Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Effect of the scale of quantitative trait data on the representativeness of a cotton germplasm sub-core collection*

Jian-cheng WANG[1,2], Jin HU[†‡1], Ya-jing GUAN[1], Yan-fang ZHU[1]

([1]*Seed Science Center, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310058, China*)

([2]*Shandong Crop Germplasm Center, Shandong Academy of Agricultural Sciences, Jinan 250100, China*)

[†]E-mail: jhu@zju.edu.cn

Received Mar. 12, 2012; Revision accepted Aug. 8, 2012; Crosschecked Jan. 3, 2013

Abstract: A cotton germplasm collection with data for 20 quantitative traits was used to investigate the effect of the scale of quantitative trait data on the representativeness of plant sub-core collections. The relationship between the representativeness of a sub-core collection and two influencing factors, the number of traits and the sampling percentage, was studied. A mixed linear model approach was used to eliminate environmental errors and predict genotypic values of accessions. Sub-core collections were constructed using a least distance stepwise sampling (LDSS) method combining standardized Euclidean distance and an unweighted pair-group method with arithmetic means (UPGMA) cluster method. The mean difference percentage (MD), variance difference percentage (VD), coincidence rate of range (CR), and variable rate of coefficient of variation (VR) served as evaluation parameters. Monte Carlo simulation was conducted to study the relationship among the number of traits, the sampling percentage, and the four evaluation parameters. The results showed that the representativeness of a sub-core collection was affected greatly by the number of traits and the sampling percentage, and that these two influencing factors were closely connected. Increasing the number of traits improved the representativeness of a sub-core collection when the data of genotypic values were used. The change in the genetic diversity of sub-core collections with different sampling percentages showed a linear tendency when the number of traits was small, and a logarithmic tendency when the number of traits was large. However, the change in the genetic diversity of sub-core collections with different numbers of traits always showed a strong logarithmic tendency when the sampling percentage was changing. A CR threshold method based on Monte Carlo simulation is proposed to determine the rational number of traits for a relevant sampling percentage of a sub-core collection.

Key words: Sub-core collection, Mixed linear model, Least distance stepwise sampling, Monte Carlo simulation, CR threshold method

doi:10.1631/jzus.B1200075 **Document code:** A **CLC number:** S32; S56

1 Introduction

Core collections provide a convenient way to preserve germplasm resources with genetic characteristics of agronomic interest. A core collection is a representative sample of the whole collection which has minimum repetitiveness and maximum genetic diversity of a plant species (Frankel and Brown, 1984). The core collection serves as a working collection to be evaluated and utilized preferentially (Silvar *et al.*, 2010; Biabani *et al.*, 2011; Pino del Carpio *et al.*, 2011; Wang *et al.*, 2011). In this way, it is possible to preserve most of the genes in large germplasm populations using limited funds.

One common approach for constructing a core collection is to group the germplasm population by growing regions, ecotypes or other classification rules,

‡ Corresponding author

* Project supported by the Special Foundation for Agro-Scientific Research in the Public Interest of China (No. 201203052), the China Postdoctoral Science Foundation (No. 2012M521184), and the Shandong Provincial Natural Science Foundation of China (No. ZR2010CQ016)

to create high levels of difference among groups. Sub-core collections are then selected from each group and combined to form the core collection (Brown, 1995; Wang *et al.*, 2008). Representativeness is the most important characteristic of a core collection (Zhang J. *et al.*, 2010). There are many parameters for measuring and validating the representativeness of core collections, such as the mean, variance, range, or coefficient of variation between the core collection and the initial population (Mei *et al.*, 2012). To construct a representative core collection, different types of data can be used (Upadhyaya *et al.*, 2010; Cheng *et al.*, 2011; Smýkal *et al.*, 2011). There are many factors that may affect the representativeness of a core collection, such as the genetic diversity of plant germplasm, data type, number of traits observed, grouping method, sampling method, and sampling percentage (Upadhyaya *et al.*, 2006; Rao *et al.*, 2011; Díez *et al.*, 2012). Quantitative traits have been used to construct core collections for a long time (Santesteban *et al.*, 2009). However, the ideal quantity of data for a quantitative trait is under debate. Quantitative traits are usually controlled by many minor genes, and their observed values are commonly affected by the environment. Therefore, efforts aimed at constructing core collections based on observed values of quantitative traits might be misleading (Hu *et al.*, 2000). The mixed linear model has been reported to be a useful tool in analyzing variance components and predicting values of random effects (Wulff, 2009; Kang *et al.*, 2010; Zhang Z. *et al.*, 2010).

Cotton is the most important natural fiber crop in the world (Campbell *et al.*, 2010). However, the extensive planting of a few closely-related breeding lines is a potential hazard to the maintenance of cotton yields, which have almost reached a plateau (Mei *et al.*, 2012). Genetic improvement of a crop is a potential way to overcome many production constraints (Zeng *et al.*, 2011). It is imperative to enhance the utilization of cotton germplasm. A core collection provides a convenient way to conduct that work. Thus, the objectives of this research were: (1) to adopt a mixed linear model approach to eliminate environmental effects from data of cotton quantitative traits; (2) to use those data to investigate the ideal quantity of quantitative trait data for core collection construction based on Monte Carlo simulations.

2 Materials and methods

2.1 Materials

One hundred and sixty-eight cotton varieties (Hu *et al.*, 2000) were planted in the same region (Liaoning, China). All the varieties are used as breeding materials by the Liaoning Economy Crop Research Institute and this collection served as a germplasm group in this study. All varieties were planted in rectangular plots of 20 m^2 for two years with two replications per year. Data for 20 quantitative traits were recorded. There were 11 agronomic traits (plant height, height of fruit branch, length of fruiting node, length of boll stalk, number of fruiting branches per plant, bolls per plant, incidence of infected plants, index of wilt disease, growth period, boll weight, and lint percentage), 5 fiber traits (fiber length, fiber uniformity, fiber strength, fiber elongation, and micronaire), and 4 seed traits (seed length, seed width, ratio of seed length to seed width, and kernel weight).

2.2 Genetic model to minimize environmental effects

A mixed linear model approach was used to predict the genotypic values of accessions to eliminate environment effects and *GE* (genotype×environment) effects. The observed values of any cotton variety could be expressed as: $Y_{hk(ij)}=\mu+E_h+R_{i(h)}+C_{j(h)}+G_{k(ij)}+GE_{hk(ij)}+\varepsilon_{hk(ij)}$, where μ is the population mean; E_h is the fixed effect of the hth environment; $R_{i(h)}$ is the fixed effect of the ith row within the hth environment; $C_{j(h)}$ is the fixed effect of the jth column within the hth environment; $G_{k(ij)}$ is the random effect of the kth genotype within the ith row and the jth column and $G_{k(ij)}\sim(0, \sigma_G^2)$; $GE_{hk(ij)}$ is the random effect of the interaction between the hth environment and the kth genotype, and $GE_{hk(ij)}\sim(0, \sigma_{GE}^2)$; $\varepsilon_{hk(ij)}$ is the residual effect and $\varepsilon_{hk(ij)}\sim(0, \sigma_\varepsilon^2)$ (Zhu and Weir, 1996). The minimum norm quadratic unbiased estimation (MINQUE) method combined with the adjusted unbiased prediction (AUP) method was adopted to predict without bias the genotypic values of the 168 cotton varieties (Zhu and Weir, 1996). Genotypic values of each trait were standardized (μ=0, σ=1, where μ is the population mean and σ is the standard deviation of the trait). Table 1 shows the phenotypic and predicted genotypic values of the twenty quantitative traits.

Table 1 Information relating to 20 quantitative traits measured among 168 accessions

Trait name	Source	Phenotypic value			Predicted genotypic value		
		Mean	Max	Min	Mean	Max	Min
Plant height	IPGRI	66.3	89.8	36.3	65.8	76.6	57.2
Height of fruit branch	Experience	5.3	6.9	3.8	5.3	6.0	4.6
Length of fruiting node	Experience	7.5	12.3	2.2	7.4	10.2	4.1
Length of boll stalk	Experience	1.9	3.4	0.5	1.9	2.4	1.5
Number of fruiting branch per plant	Experience	6.6	9.6	4.0	6.6	7.7	5.3
Bolls per plant	Experience	5.0	7.9	2.5	5.0	6.2	4.0
Incidence of infected plant	Experience	42.1	89.3	6.3	30.5	51.3	15.6
Index of wilt disease	Experience	22.6	64.2	2.6	15.7	24.6	8.7
Growth period	Experience	133	151	121	133	143	127
Boll weight	Experience	4.9	6.8	2.7	4.8	6.2	3.3
Lint percentage	IPGRI	34.6	40.4	20.5	34.2	40.3	22.3
Fiber length	IPGRI	26.5	32.4	17.5	26.5	30.0	19.8
Fiber uniformity	Experience	50.6	59.4	19.9	50.8	54.9	45.0
Fiber strength	IPGRI	20.0	24.7	11.1	20.0	23.0	15.1
Fiber elongation	IPGRI	6.1	10.9	3.1	6.1	7.0	4.2
Micronaire	IPGRI	4.1	5.9	1.9	4.0	4.7	2.6
Seed length	Experience	8.66	9.75	7.25	8.63	9.29	7.89
Seed width	Experience	5.16	6.55	4.45	5.16	5.42	4.98
Ratio of seed length to seed width	Experience	1.68	1.92	1.22	1.68	1.78	1.51
Kernel weight	Experience	1.17	1.72	0.60	1.18	1.33	1.06

IPGRI: International Plant Genetic Resources Institute

2.3 Construction and evaluation parameters of sub-core collections

The least distance stepwise sampling (LDSS) method (Wang *et al.*, 2007) was adopted to construct sub-core collections. The procedure was: (1) The genetic distances among accessions were calculated and accessions were classified by hierarchical cluster analysis based on their genetic distance; (2) One accession from a subgroup with the least distance was randomly removed and another accession of the subgroup was sampled; (3) The genetic distances among the remaining accessions were calculated, and the sampling was repeated in the same way. The stepwise samplings were performed until the percentage of the remaining accessions reached the desired sampling percentage. This method performs sampling based on the subgroup with the least genetic distance, which can efficiently eliminate redundant accessions and ignore the effect that the use of different clustering methods may have on the composition of the final sub-core collection. The standardized Euclidean distance was used as genetic distance in the LDSS method to select core accessions (Wang *et al.*, 2008).

The criteria of mean difference percentage (MD), variance difference percentage (VD), coincidence rate of range (CR), and variable rate of coefficient of variation (VR) were chosen as parameters to evaluate the representativeness of the sub-core collections (Hu *et al.*, 2000; Kang *et al.*, 2006). Those four parameters were formulated as follows:

MD=(S_t/n)×100%, where S_t is the number of traits which have a significant difference (α=0.05) between their means in the initial collection and in the core collection; n is total number of traits.

VD=(S_F/n)×100%, where S_F is the number of traits which have a significant difference (α=0.05) between their variances in the initial collection and in the core collection; n is total number of traits.

$$\mathrm{CR}=\frac{1}{n}\sum_{i-1}^{n}R_{\mathrm{C}(i)}/R_{\mathrm{I}(i)}\times 100\%,$$

where $R_{C(i)}$ is the range of the ith trait in the core collection; $R_{I(i)}$ is the range of the corresponding trait in the initial collection; n is total number of traits.

$$\mathrm{VR}=\frac{1}{n}\sum_{i-1}^{n}\frac{CV_{C(i)}}{CV_{I(i)}}\times 100,$$

where $CV_{C(i)}$ is the coefficient of variation of the ith trait in the core collection; $CV_{I(i)}$ is the coefficient of

variation of the corresponding trait in the initial collection; n is total number of traits.

2.4 Monte Carlo simulation of the number of traits and the sampling percentage

Using the standardized Euclidean genetic distance, sub-core collections were constructed from 1 to 20 quantitative traits. To perform a comprehensive analysis, the sampling percentage was varied from 10% to 30% (sampling percentages under 10% were too small to calculate evaluation parameters) for each number of traits. The parameters for evaluation were calculated from each sub-core collection. This procedure was replicated 20 times, and the trait order was randomized in each replication to homogenize the trait effect (the distribution of the variation was different among traits). The mean value (calculated using original values, not standardized values) of each evaluation parameter, considering all replications at each class according to the number of traits, served as the final value for drawing 3D maps (the sampling percentage, the number of traits and the value of evaluation parameters). Data for 3D maps were analyzed by curve fitting analysis based on a least square method, and the corresponding R^2 (coefficient of determination of fitted equations) values were calculated. R^2 provides a measure of how well future outcomes are likely to be predicted by the results of the curve fitting analysis.

2.5 Data analysis

The prediction of genotypic values by the mixed linear model approach, the LDSS procedures, the calculation of evaluation parameters, the Monte Carlo simulation, and the 3D map drawing were performed using computer code programmed by the authors based on MATLAB software (Version 6.5; the Mathworks, 2002). Curve fitting analysis was conducted using the toolbox of 'curve fitting' in MATLAB software (Version 6.5) (the Mathworks, 2002).

3 Results

3.1 Variation in evaluation parameters with changes in the number of traits and the sampling percentage

The mean value of each evaluation parameter tended to stabilize when the procedure was replicated more than ten times. The simulation results of 20 replications are summarized in Fig. 1. The values of in the number of traits for any sampling percentage

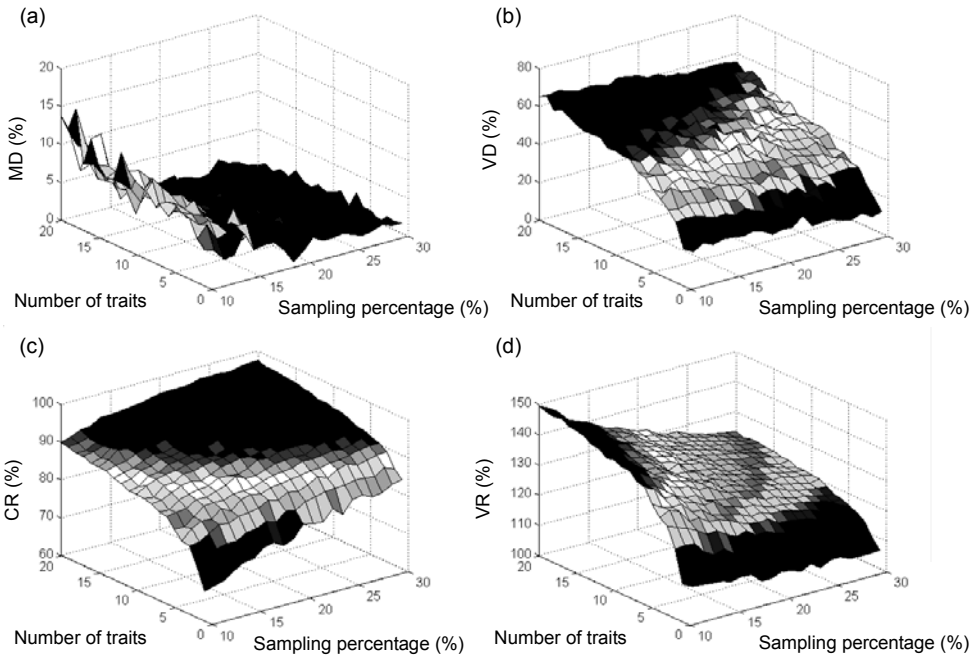

Fig. 1 Response surfaces for different parameters according to the number of traits and sampling percentages
MD: mean difference percentage; VD: variance difference percentage; CR: coincidence rate of range; VR: variable rate of coefficient of variation

VD, CR, and VR changed significantly with variation while the MD showed little change (Fig. 1). The MD fluctuated widely with variation in the number of traits at low sampling percentages but fluctuated little at high sampling percentages (Fig. 1a). The CR increased with increasing sampling percentage for each number of traits, and increased as the number of traits increased for each sampling percentage (Fig. 1c). The CR changed dramatically when the number of traits or the sampling percentage was not large, while at higher levels of those two factors, the CR changed smoothly, eventually reaching 100% (Fig. 1c). The VD and VR decreased as the sampling percentage increased for each number of traits, and increased as the number of traits increased for each sampling percentage (Figs. 1b and 1d). Like the CR, the VD and VR changed dramatically when the number of traits or the sampling percentage was not high, but changed smoothly as those two factors increased to higher levels (Fig. 1). The VD and VR showed similar variation; however, the changing trends of the VR were more significant than those of the VD (Fig. 1).

3.2 Variation in R^2 of CR curves with changes in the number of traits and the sampling percentage

The 3D map showed that the number of traits and the sampling percentage affected the values of the CR (Fig. 1). Therefore, they could act as factors affecting the representativeness of cotton sub-core collections in this research. When one factor had a fixed value, a changing CR curve was produced by the other factor. For example, if the number of traits was equal to 10, the values of CR would change with the sampling percentage, increasing from 10% to 30% (Fig. 1). Therefore, 11 CR values would be achieved and those values would change regularly and present a changing curve. The equations for the curve showing the change in CR with sampling percentage for different numbers of traits, and the equations for the curve showing the change in CR with the number of traits for different sampling percentages were fitted by a least square method, and the corresponding R^2 values were calculated (Tables 2 and 3). With an increase in the number of traits, the CR linear R^2 changed dramatically,

Table 2 Linear and logarithmic equations with relevant R^2 of values of CR changing with sampling percentage, for different numbers of traits

Number of traits	Linear equation	R^2	Logarithmic equation	R^2
1	y=0.7514x+68.287	0.9443	y=5.5800ln$x$+64.495	0.9041
2	y=0.5819x+74.844	0.9343	y=4.4425ln$x$+71.645	0.9454
3	y=0.5237x+78.655	0.9498	y=3.9296ln$x$+75.924	0.9286
4	y=0.4991x+80.457	0.9295	y=3.8454ln$x$+77.637	0.9581
5	y=0.4964x+81.506	0.9328	y=3.8043ln$x$+78.745	0.9514
6	y=0.4355x+83.370	0.9455	y=3.2538ln$x$+81.130	0.9163
7	y=0.4399x+83.932	0.9283	y=3.3617ln$x$+81.506	0.9415
8	y=0.4186x+85.110	0.9452	y=3.1726ln$x$+82.858	0.9426
9	y=0.4124x+85.803	0.9199	y=3.1954ln$x$+83.435	0.9587
10	y=0.3983x+86.474	0.9525	y=3.0241ln$x$+84.320	0.9533
Average 1–10		0.9382		0.9400
11	y=0.4208x+86.514	0.9076	y=3.3022ln$x$+84.007	0.9705
12	y=0.3969x+87.233	0.9249	y=3.0722ln$x$+84.960	0.9621
13	y=0.3852x+87.674	0.9195	y=2.9887ln$x$+85.452	0.9613
14	y=0.3591x+88.521	0.9380	y=2.7513ln$x$+86.525	0.9562
15	y=0.3613x+89.246	0.9269	y=2.8031ln$x$+87.163	0.9685
16	y=0.3707x+89.243	0.8961	y=2.9285ln$x$+86.992	0.9712
17	y=0.3770x+89.433	0.9481	y=2.8785ln$x$+87.360	0.9596
18	y=0.3770x+89.676	0.9115	y=2.9500ln$x$+87.449	0.9688
19	y=0.3619x+90.226	0.9037	y=2.8282ln$x$+88.095	0.9581
20	y=0.3418x+90.909	0.9069	y=2.6882ln$x$+88.860	0.9741
Average 11–20		0.9183		0.9650
Average 1–20		0.9283		0.9525

R^2: coefficient of determination of fitted equations; x: the number of traits; y: the value of CR (%)

Table 3 Linear and logarithmic equations with relevant R^2 of values of CR changing with the number of traits, for different sampling percentages

Sampling percentage (%)	Linear equation	R^2	Logarithmic equation	R^2
10	y=0.8657x+74.776	0.8050	y=6.9745ln$x$+69.103	0.9860
11	y=0.8451x+75.643	0.8051	y=6.7924ln$x$+70.138	0.9815
12	y=0.8306x+77.130	0.7818	y=6.7587ln$x$+71.545	0.9767
13	y=0.7953x+78.329	0.7898	y=6.4529ln$x$+73.019	0.9814
14	y=0.8207x+78.883	0.8168	y=6.5755ln$x$+73.582	0.9894
15	y=0.8134x+79.195	0.8338	y=6.4546ln$x$+74.072	0.9909
16	y=0.7398x+80.675	0.8120	y=5.9461ln$x$+75.856	0.9899
17	y=0.7510x+81.242	0.8305	y=5.9599ln$x$+76.512	0.9871
18	y=0.7917x+80.852	0.8042	y=6.3730ln$x$+75.675	0.9834
19	y=0.7707x+81.705	0.8140	y=6.1756ln$x$+76.725	0.9863
20	y=0.7052x+82.870	0.8462	y=5.5583ln$x$+78.510	0.9920
21	y=0.7007x+83.122	0.8414	y=5.5454ln$x$+78.741	0.9946
22	y=0.6916x+83.597	0.8131	y=5.5521ln$x$+79.106	0.9890
23	y=0.6438x+84.743	0.8572	y=5.0462ln$x$+80.821	0.9937
24	y=0.6943x+84.099	0.8239	y=5.5360ln$x$+79.671	0.9884
25	y=0.6347x+85.307	0.8645	y=4.9533ln$x$+81.487	0.9935
26	y=0.6105x+86.028	0.8428	y=4.8221ln$x$+82.231	0.9924
27	y=0.6516x+85.493	0.8332	y=5.1774ln$x$+81.375	0.9928
28	y=0.6392x+85.893	0.8214	y=5.1086ln$x$+81.790	0.9903
29	y=0.5870x+86.848	0.8358	Y=4.6499ln$x$+83.169	0.9896
30	y=0.5957x+86.754	0.8489	y=4.6824ln$x$+83.097	0.9897
Average 10–30		0.8428		0.9885

R^2: coefficient of determination of fitted equations; x: the sampling percentage; y: the value of CR (%)

going up and down repeatedly. The CR logarithmic R^2 also changed up and down repeatedly, but not as dramatically as for the linear R^2 (Fig. 2). The average of the CR's linear R^2 was 0.9382 when the number of traits was less than 10, and was 0.9183 when the number of traits was from 11 to 20. The average of the CR's logarithmic R^2 was 0.9400 when the number of traits was less than 10, and was 0.9650 when the number of traits was from 11 to 20 (Table 2). On the whole, as the number of traits increased, the CR's linear R^2 decreased while the logarithmic R^2 increased (Fig. 2). The CR's linear R^2 changed from 0.9443 to 0.9069, and was 0.9283 on average; the CR's logarithmic R^2 changed from 0.9041 to 0.9741, and was 0.9525 on average (Table 2).

As the sampling percentage increased, the CR's linear R^2 changed sharply, whereas the CR's logarithmic R^2 changed little (Fig. 3). On the whole, the CR's linear R^2 increased and the CR's logarithmic R^2 remained high as the sampling percentage increased (Fig. 3). The CR's linear R^2 changed from 0.8050 to 0.8489, and was 0.8248 on average; the CR's logarithmic R^2 changed from 0.9860 to 0.9897, and was 0.9885 on average (Table 3).

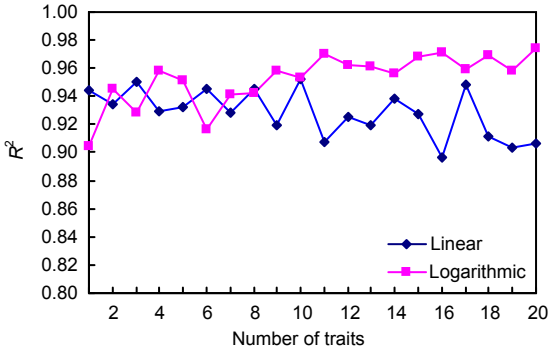

Fig. 2 Coefficient of determination (R^2) of the fitted equation of CR values changing with sampling percentage for different numbers of traits

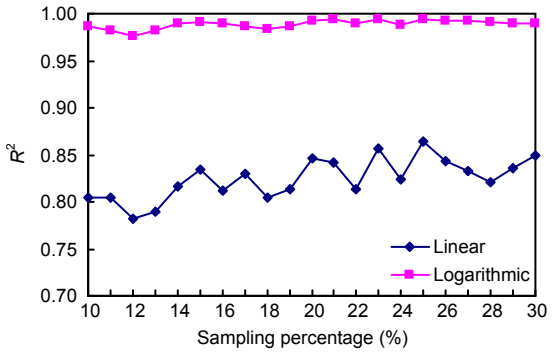

Fig. 3 Coefficient of determination (R^2) of the fitted equation of CR values changing with the number of traits for different sampling percentages

4 Discussion

Cotton breeding is handicapped by a lack of information on genetic diversity. A systematic genetic assessment of gene sources will help to reduce redundancy in the construction of core collections (Kulkarni *et al.*, 2009). Core collection studies of cotton have been conducted for many years (Xu *et al.*, 2006; Mei *et al.*, 2012). Numerous data of both quantitative and qualitative traits of cotton have been collected for core collection construction (Campbell *et al.*, 2010). In this study, twenty quantitative traits were selected. Some of the traits used for characterization (e.g., plant height, lint percentage, and micronaire) are included in the descriptors list of the International Plant Genetic Resources Institute (IPGRI). However, since there were insufficient easily measurable quantitative traits on that list for investigating the effect of the number of traits on the representativeness of a cotton sub-core collection, twenty quantitative traits were selected based on our earlier research (Hu *et al.*, 2000).

Many researchers have reported that more traits do not necessarily mean more representativeness for core collections, and that core collections constructed using all traits available might have even less representativeness than those constructed using fewer traits, if the traits are properly selected by principal component analysis (Malosetti and Abadie, 2001; Upadhyaya *et al.*, 2006; Santesteban *et al.*, 2009). Phenotypic values of traits (especially quantitative traits) are affected greatly by environmental and experimental errors. Genotypic values predicted by mixed linear models have been reported to be a more suitable dataset for core collection construction than phenotypic values (Hu *et al.*, 2000; Li *et al.*, 2004). The results of the present research showed that in sub-core collections constructed by LDSS with predicted genotypic values, more traits (greater than 10) did not reduce the representativeness of sub-core collections, and that representativeness increased steadily as the number of traits increased. The reason might be that the values of the traits used in the present research were treated using a mixed linear model to generate predicted genotypic values and most errors were eliminated by this method (Hu *et al.*, 2000; Wang *et al.*, 2007). However, the representativeness of the sub-core collections increased slowly when the number of traits was quite large. A threshold for the number of traits could be determined by evaluating parameters based on actual needs.

The results of the present research showed that the representativeness of a sub-core collection was greatly affected by two closely connected factors: the number of traits used in sub-core collection construction, and the sampling percentage in core accession sampling. Variation in the genetic diversity of sub-core collections with different sampling percentages showed a linear tendency when the number of traits was small, and a logarithmic tendency when the number of traits was large (greater than 10). However, variation in the genetic diversity of sub-core collections with different numbers of traits always showed a strong logarithmic tendency with changes in the sampling percentage. Therefore, to construct a representative sub-core collection, the following advice might be helpful. When the sampling percentage is relatively small, which may happen due to resource constraints, the number of traits needs to be increased to better sample the genetic diversity information of the population; when the sampling percentage is relatively large, the number of traits can be reduced to save time and money. Traits normally showing large variability are preferred for obvious reasons. The CR shows the extent of preservation of the trait-scope in a core collection, and has been reported to be an important parameter for evaluation of the representativeness of core collections (Frankel and Brown, 1984; Hu *et al.*, 2000; Oliveira *et al.*, 2010). For any germplasm group, a figure showing the variation in the CR in sub-core collections in response to increases in the number of traits and the sampling percentage can be made based on the methods proposed in this study, and a threshold plane of CR (usually not less than 80%) can be drawn. There is a curve of intersection between the curved surface of the CR's changing trend and a threshold plane of the CR. This curve intuitively shows the rational number of traits for the relevant sampling percentage of the sub-core collection (CR threshold method). Further research is needed to find mathematical equations or other rules for that curve.

Acknowledgements

We are grateful to Rui-xiang LI (Liaoning Institute of Economic Crops) and Jun ZHU (College of Agriculture and Biotechnology, Zhejiang University) for the use of part of their data.

References

Biabani, A., Carpenter-Boggs, L., Coyne, C.J., Taylor, L., Smith, J.L., Higgins, S., 2011. Nitrogen fixation potential in global chickpea mini-core collection. *Biol. Fertil. Soils*, **47**(6):679-685. [doi:10.1007/s00374-011-0574-0]

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

Campbell, B.T., Saha, S., Percy, R., Frelichowski, J., Jenkins, J.N., Park, W., Mayee, C.D., Gotmare, V., Dessauw, D., Giband, M., 2010. Status of the global cotton germplasm resources. *Crop Sci.*, **50**(4):1161-1179. [doi:10.2135/cropsci2009.09.0551]

Cheng, Z., Gasic, K., Wang, Z., Chen, X., 2011. Genetic diversity and genetic structure in natural populations of *Prunus davidiana* germplasm by SSR markers. *J. Agric. Sci.*, **3**(4):113-125. [doi:10.5539/jas.v3n4p113]

Díez, C.M., Imperato, A., Rallo, L., Barranco, D., Trujillo, I., 2012. Worldwide core collection of olive cultivars based on simple sequence repeat and morphological markers. *Crop Sci.*, **52**(1):211-221. [doi:10.2135/cropsci2011.02.0110]

Frankel, O.H., Brown, A.H.D., 1984. Plant Genetics Resources Today: a Critical Appraisal. *In*: Holden, J.H.W., Williams, J.T. (Eds.), Crop Gentic Resources: Conservation and Evaluation. George Allen and Unwin, London, UK, p.249-257.

Hu, J., Zhu, J., Xu, H.M., 2000. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor. Appl. Genet.*, **101**(1-2):264-268. [doi:10.1007/s001220051478]

Kang, C.W., Kim, S.Y., Lee, S.W., Mathur, P.N., Hodgkin, T., Zhou, M.D., Lee, R.J., 2006. Selection of a core collection of Korean sesame germplasm by a stepwise clustering method. *Breed Sci.*, **56**(1):85-91. [doi:10.1270/jsbbs.56.85]

Kang, H.M., Sul, J.H., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., Eskin, E., 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**(4):348-354. [doi:10.1038/ng.548]

Kulkarni, V.N., Khadi, B.M., Maralappanavar, M.S., Deshapande, L.A., Narayanan, S., 2009. The worldwide gene pools of *Gossypium arboreum* L. and *G. herbaceum* L., and their improvement. *Genet. Genom. Cotton*, **3**(1):69-97. [doi:10.1007/978-0-387-70810-2_4]

Li, C.T., Shi, C.H., Wu, J.G., Xu, H.M., Zhang, H.Z., Ren, Y.L., 2004. Methods of developing core collections based on the predicted genotypic value of rice (*Oryza sativa* L.). *Theor. Appl. Genet.*, **108**(6):1272-1276. [doi:10.1007/s00122-003-1536-1]

Malosetti, M., Abadie, T., 2001. Sampling strategy to develop a core collection of uruguayan maize landraces based on morphological traits. *Genet. Res. Crop Evol.*, **48**(4):381-390. [doi:10.1023/A:1012003611371]

Mei, Y.J., Zhou, J.P., Xu, H.M., Zhu, S.J., 2012. Development of sea island cotton (*Gossypium barbadense* L.) core collection using genotypic values. *Austr. J. Crop Sci.*, **6**(4):673-680.

Oliveira, M.F., Nelson, R.L., Geraldi, I.O., Cruz, C.D., de Toledo, J.F.F., 2010. Establishing a soybean germplasm core collection. *Field Crops Res.*, **119**(2-3):277-289. [doi:10.1016/j.fcr.2010.07.021]

Pino del Carpio, D., Basnet, R.K., de Vos, R.C.H., Maliepaard, C., Visser, R., Bonnema, G., 2011. The patterns of population differentiation in a *Brassica rapa* core collection. *Theor. Appl. Genet.*, **122**(6):1105-1118. [doi:10.1007/s00122-010-1516-1]

Rao, E.S., Kadirvel, P., Symonds, R.C., Geethanjali, S., Ebert, A.W., 2011. Using SSR markers to map genetic diversity and population structure of solanum pimpinellifolium for development of a core collection. *Plant Genet. Res.*, **10**(1):38-48. [doi:10.1017/S1479262111000955]

Santesteban, L.G., Miranda, C., Royo, J.B., 2009. Assessment of the genetic and phenotypic diversity maintained in apple core collections constructed by using either agro-morphologic or molecular marker data. *Span. J. Agric. Res.*, **7**(3):572-584.

Silvar, C., Casas, A.M., Kopahnke, D., Habekusharp, A., Schweizer, G., Gracia, M.P., Lasa, J.M., Molina-Cano, J.L., Igartua, E., Ordon, F., 2010. Screening the spanish barley core collection for disease resistance. *Plant Breed.*, **129**(1):45-52. [doi:10.1111/j.1439-0523.2009.01700.x]

Smýkal, P., Bačová-Kerteszová, N., Kalendar, R., Corander, J., Schulman, A., Pavelek, M., 2011. Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers. *Theor. Appl. Genet.*, **122**(7):1385-1397. [doi:10.1007/s00122-011-1539-2]

Upadhyaya, H.D., Gowda, C.L.L., Pundir, R.P.S., Reddy, V.G., Singh, S., 2006. Development of core subset of finger millet germplasm using geographical origin and data on 14 quantitative traits. *Genet. Res. Crop Evol.*, **53**(4):679-685. [doi:10.1007/s10722-004-3228-3]

Upadhyaya, H.D., Sarma, N., Ravishankar, C.R., Albrecht, T., Narasimhudu, Y., Singh, S.K., Varshney, S.K., Reddy, V.G., Singh, S., Dwivedi, S.L., 2010. Developing a mini-core collection in finger millet using multilocation data. *Crop Sci.*, **50**(5):1924-1931. [doi:10.2135/cropsci2009.11.0689]

Wang, C.R., Chen, S., Yu, S., 2011. Functional markers developed from multiple loci in gs3 for fine marker-assisted selection of grain length in rice. *Theor.*

Appl. Genet., **122**(5):905-913. [doi:10.1007/s00122-010-1497-0]

Wang, J.C., Hu, J., Xu, H.M., Zhang, S., 2007. A strategy on constructing core collections by least distance stepwise sampling. *Theor. Appl. Genet.*, **115**(1):1-8. [doi:10.1007/s00122-007-0533-1]

Wang, J.C., Hu, J., Huang, X.X., Xu, S.C., 2008. Assessment of different genetic distances in constructing cotton core subset by genotypic values. *J. Zhejiang University-Sci. B*, **9**(5):356-362. [doi:10.1631/jzus.B0710615]

Wulff, S.S., 2009. Evaluation of the mixed linear model with orthogonalized and studentized residuals. *J. Stat. Theory Pract.*, **3**(2):463-476. [doi:10.1080/15598608.2009.10411938]

Xu, H.M., Mei, Y.J., 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. Res. Crop Evol.*, **53**(3):515-521. [doi:10.1007/s10722-004-2032-4]

Zeng, L.H., Meredith, W.R., Boykin, D.L., 2011. Germplasm potential for continuing improvement of fiber quality in upland cotton: combining ability for lint yield and fiber quality. *Crop Sci.*, **51**(1):60-68. [doi:10.2135/cropsci2010.07.0413]

Zhang, J., Wang, Y., Zhang, X.Z., Li, T.Z., Wang, K., Xu, X.F., Han, Z.H., 2010. Sampling strategy to develop a primary core collection of apple cultivars based on fruit traits. *Afr. J. Biotechnol.*, **9**(2):123-127.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.*, **42**(4):355-360. [doi:10.1038/ng.546]

Zhu, J., Weir, B.S., 1996. Diallel analysis for sex-linked and maternal effects. *Theor. Appl. Genet.*, **92**(1):1-9. [doi:10.1007/BF00222944]