Use of Heuristic Approach for the Development of a Core Set from Large Germplasm Collection of Foxtail Millet (*Setaria italica* L.)

Jayarama Gowda[1], M Krishanappa[1], Niti Pathak[2,3], PN Mathur[3*] and A Seetharam[1]

[1]*All India Coordinated Small Millet Improvement Project, Gandhi Krishi Vignyan Kendra Campus, University of Agricultural Sciences, Bangalore-560 065, Karnataka*
[2]*National Research Centre on DNA Fingerprinting, NBPGR, Pusa Campus, New Delhi-110 012*
[3]*Bioversity International, Office for South Asia, NASC Complex, Pusa Campus, New Delhi-110 012*

(Received: 24 May 2012; Revised: 5 October 2012; Accepted: 12 October 2012)

Foxtail millet, (*Setaria italica* L.) is an ancient crop and important minor cereal. It has a long history of cultivation in India and possesses rich genetic diversity conserved in genebanks. Large germplasm collection in genebanks limit accessibility of this crop in crop breeding. It is, therefore, important to develop core sets of germplasm from large collection for conserving maximum diversity of this crop. “PowerCore” is a new approach that simplifies the generation process of a core set by significantly cutting down the number of core entries but maintaining maximum diversity. Here, we have developed a core set from 1,482 accessions of foxtail millet using data of 25 quantitative and qualitative descriptors. The newly formed core set has 59 accessions which represents 4% of the total collection. This heuristic method has proved to be an efficient tool for developing core sets from collections of unequal diversity and differentiation

Key Words: Characterization, Core collection, Foxtail millet, Minor cereal, PowerCore

Introduction

Foxtail millet [*Setaria italica* (L.) Beauv.], a member of tribe Panicoideae, is one of the oldest crops cultivated for hay, pasture and food grain. It is an important crop in China besides India, Japan and other countries in Asia and Europe. It is also grown in North and South America, Australia and Africa as a grass. It was grown fairly extensively with its cultivation extending from temperate Eurasia to tropics and sub-tropics of Asia. Presently, in India, the crop is cultivated on a very limited area in sporadic patches in many states throughout the country. Known for its drought tolerance, foxtail millet can withstand severe moisture stress and can adjust to wide range of soil condition. At the All India Coordinated Small Millet Improvement Programme (AICSMIP) at Bangalore, India, the foxtail millet collections exceeding around 1,500 accessions have been assembled and maintained, there represent good diversity from various regions within and outside the country.

Increasing size of germplasm collection in most genebanks, limits their accessibility for use in crop breeding and their subsequent management. In addition, redundant resources have become an obstacle to the effective management and utilization of these resources. Therefore, it has been proposed that a limited set of accessions be selected capturing as much genetic diversity as possible to offer a good starting point when searching for new traits and could be used for in-depth evaluation, thus increasing the knowledge of the entire collection (Knüpffer and van Hintum, 1995). Core collections improve the management and effective use of plant genetic resources (Brown, 1989). A core collection is a subset of a large germplasm collection that contains chosen accessions capturing most of the genetic variability of the whole genebank. Such a core subset provides a proper working collection for the extensive searching of desired alleles and a point of entry to the entire germplasm collection (Holbrook and Anderson, 1995; Dussert *et al.,* 1997).

Since core sets are derived from the wide spectrum of genetic diversity of the whole collection, most of this diversity is expected to be retained. The sampling strategies for choosing core entries can be divided into two approaches, simple random sampling and stratified random sampling (Brown, 1989a and 1989b; Spagnoletti Zeuli and Qualset, 1993). The most common method employed in obtaining a core collection of desired size is

[*] *Author for Correspondence: E-mail: p.mathur@cgiar.org*

that of stratified random sampling (Peeters and Martinelli, 1989; Chandra et al., 2002). Several strategies have also been suggested for determining the appropriate sampling fraction from each group or strata. These methods include proportional, log frequency and constant allocations (Brown, 1989a; Spagnoletti Zeuli and Qualset, 1993).

Though many of these methods have been used successfully in obtaining core sets of desired size, inequality in diversity of accessions and differentiation of stored accessions has been a recurring problem that skews the subset population. This problem has led to several attempts to build germplasm core collections through the maximization of allelic or phenotypic richness. More recently, the Rural Development Administration (RDA), South Korea, has used the advanced M (maximum) strategy with a heuristic search for establishing core sets and accordingly a programme known as "PowerCore" (http://genebank.rda.go.kr/powercore/) has been developed which has been used in the present study. In this paper, we discuss the development of a core set from a large germplasm collection of 1,482 accessions of foxtail millet using Heuristic Core Collection (HCC) program. The HCC program employs the advanced M-strategy using a modified heuristic algorithm. It creates subsets representing all alleles or observations classes, with the least allelic redundancy, and ensures a highly reproducible list of entries. Further, the heuristic strategy is compared with cluster analysis using Ward's method (Ward, 1963). This approach has already been used in developing core set from large rice collection (Chung *et al.,* 2009).

Thus, in establishing a good core collection it is very important to chose best sampling strategy that represents largest diversity of the entire collection. Earlier cluster analysis has been widely used as an important tool to group accessions for constructing core collection. Recent HCC programme 9 uses the advance M-strategy. This modified heuristic algorithm can be applied for the selection of gentype data (allelic richness), the reduction of redundancy and the development of a more extensive analysis in the management and utilization of large collection of plant genetic resources.

Materials and Methods

The Indian foxtail millet germplasm collection used for developing core subset consisted of 1,482 accessions. All accessions were planted during *kharif* season (July-November) in year 2001 in the experimental field at the Main Research Station, GKVK, Bangalore, which is situated at 13^0N latitude and $77^0$35'E longitude at an altitude of 890 m. The average annual rainfall is around 850 mm; which is mostly received during July to October. The soil of the experimental field was red sandy loam with an acidic pH of 5.5. The soil is low in organic carbon (0.4%) with moderate availability of nitrogen (300 kg/ha) and phosphorus (185 kg/ha) and fairly rich in potash (225 kg/ha). The material was planted using an Augmented Block Design (Federer, 1956) using three standard checks.

The passport information for all the 1,482 accessions was published in a catalogue on "Evaluation of foxtail millet (*Setaria italica*) germplasm (Gowda *et al.,* 2002)". The germplasm collections represent diversity from 15 states of India which includes: Andhra Pradesh (164), Bihar (57), Gujarat (11), Himachal Pradesh (8), Jammu and Kashmir (12), Karnataka (139), Kerala (7), Madhya Pradesh (34), Maharashtra (3), Orissa (6), Punjab (10), Rajasthan (2), Tamil Nadu (49), Uttar Pradesh (505) and West Bengal (25). In addition to diversity from India, these collection also include exotic diversity from Africa (8), Bangladesh (3), China (25), Pakistan (1), Turkey (91), USA (39) and USSR (2). There were 373 accessions for which the passport information was not known.

All the accessions were characterised for 25 important traits as outlined by IPGRI (1985). Eleven traits were measured on a quantitative scale including days to 50% flowering, plant height, of basal tillers, flag leaf length, flag leaf width, peduncle length, ear length, panicle exertion, days to maturity, grain yield per plant and 1,000-grain weight. Fourteen qualitative characters included plant pigmentation at flowering, leaf colour, blade pubescence, sheath pubescence, degree of lodging at maturity, senescence, inflorescence lobes, inflorescence bristles, lobe compactness, inflorescence shape, inflorescence compactness, fruit colour, grain shape and apical sterility in panicle. The accession level information has been published by the All India Coordinated Small Millet Improvement Programme (AICSMIP), Bangalore, India (Gowda *et al.,* 2002).

"Power Core" was used for developing core collection, which is a program that applies the advanced M-strategy with a heuristic search for establishing core or allele mining sets and thus possesses the power to represent all alleles or classes. It effectively simplifies the generation process of a core set while significantly cutting down the number of core entries, maintaining 100% of the diversity as categorical variables. The validation of core sets using "PowerCore" has also been compared with the core developed using the traditional clustering method for the same number of accessions using Wards clustering method (Ward, 1963) and the distance used was Euclidean.

Core collections are considered to well represent the genetic diversity of the initial collection if the following two criteria are met: (1) no more than 20% of the traits had different means (significant at α= 0.05) between the core collection and the initial collection and (2) Coincidence Rate (CR) was retained by the core collection in no less than 80% of the traits (Hu *et al.,* 2000). The design concept and implementation strategy of 'PowerCore" and the validation on the outcome in comparison with other methods has been well described by (Kim *et al.,* 2007). "PowerCore" by default classified the continuous variables into different categories based on Sturges rule (Sturges, 1926), which is described as: $K= 1 + \log_2 n$, where n = number of observed accessions. However, the software also allows modification of this rule to make desired number of classes for the continuous variables. Once classification of the continuous variables is performed, the software takes into account all classes, without omission of any of its variables. It thus, possesses the capability to cover all the distribution ranges of each class.

Results and Discussion

"PowerCore" successfully selected 59 accessions (Table 1) from the entire collection of 1,482 accessions, an even distribution of characters was observed in the core set when classified by this method. Most of these core collections were from Uttar Pradesh (28) followed by 5 accessions each from Andhra Pradesh and Karnataka, thereby indicating that the collections from Uttar Pradesh have more diversity than these from other parts of the country. In the present study, maximum possible number of classes for each variable was used in order to capture 100% diversity for each unique class in a core set. The core size was from 4% of the total collection. For the validation, the following statistical parameters were analysed to compare the mean and variance ratio between core and entire collections and are presented in Table 2.

Table 1. List of core accessions identified based on 25 variables

GS20	GS374	GS820	GS1315	GS1954
GS27	GS401	GS848	GS1372	GS2025
GS77	GS545	GS872	GS1377	GS2029
GS78	GS617	GS887	GS1388	GS2035
GS84	GS639	GS889	GS1404	GS2040
GS105	GS678	GS890	GS1430	GS2076
GS160	GS717	GS892	GS1500	GS2096
GS260	GS736	GS900	GS1507	GS2100
GS275	GS739	GS958	GS1618	GS2164
GS338	GS764	GS963	GS1657	GS2248
GS364	GS766	GS1242	GS1794	GS2258
GS372	GS809	GS1308	GS1929	

Following four statistical parameters were analysed using "PowerCore" to compare the mean and variance ratio between core and entire collections. The percentage of the significant difference between the core sets and the entire collection was calculated for the mean difference percentage (MD, %) and the variance difference percentage (VD, %) of traits. Coincidence rate (CR, %) and variable range (VR, %) were designed to evaluate the properties of the core set against the entire collection (Hu *et al.,* 2000).

Mean Difference Percentage (MD %) – which is estimated as:

$$\text{MD (\%)} = \frac{1}{m}\sum_{j=1}^{m}\frac{Me - Mc}{Mc}\,\text{X}\,100$$

Where, M_e = Mean of entire collection; M_c = Mean of core collection, and m = number of traits.

Variance Difference (VD %) – which is estimated as:

$$\text{VD (\%)} = \frac{1}{m}\sum_{j=1}^{m}\frac{Ve - Vc}{Vc}\,\text{X}\,100$$

Where, V_e = Variance of entire collection, V_c = Variance of core collection, and m = number of traits.

Confidence ratio (CR %) – which is estimated as:

$$\text{CR (\%)} = \frac{1}{m}\sum_{j=1}^{m}\frac{Rc}{Re}\,\text{X}\,100$$

Where, R_e = Range of entire collection, R_c = Range of core collection, and m = number of traits.

CR% indicates whether the distribution ranges of each variable in the core set are well represented when

Table 2. Comparative description of mean and variance components between entire and core accessions developed using "PowerCore" and clustering approach

Statistical parameters	Days to 50 percent flowering			Plant height (cm)			No. of basal tillers			Flag leaf length (cm)		
	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)
No. of accessions	1482	59	288	1482	59	288	1482	59	288	1482	59	288
Mean	49.25	48.51	49.38	142.40	133.42	142.56	3.91	4.03	3.89	27.68	29.51	27.94
Minimum	33	35	33	52.20	52.20	91.40	1	1	1	16	17	16
Maximum	69	69	69	184	174	184	12.20	12	12.20	47.50	45.25	47.50
Range	36	34	36	131.80	121.80	92.60	11.20	11	11.20	31.50	28.25	31.50
CV%	7.57	12.57	9.05	10.39	18.67	9.86	36.18	57.87	43.70	19.91	23.91	22.73

Statistical parameters	Flag leaf width (cm)			Peduncle length (cm)			Ear length (cm)			Panicle exertion		
	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)
Number of accessions	1482	59	288	1482	59	288	1482	59	288	1482	59	288
Mean	1.51	1.66	1.52	28.05	29.91	28.04	14.76	15	14.92	13.31	13.86	13.23
Minimum	0.78	0.85	0.85	13.40	15.25	13.40	3.60	3.60	3.60	1.50	2.70	1.50
Maximum	4.40	4.40	3.50	56.50	53	56.50	24.50	23.80	24.50	29	29	29
Range	3.62	3.55	2.65	43.10	37.75	43.10	20.90	20.20	20.90	27.5	26.3	27.50
CV%	18.57	35.64	19.08	15.81	25.59	19.37	16.57	21.07	20.71	24.89	36.38	29.78

Statistical parameters	Days to maturity			Grain yield per plant (g)			Thousand grain weight (g)		
	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)	Entire	Core (PowerCore)	Core (Cluster approach)
Number of accessions	1482	59	288	1482	59	288	1482	59	288
Mean	90.79	88.88	90.93	9.74	10.29	10.06	3.06	3.05	3.02
Minimum	72	72	72.00	2.10	2.10	2.10	1.90	2	2
Maximum	110	110	110	23.80	23.80	20.20	4	3.90	4
Range	38	38	38	21.70	21.70	18.10	2.10	1.90	2
CV%	7.40	8.93	7.81	39.98	43	42.25	11.86	14.30	12.91

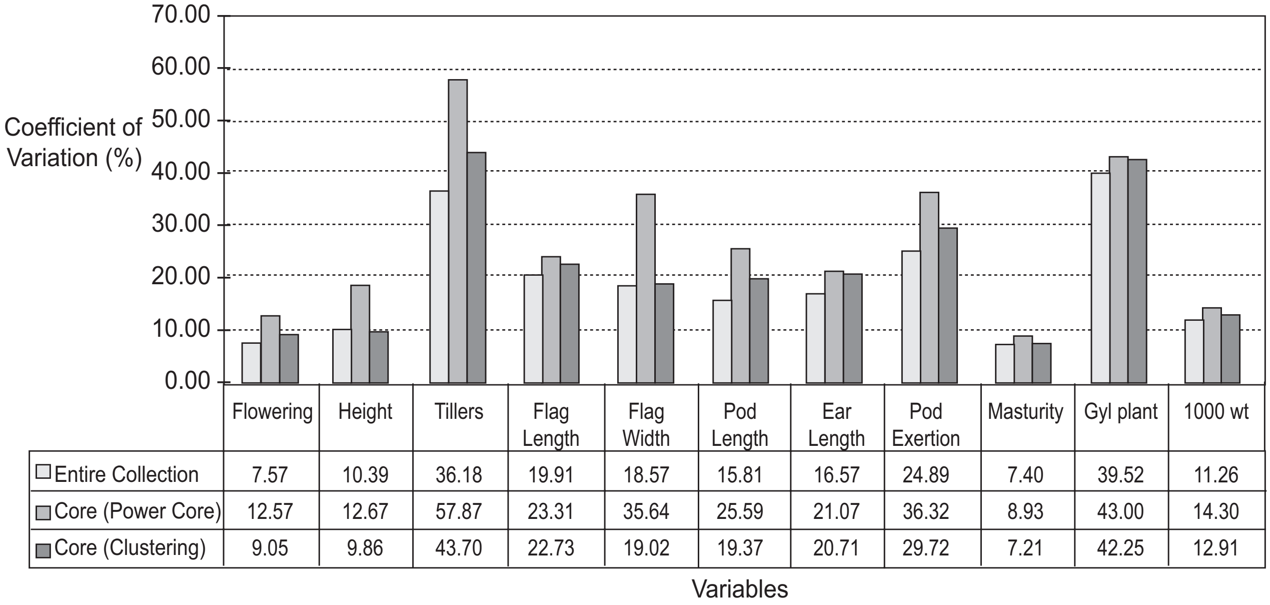
compared to the entire collection.

Variable Rate (VR %) – which is estimated as:

$$CR\ (\%)\ =\frac{1}{m}\sum_{j=1}^{m}\frac{Rc}{Re}\text{X}\ 100$$

Where, CV_e = Coefficient of variation of entire collection, CV_c = Coefficient of variation of core collection, and m = number of traits. VR% allows a comparison between the coefficient of variation values existing in the core collections and the entire collections and determines how well it is being represented in the core sets.

The results showed that there was no significant difference (α=0.05) for the means of all traits between core and entire collections. The estimated values for MD% was -2.2, which indicated that there is no difference in the mean values of entire and core collections. VD% was estimated to be 30.56, the VD values indicated that the variance for the entire and the core populations are not the same. The CR% obtained was 94.83, which indicated that the core has captured all accessions from all the classes and, thus, is a representative of the entire collection. High VR % (71.04) indicated that the coefficient of variation in the core set is higher compared to entire collections for all the variables.



	Flowering	Height	Tillers	Flag Length	Flag Width	Pod Length	Ear Length	Pod Exertion	Masturity	Gyl plant	1000 wt
□ Entire Collection	7.57	10.39	36.18	19.91	18.57	15.81	16.57	24.89	7.40	39.52	11.26
▨ Core (Power Core)	12.57	12.67	57.87	23.31	35.64	25.59	21.07	36.32	8.93	43.00	14.30
■ Core (Clustering)	9.05	9.86	43.70	22.73	19.02	19.37	20.71	29.72	7.21	42.25	12.91

Fig. 1. Comparison of CV between entire collection, core collection developed using PowerCore and clustering approach (Wards method) for quantitative variables

In order to compare the efficiency of "PowerCore" for developing core collection over other clustering method, mean and statistical parameters for entire population, core developed using "PowerCore" and core developed using clustering method were compared. This indicated that the mean components (mean, minimum, maximum and range) are the same for the two core sets developed, except for plant height where, core developed by clustering method does not include all the classes (Table 2). The only difference between these two core sets was that the "PowerCore" core was based on only 4% of the entire population, whereas, the clustering core was based on 15.64% of entire collection. The variance in core developed using clustering method was either close to the entire collections or was higher for all the descriptors, but was never higher to the core developed using "PowerCore". The Coefficient of Variation was also either close or higher to the entire collection for all descriptors for core developed based on clustering method, but never higher to core developed using "PowerCore" for these descriptors. Histogram comparing CV for the entire and core sets is shown in Fig. 1. High value obtained for CR % (94.83) suggests that the core attained using the heuristic approach method could be adopted as a representative of the whole collection.

A comparison of Shannon-Weiner (Shannon and Weaver, 1949) diversity index for the entire collection, core developed using "PowerCore" and core developed using clustering method also indicated a high diversity for all the qualitative traits in core developed using "PowerCore" compared to core developed using clustering approach, except for a few variables, where it was observed at par (Table 3).

Table 3. Comparison of Shannon-weiner diversity index values for entire collection, core collection developed using "PowerCore" and clustering method (Wards method)

Variables	Core collections (powercore)	Core collections (cluster approach)	Entire collections
Plant pigmentation at flowering	0.505	0.311	0.379
Leaf colour	1.102	0.756	0.761
Blade pubescence	0.909	0.728	0.714
Sheath pubescence	0.341	0.064	0.045
Degree of lodging at maturity	0.690	0.634	0.539
Senescence	0.364	0.444	0.342
Inflorescence lobes	1.205	1.037	0.934
Inflorescence bristles	1.520	1.525	1.393
Lobe compactness	1.111	0.941	0.913
Inflorescence shape	1.473	1.368	1.349
Inflorescence compactness	0.587	0.292	0.251
Fruit colour	1.385	0.942	0.884
Grain shape	0.671	0.462	0.466
Apical sterility in panicle	0.689	0.692	0.692

Conclusions

"PowerCore" is a new and a faster approach for developing core collection, which effectively simplifies the generation process of a core set with reduced number of core entries but maintaining high percent of diversity compared to other methods used. The efficiency of "PowerCore" when compared with other clustering methods showed that in most cases the mean component for the two core sets were at par and the variance components were higher in core collection developed using "PowerCore". Thus, it can be concluded that the core sets identified using "Power Core" are small in size with greater diversity captured compared to traditional clustering method.

Acknowledgements

The authors acknowledge the support of Rural Development Administration (RDA), South Korea for use of "PowerCore" software and also to Bioversity International for financial assistance.

References

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, DR Marshal, JT Williams (eds) *The Use of Plant Genetic Resources*. Cambridge University Press, Cambridge, pp 136-156.

Chandra S, Z Huaman, S Hari Krishna, 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:** 325-1334.

Chung HK, KW Kim, JW Chung, JR Lee, SY Lee, A Dixit, HK Kang, W Zhao, KL McNally, RS Hamilton, JG Gwag, and YJ Park (2009) Development of a core set from a large rice collection using a modified heuristic algorithm to retain maximum diversity. *J. Interg. Plant Biol.* **51:** 1116-1125.

Dussert S, N Chabrillange, F Anthony, F Engelmann, C Recalt, and S Hamon (1997) Variability in storage response within a coffee (*Coffea* spp.) core collection under slow growth conditions. *Plant Cell Rep*. **16**: 344-348.

Federer, WT (1956) Augmented (or hoonuiaku) designs. Hawaii. *Plant Rec*. **2**: 191-208.

Gowda J, BH Halaswamy, VK Magar, BM Ramakrishna, M Krishnappa, KR Vasanth, K Seenappa, G Somu and A Seetharam (2002) *Catalogue on Evaluation of Foxtail Millet* [*Setaria italica* (*L.*) Beauv] Germplasm, pp 1-80.

Holbrook CC and WF Anderson (1995) Evaluation of a core collection to identify resistance to late leaf spot in peanut. *Crop Sci*. **35**: 1700-1702.

Hu J, J Zhu, XM Xu (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor. Appl. Genet*. **101**: 264-268.

IPGRI (1985) *Descriptor for Setaria italica and S. pumilia*. International Plant Genetic Resources Institute, Rome, Italy.

Knüpffer H and JL Van Hintum (1995) The barley core collection: an international effort. *In:* T Hodgkin, ADH Brown, TJL van Hintum and EAV Morales, (eds.) *Core Collection of Plant Genetic Resources*. John Wiley and Sons, Chichester, UK, pp 171-178.

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

Peeters JP and JA Martinelli (1989) Hierarchical cluster analyses as a tool to manage variation in germplasm collections. *Theor. Appl. Genet.* **78:** 42-48.

Shannon CE and W Weaver (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Spagnoletti Zeuli PL and CO Qualset (1993) Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. *Theor. Appl. Genet.* **87:** 295-304.

Sturges HA (1926) The choice of a class-interval. *J. Amer. Statist. Assoc.* **21:** 65-66.

Ward J (1963) Heirarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* **58:** 236-244.