Plant Breeding **131,** 110—117 (2012)
doi:10.1111/j.1439-0523.2011.01896.x

Efficiency of PowerCore in core set development using amplified fragment length polymorphic markers in mungbean

Kyaw Thu Moe[1], Jae-Gyun Gwag[2] and Yong-Jin Park[1,3,4]

[1]Department of Plant Resources, College of Industrial Sciences, Kongju National University, Yesan 340-702, Korea; [2]Rural Development Administration (RDA), Suwon 441-100, Korea; [3]Legume Bio-Resource Center of Green Manure (LBRCGM), Kongju National University, Yesan 340-702, Korea; [4]Corresponding author, E-mail: yjpark@kongju.ac.kr

With 2 figures and 5 tables

Received February 10, 2011/Accepted July 4, 2011
Communicated by W. Link

Abstract

The mungbean [*Vigna radiata* (L.) Wilczek] is a member of the Fabaceae, consists of around 20 000 species. With the rapid increase in the number of germplasm collections, many gene banks face problems of redundant resources. To cope with this problem, the development of an allele-mining set is especially important. Our present study supports the efficiency of PowerCore in the development of core set from 705 collected accessions using AFLP markers. The result demonstrated the higher allele (fragment) capturing efficiency of PowerCore than other strategies (distance-based, stratified random and random) tested at any level of sample size (5%, 10%, 15%, 20% of total accessions) selected. Highly significant correlation ($r = 0.96$) was observed between allele frequency distribution of entire collections and that of core set developed by PowerCore. The resulted core set was confirmed with 15 SSR data. The result will be useful, especially for reducing the redundant resources in the gene bank and allele mining from a large germplasm collection.

Key words: mungbean — PowerCore — efficiency of allele mining — AFLP — SSR

The mungbean [*Vigna radiata* (L.) Wilczek], a member of the Fabaceae comprising around 20 000 species, is an important economic crop (Tangphatsornruang et al. 2010). It probably originated in India (De Candole 1886, Zhukovsky 1950, Bailey 1970) or the Indo-Burmese region (Vavilov 1951), and it is widely grown in South-east Asia, Africa, South America and Australia. It was stated that the mungbean was apparently grown in the United States as early as 1835 as Chickasaw pea. Mungbean is also referred to as green gram, golden gram and chop suey bean. It is grown widely for a variety of purposes, as human food (dry beans or fresh sprouts), a green manure crop and forage for livestock.

Early genetic studies of mungbean have been conducted in India since 1939 (Bose 1939). A great deal of research has been conducted at the Asian Vegetable Research and Development Center (AVRDC) in Taiwan, which is an international centre responsible for mungbean research worldwide. The AVRDC was designated as the base collection centre for mungbean by the International Board for Plant Genetic Resources (IBPGR). Many accessions of *V. radiata* are maintained in different countries, but a considerable amount of duplication of accessions exists among these collections. The major collections of mungbean germplasm are maintained at the AVRDC in Shanhua, Taiwan; University of the Philippines in Los Banos; the US Department of Agriculture in Georgia; and the Indian Agricultural Research Institute in New Delhi (Anishetty and Moss 1988, Poehlman 1991). To date, the AVRDC has 10 733 accessions of *Vigna* and related species, including 5900 of *V. radiata* var. *radiata* obtained from various laboratories and sources (Shanmugasundaram et al. 2009).

With the rapid increase in the number of accessions in crop germplasm collections, many gene banks face problems of redundant resources and cost of maintaining these collections, which may be an obstacle for their full exploitation, evaluation and utilization (Holden 1984). To cope with problems of management, research and application, Frankel and Brown (1984) proposed the concept of the core set. The design of a core set should include the maximum possible genetic diversity contained in the entire collection with minimum repetition. The information obtained from such a core set can aid in the judicious use of the entire collection. To date, most core sets for crops have been developed based on passport data, giving the geographic origin, morphological and phenotype traits, and biochemical or molecular markers (Perry et al. 1991, Joe and Orlando 1996, Hokanson et al. 1998, Ortiz et al. 1998, Upadhyaya and Ortiz 2001, Chandra et al. 2002). A good core set should minimize redundant entries and should be sufficiently large so as to provide reliable conclusions for the entire collection (Brown 1989). The sampling proportion and variation representation of the entire collection are important in construction of the core set to retain the greatest degree of genetic diversity.

Many different methodologies are available for building sampling strategies (Zhao et al. 2010a). Schoen and Brown (1993) addressed the issue of how to use genetic markers to sample collections of wild crops while maximizing allelic richness. They have compared different strategies such as H strategy, M strategy (marker allele richness), C strategy (constant number of accessions per region), P strategy (proportion to the number of accessions available per region), L strategy (proportion to the logarithm of the number of accessions available per region) and R strategy (random sampling accessions) and pointed out that the target allele retention was maximized under the M strategy. Bataillon et al. (1996) used computer simulation and found that the maximization strategy (M strategy) was more effective in retaining widespread and low-frequency neutral alleles compared with other sampling strategies. In some studies (Zhang et al. 2000, Upadhyaya et al. 2002, Kang et al. 2006), the cluster algorithms of Ward's clustering method have been used to sample

the core members from the whole collection. At least one accession was selected randomly in each cluster group based on a dendrogram with proper threshold values, although the issue of how to theoretically ascertain threshold values of genetic distance for classification criteria after cluster analysis has not been fully resolved yet (Hu et al. 2000).

Gouesnard et al. (2001) devised the MSTRAT algorithm by implementing the M strategy for selecting accessions, and Kim et al. (2007) developed PowerCore software: a programme applying the advanced M strategy with heuristic searching to establish allele mining or core set. Using 1000 virtual accessions of rice, a comparison was carried out between PowerCore (http://genebank.rda.go.kr/eng/PowerCore/powercore.jsp) and MSTRAT by Kim et al. (2007). They pointed out that PowerCore retains all classes in the core collection with a minimum number of accessions. It gives the user the ability to perform preferential selection by placing the symbol '∼' in front of accessions that the user wishes to retain. Chung et al. (2009) found that the heuristic core collection method was more efficient than the proportional core collection and random core collection (RCC) strategies in developing a core set. Zhao et al. (2010a) found a similar result in the development of a core set for rice, whereby the use of a modified heuristic search in the core set was better than stratified random sampling and a random sampling method to capture maximum alleles with minimum redundancy. The development of a mini-core set for finding new alleles from large entries using genomic tools is of great interest for researchers. These developments will prove advantageous for plant breeders trying to increase yields and create new varieties that are resistant to diseases, pests, drought and salinity, and/or have improved nutritional quality (Latha et al. 2004).

Recently, molecular genetic markers have been widely used to characterize gene bank collections and for diversity analysis (Chung and Park 2010, Zhao et al. 2010b,c). Our present study developed a core set using PowerCore programme that uses a heuristic approach and amplified fragment length polymorphism (AFLP) data from an entire collection of 705 mungbean accessions conserved in the National Genebank of Rural Development Administration, Korea (RDA-genebank), and evaluated the efficiency of allele mining comparing with other sets of allele-mining strategy for the development of a core set.

Materials and Methods

Plant material and DNA extraction: The 705 mungbean accessions from 26 countries were taken from the Rural Development Administration (RDA), Korea (Table 1). The plants of each accession were grown in greenhouse, and DNA was extracted from fresh leaves of 15-day-old seedlings using a DNA extraction kit (Qiagen, Hilden, Germany). The relative purity and concentration of extracted DNA was estimated with the NanoDrop ND-1000 (Dupont Agricultural Genomics Laboratory, Wilmington, DE, USA). The final concentration of each DNA sample was adjusted to 20 ng/μl.

Genotyping using AFLP markers: AFLP analysis was carried out essentially as described by Vos et al. (1995) with minor modifications. Genomic DNA (100 ng) was digested with 1 U each of *Eco*RI and *Mse*I restriction enzymes. The *Eco*RI and *Mse*I adapters were ligated to the ends of restricted fragments. The digested and ligated template DNA was pre-amplified using *Eco*RI-(EP;5′-GACTGCGT ACCAATTCA-3′) and *Mse*I-(MP;5′-GATGAGTCCTGAGTAAC-3′)-directed primers. These primers contained a core sequence, restriction site and an additional selective nucleotide at the 3′ end.

Table 1: Origin of 705 accessions of mungbean used in this study

Origin			
Region	Country	No. of Acc.	Total
Africa	Kenya	1	7
	Madagascar	1	
	Nigeria	5	
East Asia	Korea	427	455
	Japan	5	
	China	12	
	Taiwan	11	
South East Asia	Indonesia	2	87
	Malaysia	2	
	the Philippines	50	
	Thailand	22	
	Vietnam	11	
Central Asia	Iran	20	56
	Uzbekistan	21	
	Afghanistan	15	
South West Asia	India	29	51
	Pakistan	11	
	Sri Lanka	1	
	Nepal	10	
Europe	England	3	17
	the Netherlands	1	
	Russia	4	
	Turkey	9	
Oceania	Australia	13	13
America	USA	18	19
	Guatemala	1	
Total			705

The pre-amplification was performed in a total volume of 20 μl containing approximately 2 ng of restricted ligated template DNA, 10 ng each of primers (EP and MP), 0.2 mM of dNTPs, 0.2 mM of 1 U of *Taq* polymerase and 2 mM of 10× PCR buffer containing 1.5 mM $MgCl_2$. The amplification was performed in a BIO-RAD S1000 thermocycler. The pre-amplification profile was as follows: one cycle of 95°C for 1 min and 20 cycles of 94°C for 30 s, 56°C for 1 min and 72°C for 1 min. The pre-amplification products were then diluted 20-fold with 10 mM Tris–HCl, 0.1 mM EDTA, pH 8.0.

Selective amplification was carried out using 6 primer pair combinations. Six *Eco*RI (EP63, 64, 65, 75, 76, 77 with EP + GAA, EP + GAC, EP + GAG, EP + GTA, EP + GTC and EP + GTG sequences, respectively) and four *Mse*I-end-directed (MP40, 49, 66, 61 with MP + AGC, MP + CAG, MP + GAT and MP + CTG sequences, respectively) primers (Operon). The selective amplification reaction was essentially the same as that for pre-amplification except that 2 μl of 20-fold diluted pre-amplification products was used as template, and 2 ng of *Eco*RI and 6 ng of *Mse*I M13-tail-attached primers were used. M13-tail PCR method described by Schuelke (2000) was used to measure the size of the amplified products (Schuelke, 2000). Conditions of the PCR amplification were as follows: 94°C for 3 min, 30 cycles each at 94°C for 30 s, 55°C (varied) for 45 s, 72°C for 1 min, followed by 10 cycles of 94°C for 30 s, 53°C for 45 s, 72°C for 1 min and a final extension at 72°C for 10 min. The fragments were resolved on a 3130 × 1 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) using GeneMapper 4.0 software and sized precisely using GeneScan 500 ROX (6-carbon-X-rhodamine) molecular size standards (35–500 bp). The bands were scored from 50 to 480 in every 2-bp difference. Among them, only 15 polymorphic fragments were analysed.

Genotyping using microsatellite or simple sequence repeat (SSR) markers: A set of 15 primer pairs, newly developed and presented by Gwag et al. (2010), was selected to use in the present study. The size of polymorphic polymerase chain reaction (PCR) product was analysed following the M13 tail PCR method described by Schuelke (2000), including a universal M13 oligonucleotide (TGTAAAACGACGGC-

CAGT) labelled with one of the fluorescent dyes, 6-FAM, NED or HEX. Microsatellite alleles were resolved on an ABI Prism 3100 DNA sequencer (Applied Biosystems).

Development of an allele-mining set: The advanced M strategy by a modified heuristic algorithm implemented in the PowerCore software by Kim et al. (2007) was used to develop the allele-mining set. The PowerCore software maximizes the number of alleles with the least redundancy (Kim et al. 2007). In the PowerCore software, the A* algorithm, a heuristic algorithm that finds the optimum path from the initial to the final stages, was used:

$$f(n) = g(n) + h(n)$$

Here, with $g(n)$ as the number of accessions inserted into the frequency table and $h(n)$ as the maximum number of empty cells within each column, this algorithm expands the paths that have the lowest value for $g(n) + h(n)$, where $g(n)$ is the cost for the path from the initial state to the current node and $h(n)$ serves as an estimate of the cost for the cheapest path from that node to the designated node. When expanding each of the steps, the sum of $g(n)$ and $h(n)$ will be evaluated and the accession with the lowest value will be chosen. If $h(n)$ is admissible without overestimating the costs of reaching the goal, then A* will always find an optimal solution (http://genebank.rda.go.kr/PowerCore/) (Kim et al. 2007).

Data analysis: Parameters such as total number of alleles per locus, number of rare alleles per locus (i.e., alleles with frequency $< 5\%$), number of unique alleles per locus (alleles occurring in only one accession), Shannon–Weaver diversity index (I) (Shannon and Weaver 1949), Nei's gene diversity index (H) (Nei 1973) and polymorphic information content (PIC) per locus were calculated for the entire set and core accessions using PowerCore software (Kim et al. 2007) and PowerMarker 3.25 (Liu and Muse 2005) software based on Rogers' distance (Rogers 1972). All indices were calculated independently in both the entire and the core set to determine whether the diversity for each locus was retained in the core set. Allele frequencies were analysed with PowerMarker 3.25, and frequency distributions for each locus were determined using Excel 2007 software (Microsoft, Redmond, WA, USA).

The Shannon–Weaver diversity index (I) as presented was estimated using:

$$I = \sum_{i=0}^{n} pi \log_e pi,$$

where pi is the frequency of the phenotypic class.

Nei's gene diversity (H) was calculated based on the formula:

$$H = 1 - \sum_{i=0}^{n} \left(\frac{ni}{N}\right)^2,$$

where ni is the allele frequency at the ith locus, n is the number of alleles at this locus and N is the total number of accessions. The PIC for each marker was calculated based on the formula:

$$\text{PIC} = 1 - \sum_{i=0}^{n} pi^2 - \sum_{i=0}^{n} \sum_{j=i+1}^{n} 2pi^2 pj^2,$$

where p is the relative frequency of the jth pattern for marker i (Botstein et al. 1980).

Evaluating the efficiency in the development of core set: To compare the allele capturing efficiency of PowerCore (PoCC), we also constructed dendrogram using PowerMarker 3.25 software programme. Then a distance-based core set was collected based on distance (Figure S1) resulted by diversity analysis. The other methods that can be used for the development of core set such as stratified random core collection (SRC) and random core collection (RCC) were also compared at different levels of sample size.

Validating the core set: The use of the core set may improve the efficiency of germplasm evaluation by reducing the number of accessions evaluated to increase the probability of finding genes of interest. To see the effectiveness of the core set in this study, 15 SSR markers were used on the entire and the core set.

Results

Development of a core set using AFLP data

A total of 222 (31.49%) accessions as a core set were retained by PowerCore programme. The total number of fragments,

Table 2: Total number of fragments, genetic diversity index and polymorphic information content (PIC) for AFLP combinations in the entire accessions (705) and core set 31.5% (222) of mungbean

		Entire collection						Core collection			
Analysed event	Maximum size	Fragment	R.F.	S.F.	I^1	H^2	PIC[3]	Fragment	I^1	H^2	PIC[3]
1	243	50	37	24	1.843	0.620	0.6092	50	2.755	0.851	0.8431
2	242	57	37	21	2.574	0.844	0.8296	57	3.238	0.924	0.9199
3	198	52	33	19	2.537	0.854	0.8407	52	3.143	0.924	0.9199
4	228	54	37	19	2.377	0.804	0.7877	54	3.137	0.919	0.9148
5	226	72	55	25	2.752	0.849	0.8392	72	3.595	0.944	0.9421
6	232	50	39	25	1.774	0.596	0.5860	50	2.790	0.853	0.8461
7	233	56	35	26	2.574	0.844	0.8309	56	3.194	0.921	0.9172
8	204	34	23	12	1.923	0.783	0.7518	34	2.358	0.837	0.8199
9	192	53	41	25	2.147	0.779	0.7535	53	2.932	0.896	0.8885
10	210	43	33	24	1.410	0.479	0.4717	43	2.329	0.743	0.7354
11	213	42	31	16	2.430	0.858	0.8454	42	2.806	0.890	0.8826
12	208	48	34	17	2.094	0.749	0.7217	48	2.896	0.880	0.8723
13	224	39	29	17	2.024	0.790	0.7631	39	2.554	0.860	0.8468
14	224	21	13	10	0.741	0.258	0.2545	21	1.463	0.537	0.5262
15	192	24	16	12	0.839	0.294	0.2894	24	1.370	0.491	0.4813
Total		695	493	292	30.039	10.401		695	40.56	12.47	
Prime (Mean/Max)					0.7277	0.8082	0.8023		0.7522	0.8806	0.8744
Average		46.33					0.6783	46.33			0.8237

[1]Shannon–Weaver diversity index.
[2]Nei's genetic diversity.
[3]Polymorphic information content (PowerMarker).
R.F., rare fragments; S.F., specific fragments.

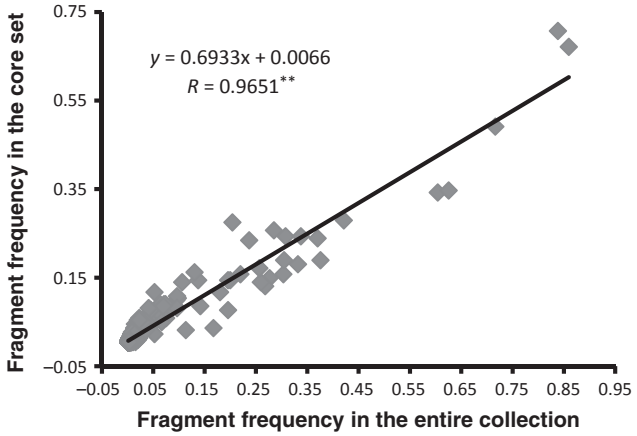

Fig. 1: Frequency distribution of the fragment size of the 31.48% core set (222 individuals) vs. the entire collection (705 individuals) using AFLP markers

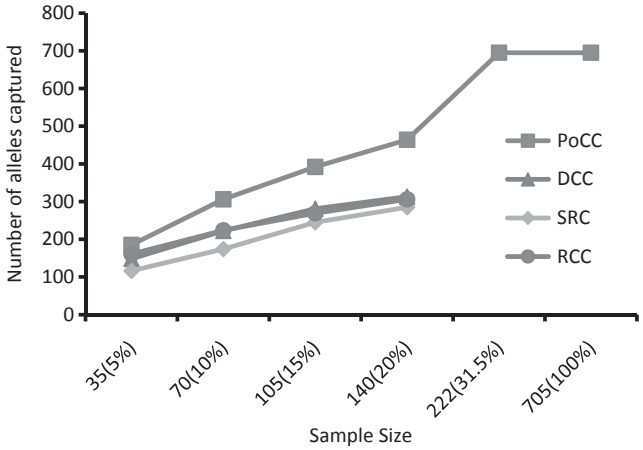

Fig. 2: Number of fragments captured with respect to accession sample size in four collection strategies. PowerCore (PoCC), distance-based core collection (DCC), stratified random core collection (SRC) and random core collections (RCC) represent the total number of fragments captured using a modified heuristic algorithm (PowerCore), distance-based core collection (dandrogram), stratified random collection (regional stratified) and RCC (random table) method, respectively

genetic diversity indices and polymorphic information content by AFLP fragments in all accessions (705) and in the core set (222) of mungbean are summarized in Table 2. All fragments (695) identified were conserved in the core set. A higher Shannon–Weaver diversity index (40.56), Nei's genetic diversity (12.47) and polymorphic information content (0.8237) were found in the core set. The frequency distribution of fragments recovered with the core set (222 accessions) was compared with the entire collection (705) of mungbean (Fig. 1). Most of the fragments were concentrated at a low frequency (< 10). When frequencies of all fragments of the core set were plotted against that of the entire collection, a highly significant correlation ($R = 0.96$) was observed.

Evaluating the allele capturing efficiency

The construction of a so-called 'core set' or 'allele-mining set' from a large germplasm collection is a situation, where allelic richness is a relevant measure of diversity (Schoen and Brown 1993, Bataillon et al. 1996), because as many alleles as possible should be retained in the allele-mining set, where they would be available for phenotypic screening and breeding programmes (Zhao et al. 2010a). The allele capturing efficiency of PowerCore was compared with the other strategies used for the development of core set such as distance-based core collection (DCC), SRC and RCC for different levels of sample size (Fig. 2). Core set developed by PoCC gave rise to higher efficiency than all other methods used in our study at all levels of sample sizes. PoCC captured 185, 306, 392 and 464 alleles, while DCC captured 148, 222, 281 and 313 alleles at 5%, 10%, 15% and 20% core sets, respectively. The lowest quantity of allele was captured by SRC (116, 174, 245 and 285), and RCC gave similar allele capturing efficiency (161, 224, 269 and 306) as that of DCC at respective levels of sample size.

Validation of the core set

To validate the heuristic approach, the same accessions in this study were assessed in a set of 15 SSR markers between the same entire collection and core sets. Statistics describing the allelic diversity of these 705 accessions for 15 SSR markers are summarized in Table 3; 66 alleles were detected with the 15 SSR markers in the entire collection. The number of alleles per locus ranged from 2 to 9 with an average of 4.4 in entire collection, but in 20% core set, it ranged from 2 to 6 with an average of 3.13 alleles per locus. For these 15 markers, PIC ranged from 0.0782 to 0.5424, with an average of 0.2945 for entire collections and ranged from 0.0540 to 0.8082, with an average of 0.2864 for 20% core set.

The correlation coefficients (r) of the mean diversity index between the core set and the entire collection were highly significant at 1% level for the Shannon–Weaver diversity index (I) in AFLP analysis and were significant at 5% level for Nei's genetic diversity (H) and PIC in AFLP analysis, for Shannon–Weaver diversity index (I) and Nei's genetic diversity index (H) in SSR analysis. There was no significant correlation between the entire collection and core set for PIC in SSR analysis (Table 4).

Table 5 summarizes the total number of DNA fragments and alleles detected for the two types of markers. For the second set of 15 SSR markers, 71% of the alleles observed in the 705 accessions were captured in the core set. There are four unique alleles and 16 rare ($< 5\%$ frequency) alleles in the second set of SSRs; all unique alleles were kept in the core set, but only 15 rare alleles were kept in the core set. One rare allele is lost because 20% (140) of accessions were selected according to AFLP result not by the PowerCore programme itself. Even if not all useful alleles were captured, the heuristic approach also does better than other sampling strategies (Table 5). Allele capturing efficiency was the highest in heuristic approach (PoCC), 0.23 (44% of total allele captured) at 10% core set and 0.3 (67% of total allele captured) at 20% core set.

Discussion

In mungbean, the range of genetic variability for characters of economic importance has been studied for large collections of accessions. In general, the variability is reported to be sufficiently extensive for progress in mungbean breeding programmes. As knowledge of genetic variability for a specific trait is highly valuable for the utilization of germplasm

Table 3: Total number of alleles, genetic diversity index and polymorphic information content (PIC) for 15 SSR loci in the entire accessions (705) and 20% (140) core set of mungbean

		Entire collection						Core collection			
SSR locus	Size range (bp)	Allele	Rare allele	Specific allele	I^1	H^2	PIC[3]	Allele	I^1	H^2	PIC[3]
GB-VR-7	270–310	9	4	1	0.335	0.165	0.1424	5	1.267	0.651	0.0980
GB-VR-13	154–181	2	0	0	0.482	0.304	0.2044	2	0.637	0.444	0.2071
GB-VR-14	251–256	2	0	0	0.688	0.497	0.3714	2	0.694	0.54	0.3697
GB-VR-17	145–163	4	0	0	0.158	0.066	0.2080	4	0.557	0.29	0.2622
GB-VR-38	124–142	5	3	1	0.414	0.253	0.1311	2	0.787	0.475	0.0787
GB-VR-77	301–313	3	1	0	0.697	0.497	0.3737	3	0.787	0.475	0.3716
GB-VR-87	263–275	2	0	0	0.587	0.398	0.3216	2	0.668	0.475	0.3318
GB-VR-91	151–167	9	1	0	1.114	0.536	0.4987	6	1.985	0.827	0.4876
GB-VR-93	110–125	6	0	0	1.149	0.593	0.5424	5	1.533	0.741	0.6082
GB-VR-113	147–241	5	1	0	0.239	0.099	0.0782	4	0.761	0.377	0.0708
GB-VR-142	224–260	6	1	0	1.058	0.57	0.4491	3	1.565	0.753	0.4132
GB-VR-172	226–235	4	2	1	0.492	0.307	0.2582	2	0.787	0.475	0.2954
GB-VR-180	253–295	3	1	1	0.296	0.174	0.1465	2	0.396	0.293	0.0540
GB-VR-184	271–279	2	0	0	0.58	0.405	0.3119	2	0.667	0.515	0.2759
GB-VR-198	223–303	4	2	0	0.719	0.504	0.3805	3	0.978	0.543	0.3716
Total		66	16	4	9.008	5.368		47	14.069	7.874	
Prime (Mean/Max)					0.523	0.603	0.5430		0.473	0.635	0.4708
Average		4.4					0.2945	3.13			0.2864

[1]Shannon–Weaver diversity index.
[2]Nei's genetic diversity.
[3]Polymorphic information content (PowerMarker).

Table 4: *T*-test results between the entire collection and the core set

Source of variables	Mean	Standard deviation	Different mean	Different Standard deviation	*t*-value	P (two tail)	*r*
AFLP							
I^1							
Entire collection	2.0026	0.6114	0.7014	0.0103	3.1150	0.0040	0.9547**
Core set	2.7040	0.6217					
H^2							
Entire collection	0.6934	0.2022	0.1379	0.0640	2.1810	0.0388	0.9459*
Core set	0.8313	0.1382					
PIC							
Entire collection	0.6783	0.1974	0.1454	0.0573	2.3280	0.0283	0.9460*
Core set	0.8237	0.1400					
SSR							
I^1							
Entire collection	0.6005	0.1820	0.3374	0.0054	2.4018	0.0240	0.7814*
Core set	0.9379	0.1766					
H^2							
Entire collection	0.3579	0.1718	0.1671	0.0044	2.7247	0.0109	0.7275*
Core set	0.5249	0.1674					
PIC							
Entire collection	0.2945	0.1634	0.0081	0.0064	2.1447	0.4572	0.7606[ns]
Core set	0.2864	0.1570					

[ns]Correlation is not significant, *Correlation is significant at 0.05 level, **Correlation is significant at 0.01 level.
[1]Shannon–Weaver diversity index.
[2]Nei's genetic diversity.
PIC, Polymorphic information content.

resources in breeding programmes, a set of descriptors was prepared for mungbean. In recent years, several markers [e.g., AFLP, restriction fragment length polymorphism (Young et al. 1992), random amplified polymorphic DNA and SSR markers for mungbean (Lakhanpaul et al. 2000, Kumar et al. 2002, Betal et al. 2004, Sangiri et al. 2007, Moe et al. 2010] have been used for genetic diversity analysis and verification.

However, our basic knowledge of the extent of allelic variation within this species is still insufficient. Considering the huge number of accessions held collectively in gene banks, the germplasm collections are thought to harbour a wealth of hidden allelic variants. An allele-mining set is essential, especially if the gene pool is so large that conducting controlled genetic crosses is not feasible to determine the function of each allele. Developing such an allele-mining set has been proposed as a means of increasing the economical use of germplasm (Frankel 1984). Brown et al. (1987) recommended that the number of collections in the core set should account for 5–10% of the base collection, and that the core set should represent at least 70% of the genetic diversity in the base collection. Diwan et al. (1995) suggested that the core set sampling should always be >10%, while van Hintum (1995) suggested that the sampling proportion should depend on the particular objective of the core set and should be 5–20% of the base collection.

Although the correlation coefficient of the allelic frequency distribution of the entire and the allele-mining set was highly

Table 5: Capturing total number and proportion of alleles (fragments) in the same entire accessions and allele-mining set by two types of markers

Collection set of markers	Entire collection	PoCC[1]			DCC[2]		SRC[3]		RCC[4]	
Accessions	705	222	140	70	140	70	140	70	140	70
AFLP	695	695	464	306	313	222	285	174	306	224
Percentage of total[5] (%)	100	100	67	44	45	32	41	25	44	32
Capturing efficiency[6]	One fragment per 1.01	0.32	0.30	0.23	0.45	0.32	0.49	0.40	0.46	0.31
15 SSRs	66	66	47		36		29		34	
Percentage of total (%)	100	100	71		55		44		52	
Capturing efficiency	One allele per 10.68	3.36	2.98		3.89		4.83		4.12	

[1]Allele-mining set constructed using heuristic approach (PowerCore).
[2]Allele-mining set constructed using distance-based core set (PowerMarker).
[3]Allele-mining set constructed using stratified random sampling approach (origin strata).
[4]Allele-mining set constructed using random sampling approach (random table).
[5]Percentage of total alleles (fragments) captured in the core set/alleles (fragments) in the entire accessions.
[6]The accession number of capturing one allele (fragment).
DCC, distance-based core collection; PoCC, PowerCore; RCC, random core collections; SRC, stratified random core collection.

significant (Fig. 2) and an allele-mining set of 31.49% of accessions by AFLP can cover 100% of fragments captured and diversity as in entire collections, the proportion of accession was still high. So, we reduced to 20, 15, 10 and 5% of total accessions and compared the allele capturing efficiency of PowerCore, against other strategies. The result demonstrated the higher efficiency of PowerCore at any level of sample size collected (Fig. 2). Furthermore, the high proportion (493, 70.93%) of rare fragments (< 5% frequency) and specific fragments (292, 30.26%) found in our sample indicates that conversely, many informative alleles remain to be mined in our mungbean collections (Table 2). The comparatively lower proportion (16, 28.07%) of rare alleles and specific alleles (4, 7.01%) found in our sample indicates that less informative alleles could be discerned in mungbean collections using SSR than AFLP markers. It is because of less polymorphism of SSR markers. In general, the number of accessions retained in the allele-mining set depends on the polymorphic efficiency of the markers used. To retain maximum genetic diversity in the allele-mining set, an increase in the number of markers, especially the allele number, will be needed and the size of the allele-mining set will also increase correspondingly (Zhao et al. 2010a).

Although highly significant correlation ($r = 0.95$) was found between the entire and the allele-mining set for the Shannon–Weaver diversity index with AFLP, other indices were significant only at 0.05 level for both AFLP and SSR (Table 4). The total genetic diversity revealed by the Shannon–Weaver diversity index (I), Nei's gene diversity (H) and PIC was higher in the allele-mining set than in the entire collection because they were of unequal size (Tables 2 and 3). Therefore, the use of indices such as I and H may be disputed (Hennink and Zeven 1991).

Hennink and Zeven (1991) proposed relative indices, defined as $H' = H_{mean}/H_{max}$ and $I' = I_{mean}/I_{max}$, respectively. In comparison, we found that H', I, and PIC′ ($= PIC_{mean}/PIC_{max}$) in the allele-mining set were similar to those for the entire set, indicating that these indices of genetic diversity can be better used as parameters to evaluate the quality of the allele-mining set.

To devise plant breeding strategies for crop improvement, a breeder would ideally know the relative value of all alleles for genes of interest in the primary germplasm, an unlikely prospect. However, information can be gathered by establishing the allele-mining set (Varshney et al. 2005). So, the development of an allele-mining set, which represents the genetic diversity of a crop with minimal redundancy and increases utility of the collection as a whole, is especially important as the funding for germplasm collections decreases (Marita et al. 2000). Many core sets were successfully developed after Frankel proposed the theory of the core set in 1984, but the selection of an appropriate sampling strategy is still important in the construction of a core set. Given the nature of the allele-mining set, it is impossible to guarantee the complete capture of all useful alleles (McKhann et al. 2004). However, the core set using the heuristic approach here eliminated the redundancy in the mungbean collection. We successfully developed an allele-mining set using genotype data and a heuristic approach with least redundancy in mungbean collection. Our present report supported the efficiency of PowerCore in the development of core set for large germplasm collections.

Acknowledgements

This work was carried out with the support of Cooperative Research Program for Agriculture Science and Technology Development (Project No. 200908FHT020609001), Rural Development Administration, Korea.

References

Anishetty, N. M., and H. Moss, 1988: Vigna genetic resources: Current status and future plans. Mungbean: Proceedings of the Second International Symposium. Asian Vegetable Research and Development Center, Shanhua, Taiwan, 13—18.

Bailey, L. H. 1970: Manual of Cultivated Plants. (Rev.) MacMillan, New York, 1116.

Bataillon, T. L., J. L. David, and D. J. Schoen, 1996: Neutral genetic markers and conservation genetics: simulated germplasm collections. Genetics **144,** 409—417.

Betal, S., P. R. Chowdhury, S. Kundu, and S. S. Raychaudhuri, 2004: Estimation of genetic variability of Vigna radiate cultivars by RAPD Analysis. Biol. Plant. **48,** 205—209.

Bose, R. D., 1939: Studies in Indian pulses. IX. Contributions to the genetics of mung (Phaseolus radiates Linn., phaureus Roxb.). Indian J. Agr. Sci. **9,** 575—594.

Botstein, D., R. L. White, M. Skolnick, and R. W. Davis, 1980: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. **32,** 314—331.

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

Brown, A. H. D., J. P. Grace, and S. S. Speer, 1987: Designation of a core collection of perennial glycine. Soyb. Genet. Newsl. **14,** 59—70.

Chandra, S., Z. Huaman, S. Hari Krishna, and R. Ortiz, 2002: Optimal sampling strategy and core collection size of Andean tetraploid potato based on isozyme data – a simulation study. Theor. Appl. Genet. **104,** 1325—1334.

Chung, J. W., and Y. J. Park, 2010: Population structure analysis reveals the maintenance of isolated sub-populations of weedy rice. Weed Res. **50,** 606—620.

Chung, H. K., K. W. Kim, J. W. Chung, J. R. Lee, S. Y. Lee, A. Dixit, H. K. Kang, W. Zhao, K. L. McNally, R. S. Hamilton, J. G. Gwag, and Y. J. Park, 2009: Development of a core set from a large rice collection using a modified heuristic algorithm to retain maximum diversity. J. Integr. Plant Biol. **51,** 1116—1125.

De Candole, A., 1886: Origin of Cultivated Plants. Hafner Publ. Co., New York, N.Y. (Reprint of 2nd edn 1959).

Diwan, N., M. S. McIntosh, and G. R. Bauchan, 1995: Methods of developing a core collection of annual Medicago species. Theor. Appl. Genet. **90,** 755—761.

Frankel, O. H., 1984: Genetic perspectives of germplasm conservation. In: W. Arber, K. limensee, W. J. Peacock, and P. Starlinger (eds), Genetic Manipulation: Impact on Man and Society, 161—171. Cambridge University Press, Cambridge.

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

Gouesnard, B., T. M. Bataillon, G. Decoux, C. Rozale, D. J. Schoen, and J. L. David, 2001: Mstrat: an algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J. Hered. **92,** 93—94.

Gwag, J. G., A. Dixit, Y. J. Park, K. H. Ma, S. J. Kwon, G. T. Cho, G. A. Lee, S. Y. Lee, H. K. Kang, and S. H Lee, 2010: Assessment of genetic diversity and population structure in mungbean. Genes Genomics **32,** 299—308.

Hennink, S., and A. C. Zeven, 1991: The interpretation of Nei and Shannon-Weaver within population variation indices. Euphytica **51,** 235—240.

van Hintum, T. J. L., 1995: Hierarchical approaches to the analysis of genetic diversity in crop plants. In: T. Hodgkin, A. H. D. Brown, and T. J. L. van Hintum (eds), Core Collections of Plant Genetic Resources, 23—34. Wiley, Chichester.

Hokanson, S. C., A. K. Szewc-McFadden, W. F. Lamboy, and J. R. McFerson, 1998: Microsatellite (SSR) markers reveal genetic identities, genetic diversity and relationships in a *Malus domestica* borkh core subset collection. Theor. Appl. Genet. **97,** 671—683.

Holden, J. H. W., 1984: The second ten years. In: J. H. W. Holden, and J. T. Williams (eds), Crop Genetic Resources: Conservation and Evaluation, 277—285. Allen and Unwin, Winchester.

Hu, J., J. Zhu, and H. M. Xu, 2000: Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotype values of crops. Theor. Appl. Genet. **101,** 264—268.

Joe, T., and G. D. Orlando, 1996: AFLP analysis of gene pools of a wild bean core collection. Crop Sci. **36,** 1375—1384.

Kang, C. W., S. Y. Kim, S. W. Lee, P. N. Mathur, T. Hodgkin, M. D. Zhou, and J. R. Lee, 2006: Selection of a core collection of Korean sesame germplasm by a stepwise clustering method. Breed. Sci. **56,** 85—91.

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

Kumar, S. V., G. Tan, S. C. Quah, and K. Yusoff, 2002: Isolation of microsatellite markers in mungbean, *Vigna radiate*. Mol. Ecol. Notes **2,** 96—98.

Lakhanpaul, S., C. Sonia, K. V. Bhat, and S. Chadh, 2000: Random amplified polymorphic DNA (RAPD) analysis in Indian mungbean (*Vigna radiata* L. Wilczek) cultivars. Genetica **109,** 227—234.

Latha, R., L. Rubia, J. Bennett, and M. S. Swaminathan, 2004: Allele mining for stress tolerance genes in Oryza species and related germplasm. Mol. Biotechnol. **27,** 101—108.

Liu, K., and S. V. Muse, 2005: PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics **21,** 2128—2129.

Marita, J. M., J. M. Rodriguez, and J. Nienhuis, 2000: Development of an algorithm identifying maximally diverse core collections. Genet. Resour. Crop Evol. **47,** 515—526.

McKhann, H. I., C. Camilleri, A. Berard, T. Bataillon, J. L. David, X. Reboud, V. L. Corre, C. Caloustian, I. G. Gut, and D. Brunel, 2004: Nested core collections maximizing genetic diversity in *Arabidopsis thaliana*. Plant J. **38,** 193—202.

Moe, K. T., J. W. Chung, Y. I. Cho, J. K. Moon, J. H. Ku, J. K. Jung, J. Lee, and Y. J. Park, 2010: Sequence information on simple sequence repeats and single nucleotide polymorphisms through transcriptome analysis of mungbean. J. Integr. Plant Biol. **53,** 63—73.

Nei, M., 1973: Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70,** 3321—3323.

Ortiz, R., E. N. Ruiz-Tapia, and A. Mujica-Sanchez, 1998: Sampling strategy for a core collection of Peruvian quinoa germplasm. Theor. Appl. Genet. **96,** 475—483.

Perry, M. C., M. S. Mclntosh, and A. K. Stoner, 1991: Geographical patterns of variation in the USDA soybean germplasm collection: II. allozyme frequencies. Crop Sci. **31,** 1356—1360.

Poehlman, J. M., 1991: The Mungbean. Oxford and IBH Publishing Co., New Delhi.

Rogers, J. S., 1972: Measures of genetic similarity and genetic distance. Stud. Genet. VII Univ. Tex Publ. 7213, 145—153.

Sangiri, C., A. Kaga, N. Tomooka, D. Vaughan, and P. Srinives, 2007: Genetic diversity of the mungbean (*Vigna radiata*, Leguminosae) gene pool on the basis of microsatellite analysis. Aust. J. Bot. **55,** 837—847.

Schoen, D. J., and A. H. D. Brown, 1993: Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc. Natl. Acad. Sci. USA **90,** 10623—10627.

Schuelke, M., 2000: An economic method for the fluorescent labeling of PCR fragments. Nat. Biotechnol. **18,** 233—234.

Shanmugasundaram, S., J. D. H. Keatinge, and J. A. Hughes, 2009: The mungbean transformation, diversifying crops, defeating malnutrition. IFPRI Discussion Paper 00922 on project, 2020 Vision Initiative, supported by the CGIAR.

Shannon, C. E., and W. Weaver, 1949: The mathematical Theory of Communication. University of Illinois Press, Urbana.

Tangphatsornruang, S., D. Sangsrakru, J. Chanprasert, P. Uthaipaisanwong, T. Yoocha, N. Jomchai, and S. Tragoonrung, 2010: The chloroplast genome sequence of mungbean (*Vigna radiata*) determined by high-throughput pyrosequencing: structural organization and phylogenetic relationships. DNA Res. **17,** 11—22.

Upadhyaya, H. D., and R. Ortiz, 2001: A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theor. Appl. Genet. **102,** 1292—1298.

Upadhyaya, H. D., P. J. Bramel, R. Ortiz, and S. Singh, 2002: Developing a mini core of peanut for utilization of genetic resources. Crop Sci. **42,** 2150—2156.

Varshney, R. K., G. A. Andreas, and M. E. Sorrells, 2005: Genomics-assisted breeding for crop improvement. Trends Plant Sci. **10,** 621—630.

Vavilov, N. I., 1951: The origin, variation, immunity, and breeding of cultivated plants (Translation by K.S. Chester). Chron. Bot. **13,** 1—364.

Vos, P., R. Hogers, M. Bleeker, M. Reijians, T. Van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau, 1995: AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23,** 4407—4414.

Young, N. D., L. Kumar, D. Menancio-Hautea, D. Danesh, N. S. Talekar, S. Shanmugasundaram, and D. H. Kim, 1992: RFLP

mapping of major bruchid resistance gene in mungbean (*Vigna radiata*). Theor. Appl. Genet. **84,** 839—844.

Zhang, X. R., Y. Z. Zhang, Y. Cheng, F. Xiangyun, G. Qingyuan, Z. Mingde, and T. Hodgkin, 2000: Establishment of sesame germplasm core collection in China. Genet. Resour. Crop Evol. **47,** 273—279.

Zhao, W. G., G. T. Cho, K. H. Ma, J. W. Chung, J. G. Gwag, and Y. J. Park, 2010a: Development of an allele-mining set in rice using a heuristic algorithm and SSR genotype data with least redundancy for the post-genomic era. Mol. Breed. **26,** 639—651.

Zhao, W. G., J. W. Chung, G. A. Lee, K. H. Ma, H. H. Kim, K. T. Kim, I. M. Chung, J. K. Lee, N. S. Kim, S. M. Kim, and Y. J. Park, 2010b: Molecular genetic diversity and population structure of a selected core set in garlic and its relatives using novel SSR markers. Plant Breed. **130,** 46—54.

Zhao, W. G., J. W. Chung, Y. I. Cho, W. H. Rha, G. A. Lee, K. H. Ma, S. H. Han, K. H. Bang, C. B. Park, S. M. Kim, and Y. J. Park, 2010c: Molecular genetic diversity and population structure in Lycium accessions using SSR markers. C. R. Biol. **333,** 793—800.

Zhukovsky, P. M., 1950: Cultivated plants and their wild relatives (Translation by Hudson, P.S., 1962: Common wealth Agricultural Bureau, Farnham Royal, England, 107).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Dendrogram for 705 mungbean accessions by Ward minimum variance clustering method using AFLP markers; blue colour represents 5% core collection, red represents 10% core collection, blue + red represents 15% core collection and blue + red + green represents 20% core collection.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.