ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization; 1–9
doi:10.1017/S1479262115000556

Integrating principal component score strategy with power core method for development of core collection in Indian soybean germplasm

C. Gireesh*†, S. M. Husain, M. Shivakumar, G. K. Satpute, Giriraj Kumawat, Mamta Arya, D. K. Agarwal and V. S. Bhatia

ICAR-Directorate of Soybean Research, Indore, Madhya Pradesh, India

Received 18 July 2015; Revised 2 October 2015; Accepted 13 October 2015

Abstract

Soybean is a leading oilseed crop in India, which contains about 40% of protein and 20% of oil. Core collection will accelerate the management and utilization of soybean genetic resources in breeding programmes. In the present study, eight agromorphological traits of 3443 soybean germplasm were analysed for the development of core collection using the principal component score (PCS) strategy and the power core method. The PCS strategy yielded core collection (CC1) of 576 accessions, which accounted for 16.72% of the entire collection (EC). The analysis based on the power core programme resulted in CC2 of 402 accessions, which accounted for 11.67% of the EC. Statistical analysis showed similar trends for the mean and range estimated in both core collections and EC. In addition, the variance, standard deviation and coefficient of variance were in general higher in core collections than in the EC. The correlations observed in the EC in general were preserved in core collections. A total of 311 and 137 unique accessions were found in CC1 and CC2 in addition to 265 accessions that were found to be common in both core collections. These 265 common accessions were the most diverse core sets, which accounted for 7.64% of the EC. We proposed to constitute an integrated core collection (ICC) by integrating both common and unique accessions. The ICC comprised 713 accessions, which accounted for about 20.62% of the EC. Statistical analysis indicated that the ICC captured maximum variation than CC1 and CC2. Therefore, the ICC can be extensively evaluated for a large number of economically important traits for the identification of desirable genotypes and for the development of mini core collection in soybean.

Keywords: entire collection; germplasm; principal component analysis; variability

Introduction

Soybean [*Glycine max* (L.) Merrill] is a 'miracle crop' that contains about 40% of protein and 20% of oil. It is a major source of edible oil and protein in the world, and cultivated widely in the USA, Brazil, China, India and Argentina. Remarkable progress made in plant genetic resource management in recent days has resulted in the collection of a huge set of plant germplasm that hinders the very purpose for which they exist (Odong *et al.*, 2013). Frankel (1984) proposed the concept of core collection that could be established from an existing collection for better management and utilization of plant genetic resources. Core collection can be defined as a minimum

* Corresponding author. E-mail: giri09@gmail.com
† Present address: ICAR-Indian Institute of Rice Research, Hyderabad, India.

set of accessions representing maximum genetic diversity of the whole collection with increased diversity and reduced redundancy among core lines. Core collection includes cultivars, breeder lines, landraces and wild species. However, core collection is no substitute for whole collection (van Hintum *et al.*, 2000). Recently, core collection has become a powerful tool for evaluation of germplasm, identification of trait-specific accessions, gene discovery through association mapping, allele mining, genomic study, marker development and molecular breeding (Qiu *et al.*, 2013).

Drawing representative samples from whole collection for the constitution of core collection is the heart of core collection that determines its quality. Brown (1989a) initially developed several methods of core collection, which includes random sampling strategy of 10% of base collection that represents more than 70% of genetic variation, and suggested that core collection should be optimally 10 to 20% of the entire collection (EC) (Brown, 1989b). Noirot *et al.* (1996) developed the principal component score (PCS) strategy that employs principal component analysis (PCA) to eliminate collinearity between variables and selects individuals based on their cumulative relative contribution. The PCS strategy has been successfully used in the establishment of core collection in coffee (Hamon *et al.*, 1995), mungbean (Bisht *et al.*, 1998), groundnut (Upadhyaya *et al.*, 2003), ragi (Upadhyaya *et al.*, 2006) and sesamum (Mahajan *et al.*, 2007). Another approach known as the power core method of core collection was developed by Kim *et al.* (2007), which utilizes the advanced M (maximization strategy) implemented through the modified heuristic algorithm for the development of core collection . The power core programme is employed in the development of core collection in barnyard millet (Jayarame Gowda *et al.*, 2009) and ragi (Chandrashekhar *et al.*, 2012). In soybean, core collection has been developed in Chinese germplasm collection, USDA collection and Korean soybean germplasm (Zhang *et al.*, 2003; Cho *et al.*, 2008; Oliveira *et al.*, 2010; Guo *et al.*, 2014).

In the present study, two different core collection approaches, namely PCS strategy and power core method, were employed for the development of core collection in Indian soybean germplasm. Initially, two independent core sets were developed by the two approaches, and these core sets were further compared to identify diverse and unique core set accessions for the constitution of integrated core collection (ICC) in soybean.

Materials and methods

In the present study, a total of 3443 accessions of soybean germplasm maintained at National Active Germplasm Site for Soybean, ICAR-Directorate of Soybean Research, Indore, Madhya Pradesh (India) were used for the development of soybean core collection. The soybean germplasm used in the study was comprised of exotic collections, indigenous collections, landraces and local cultivars. We characterized the soybean germplasm for eight agromorphological traits: days to 50% flowering (DF); days to pod initiations (DPI); days to maturity (DM); plant height (PH); number of branches per plant (Br/pl); number of nodes per plant (Nodes/pl); number of pods per plant (Pods/pl); yield per plant (Yield/pl). Two different approaches were employed for the development of core collection in soybean: PCS strategy (Noirot *et al.*, 1996) and power core method (Kim *et al.*, 2007). The resultant core sets were compared to identify unique and diverse core set accessions to constitute an ICC.

The PCS strategy described by Noirot *et al.* (1996) is a multivariate analysis used for the identification of core sets from germplasm. The PCS strategy employs PCA to eliminate the collinearity of multiple variables and select diverse accessions based on the relative contribution of each accession. Eight agromorphological data of 3443 soybean germplasm were subjected to PCA, and generalized sum of square (GSS) for all the accessions was calculated from the PCS (Lebart *et al.*, 1977). The contribution P_i of the individual i to the total GSS was calculated as

$$P_i = \sum_{j=1}^{K} x_{ij}^2.$$

Furthermore, the PCSs for those principal components (PCs) whose eignevalue was more than 1 were used for the estimation of CR_i (relative contribution of individuals). The CR_i of the individual i to the total GSS was calculated as

$$CR_i = P_i/(NK),$$

where N is the number of individuals and K the variables.

The genotype with the highest CR_i (in %) to the total GSS was considered for the core set. This was repeated up to 50% of GSS, and all the genotypes with the highest CR_i to 50% of GSS were considered for the final core set.

In another approach, the advanced M (maximization) strategy implemented through a modified heuristic algorithm described by Kim *et al.* (2007) was also carried out using Power core v1.0 software to constitute a core set. The M strategy with random and heuristic searches selects the most diverse accessions to represent the entire variability of the whole collection. Power core creates the number of classes for any quantitative character as a default value based on Sturge's rule, which provides opportunity to increase the number of classes for quantitative traits. The power core programme

minimizes the loss of useful alleles and, hence, effectively selects accessions with the highest diversity, reducing the repeated alleles.

Quantitative data of core collections and entire collection were further subjected to statistical analysis to estimate range, mean, standard deviation, variance and coefficient of variance (CV) to validate the core sets. Means of the EC and core collections were compared for all the quantitative traits using the *t* test.

Results

Development of core collection

The eight agromorphological data of 3443 soybean accessions were subjected to PCA and power core software for the development of core sets. The PCA revealed six PCs, of which four PCs had more than one eigenvalue, together explaining 83.7% of the cumulative variance (Table 1). The PCS for those PCs whose eigenvalue was more than 1 was further used to estimate CR_i for each accession. Accessions with the maximum CR_i to 50% of the total CR_t were selected for the constitution of core collection. Thus, the PCS method identified 576 accessions for core collection, which accounted for 16.72% of the EC. In contrast, power core software identified 403 accessions for core collection, which accounted for about 11.67% of the EC.

Contribution of the variables in PCA

The contribution of variables to each PC was estimated using eigenvector values (Table 1). DF (30.48%), DPI (31.89%), PH (16.23%) and DM (10.78%) contributed 89.39% of the total variance to PC1. While in PC2, Yield/pl (33.89%), Pods/pl (33.87) and Br/pl (25.90) contributed 9.65% of the total variance. DM (37.12%), PH (24.35%) and Br/pl (15.87%) contributed 77.34% of the total variance to PC3. In PC4, Nodes/pl alone contributed 89.75% of the total variance.

Validation of core collections

Eight agromorphological data of the two core collections (CC1 and CC2) and EC were subjected to statistical analysis to estimate range, mean, variance, standard deviations, CV (Table 2) and correlation coefficient (Tables 3 and 4). The mean DF was 44.8 (EC), 41.6 (CC1) and 46.08 (CC2), with a range of 24–94 in both core collections and EC. DPI ranged from 30 to 98 in both core collections and EC, with a mean of 54.7 (EC), 52.1 (CC1) and 56.65 (CC2). PH ranged from 5.4 to 118.8 in EC, with a mean of 60.3, while it ranged from 12 to 118.8 in CC1, with a mean of 48.1, and from 5.4 to 118.8 in CC2, with a mean of 55.57. Br/pl ranged from 0.33 to 22 in EC, CC1 and CC2), with a mean of 4.3 (EC), 5.2 (CC1) and 5.62 (CC2). Nodes/pl ranged from 1 to 157.67 in EC and CC2 and from 1.67 to 139.6 in CC1, with a mean value of 14 (EC), 18.4 (CC1) and 23.74 (CC2). Pods/pl ranged from 1.33 to 301 in EC, CC1 and CC2, with a mean of 43.6 (EC), 45.9 (CC1) and 49.78 (CC2). Yield/pl (g) ranged from 0.06 to 64.7 in both core collections and EC, with a mean of 6.1 (EC), 7.0 (CC1) and 7.89 (CC2).

The difference between the means of CC1 and EC was significant for DF, DPI, DM, PH, Br/pl, Nodes/pl and Yield/pl (g), whereas the difference between the means of CC2 and EC was significant for DF, DPI, PH, Br/pl, Nodes/pl, Pods/pl and Yield/pl (g). However, the mean difference between CC1 and EC was not significant for

Table 1. Principal component analysis of the eight quantitative traits of 3443 soybean germplasm

Variables	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Eigenvalue	2.6871	1.7222	1.285	1.0018	0.4986	0.3862	0.3154	0.1038
Proportion	0.336	0.215	0.161	0.125	0.062	0.048	0.039	0.013
Cumulative	0.336	0.551	0.712	0.837	0.899	0.948	0.987	1.00
Contribution of variables								
DF	30.48	1.80	0.54	0.57	1.14	18.52	1.00	45.95
DPI	31.89	1.02	0.01	0.02	2.58	11.97	0.00	52.51
DM	10.78	1.60	37.12	0.27	17.09	21.82	10.37	0.94
PH	16.23	1.35	24.35	1.71	1.16	46.33	8.85	0.02
Br/pl	4.24	25.90	15.87	5.15	21.12	0.00	27.72	0.01
Nodes/pl	0.94	0.58	2.90	89.75	0.14	0.76	4.40	0.54
Pods/pl	3.80	33.87	9.68	2.31	10.88	0.24	39.18	0.04
Yield/pl (g)	1.64	33.89	9.53	0.23	45.89	0.35	8.47	0.00

PC, principal component; DF, days to 50% flowering; DPI, days to pod initiation; DM, days to maturity; PH, plant height; Br/pl, number of branches per plant; Nodes/pl, number of nodes per plant; Pods/pl, number of pods per plant; Yield/pl, yield per plant (g).

Table 2. Comparison of mean, range, variance, standard deviation, coefficient of variance between core collections and entire collections for the eight quantitative traits

Parameters	DF	DPI	DM	PH	Br/pl	Nodes/pl	Pods/pl	Yield/pl (g)
Mean								
EC	44.8	54.7	96.3	60.3	4.3	14	43.6	6.1
CC1	41.6*	52.1*	94.6*	48.1*	5.2*	18.4*	45.9NS	7*
CC2	46.08*	56.65*	97.31NS	55.57*	5.62*	23.74*	49.78*	7.89*
ICC	43.11*	52.91*	95.76NS	50.01*	5.18*	18.98*	45.04NS	6.96*
Range								
EC	24–94	30–98	73–137	5.4–118.8	0.33–22	1–157.67	1.33–301	0.06–64.7
CC1	24–93	30–98	73–137	12–118.8	0.33–22	1.67–139.6	1.33–301	0.06–64.7
CC2	24–94	30–98	73–137	5.4–118.8	0.33–22	1–157.67	1.33–301	0.06–64.7
ICC	24–94	30–98	73–137	5.4–118.8	0.33–22	1–157.6	1.33–301	0.06–64.7
Variance								
EC	63.8	73.4	68	347	3.3	75.2	580	17.3
CC1	117.38	176.54	140.59	537.73	9.44	296.49	1610.42	44.22
CC2	137.62	144.6	144.79	490.62	9.54	463.97	2054.81	60.36
ICC	137.28	170.8	150.68	522.78	8.71	316.38	1507.01	41.65
Standard deviation								
EC	8	8.6	8.2	18.6	1.8	8.7	24.1	4.2
CC1	10.8	13.3	11.9	23.2	3.1	17.2	40.1	6.7
CC2	11.73	12.03	12.03	22.15	3.09	21.54	45.33	7.77
ICC	11.72	13.07	12.28	22.86	2.95	17.79	38.82	6.45
Coefficient of variance								
EC	17.8	15.7	8.6	30.9	42.5	62.2	55.2	68.6
CC1	26	25.5	12.5	48.2	58.9	93.4	87.5	94.9
CC2	25.46	21.23	12.37	39.86	54.94	90.73	91.06	98.47
ICC	27.17	24.69	12.81	45.72	56.92	93.69	86.19	92.77

DF, days to 50% flowering; DPI, days to pod initiation; DM, days to maturity; PH, plant height; Br/pl, number of branches per plant; Nodes/pl, number of nodes per plant; Pods/pl, number of pods per plant; Yield/pl, yield per plant (g); EC, entire collection; CC1, core collection derived from the PCS method; NS, non-significant at the 5% level; CC2, core collection derived from the power core method; ICC, integrated core collection.
*Significant at the 5% level.

Pods/pl, while that between CC2 and EC was not significant for DM (Table 2).

Correlation studies

The core collections developed in the present study was further validated using correlation studies. Correlation coefficient was estimated among the eight traits in both core collections and EC (Tables 3 and 4). The correlation between DF and DPI was highly significantly positive in EC (0.877) as well as in CC1 (0.871) and CC2 (0.799). DM showed a highly significant positive association with DF (EC 0.339, CC1 0.573 and CC2 0.616) and DPI (EC 0.437, CC1 0.627 and CC2 0.573) in both core collections and entire collection. PH showed a highly significant positive association with DF (EC 0.557, CC1 0.633 and CC2, 0.401) and DPI (EC 0.527, CC1 0.619 and CC2 0.453) in both core collections and entire collection, while it showed a highly significant positive association with DM in CC1 (0.271) and CC2 (0.173), but it was not significant in EC (0.017). Br/pl showed a highly significant positive association with DF (EC 0.122, CC1 0.299 and CC2 0.212), DPI (EC 0.188, CC1 0.358 and CC2 0.23) and DM (EC 0.437, CC1 0.531 and CC2 0.408) in both core collections and EC, but it showed a significant negative association with plant height in EC (−0.091), while it showed a significant positive association in CC1 (0.194) and CC2 (0.145).

Nodes/pl showed a highly significant positive association with DF (0.101), DPI (0.157), PH (0.102) and Br/pl (0.074), and a significant negative association with DM (−0.069) in EC. However, the association of Nodes/pl with DF (0.161), DPI (0.24), PH (0.193) was significantly positive, while its association with DM (−0.057) and Br/pl (−0.03) was non-significantly negative in CC1; however, it showed a significant negative association with DM (−0.221) and Br/pl (0.1211), a non-significant positive association with DPI (0.155) and PH (0.027) and a non-significant negative association with DF (−0.056) in CC2.

Pods/pl showed a significant positive association with DF (EC 0.171, CC1 0.258 and CC2 0.105), DPI (EC 0.174, CC1 0.262 and CC2 0.169), DM (EC 0.043,

Table 3. Comparison of correlation coefficient (*r*) between core collections and entire collection for the quantitative traits

	DF				DPI				DM			
	EC	CC1	CC2	ICC	EC	CC1	CC2	ICC	EC	CC1	CC2	ICC
DPI												
r	0.877	0.871	0.799	0.87								
P	0	0	0	0								
DM												
r	0.339	0.573	0.616	0.64	0.437	0.627	0.573	0.608				
P	0	0	0	0	0	0	0	0				
PH												
r	0.557	0.633	0.401	0.557	0.527	0.619	0.453	0.573	0.017	0.271	0.173	0.252
P	0	0	0	0	0	0	0	0	0.332	0	0.001	0
Br/pl												
r	0.122	0.299	0.212	0.287	0.188	0.358	0.23	0.326	0.437	0.531	0.408	0.494
P	0	0	0	0	0	0	0	0	0	0	0	0
Nodes/pl												
r	0.101	0.161	−0.056	0.112	0.157	0.24	0.074	0.211	−0.069	−0.057	−0.221	−0.09
P	0	0	0.273	0.003	0	0	0.155	0	0	0.172	0	0.016
Pods/pl												
r	0.171	0.258	0.105	0.191	0.174	0.262	0.169	0.238	0.043	0.277	0.163	0.205
P	0	0	0.042	0	0	0	0.001	0	0.011	0	0.001	0
Yield/pl (g)												
r	−0.247	−0.026	−0.114	−0.075	−0.23	−0.022	−0.079	−0.036	−0.133	0.013	−0.044	−0.035
P	0	0.528	0.028	0.05	0	0.595	0.131	0.358	0	0.758	0.389	0.358

DF, days to 50% flowering; DPI, days to pod initiation; DM, days to maturity; EC, entire collection; CC1, core collection derived from the PCS method; CC2, core collection derived from the power core method; ICC, integrated core collection; PH, plant height; Br/pl, number of branches per plant; Nodes/pl, number of nodes per plant; Pods/pl, number of pods per plant; Yield/pl, yield per plant.

Table 4. Comparison of correlation coefficient (*r*) between core collections and entire collection for the quantitative traits

	PH				Br/pl				Nodes/pl				Pods/pl			
	EC	CC1	CC2	ICC	EC	CC1	CC2	ICC	EC	CC1	CC2	ICC	EC	CC1	CC2	ICC
Br/pl																
r	−0.091	0.194	0.145	0.188												
P	0	0	0.004	0												
Nodes/pl																
r	0.102	0.193	0.027	0.164	0.074	−0.03	−0.111	−0.014								
P	0	0	0.592	0	0	0.469	0.027	0.713								
Pods/pl																
r	0.258	0.334	0.28	0.313	0.384	0.51	0.45	0.487	−0.068	−0.169	−0.274	−0.176				
P	0	0	0	0	0	0	0	0	0	0	0	0				
Yield/pl (g)																
r	−0.086	0.033	0.02	0.036	0.175	0.303	0.209	0.255	−0.003	−0.049	−0.111	−0.049	0.456	0.557	0.553	0.555
P	0	0.434	0.697	0.348	0	0	0	0	0.879	0.244	0.029	0.199	0	0	0	0

PH, plant height; Br/pl, number of branches per plant; Nodes/pl, number of nodes per plant; Pods/pl, number of pods per plant; EC, entire collection; CC1, core collection derived from the PCS method; CC2, core collection derived from the power core method; ICC, integrated core collection; Yield/pl, yield per plant.

CC1 0.277 and CC2 0.163), PH (EC 0.258, CC1 0.334 and CC2 0.28) and Br/pl (EC 0.384, CC1 0.51 and CC2 0.45), while it showed a significant negative association with Nodes/pl (EC −0.068, CC1 −0.169 and CC2 −0.274) in both core collections and EC.

Yield/pl showed a significant positive association with Br/pl (0.175) and Pods/pl (0.456), a significant negative association with DF (−0.247), DPI (−0.23), DM (−0.133) and PH (−0.086), and a non-significant negative association with Nodes/pl (−0.003) in EC. However, it showed a significant positive association with Br/pl (0.303) and Pods/pl (0.557), a non-significant positive association with DM (0.013) and PH (0.033), and a non-significant negative association with DF (−0.026), DPI (−0.022) and Nodes/pl (−0.049) in CC1. Furthermore, it showed a significant positive association with Br/pl (0.209) and Pods/pl (0.553), a non-significant positive association with PH (0.02), a significant negative association with DF (−0.23) and Nodes/pl (−0.111), and a non-significant negative association with DPI (−0.079) and DM (−0.044) in CC2.

ICC in soybean

In the present study, the PCS method identified 576 accessions and power core identified 402 accessions for the constitution of core collection. When both core collections were compared, interestingly, 264 accessions were found to be common in both core collections. In addition, there were 311 and 137 unique accessions derived from the PCS and power core methods, respectively. The ICC was constituted by integrating both common and unique accessions from the two core collections. This integrated method yielded 713 accessions, which accounted for about 20.62% of the EC. However, the 265 common accessions in both core collections were the most diverse core set accessions, which accounted for 7.64% of the EC.

The quantitative data of the ICC were analysed to estimate mean, range, variance, standard deviation and CV (Table 2). The mean difference between ICC and EC was significant for DF, DPI, PH, Br/pl, Nodes/pl and Yield/pl. However, the *t* test was not significant for DM and Pods/pl. The range for all the eight traits of ICC was similar to that of EC. The variance, standard deviation and CV of all the traits were higher in ICC than in EC.

Correlations studies in ICC (Tables 3 and 4) showed that yield was negatively associated with DF (−0.0175), DPI (−0.036), DM (−0.035) and Nodes/pl (−0.049), whereas it was significantly positively associated with PH (0.036), Br/pl (0.255) and Pods/pl (0.555). Among the traits, DPI showed a significant positive association with DF (0.87), DM (0.608), PH (0.573), Br/pl (0.326),

Nodes/pl (0.211) and Pods/pl (0.238). DM was significantly positively associated with PH (0.252), Br/pl (0.494) and Pods/pl (0.205), but negatively associated with Nodes/pl (−0.09). PH was significantly positively associated with Br/pl (0.188), Nodes/pl (0.164) and Pods/pl (0.313). Br/pl was negatively associated with Nodes/pl (−0.014) and Pods/pl (−0.487), whereas Nodes/pl was significantly negatively associated with Pods/pl (−0.176).

Discussion

Development of core collection in soybean will accelerate the evaluation process for the identification of desirable genotypes for economically important traits. The PCS and power core methods employed for the development of core collection in the present are statistically different approaches. The PCS method uses PCA to identify diverse individuals for the development of core collection. In contrast, the power core strategy uses a heuristic search tool to identify diverse individuals for the development of core collection.

A core collection of 576 accessions was developed from 3443 soybean accessions using PCA. CC1 accounted for 16.72% of the EC. The PCA was used to measure the independent impact of the variables on the total variance, wherein the degree of contribution of each original variable with which each principal component was associated was determined using coefficient of vectors. The eigenvector indicated the degree of contribution of a variable with which each PC was associated. Using the eigenvector, we estimated the percentage contribution of the variable in each PC. The following variables were found to be the significant contributor of variance in PCA: DF, DPI, PH and DM in PC1; Yield/pl, Pods/pl and Br/pl in PC2; DM, PH, Br/pl PC3; Nodes/pl in PC4.

The two-sample *t* test showed a significant difference between the means of CC1 and EC for all the traits except Pods/pl. The range for all the traits, except for PH and Nodes/pl, was similar in both CC1 and EC. This indicates that the chosen accessions in CC1 are representative of EC, and that the variation in the EC is preserved in the core collection except for PH and Nodes/pl. Statistical parameters such as variance, standard deviation and CV were higher in CC1 than in the EC, indicating the presence of adequate variability in core collections for all the quantitative traits (Hamon *et al.*, 1995; Bisht *et al.*, 1998; Upadhyaya *et al.*, 2003, 2006; Mahajan *et al.*, 2007). Upadhyaya *et al.* (2003) analysed the taxonomical, geographical and morphological descriptor data of 14,310 accessions in groundnut for the development of core collection consisting of 1704 accessions by using the PCS strategy.

Another core collection of 402 accessions was developed using power core software, which accounted for 11.67% of the EC. Similar to CC1, the *t* test showed a significance difference between the means of CC2 and EC for all the traits except for DM. The range for all the traits in CC2 was similar to that in the EC, indicating that CC2 explained 100% of the variance in the EC for all the traits. The variance, standard deviation and CV of CC1 were higher than that of EC, indicating that CC2 captured adequate variability from the EC (Jayarame Gowda *et al.*, 2009; Chandrashekhar *et al.*, 2012). Jayarame Gowda *et al.* (2009) employed the power core programme for the development of core collection in barnyard millet. They subjected 24 agromorphological data of 729 barnyard millet accessions for the development of core collection consisting of 50 accessions. In another study conducted by Chandrashekhar *et al.* (2012), 12 morphological data of 1000 ragi accessions were used in the power core programme for the development of core collection consisting of 77 accessions.

Except for DPI and PH, variance and standard deviation were higher in CC2 compared with CC1, indicating that higher variability is present in CC2 than in CC1. However, CV was higher in CC1 than in CC2 except for Pods/pl and Yield/pl.

While developing core collection, it is important to have a proper and adequate sampling strategy to preserve character associations that occur due to co-adopted gene complexes (Ortiz *et al.*, 1998). Among the 28 correlation combinations studied, most of the correlations observed in the EC were preserved in core collections, except for seven combinations (Yield/pl and DF, Yield/pl and DPI, DM and PH, DM and Nodes/pl, Yield/pl and DM, Yield/pl and PH, Br/pl and Nodes/pl) in CC1 and eight combinations (DF and Nodes/pl, DPI and Nodes/pl, Yield/pl and DPI, DM and PH, Yield/pl and DM, PH and Nodes/pl, Yield/pl and PH, Yield/pl and Nodes/pl) in CC2. However, when considering both core collections together, of the 28 correlations, only four correlations (Yield/pl and DPI, Yield/pl and DM, Yield/pl and PH, PH and DM) were not found to be similar. This indicates that phenotypic correlations observed in the EC in general were preserved in the core collection of soybean developed in the present study. The correlation coefficient value more than 0.71 is considered to be meaningful, because more than 50% of the variation in one trait is predicted by the other trait (Skinner *et al.*, 1999; Upadhyaya *et al.*, 2003). In our study, the presence of a higher correlation coefficient value between DF and DPI ($r > 0.799$) shows that one of these traits need not be measured all the time during germplasm evaluation in future studies. The lower correlation coefficient value of two important traits (DF and Br/pl) influences the yield of soybean. DF showed a

negative correlation with yield ($r = -0.247$ in EC, $r = -0.026$ in CC1 and $r = -0.114$ in CC2), while Br/pl showed a positive correlation ($r = 0.175$ in EC, $r = 0.303$ in CC1 and $r = 0.209$ in CC2), suggesting that either of these traits may be a useful measure in choosing newer accessions for further evaluation of protein contents.

A different statistical approach of core collection provides different core sets, although the aim of every method is to draw representative samples from the EC. In our study, when we compared the two core collections (CC1 and CC2) with EC, we could find unique accessions and common accessions derived from both core collections. Both unique accessions and common accessions were included for the constitution of ICC, which comprised 713 accessions. The ICC accounted for 20.62% of the EC. The two-sample *t* test showed a significant difference between the means of CC1 and EC for all the traits except Pods/pl. However, the *t* test showed significance differences between the means of CC2 and EC for all traits except DM. However, in the case of ICC, the mean difference between ICC and EC was significant for all traits except DM and Pods/pl. In addition, the variance and standard deviation of DM and the CV of DF, DM and Nodes/pl were highest in ICC when compared with EC, CC1 and CC2. This clearly shows that maximum variability was captured in ICC for DF, DM and Nodes/pl compared with CC1 and CC2. In the EC, of the 28 correlation combinations, 26 were found to be significant at the 5% probability level. Among the 26 significant correlation combinations in EC, 23 were well preserved in ICC. This indicates that co-adopted gene complexes were well preserved in ICC than in CC1 and CC2. Therefore, the ICC developed in the present study captures more variations than CC1 and CC2, representing the entire genetic variability of EC. However, 265 accessions were found to be common in both core collections, which accounted for 7.64% of the EC. These common accessions were identified by the PCS method and power core software; therefore, they are the most diverse core set accessions.

Cho *et al.* (2008) characterized 2765 accessions of soybean landraces with six SSR markers, and identified a final core set of 260 accessions based on marker allele stratification. This core set revealed nearly the same diversity as the other results on morphological traits of Korean soybean landraces with respect to mean, range, standard deviation and CV. Guo *et al.* (2014) developed an integrated applied core collection (IACC) of soybean based on the evaluation data of agronomic, nutritional traits, biotic and abiotic stress tolerance in soybean germplasm resources. Molecular characterization with SSR markers and phenotypic data show that at the molecular level, soybean IACC harbours a similar level of genetic diversity to the established mini core collection (MCC), and that at the phenotypic level, the IACC encompasses more accessions with desirable traits than does the established MCC.

The soybean core collection developed in the present study will serve as important genetic resources for soybean breeders and researchers for the initial screening of soybean germplasm and for the identification of desirable genotypes for economically important traits. Screening of soybean core collection for various diseases, pests, moisture stress, salinity and temperature stress will enable to identify novel resistant types in a short span of time. Development of core collection will also help in tackling challenges that emerge out of climate changes because core collection represents the genetic variability of the EC and desirable genotypes can be readily identified. Owing to limited available resources, evaluating the EC may not be practically feasible; therefore, core collection can act as a working collection for breeders to be used in evaluation and breeding programmes. The soybean core collection developed in the present study can also be used in association mapping studies for the identification of genes/QTLs associated with various economically important traits. The present soybean core collection needs to be revised periodically as and when new accessions of soybean germplasm are collected and new data are generated.

Acknowledgements

The authors greatly acknowledge the support provided by the ICAR-Directorate of Soybean Research, Indore, Madhya Pradesh for the conduct of the experiments. The authors also express their sincere gratitude to the ICAR-National Bureau of Plant Genetic Resources, New Delhi for providing the soybean genetic resources.

References

Bisht IS, Mahajan RK and Patel DP (1998) The use of characterisation data to establish the Indian mungbean core collection and assessment of genetic diversity. *Genetic Resources and Crop Evolution* 45: 127–133.

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

Brown AHD (1989b) The case for core collections. In: Brown AHD, Frankel OH, Marshall DR and Williams JT (eds) *The Use of Plant Genetic Resources*. Cambridge: Cambridge University Press, pp. 136–155.

Chandrashekhar H, Gowda J and Ugalat J (2012) Formation of core set in Indian and African finger millet [*Eleusine coracana* (L.) Gaertn] germplasm accessions. *Indian Journal of Genetics and Plant Breeding* 72: 358–363.

Cho GT, Yoon M-S, Lee J, Baek H-J, Kang J-H, Kim T-S and Paek N-C (2008) Development of a core set of Korean soybean

landraces [*Glycine max* (L.) Merr.]. *Journal of Crop Science Biotechnology* 11: 157–162.

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber WK, Llimensee K, Peacock WJ and Starlinger P (eds) *Genetic Manipulation: Impact on Man and Society*. Cambridge: Cambridge University Press, pp. 161–170.

Guo Y, Li Y, Hong H and Qiu L-J (2014) Establishment of the integrated applied core collection and its comparison with mini core collection in soybean (*Glycine max*). *The Crop Journal* 2: 38–45.

Hamon S, Noirot D and Anthony F (1995) Developing a coffee core collection using the principal components score strategy with quantitative data. In: Hodgkin T, Brown AHD, van Hintum TJL and Morales EAV (eds) *Core Collections of Plant Genetic Resources*. Chichester, UK: IPGRI-Wiley & Sons, pp. 117–126.

Jayarame Gowda, Bharathi S, Somu G, Krishnappa M and Rekha D (2009) Formation of core set in barnyard millet [*Echinochloa frumentacea* (Roxb.) Link] germplasm using data on twenty four morpho-agronomic traits. *International Journal of Plant Sciences* 4: 1–5.

Kim KW, Chung HK, Cho GT, Ma KH, Chandrabalan D, Gwag JG, Kim TS, Cho EG and Park YJ (2007) PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23: 2155–2162.

Lebart L, Morineau A and Tabart N (1977) *Techniques de la Description Statistique: Méthodes et Logiciels pour l'Analyse des Grands Tableaux*. Paris, France: Dunod.

Mahajan RK, Bisht IS and Dhillon BS (2007) Establishment of a core collection of world sesame (*Sesamum indicum* L.) germplasm accessions. *SABRAO Journal of Breeding and Genetics* 39: 53–64.

Noirot M, Hamon S and Anthony F (1996) The principal component scoring: a new method of constituting a core collection using quantitative data. *Genetic Resources and Crop Evolution* 43: 1–6.

Odong TL, Jansen J, van Eeuwijk FA and van Hintum TJL (2013) Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. *Theory and Applied Genetics* 126: 289–305.

Oliveira MF, Nelson RL, Geraldi IO, Cruz CD and de Toledo JFF (2010) Establishing a soybean germplasm core collection. *Field Crops Research* 119: 277–289.

Ortiz R, Ruiz-Tapia EN and Mujica-Sanchez A (1998) Sampling strategy for a core collection of Peruvian quinoa germplasm. *Theory and Applied Genetics* 96: 475–483.

Qiu L-J, Xing L-L, Guo Y, Wang J, Jackson SA and Chang RZ (2013) A platform for soybean molecular breeding: the utilization of core collections for food security. *Plant Molecular Biology* 83: 41–50.

Skinner DZ, Bauchan GR, Auricht G and Hughes S (1999) A method for the efficient management and utilization of large germplasm collections. *Crop Science* 39: 1237–1242.

Upadhyaya HD, Gowda CLL, Pundir RPS, Gopal Reddy V and Singh S (2006) Development of core subset of finger millet germplasm using geographical origin and data on 14 quantitative traits. *Genetic Resources and Crop Evolution* 53: 679–685.

Upadhyaya HD, Ortiz R, Bramel PJ and 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 TJL, Brown AHD, Spillane C and Hodgkin T (2000) Core collections of plant genetic resources. *IPGRI Technical Bulletin No. 3*. Rome, Italy: International Plant genetic Resources Institute.

Zhang B, Qiu L and Chang R (2003) Advance on genetic diversity and core collection establishment for soybean. *Crop Journal* 3: 46–48.