2.86	28.90
	D_2	1.29	29.35	1.28	26.47	0.91	19.68	0.97	20.34	0.85	19.09
	D_3	2.73	36.50	2.67	39.29	3.07	34.65	2.84	32.17	2.69	27.93

The ANOVA results show that each of the statistical methods and the sample fraction modified the representativeness of the core collections that were established as measured in terms of the MD% and the VD% indices (Table 2). The interactions between the cluster methods and the sample allocation methods were also significant for both indices. This indicates that the MD% and VD% were affected by the sample allocation methods in different ways, depending on the cluster methods used. The other interactions were not significant for the both sample representativeness indices, showing consistent relationships between the indices and the cluster methods for all the sample fractions and also between the indices and the sample allocation methods for all the sample fractions.

Table 2. Means squares (MS) and *p*-values (*P*) from analysis of variance for MD% and the VD% obtained using the 50 sampling strategies that were used to develop a core collection of the Polish spring triticale germplasm resources.

		MD%		VD%	
Source of variation	df	MS	*P*	MS	*P*
Sample fraction (A)	4	0.57	0.0165	1086.86	<.0001
Cluster method (B)	1	49.94	<.0001	453.83	0.0011
Allocation method (C)	4	25.06	<.0001	2631.51	<.0001
A × B	4	0.08	0.7906	104.18	0.0437
A × C	16	0.11	0.8869	21.06	0.9424
B × C	4	5.30	<.0001	101.29	0.0487
A × B × C	16	0.13	0.7601	50.90	0.2532
Error	200	0.18		41.64	

Additionally, each of the two-way interactions (or the lack of this interaction) between the two statistical methods is repeatable across all of the considered sample fractions.

The resulting comparison of the mean values using the two indices for the five different sample fractions (Figure 1) suggests that when the number of accessions in the core collection increases, the representations of the diversity of the entire collection also increases. The VD% for the 20% sample fraction was significantly smaller than that of the 10% and 15% sample fraction. There were no significant differences between the procedures using 20% of the accessions and the 25% or the 30% samples according to Tukey's procedure.

The core collections in which Ward's cluster analysis method was used to create the sample had significantly lower values for the MD% and the VD% indices than the sampling methods that used the UPGMA methods (Figure 2). Using Ward's method was more effective in the reduction of the MD% than the UPGMA method as compared to the VD%. This suggests that the consistency of the trait means in the core and in the entire collection was improved more by Ward's method in relation to UPGMA when the consistency of the trait variances in both collections was examined.

The logarithmic strategy favors small groups, compared with the proportional strategy where a higher percentage of entries will be selected from small groups compared to larger groups (Brown et al., 1987; Li et al., 2005). Using the logarithmic, D_1 and the D_3 sample allocation methods to develop a core collection was least effective. This sample allocation method has the highest values for the MD% and the VD% indices (Figure 3). In the present analysis, two of the five sample allocation methods (the proportional and the D_2) were characterized by a high level of effectiveness, and these two methods also had relatively smaller values of the indices of validities – MD% and VD%.

The ANOVA shows that the interactions between the cluster methods and the sample allocation methods for both of the goodness indices were also significant. Figure 4 presents this interaction plot for the MD% and the VD% indices. The proportional (Pro) and the D_2 sample allocation methods proved to be equivalently the most effective, when using Ward's or UPGMA cluster methods, in developing core collections of any size within the studied range for the spring triticale germplasm collection. When using the UPGMA method of cluster analysis the effectiveness of other sample allocation methods is more affected compared to using Ward's method at each sample fraction.

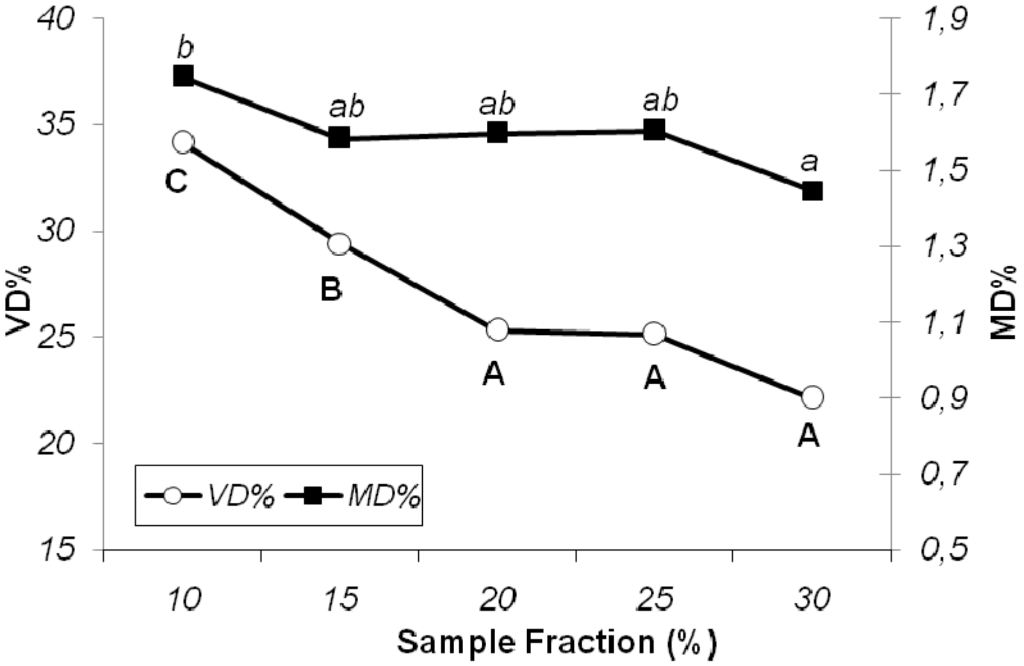

Figure 1. Comparison of mean values of MD% and VD% for five sampling sizes. Data points with different letters indicate that a significant difference between sampling size, based on the Tukey's procedure at 0.05 probability level. Lowercase letters give the results of the Tukey's procedure for MD% and capital letter for VD%,

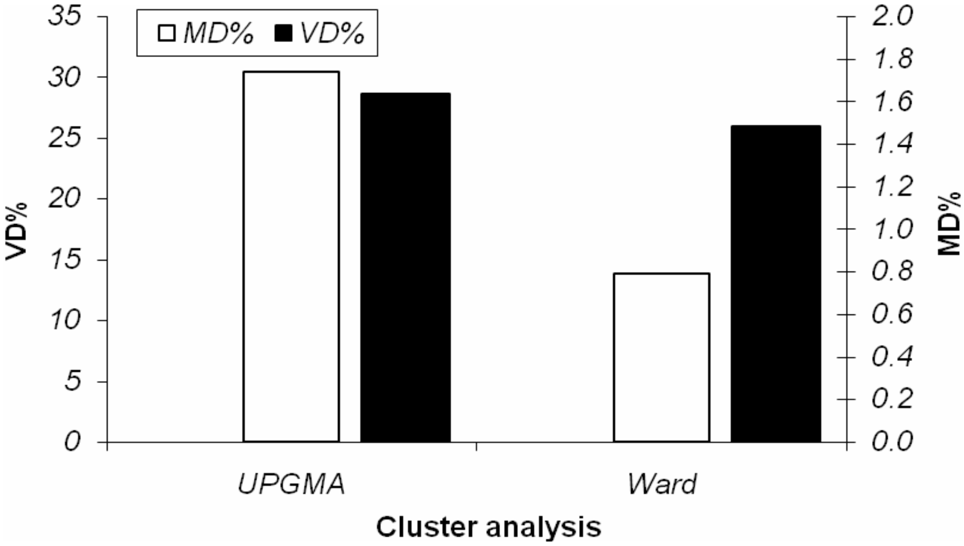

Figure 2. Comparison of the means value of MD% and VD% for two cluster analysis methods (UPGMA, Ward).

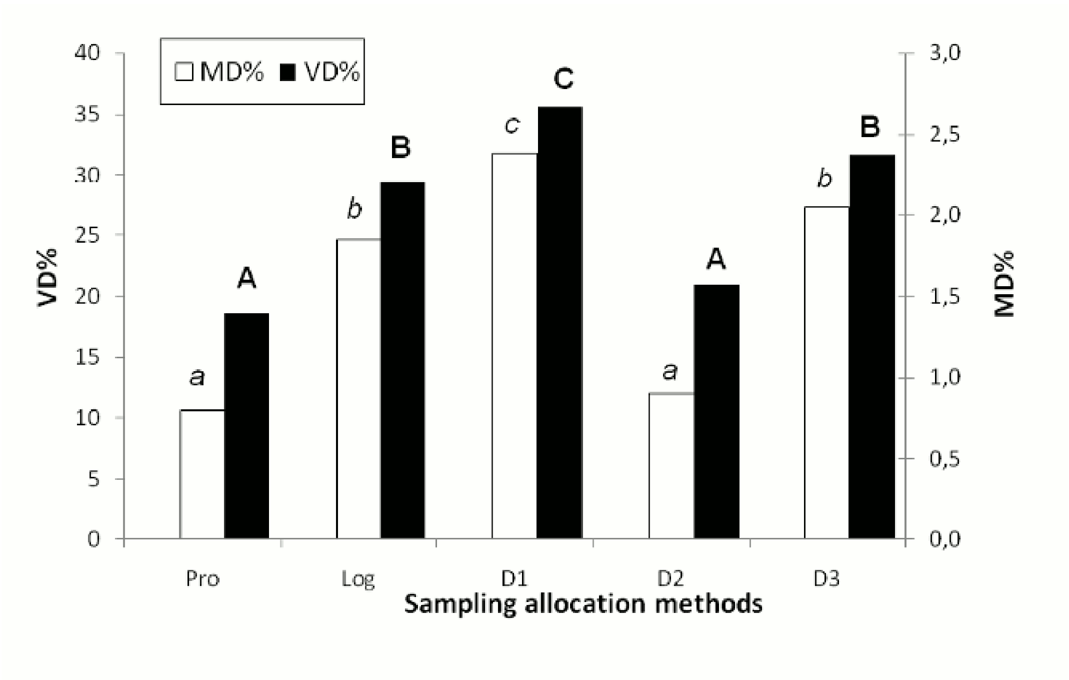

Figure 3. The comparison of MD% and VD% mean values for five sample allocation methods. Different letters indicate a significant difference between sample allocation methods, based on the Tukey's procedure at 0.05 probability level. Lowercase letter gives the results of the Tukey's procedure for MD%, and capital letter for VD%.

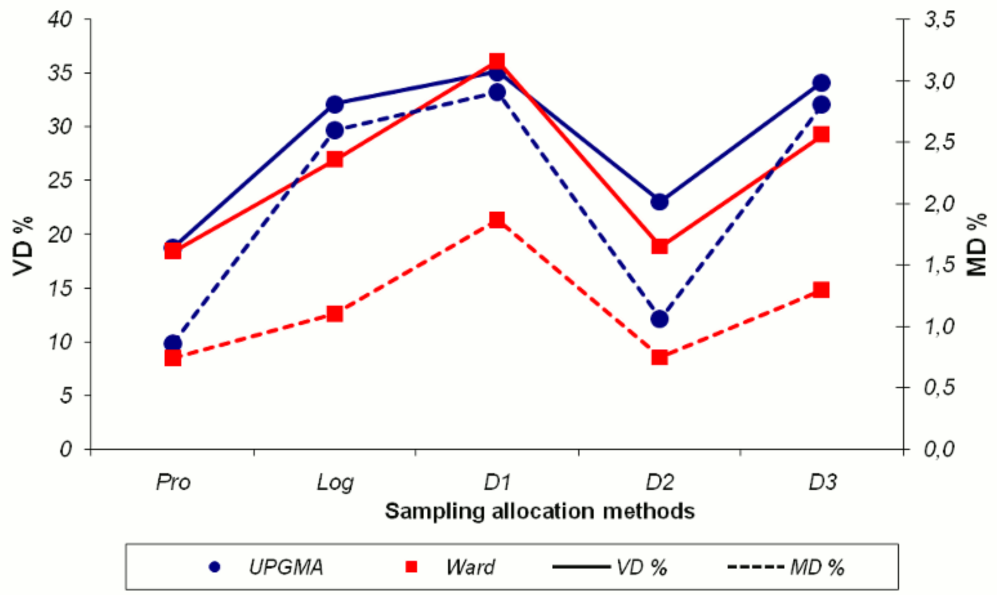

Figure 4. Interaction plot of the sample allocation methods and cluster analysis methods for MD% and VD%.

CONCLUSIONS

Increasing sample fraction improved representativeness of phenotypic diversity of the core collections for spring triticale consistently for both cluster methods and sampling strategies. Although the representativeness of the core collection was affected by its size and was more sensitive as measured by VD% than by MD%, the core collection including at least 20% of the entire collection should be sufficient to provide representative information.

Ward's method of cluster analysis enables the sample to stratify the entire spring triticale collection in such a way that draws more representative core collections than when using the UPGMA method. Additionally, when using the UPGMA method the effectiveness of the Pro, Log, D_2 and D_3 sample allocation methods is more affected as compared to using Ward's method at each sample fraction.

Two sample allocation methods, proportional (Pro) and D_2, proved to be equivalently most effective, when using Ward's or the UPGMA cluster methods for developing as much representative as possible core collections of any size within the studied range for the spring triticale entire collection.

References

Brown, A.H.D. (1989). Core collections: a practical approach to genetic resources management. *Genome* 31, 818–824.

Brown, A.H.D., Grace, J.P., Speer, S.S. (1987). Designation of a "core" collection of perennial Glycine. *Soybean Genetics Newsletter*14, 59–70.

Brown, A.H.D. 1995. The core collection at the crossroads. In: Hodgkin, T., Brown, A.H.D., van Hintum, Th.J.L., Morales, E.A.V., (Eds.). *Core Collections of Plant Genetic Resources*. John Wiley and Sons, UK, 3-19.

Charmet, G., Balfourier, F. (1995). The use of geostatistics for sampling a core collection of perennial ryegrass populations. *Genetic Resources and Crop Evolution* 42, 303–309.

Diwan, N., Mclntosh, M.S., Bauchan, G.R. (1995). Methods of developing a core collection of annual *Medicago* species. *Theoretical and Applied Genetics* 90, 755–761.

Franco, J., Crossa, J., Taba, S., Shands, H. (2005). A sampling strategy for conserving genetic diversity when forming core subsets. *Crop Science* 45, 1035–1044.

Franco, J., Crossa, J., Warburton, M.L., Taba, S. (2006) Sampling strategies for conserving maize diversity when forming core subsets using genetic markers. *Crop Science* 46, 854–864.

Frankel, O.H. (1984) Genetic perspectives of germplasm conservation. In: Arber W., Llimensee K., Peacock W.J., Starlinger P. (Eds.). *Genetic manipulation: impact on man and society*. Cambridge University Press, Cambridge, 161–170.

Frankel, O.H., Brown, A.H.D. (1984). Plant genetic resources today: A critical appraisal. In: Holden, J.H.W., Williams, J.T., (Eds.). *Crop Genetic Resources: Conservation and Evaluation*. Allen and Unwin, London, 249-257.

Hartung, K., Piepho, H.P. (2005). A threshold model for multiyear genebank data based on different rating scales. *Crop Science* 45, 1045–1051.

Jahufer, M.Z.Z., Cooper, M., Harch, B.D. (1997). Pattern analysis of the diversity of morphological plant attributes and yield in a world collection of white clover (*Trifolium repens* L.) germplasm characterised in a summer moisture stress environment of Australia. *Genetic Resources and Crop Evolution* 44, 289–300.

Jansen, J., van Hintum, T. (2007). Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. *Theoretical and Applied Genetics* 114, 421–428.

Kim, K.W., Chung, H.K., Cho, G.T., Ma, K.H., Chandrabalan, D., Gwag, J.G., Kim, T.S., Cho, E.G., Park, Y.J. (2007). PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23, 2155–2162.

Li, Y., Shi, Y., Cao, Y., Wang, T. (2005). Establishment of a core collection for maize germplasm preserved in Chinese National Genebank using geographic distribution and characterization data. *Genetic Resources and Crop Evolution* 51, 845–852

Marita, J., Rodriguez, J.M., Nienhuis, J. (2000). Development of an algorithm identifying maximally diverse core collections. *Genetic Resources and Crop Evolution* 47, 515–526.

Mergoum, M., Gómez-Macpherson, H. (2004). *Triticale improvement and production*. Food and Agriculture Organization (FAO), Rome.

Piepho, H.P., Mohring, J. (2005). Best Linear Unbiased Prediction of cultivar effects for subdivided target regions. *Crop Science* 45, 1151–1159

SAS Institute Inc. (2004). SAS OnlineDoc® 9.1.3. Cary, NC URL http://support.sas.com/onlinedoc/913; verified 16 November 2010

Spagnoletti Zeuli, P.L., Qualset, C.O. (1993). Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. *Theoretical and Applied Genetics* 87, 295–304

Upadhyaya, H., Dwivedi, S., Gowda, C., Singh, S. (2007). Identification of diverse germplasm lines for agronomic traits in a chickpea (*Cicer arietinum* L.) core collection for use in crop improvement. *Field Crops Research* 100, 320–326

Upadhyaya, H., Ortiz, R., Bramel, P.J., Singh, S. (2003). Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. *Genetic Resources and Crop Evolution* 50, 139–148

van Hintum, T., Brown, A., Spillane, C., Hodgkin, T. (2000). *Core collections of plant genetic resources*. *PGRI Technical Bulletin No. 3*. International Plant Genetic Resources Institute, Rome

Wang, J.C., Hu, J., Xu, H.M., Zhang, S. (2007) A strategy on constructing core collections by least distance stepwise sampling. *Theoretical and Applied Genetics* 115, 1–8

Williams, W.T. (1976). *Pattern analysis in agricultural science*. Elsevier, New York

Xu, H., Mei, H., Hu, J., Zhu, J., Gong, P. (2006). Sampling a core collection of Island cotton (*Gossypium barbadense* L.) based on the genotypic values of fiber traits. *Genetic Resources and Crop Evolution* 53, 515–521