Australian Journal of Crop Science AJCS

AJCS 6(4):673-680 (2012) ISSN:1835-2707

Development of Sea Island cotton (*Gossypium barbadense* L.) core collection using genotypic values

Yongjun Mei[1], Jiaping Zhou[2,3], Haiming Xu[2,3*] and Shuijin Zhu[3*]

[1]College of Plant Science and Technology, Tarim University, Alar, Xinjiang 843300, PR China
[2]Institute of Bioinformatics, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, PR China
[3]Department of Agronomy, College of agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, PR China

[*]Corresponding author: hmxu@zju.edu.cn, shjzhu@zju.edu.cn

Abstract

A core collection usually provides an effective entry to access to entire genetic resource. It could simplify screening potential of breeding materials in a manageable size. In order to phenotypic evaluations, a completely randomized block designed experiment was carried out for 265 Sea Island cotton varieties. A mixed linear model approach was employed to predict genotypic values of two fiber quality traits and eight agronomic traits. Based on the genotypic values, some candidate core subsets were constructed using five hierarchical clustering methods combined with preferred and deviation sampling at three sampling proportions. The genetic variations captured by the subsets were compared in means, variances, ranges and coefficients of variation. The result revealed that core subset of 27 accessions, based on UPGMA clustering method combined with the deviation sampling strategy at 10% proportion (C4S2-10), exhibited maximum VR%, VD% and invariant MD% and CR%. Therefore, this subset was determined as the core collection of the Sea Island cotton. The representative and validation of the core collection was further examined by the accession distribution pattern plotted by the first two principal components, as well as the correlation coefficients. The core accessions with high fiber quality of lint cotton and yield of pre-frost cotton, being as important potential materials for quality or yield improvement, are worthy to be further studied.

Keywords: Clustering method, genetic diversity, germplasm resource, *Gossypium barbadebse* L., sampling strategy.

Abbreviations: MD%-Mean difference percentage; VD%-Variance difference percentage；CR%-Coincidence rate of range; VR%-Variable rate in variation coefficient; UPGMA- Unweighted pair group method with arithmetic average.

Introduction

Sea Island cotton (*Gossypium barbadebse* L.) is one of the four commercially-cultivated species of cotton which is highly desirable to textile industry due to its excellent fiber quality. Like many globally important commercial crops, continued genetic improvement of cotton is requisite to increase both the quality and quantity of cotton production. Since 1990, numerous genetic improvement studies have been conducted to quantify levels of cotton (Calhoun and Bowman, 1999). Despite many changes in cotton technology, it is evident that cotton yields are at a plateau. The year to year variation in cotton yield within the last 20 years is four times greater than the previous-20 year period (Meredith, 2000). Besides weather, management and pest, the extensive planting of few closely-related breeding lines is a potential hazard to sustainable increasing of yield and improvement of the fiber quality. He et al. (2002) analyzed 14 varieties of Sea Island cotton and found 13 varieties derived from same progenitor, indicating a close relationship and narrow genetic base. Genetic improvement of crop has the potential to overcome many production constraints (Furat and Uzun, 2010). Thus, it is imperative to enhance the utilization of the germplasm resources for breeding elite cotton varieties with excellent environmental adaptation. Development of a core collection has been suggested as a means to enhance use of genetic resources in the crop improvement programs. The core collection proposed by Frankel (1984) and is usually defined as a representative subset of an entire germplasm collection with minimum genetic redundancy and maximum genetic diversity of a crop species and its relatives (Brown, 1989; Frankel and Brown, 1984a; Frankel and Brown, 1984b). It is an effective entry way to access the germplasm resources, which could alleviate the burden in management of germplasm collection for curators. It also can simplify screening of exotic material for plant breeders benefiting from reduced size of surveyed materials. In the practice of core collection studies in most crops, diverse types of data, such as morphological (Balakrishnan et al., 2000), agronomic and eco-geographical descriptors (Ghamkhar et al., 2008), as well as molecular makers like AFLPs in barley (Van Treuren et al., 2006), microsatellite markers in peanut (Kottapalli et al., 2007), SNP in grape (Le Cunff et al., 2008), have been employed in measuring genetic similarity. Hu et al. (2000) compared variations captured by subsets based on genotypic and phenotypic values, and concluded that a core collection of genotypic values has larger genetic variation of traits and is more representative subset of the initial collection. That is because of quantitative inherent of traits which are controlled by genotype and environments and their interactions (G×E). The phenotypic similarities of individuals do not usually reflect their genetic similarities or variation especially under affection of environments. The important challenge in selection of a core collection is how to reduce its size while capturing the genetic diversity as much as possible. It is crucial to employ an appropriate sampling strategy. There are many different methods proposed and applied in sampling a representative core collection. Most methods used phenotypic data to measure genetic similarity between accessions. Hu et al. (2000) proposed the stepwise clustering method based on genotypic values. Chung (2009) developed a core set from a large rice collection using a Modified Heuristic Algorithm. Ghamkhar et al. (2008) compared ten sampling strategies using ecological and agro-morphological data and found that maximizing strategy best represents the whole collection. One common approach for constructing a core collection is stratifying the whole collection by regions or ecotypes and then selecting representative core accessions from each stratum. Xu et al. (2006) compared four clustering methods and three sampling strategies, by which an optimal sampling strategy and proportion were screened and used to develop a core collection using five fiber quality traits of cotton. The current study was conducted (1) to investigate the genetic diversity of two fiber quality and eight agronomic traits among 265 Sea Island cotton accessions covering ~85% of germplasm used in China, and (2) to develop a core collection for promoting utilization of Sea Island resources in cotton breeding programs.

Results and Discussion

Variance analysis for ten cotton traits

Most phenotypic traits in plants are quantitative and controlled not only by genotypes, but also by environments and genotype × environment interaction. The phenotypic variance of quantitative traits can be partitioned into genetic variance, environmental variance, GE interaction variance and residual variance (Table 1). The variance analysis for ten traits of Sea Island cotton revealed significant genotypic variations in reflectivity, yellowness, boll-opening stage, boll weight and lint percentage. For yellowness, the genotypic variance accounted more than 50% of total phenotypic variations. The other traits such as flowering stage, boll stage, boll-opening stage, pre-frost boll number and pre-frost lint cotton, were mainly affected by environmental effects. The environmental effects on boll stage and boll-opening stage even consisted of 81% and 64% of total variances, respectively. The extent of interaction variance for pre-frost boll number percentage reached to 48%, while no interaction effects detected for the other nine traits, indicating that expression of pre-frost boll number percentage genes are easily influenced by environments (years). Table 1. also showed significant block effects for two fiber and four agronomic traits except pre-frost boll number, boll weight, pre-frost boll number percentage and pre-frost lint cotton. All residual variances were significant. In particular, relatively larger variance proportions were observed for boll weight, lint percentage and pre-frost lint cotton. These results demonstrated that phenotypic observations are strongly influenced by environment, as well as sampling errors. Therefore, it is more appropriate to use genotypic values to measure genetic similarity between accessions and develop core collection.

Comparison of different core subsets

Ten core collections were developed by five clustering methods as follows: C1: single linkage (Sibson, 1973); C2: complete linkage (Sorensen, 1948); C3: centroid method (Sokal and Michener, 1958); C4: the unweighted pair group method with arithmetic average (UPGMA) (Sokal and Michener, 1958); and C5: the Ward's method (Ward, 1963) combined with the preferred sampling (S1) and the deviation sampling (S2) (Hu et al., 2000) at proportion of 15%. Table 2. compared the genotypic differences in mean, variance, range and coefficient of variation between the subsets and the initial collection. According to the criterion mentioned in methods, core collection with a larger variance difference percentage (VD%) and variable rate in variation coefficient (VR%) is supposed to provide a good representativeness in the genetic diversity of the initial collection. Under the same sampling strategy, the subset C4S1 and C4S2 had relative larger VD% and VR% while the zero mean difference percentage (MD%) and coincidence rate of range (CR%) were 100%. Thus, it could be inferred that the UPGMA clustering method which is mostly employed in clustering analysis is the best choice in stepwise clustering methods for constructing core collection in our study. Similarly, the effectiveness of the different sampling methods for core collections were compared under the same clustering method. The core subsets, sampled by the deviation sampling method, obtained larger VD% and VR% than those by the preferred sampling method. Moreover, the VD% and VR% of C4S2 tended to be larger than that of C4S1. In addition, when using the deviation sampling strategy, CR% of 100% and zero MD% were obtained. The core collection developed by the deviation sampling was considered as more representative in genetic variation of the initial collection, because the deviation sampling methods select accessions which could maximize the value of variance (Pkania et al., 2007). As a result of the preceding analysis, the UPGMA clustering method combined with the deviation sampling strategy (C4S2) could be regarded as the best for constructing a core of the cotton out of ten sampling strategies.

Evaluation of different sampling proportions

Ascertaining an adequate sampling proportion is essential in developing a representative core collection. The subsets at sampling proportion of 10%, were investigated on available genetic variation (Table 2). The results clearly showed that the subsets at 10% were more representative than 15% and 20%. The MD% at 10% sampling proportion was zero except for the C3S1 (20%), while the MD% of some subsets at 15% and 20% reached up to 20%, which did not meet the requirement for a core collection. Moreover, VD% of subsets at 10% was higher than that of 15% and 25%, except for the C2S2 (20%). All subsets at 10% exhibited higher VR% than those at the other two sampling proportions. Hence, although subsets at 15% and 20% meet the requirements for core collection, C4S2_10 (sampling proportion of 10%) is more appropriate to represent the initial collection in genetic than others.

Evaluation of core collection

According to the preceding results, the potential core subset C4S2_10 was screened as the final core for Sea Island cotton.

Table 1. Variance component proportions of two fiber traits and eight agronomic traits of Sea Island cotton in original collection.

Traits	V_G/ V_P[1]	V_E/ V_P	V_{GE}/ V_P	V_B/ V_P	V_e/ V_P
Reflectivity (%)	0.20**	0.00	0.24	0.12 **	0.44 **
Yellowness	0.55 **	0.08 **	0.00	0.09**	0.28**
Flowering stage (day)	0.07	0.51 **	0.00	0.06 **	0.37 **
Boll stage (day)	0.01	0.81 **	0.00	0.08 **	0.11 **
Boll-opening stage (day)	0.11 **	0.64 **	0.00	0.16 **	0.09 **
Pre-frost boll number	0.02	0.47 **	0.00	0.04	0.47 **
Boll weight (g)	0.21 **	0.02	0.00	0.04	0.73 **
Lint percentage (%)	0.31 **	0.04	0.00	0.02 *	0.63 **
Pre-frost boll number percentage (%)	0.01	0.16	0.48 **	0.00	0.35 **
Pre-frost lint cotton (g)	0.01	0.47**	0.00	0.00	0.52**

[1] V_G/V_P, V_E/V_P, V_{GE}/V_P, V_B/V_P, Ve/V_P, are ratios of genotypic variance, environmental variance, genotype × environment interaction variance, block variance and residual variance to phenotypic variance, respectively. *, ** indicate significance at $p \leq 0.05$and $p \leq 0.01$ respectively.

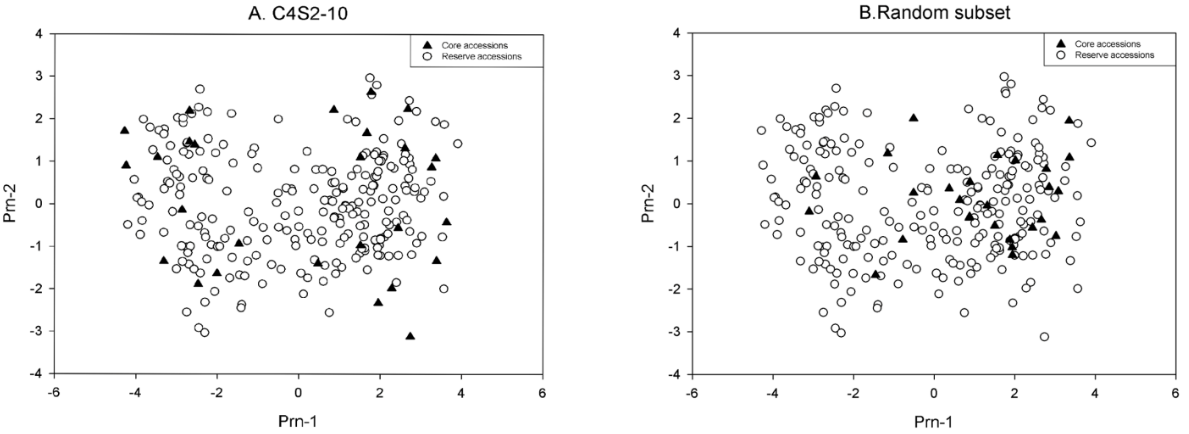

Fig 1. Principal component plots of the reserve and core accessions at 10% sampling proportion; filled triangles denoted core accessions and hollow denoted circle reserved accessions; (A).Plot for core collection (C4S2_10) sampled by the unweighted pair-group average clustering method combined with the deviation sampling; (B). Plot for core subset sampled by complete random without clustering.

In order to validate this core collection, another random subset, by complete random sampling at 10% proportion and without clustering, was constructed and compared with the C4S2_10 in genetic variation. Table 3 presented genotypic means, variances, ranges, coefficient of variation (*CVs*) and Shannon-Weaver diversity index (*H*) of the ten traits for the initial, C4S2_10 and a random subset. The means of three subsets did not exhibit any significant difference for all of ten traits, whereas significant differences were detected for the means of boll stage and boll-opening stage between the whole collection and random subset. The means of core collection were closer with initial collection than that of random subset. *F*-test detected significantly increasing variances of C4S2_10 for all surveyed traits compared with the initial collection, whereas significantly decreasing variance of the random subset for the reflectivity, flowering stage, boll-opening stage were observed. All ranges of the initial collection remained unchanged in the C4S2_10 whereas reduced in the random subset for all traits. This indicates that more particular accessions in performance of traits have been included in the C4S2_10 which conserved larger genetic diversity in initial collection than the random subset. Similar pattern could also be found in the *CV* and *H*, suggesting the genetic diversity of C4S2_10 increased after eliminating the redundant accessions. Furthermore, C4S2_10 and random subset were compared in pattern of genetic variation by principal component analysis (PCA) (Pearson, 1901). The distribution of accessions were approximately characterized by the first two principal components, which could account for 65.7% of the total observed genetic variation in the initial collection, with the first and second axes explaining 50.1% and 15.6% of the total variation, respectively. Fig 1. clearly illustrated that many overlapped accessions in central area have higher genetic similarities. Some redundant accessions should be excluded to make the subset more representative. With regard to core accessions, marked by triangle, wider range of accessions were sampled in periphery of plot A (C4S2-10) than plot B (random subset) (Fig 1). As a result, much better coverage or distribution pattern of the initial population could be conserved in C4S2-10 than random core. Therefore, it could be concluded that the genetic structure and variation of the initial cotton population are well represented by the C4S2-10. An adequate core collection should maintain genetic associations arising out of co-adapted gene complexes in entire collection (Ortiz et al., 1998). Comparison of genetic correlation coefficients among traits based on predicted genotypic values were conducted for all quantitative traits in the entire collection and core collection, separately (Table 4). Obviously, there is significant negative relation between reflectivity and yellowness, both in the core collection and the initial collection. For agronomic traits of crops, it was believed that early growth stage would be highly correlated with the later period. Twenty two out of 28 significant genetic correlations were detected in the initial collection, which

Table 2. Comparison of genetic diversity between the initial collection and subsets under ten sampling strategies at proportions of 10%, 15% and 20%.

	Sampling proportion											
	10%				15%				20%			
	MD%[2]	VD%[3]	CR%[4]	VR%[5]	MD%	VD%	CR%	VR%	MD%	VD%	CR%	VR%
C1S1[1]	0	30	100	123.13	0	10	100	115.54	0	0	100	112.04
C2S1	0	20	100	125.01	0	10	100	117.54	0	0	100	113.27
C3S1	20	20	100	124.36	10	10	100	115.14	10	0	100	112.48
C4S1	0	50	100	123.66	0	40	100	118.25	0	10	100	114.43
C5S1	0	40	100	127.44	0	10	100	116.67	0	10	100	114.51
C1S2	0	20	100	125.82	0	30	100	118.14	0	10	100	112.88
C2S2	0	60	100	128.23	20	20	100	114.81	20	10	100	107.76
C3S2	0	60	100	127.91	0	30	100	116.84	0	0	100	111.21
C4S2	0	70	100	129.11	0	40	100	120.35	20	10	100	111.86
C5S2	0	60	100	128.84	0	30	100	119.18	20	10	100	110.52

[1] CiSj (i=1~5, j=1,2), subsets sampled, respectively by the single linkage (C1), the complete linkage (C2), the centroid method (C3), the unweighted pair-group average (C4), the Ward's method (C5), in combination with the preferred sampling (S1) and the deviation sampling (S2). [2] MD%, the percentage of significant difference (α =0.05) between core collection and initial collection in mean of trait. [3] VD%, the percentage of significant difference (α =0.05) between core collection and initial collection in variance of trait. [4] CR%, coincidence rate of range. [5] VR%, variable rate.

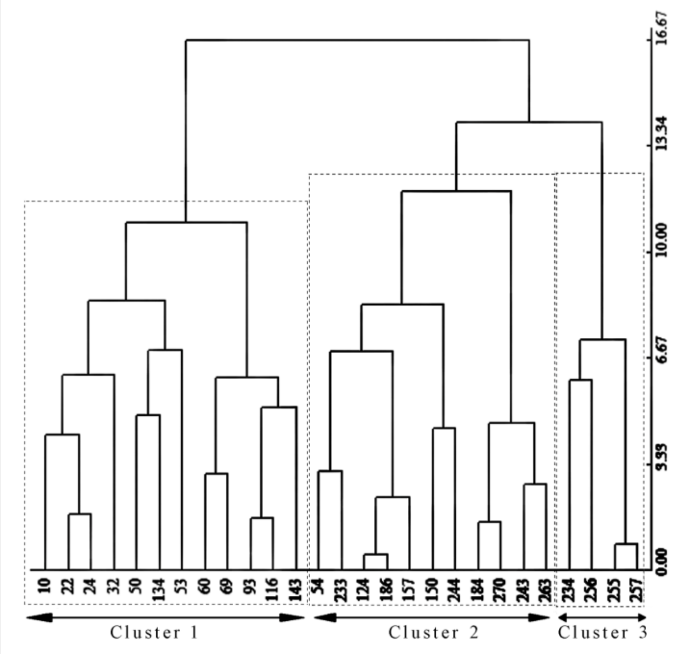

Fig 2. Dendogram of 27 Sea Island cotton accessions generated by UPGMA clustering method using genotypic values of two fiber traits and eight agronomic traits based on Mahalanobis distance.

Table 3. Comparison of genetic variation of ten cotton traits between initial collection and core collection (C4S2) or the random subset at the sampling proportion of 10%.

Traits	Population[1]	Mean	Variance	Range	*CV*[2]	*H*[3]
Reflectivity (%)	Initial	65.43	12.35	21.83	0.05	2.01
	C4S2	65.30	21.78^{*}	21.83	0.07	2.06
	Random	65.31	7.15^{*}	8.94	0.04	1.71
Yellowness	Initial	11.34	0.26	2.55	0.05	2.05
	C4S2	11.25	0.44^{*}	2.55	0.06	2.13
	Random	11.36	0.17	1.93	0.04	1.71
Flowering stage (day)	Initial	78.57	37.35	32.19	0.08	1.92
	C4S2	78.51	57.47^{*}	32.19	0.10	1.93
	Random	81.57^{**}	16.94^{**}	22.15	0.05	1.54
Boll stage (day)	Initial	83.29	187.55	59.25	0.16	1.94
	C4S2	81.06	313.35^{*}	59.25	0.22	1.95
	Random	89.82^{*}	139.14	50.00	0.13	1.72
Boll-opening stage (day)	Initial	161.84	366.60	69.06	0.12	1.81
	C4S2	159.50	583.72^{*}	69.06	0.15	1.63
	Random	171.63^{**}	194.37^{*}	54.41	0.08	1.57
Pre-frost boll number	Initial	8.32	14.34	16.46	0.46	1.99
	C4S2	7.50	21.11	16.46	0.61	2.03
	Random	6.82	13.62	14.25	0.54	1.82
Boll weight (g)	Initial	3.04	0.11	1.89	0.11	2.01
	C4S2	3.04	0.212^{**}	1.89	0.15	2.06
	Random	3.13	0.14	1.28	0.12	1.80
Lint percentage (%)	Initial	33.20	5.94	15.87	0.07	2.02
	C4S2	33.30	9.308^{*}	15.87	0.09	1.88
	Random	32.59	6.86	10.72	0.08	1.87
Pre-frost boll number percentage (%)	Initial	44.98	248.72	61.16	0.35	2.00
	C4S2	41.04	289.68	61.16	0.42	1.96
	Random	40.76	170.51	43.00	0.32	1.77
Pre-frost lint cotton (g)	Initial	8.25	13.00	16.83	0.44	2.01
	C4S2	7.31	16.97	16.83	0.56	1.92
	Random	6.87	12.94	14.15	0.52	1.89

[1] Initial, the initial collection; C4S2: the core collection developed by the unweighted pair-group average clustering method combined with the deviation sampling at 10%; Random: the subset sampled in complete random method without clustering at 10%. [2] *CV*, coefficient of variation. [3] *H*, Shannon-Weaver diversity index. $^{*,\,**}$ indicate significant difference detected between the subset and the initial population at 5%, 1% probability.

showed close relationships between the agronomic traits. However, 18.1% of correlations were not significan between agronomic traits (Table 4). No significant correlations were detected between agronomic and fiber traits in the core collection C4S2-10, which was probably caused by the different degrees of freedom in statistic testing because of different population sizes for the initial and the core collections. Significant values of linear coefficients (*r*) become smaller as the population size increases (Little and Hill, 1978). The *r* computed from a selected germplasm sample can be smaller than that from a world germplasm collection (Gomez and Gomez, 1984). Thus, some significant correlation in the initial collection would turn out to not significant. The correlation coefficients with an absolute value larger than 0.707 have been suggested to be biologically meaningful (Skinner et al., 1999). Since in this rate, more than 50% of the variation in one trait would be predicted by the other (Snedecor and Cochran, 1980). Table 4. presented such meaningful relationships both in the entire and in the core collections. Some pairs of traits have significant positive correlations, such as flowering stage and boll-opening stage (*r*=0.82), boll stage and boll-opening stage (*r*=0.97), pre-frost boll number and pre-frost lint cotton (*r*=0.95), while others show negative correlations, i.e. boll stage and pre-frost boll number percentage (*r*=-0.73), boll-opening stage and pre-frost boll number percentage (*r*=-0.78), boll-opening stage and pre-frost lint cotton (*r*=-0.72), which proves some genetic associations, arising out of co-adapted gene complexes, are maintained in the core collection. In addition, three groups were determined by hierarchical cluster analysis on the core collection based on Mahalanobis distance (Fig 2). Compared with original data of the 27 accessions from the core collection, it is interesting to note that the 27 cotton varieties are grouped together based on their growth period and their corresponding pre-frost boll number percentage. The accession from Cluster 1 have longest flower and boll period, as well as lowest boll number percent before frost, followed by Cluster 2. Accessions in Cluster 3 are characterized by the shorter growth period, higher boll number percent before frost, and better fiber characters, which could be used as potential breeding materials.

Materials and methods

Experiment and traits investigation

All 265 cotton varieties in present study were planted in the experimental farm of Tarimu University of Agricultural Reclamation, Alar, Xinjiang Province, China during three consecutive years (1990-1992). A randomized complete block design with two replications in each year was carried out. Each block consisted of two rows with a space of 0.13 m. Each row was 2 m long and 0.6 m wide on a ridge. The fields were managed with local optimal fertilizer and cultivation measures. For each variety, eight cotton plants were sampled to

Table 4. Correlation coefficients between traits in the entire collection (above diagonal) and the core collection (below diagonal)

	Reflectivity (%)	Yellowness	Flowering stage (day)	Boll stage (day)	Boll-opening stage (day)	Pre-frost boll number	Boll weight (g)	Lint percentage (%)	Pre-frost boll number percentage (%)	Pre-frost lint cotton (g)
Reflectivity (%)	-	-0.52**	-0.10	-0.16**	-0.15*	0.07	-0.19**	-0.06	0.14*	-0.52
Yellowness	-0.57**	-	0.14	0.20**	0.19**	-0.16**	0.14*	0.09	-0.20**	-0.11
Flowering stage (day)	0.05	0.29	-	0.65**	0.82**	-0.70**	0.30**	-0.11	-0.69**	-0.67**
Boll stage (day)	-0.07	0.26	0.59**	-	0.97**	-0.69**	0.27**	-0.08	-0.73**	-0.66**
Boll-opening stage (day)	-0.04	0.29	0.78**	0.96**	-	-0.75**	0.31**	-0.10	-0.78**	-0.72**
Pre-frost boll number	0.02	-0.42*	-0.73**	-0.75**	-0.82**	-	-0.35**	0.11	0.81**	0.95**
Boll weight (g)	-0.08	0.08	0.47*	0.35	0.43*	-0.42	-	-0.08	-0.37**	-0.13*
Lint percentage (%)	-0.39*	0.23	0.01	0.11	0.09	-0.01	0.04	-	0.10	0.24**
Pre-frost boll number percentage (%)	0.23	-0.37	-0.74**	-0.75**	-0.83**	0.70**	-0.50**	-0.04	-	0.76**
Pre-frost lint cotton (g)	-0.05	-0.39*	-0.66**	-0.68**	-0.74**	0.94**	-0.15	0.12	0.59**	-

$^{*, **}$ indicate significant difference detected between the subset and the initial population at 5%, 1% probability.

investigate two fiber quality traits (reflectivity (%) and yellowness) and eight agronomic traits, viz., flowering stage (day), boll stage (day), boll-opening stage (day), pre-frost boll number, boll weight (g), lint percentage (%), pre-frost boll number percentage (%), and pre-frost lint cotton (g).

Prediction of genotypic values

The genotypic values of traits were predicted by the adjusted unbiased prediction (AUP) method (Zhu, 1993; Zhu and Weir, 1996). The observed value of the j^{th} accession in the k^{th} block within the i^{th} year could be expressed as:

$$Y_{ijk} = \mu + E_i + G_j + GE_{ij} + B_{k(i)} + \varepsilon_{ijk}$$

Where, μ is the population mean; E_i is the random effect of the i^{th} environment (year), $E_i \sim (0, \sigma_E^2)$, $i = 1,2,3$; G_i is the random effect of the j^{th} accession, $G_j \sim (0, \sigma_G^2)$, $j = 1,2,...,265$; GE_{ij} is the random effect of $G_j \times E_i$, $GE_{ij} \sim (0, \sigma_{GE}^2)$; $B_{k(i)}$ is the random effect of the k^{th} block within the i^{th} environment, $B_{k(i)} \sim (0, \sigma_B^2)$, $k = 1,2$; ε_{ijk} is the residual effect, $\varepsilon_{ijk} \sim (0, \sigma_\varepsilon^2)$.

Construction of the core collection

Based on the above predicted genotypic values, the Mahalanobis distances (Mahalanobis, 1936) were calculated between accessions and used in cluster analysis of the accessions. The procedure of stepwise clustering proposed by Hu et al. (2000) was employed in developing a core collection of the Sea Island cotton. To screen optimal sampling strategy, ten potential core subsets were constructed by ten diverse combining schemes of five clustering methods including single linkage, complete linkage, centroid method, UPGMA and the Ward's method, and two sampling strategies, viz., preferred sampling, and deviation sampling. A homogeneity test (*F*-test) for variances and a *t*-test for means (α=0.05) were performed to determine the difference of traits between core subsets and the initial collection. Then MD%, VD%, CR% and VR% (Hu et al., 2000) were calculated, which were used to evaluate representativeness of each core subset. The criterion to judge the representative of a core collection is summarized as follows: no more than 20% of the traits have different means (significant at α=0.05) between the core collection and the initial collection; And the VR% of the core collection is larger than 80%. A core collection with a larger VD% and VR% is regarded to have better representativeness in genetic diversity of the initial collection.

Validation of the core collection

A complete random subset was developed at the optimal sampling proportion to evaluate the representativeness of the core collection. The differences in mean between the core collection or the random subset and the initial collection were tested by Newman–Keuls procedure (Newman, 1939) for all the traits, and the homogeneity of variances by Levene's test (Levene, 1960). The ranges, coefficients of variation and Shannon – Weaver diversity indexes (Shannon and Weaver, 1949) were also calculated and used as a measure of genetic diversity of each trait for different populations. Principal component analysis (PCA) of the data set was performed for core collection and random subset to illustrate whether the distribution pattern of initial population was well captured by the core collection.

An adequate core collection should maintain genetic associations of traits in entire collection as well. In current study, the Pearson's correlation coefficients between traits in the core collection and whole collection were calculated by the SAS software in order to investigate the genetic relationship of traits retained in core collection. Clustering analysis was also performed based on the Mahalanobis distance for screening superior accessions with potential value in breeding program or further study on diversity in molecular methods.

Conclusion

Selection of material for breeding programs is always a challenge for breeders, especially for crops with a large germplasm. The sampled core collection provides cotton breeders a good entry to the Sea Island cotton germplasm resources. The core accessions with high fiber quality of lint cotton and yield of pre-frost cotton, being as important potential material for quality or yield breeding of cotton, are worthy to be further studied.

Acknowledgments

This work is supported in part by the National Basic Research Program of China (973 Program 2010CB126006), the National Special Program for Breeding New Transgenic Variety (2008ZX08005-005) and the Scientific Research Program of the Higher Education Institution of Xinjiang, China (XJEdu2006I59).

References

Balakrishnan R, Nair NV, Sreenivasan TV.(2000) A method for establishing a core collection of *Saccharum officinarum* L. germplasm based on quantitative-morphological data. Genet Resour Crop Ev 47(1):1-9.

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

Calhoun DS, Bowman DT (1999) Techniques for development of new cultivars. In: C. W. Smith and J. T. Cothren (eds.), Cotton:Origin, history, technology, and production, John Wiley&Sons, New York.

Chung H, Kim K, Chung J, Lee J, Lee S, Dixit A, Kang H, Zhao W, McNally KL, Hamilton RS, Gwag J, Park Y (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(12):1116-1125.

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: W. Arber, et al. (eds.), Genetic Manipulation: Impact on Man and Society, Cambridge University Press Cambridge, UK.

Frankel OH, Brown AHD (1984a) Plant genetic resources today: a critical appraisal. In: J. H. W. Holden and J. T. Williams (eds.), Crop Genetic Resources: Conservation and Evaluation, George Allen & Urwin Ltd, London.

Frankel OH, Brown AHD (1984b) Current plant genetic resources: a critical appraisal, Genetics, new frontiers, Oxford & IBH Publishing Co, New Delhi, India.

Furat S, Uzun B (2010) The use of agro-morphological characters for the assessment of genetic diversity in sesame (*Sesamum indicum* L). Plant Omics J 3(3):85-91.

Ghamkhar K, Snowball R, Wintle BJ, Brown AHD (2008) Strategies for developing a core collection of bladder clover (*Trifolium spumosum* L.) using ecological and agro-morphological data. Aust J Agr Res 59(12):1103-1112.

Gomez KA, Gomez AA (1984) Statistical procedures for agricultural research, 2nd edn. John Wiley &Sons, New York.

He L, Zheng D, Zhang X, Zhang L, Cao X, Xiong R (2002) Studies on pedigree relative and main character evolution in South Xinjiang. Acta Botanica Sinica 29(6):2.

Hu J, Zhu J, Xu HM (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.

Kottapalli KR, Burow MD, Burow G, Burke J, Puppala N (2007) Molecular characterization of the US peanut mini core collection using microsatellite markers. Crop Sci 47(4):1718-1727.

Le Cunff L, Fournier-Level A, Laucou V, Vezzulli S, Lacombe T, Adam Blondon AF, Boursiquot JM, This P (2008) Construction of nested genetic core collections to optimize the exploitation of natural diversity in *Vitis vinifera* L. subsp. sativa. BMC Plant Biol 8(31):1-12.

Levene H (1960) Robust tests for equality of variances. In: I. Olkin, et al. (eds.), Contributions to Probability and Statistics:Essay in Honour of Harold Hotelling, Stanford University Press, Stanford.

Little TM, Hill J (1978) Agricultural experimentation:design and analysis John Wiley and Sons, New York.

Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2:49-55.

Meredith WR (2000) Cotton yield progress-Why has it reached a plateau? Better Crops 84(4):6-9.

Newman D (1939) The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. Biometrika 31:20-30.

Ortiz R, Ruiz-Tapia EN, Mujica-Sanchez A (1998) Sampling strategy for a core collection of Peruvian quinoa germplasm. Theor Appl Genet 96(3-4):475-483.

Pearson K (1901) On lines and planes of closest fit to systems of points in space. Philos Mag 2(6):559-572.

Pkania CK, Wu J, Xu H, Shi C, Li C (2007) Addressing rice germplasm genetic potential using genotypic value to develop quality core collections. J Sci Food Agr 87(2):326-333.

Shannon CE, Weaver W (1949) The mathematical theory of communication. Univ.Illinios Press, Urbana.

Sibson R (1973) SLINK: an optimally efficient algorithm for single-link cluster method. Comput J 16(1):30-34.

Skinner DZ, Bauchan GR, Auricht G, Hughes S (1999) A method for the efficient management and utilization of large germplasm collections. Crop Sci 39(4):1237-1242.

Snedecor GW, Cochran WG (1980) Stastical Methods, 7 Edn edn. lowa State University Press, Amis.

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kans Sci Bull 38:1409-1438.

Sorensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skrif 5:1-34.

van Treuren R, Tchoudinova I, van Soest LJM, van Hintum TJL (2006) Marker-assisted acquisition and core collection formation: A case study in barley using AFLPs and pedigree data. Genet Resour Crop Ev 53(1):43-52.

Ward JH Jr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236-244.

Xu H, Mei Y, Hu J, Zhu J, Gong P (2006) Sampling a core collection of Island Cotton (*Gossypium barbadense* L.) based on the genotypic values of fiber traits. Genet Resour Crop Ev 53(3):515-521.

Zhu J (1993) Methods of predicting genotype value and heterosis for offspring hybrids. J BioMath 8:32-44.

Zhu J, Weir BS (1996) Diallel analysis for sex-linked and maternal effects. Theor Appl Genet 92(1):1-9.