RESEARCH ARTICLE

Genetic diversity, population structure and marker trait associations for seed quality traits in cotton (*Gossypium hirsutum*)

ASHOK BADIGANNAVAR[1,2]* and GERALD O. MYERS[1]

[1]*Louisiana State University Agricultural Center, School of Plant, Environmental, and Soil Sciences, 104 M. B. Sturgis Hall, Baton Rouge, LA 70803, USA*

[2]*Present address: Nuclear Agriculture and Biotechnology Division, Bhabha Atomic Research Center, Trombay, Mumbai 400 085, India*

Abstract

Cottonseed contains 16% seed oil and 23% seed protein by weight. High levels of palmitic acid provides a degree of stability to the oil, while the presence of bound gossypol in proteins considerably changes their properties, including their biological value. This study uses genetic principles to identify genomic regions associated with seed oil, protein and fibre content in upland cotton cultivars. Cotton association mapping panel representing the US germplasm were genotyped using amplified fragment length polymorphism markers, yielding 234 polymorphic DNA fragments. Phenotypic analysis showed high genetic variability for the seed traits, seed oil range from 6.47–25.16%, protein from 1.85–28.45% and fibre content from 15.88–37.12%. There were negative correlations between seed oil and protein content. With reference to genetic diversity, the average estimate of F_{ST} was 8.852 indicating a low level of genetic differentiation among subpopulations. The AMOVA test revealed that variation was 94% within and 6% among subpopulations. Bayesian population structure identified five subpopulations and was in agreement with their geographical distribution. Among the mixed models analysed, mixed linear model (MLM) identified 21 quantitative trait loci for lint percentage and seed quality traits, such as seed protein and oil. Establishing genetic diversity, population structure and marker trait associations for the seed quality traits could be valuable in understanding the genetic relationships and their utilization in breeding programmes.

[Badigannavar A. and Myers G. O. 2015 Genetic diversity, population structure and marker trait associations for seed quality traits in cotton (*Gossypium hirsutum*). *J. Genet.* **94**, 87–94]

Introduction

Cotton is grown worldwide for its natural fibre and source of oil, and the cottonseed meal is used as feed for ruminant livestock (Wallace *et al*. 2009). Cottonseed oil is a versatile vegetable oil derived from the seeds of the cotton plant after the cotton lint has been removed, and comprises about 16% of a seed by weight. It is typically composed of about 26% palmitic acid (C16:0), 15% oleic acid (C18:1) and 58% linoleic acid (C18:2). The cottonseed meal is the byproduct after oil extraction and used as a source of fodder protein in the livestock industry, but its use in agriculture is limited. Constituting nearly half of a seed's weight, the meal contains 23% high biological-value protein. Edible cottonseed contains 64 g of protein per 100 g of edible cottonseed and a source of nine essential amino acids, potassium and complex carbohydrates (Heuzé *et al.* 2013). To balance the oil, protein and fibre content in the existing germplasm / cultivars, there is a need to survey the genome to identify genes/controlling elements responsible for these metabolic pathways. Improving the overall productivity of cotton and value of cottonseed by manipulation of quality of cotton seed oil, protein content and removal of toxic gossypol may contribute to increasing the value of cotton both as a fibre and food crop (Mansoor and Paterson 2012).

Protein and oil concentration, kernel index and percentage in cotton are controlled by multiple genes (Singh *et al.* 1985; Dani and Kohel 1989; Ye *et al.* 2003), and are strongly influenced by the environment (Kohel and Cherry 1983). Seed traits may be simultaneously controlled by seed nuclear genes, cytoplasmic genes and maternal nuclear genes (Ye *et al.* 2003). Previous studies have shown significant negative associations between oil and protein content (Kohel and Cherry 1983; Chen *et al.* 1986; Sun *et al.* 1987). Such factors may hinder progress in the simultaneous improvement of

*For correspondence. E-mail: ashokmb1@gmail.com.

Keywords. seed oil; protein; AFLP; upland cotton; association mapping.

these traits in conventional cotton breeding programmes. Genetic mapping provides a useful tool to understand the architecture of quantitative traits at the molecular level. DNA markers linked to QTL controlling seed protein content have been identified in soybean (Chung *et al.* 2003; Panthee *et al.* 2005), rice (Tan *et al.* 2001), barley (See *et al.* 2002) and field pea (Tar'an *et al.* 2004). DNA markers associated with loci controlling seed oil content or fatty acid composition have been identified in soybean (Kianian *et al.* 1999), rapeseed (Zhao *et al.* 2006), sunflower (Bert *et al.* 2003), oilseed mustard (Gupta *et al.* 2004) and canola (Hu *et al.* 2006). In cotton, 11 single QTLs have been associated with oil and protein content (Song and Zhang 2007). Amino acid-specific epistatic QTLs have also been detected, which explain 4.43–9.55% of the phenotypic variation. Using chromosome substitution lines, chromosome 4 (from the *G. barbadense* genotype 3-79, introgressed into a *G. hirsutum* TM-1 background) was associated with seed oil, protein and fibre percentage (Wu *et al.* 2009). A backcross inbred line (BIL) population involving *G. hirsutum* (as recurrent parent) and *G. barbadense* identified 17 QTLs for oil content, 22 for protein and three for gossypol content (Yu *et al.* 2012). Most of the QTL studies on cotton, using traditional mapping methods, are unique to a specific genetic background (biparental cross populations). Association mapping (AM) identifies QTLs that are detected in random set of genotypes from a diverse genetic background (Gupta *et al.* 2005). The concept of AM has been known for many years, but the increased availability of molecular markers and the refinement of statistical tools have created renewed interest in this approach (Achleitner *et al*. 2008). Cotton provides a good platform for genome-wide AM to catalogue genes for fibre traits owing to its vast genetic variation. Few recent reports demonstrated the feasibility of conducting linkage disequilibrium (LD) based AM for fibre traits in tetraploid (Abdurakhmonov *et al.* 2008; Zeng *et al.* 2009) and diploid (Kantartzi and Stewart 2008) accessions. This study was conducted to identify and map genomic regions associated with seed protein, seed oil and fibre content in diverse collection of upland cultivars using AM principles.

Materials and methods

Plant material

The plant material consist of 75 upland cotton germplasm lines derived from different geographical regions, namely, Louisiana (25), Arkansas (17), South Eastern (SE) (22), Delta (4), and Texas/southwest (SW) (7) (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet. Most of the genotypes were selected from advanced breeding lines tested in the Regional Breeder's Trial Network (RBTN), a multistate testing programme of public breeding lines covering different cotton producing regions (http://www.cottonrbtn.com). Plants were field grown as per the Louisiana Cooperative Extension Service guidelines at the Dean Lee Research Station in Alexandria, LA. Leaf samples from the representative plants were collected and bulked for DNA extraction. Phenotypic data on yield were obtained from the RBTN trial website (http://www.cottonrbtn.com) of the USA. Replicate data on lint percentage was averaged to calculate variances using SAS 2009 (SAS 9.1.3, SAS Institute, Cary, USA). Deltapine, DP 393 (PVP 200400266) was considered as the check variety and all the comparisons were made in relation with the performance of this cultivar. LP values of other cultivars in the panel were adjusted based on the relative performance of the check variety, DP 393.

From remnant planting seeds, 10 g of acid delinted seeds for each cultivar were sent to the Department of Agricultural Chemistry, LSU AgCenter, Baton Rouge, Louisiana, to determine total oil, protein and fibre contents. The seed quality traits were determined following the modified American Oil Chemist's Society (AOCS) methods of analysis protocols. Seed protein was estimated using the nitrogen combustion method (AOAC 990.03) (AOAC 1999); crude fat/oil content by petroleum ether as solvent using Soxtec System HT6; and crude fibre content by AOCS 962.09. Two replications were run and averaged over each cultivar. Correlation analysis between LP and seed traits was performed using PROC CORR in SAS.

Genotyping with amplified fragment length polymorphism (AFLP) markers

Sixty-four primer combinations were used to generate AFLP data (see table 1 in electronic supplementary material) following the procedure given by Vos *et al.* (1995) with minor modifications. Sample DNA was digested with *Eco*RI and *Mse*I restriction enzymes and oligonucleotide adapters specific to these restriction sites were ligated to the resulting fragments through incubation (37°C for 180 min) with DNA ligase. Preamplifications were done using *Eco*R I+A and

Table 1. Univariate analysis of LP and seed quality traits in upland cotton germplasm lines.

Trait	Min.	Max.	Mean	SE	Variance	SD	Median
Protein	18.05	28.45	23.80	0.31	7.28	2.69	24.30
Oil	6.47	25.16	18.09	0.32	7.80	2.79	18.02
Fibre	15.88	26.53	19.83	0.23	4.20	2.05	19.37
LP	35.67	57.35	42.97	0.57	22.67	4.76	41.53

SE, standard error; SD, standard deviation; LP, lint percentage.

Mse I+C oligo primers and selective amplification was carried out using the IR dye labelled *Eco*RI+ANN oligo primers (MWG Biotech, Ebersberg, Germany). Touchdown polymerase chain reaction (PCR) was used for selective amplifications using the following profile: initial denaturing step at 94°C for 2 min followed by initial 12 cycles at 94°C for 30 s, 65°C for 30 s (with 0.7°C decrement at every cycle) and 72°C for 1 min, then followed by 23 cycles at 94°C for 30 s, 56°C for 30 s, and 72°C for 1 min with a final extension step at 72°C for 2 min. The PCR amplified products were run on a LI-COR 4300 Sequencer (LI-COR, Lincoln, USA) and scored. The nomenclature of AFLP loci was followed according to Lacape *et al.* (2003), Myers *et al.* (2009) and Badigannavar *et al.* (2012), indicating the enzyme primer combinations with band size.

Molecular diversity and cluster analysis

For each marker used, subpopulationwise diversity statistics including the number of observed and effective alleles, Nei's genetic distances (Nei and Li 1979), expected heterozygosity and Shannon's information index were calculated using GenAlEx 6.5 software (Peakall and Smouse 2012). Allelic diversity at a given locus can be determined by polymorphic information content (PIC) and was calculated using the formula, 'PIC$= 1 - \sum f_i^2$, where f_i is the frequency of the ith allele (Weir 1996). PROC ALLELE was used to calculate PIC values and frequency estimate was done using PROC FREQ (SAS 9.1.3, SAS Institute, Cary, USA). Genetic differentiation among the subpopulation was estimated using hierarchical analysis of molecular variance (AMOVA; Excoffier *et al.* 2005) in GenAlEx 6.2. Dice similarity coefficient was calculated using the formula $D = 2a/(2a + b + c)$, where a is the number of fragments present in both accessions, b and c are the numbers of fragments that are present in either accession, respectively (Sneath and Sokal 1973). From the similarity data, genetic distance was calculated for each pair of germplasm lines and dendrogram was generated using the neighbour joining (NJ) analysis in MEGA 4.0 (Kumar *et al*. 2004) software.

Population structure and association analysis

A Bayesian model based clustering was performed using the software program 'STRUCTURE' 2.3.2 (Pritchard *et al.* 2000). The admixture model was selected in the software and allele frequencies among populations were assumed to be correlated. Each run was carried out using 10,000 iterations and 10,000 replications. A total of 2–10 k clusters were evaluated and the optimum number of cluster was determined by LnP(D) probabilities (Evanno *et al.* 2005). A graphical display (bar plot) of the population structure was generated using DISTRUCT software (Rosenberg *et al*. 2002). The pairwise kinship (K matrix) was calculated using SPAGeDi software (Hardy and Vekemans 2002). The K matrices and Q matrix describing the assignment of each genotype to specific clusters were used in mixed linear model association analyses. The mean phenotypic data was used for association analysis. Significant marker trait associations were tested using two different models, a general linear model (GLM) and a mixed linear model (MLM) in TASSEL 2.1 software (Bradbury *et al.* 2007). In GLM model, population substructure of cotton mapping panel was incorporated as covariates. In MLM, association was estimated by simultaneous accounting of multiple levels of population structure (Q matrix), relative kinship among the individuals (K matrix) and eigenvectors of PCoA as described by Yu *et al.* (2006).

Results

Phenotypic analyses

Of the upland cotton germplasm lines studied, 69 were developed by public breeding programmes and six by commercial seed companies which originate from five relatively distinct geographical regions. Among the traits analysed, LP varied from 35.67 to 57.35% (average 42.97%) (table 1). Among the seed traits, seed protein content ranged from 18.05 to 28.45% (average 23.8%), oil content ranged from 6.47 to 25.16% (average 18.09%), and fibre content varied from 15.88 to 26.53% (average 19.83%). LP showed greater genetic variability followed by protein as compared to other traits. High heritability values (up to 90%) were observed for quality parameters, while LP showed moderate heritability (49%).

The correlations among the yield and quality traits are presented in table 2. There were significant negative correlations between fibre content with oil and protein percentage ($P = 0.0001$). While not significant, protein and oil percentages were negatively correlated (-0.224) which complicates

Table 2. Correlation coefficients among LP and seed quality traits.

Trait	LP	Protein	Oil
LP	1		
Protein	−0.240	1	
Oil	0.027	−0.224	1
Fibre content	0.033	−0.340*	−0.61**

* Significant at $P \leq 0.05$; ** significant at $P \leq 0.01$.

Table 3. Genetic diversity parameters among cotton subgroups.

Subgroup	N_a*	N_e	I	H_e	UH_e
LA	1.885	1.357	0.369	0.231	0.240
ARK	1.829	1.395	0.386	0.248	0.269
SE	1.966	1.517	0.477	0.313	0.328
DELTA	1.560	1.436	0.372	0.244	0.325
SW/T	1.910	1.557	0.492	0.329	0.362

*N_a, no. of different alleles; N_e, no. of effective alleles; I, Shannon's index; H_e, expected heterozygosity; UH_e, unbiased expected heterozygosity; LA, Louisiana; ARK, Arkansas; SE, South Eastern; SW/T, southwest/Texas.

their simultaneous improvement into a single cultivar. All other correlations, particularly those between LP with seed quality traits were not significant.

Genetic analyses

A total of 234 polymorphic loci were obtained when screened across 75 cotton germplasm lines. The Shannon index, a measurement used to compare diversity between two or more subpopulations ranged between 0.35 and 0.49 (table 3). The number of effective alleles was highest for SW/T (1.57) while lowest for LA genotypes (1.36). The heterozygosity for the AFLP markers ranged from 0.23 (LA) to 0.32 (SW/T).

The frequency distribution values for relative kinship revealed that the genetic relatedness ranged from 0 to 0.9 (figure 1). Although 60% of the pairwise kinship estimates were below 0.5, there were moderate peaks around 0.7 and 0.8. Genetic relatedness is often prominent among elite genotypes, as they share common genotypes in their breeding programmes. The PIC measures how different

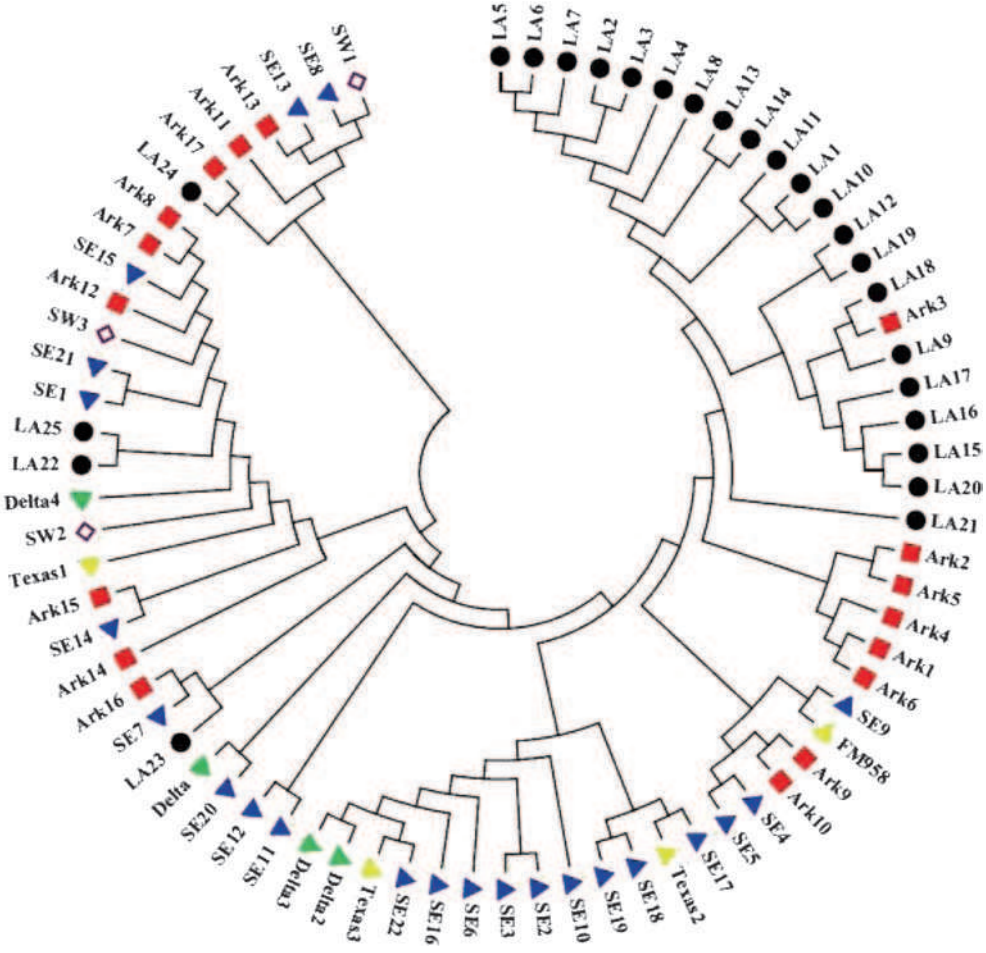

Figure 1. Dendrogram representing genetic relationship among upland cotton germplasm lines generated by NJ analysis.

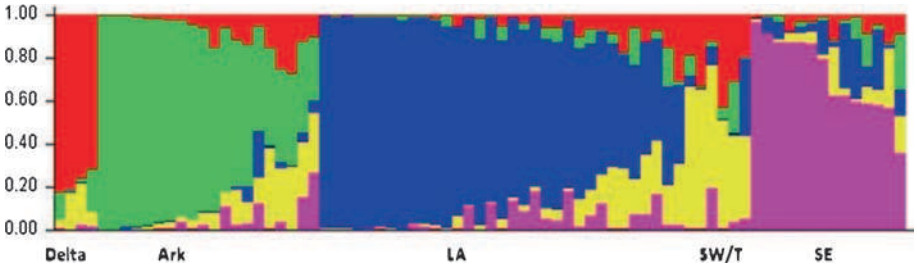

Figure 2. Bar plot representing population structure of upland cotton lines grouped into five subpopulations. Each individual genotype is represented by a line partitioned in five coloured segments that represent the estimated membership fractions to each one of the five subgroups. The bar plot was generated using Structure (Pritchard *et al.* 2000) software following the admixture model.

populations are distinguished based on probability of randomly chosen alleles. The frequency distribution for PIC using AFLP markers ranged from 0 to 0.40 with more than 90% of them falling between 0.16 and 0.40.

Population structure and genetic diversity study

Using the entire upland cotton AM panel, population structure was analysed using software 'STRUCTURE'. The parameters used for this analysis produced the highest log-likelihood score when K was 5. The identified subgroups highly correspond to the five geographical regions from where these lines have been derived. The bar plot indicated LA genotypes showing uniformity with fewer admixtures, mainly from Delta, SW/T and Ark ancestral genes (figure 2). Substantial amount of admixture was seen to occur among the clusters.

To estimate genetic diversity within and among the predefined subpopulations, Wright's F_{ST} index (Wright 1951) was calculated (table 4). Based on the pairwise F_{ST} estimates, SW/T (southwest/Texas) and Delta were closely related (0.0078), while Delta and LA group were highly diverse (0.141). The average estimate of F_{ST} was 0.052 indicating a low level of genetic differentiation among the groups. Estimates of molecular variation present in the upland genotypes revealed that although 94% of the genetic diversity was attributable to differences within populations, still there was 6% variation among groups ($P = 0.001$) (table 5). NJ analysis of genotypic data for the upland germplasm lines identified five major clusters as per their geographical distribution (figure 1). To correlate STRUCTURE results and that of NJ analysis, we compared the assignment of each of these germplasm lines into a defined cluster. Except for few cases, overall, there was good agreement for the phylogenetic relationships between the two estimates. Germplasm lines originated from Louisiana formed one major cluster, while Arkansas formed two major clusters. SE also formed a major cluster and SW, Texas and Delta germplasm lines were highly diverse and were found scattered.

Association analyses

AM was performed for LP and seed quality traits using GLM and MLM models using TASSEL software. The effectiveness of these models in controlling false positives was determined by monitoring the partial R^2 values. GLM was tested to identify single marker effects on quantitative traits. Partial R^2 was least for GLM models across all the traits. But models using PCA eigenvectors explained more variation (16–30%) than the models with structure (12–25%). The naïve MLM model, which included the kinship matrix, explained more genetic variation (up to 50%) compared to naïve GLM model (15%). LP showed high amount of partial R^2 across all the models compared to seed quality traits. Using MLM (Q+K model) 21 significant QTLs were found associated with four traits (table 6). Five each QTLs were associated with LP, protein and fibre content, while six QTLs were associated with seed oil. E3M6_260 and E4M4_242 markers were found to be significantly associated with seed oil and fibre content. The partial R^2 values ranged from 28.47 to 88.90%. Seed protein was significantly associated with five QTLs, among them, E6M2_640 recorded the high partial R^2 value (91.8%).

Table 4. Pairwise F_{ST} values estimated for cotton subgroups.

F_{ST}	LA*	ARK	SE	DELTA
ARK	0.0823	0		
SE	0.0909	0.017	0	
DELTA	0.141	0.0492	0.0174	0
SW/T	0.0983	0.0212	0.0095	0.0078

* F_{ST}, Wright's fixation index.

Table 5. Analysis of AMOVA for upland cotton germplasm lines between and within five subgroups.

Source of variation	*df*	Sum of squares	Estimated variance	% variance	*P* value
Among pops	4	268.47	2.33	6	0.001
Within pops	70	2409.77	34.42	94	0.001
Total	74	2678.41	36.75		

Discussion

An appropriate association mapping panel should encompass diverse genetic background such that efficient marker system could be employed to infer true associations (Flint-Garcia *et al.* 2005). In this study, phenotypic data on LP and seed quality traits suggested wide variability for protein, oil and fibre content. Previously *G.hirsutum* and *G. arboreum* cotton accessions also reported wide variability for oil and seed weight (Kohel 1978; Song and Zhang 2007). We noticed 6.47% of seed oil in *G. herbaceum* and 25.65% in *G.hirsutum* accessions.

Seed quality traits are directly influenced by the lint percentage, seed cotton yield, seed number, seed weight, seed coat content, moisture level and environmental factors. We observed positive correlations between LP and oil, while negative correlations between LP and protein content were also noticed. Typically, high yielding plant has a high LP which is most easily achieved by decreasing seed size. In this study, fibre content was determined from hulled seeds. The hull is expected to be higher in fibre than the embryo, such that seed size decreases with increase in per cent fibres. Similarily, since a majority of seed protein is in the embryo, with the increase in lint percentage (smaller seed), protein percentage is expected to decrease. Simultaneous improvement of oil and protein is complicated, owing to their negative correlation. According to Kohel *et al.* (1985) and Gotmare *et al.* (2004), the relationship between percentage of protein and oil are significantly negative. Oil and protein percentages in seed also decrease with harvest date, but the greatest change is in the amount of oil (Kohel and Cherry 1983). Several studies have been conducted to understand the inheritance pattern and gene action governing quality traits (Mert *et al.* 2004). Seed index was found to be predominantly under the control of genes acting additively. This trait could easily be manipulated through selection for the production of pure line varieties. Oil content is governed by dominant genes (Singh *et al.* 1985), while significant epistatic interaction was observed for oil percentage and seed index (Dani and Kohel 1989). Although the effects of environment and genotype on oil and protein content are well documented and relationships between yield, seed quality and fibre properties in cotton have been identified, studies on the inheritance and genetic factors governing these traits have not been widely addressed. This may be due to the lack of understanding of the complex pathways and multiple genes interacting in an epistatic manner controlling these traits.

Table 6. Significant QTL ($P < 0.05$) for LP and seed quality traits in upland cotton identified using MLM in TASSEL.

Trait	AFLP marker	*P* value	R^{2*}
LP	E3M6_260	0.0009	58.53
	E4M1_365	0.0032	53.53
	E4M4_242	0.0001	46.12
	E6M4_249	0.004	34.50
	E5M3_65	0.016	39.10
Oil	E3M2_145	0.002	28.47
	E5M7_180	0.004	34.90
	E5M7_195	0.005	35.03
	E4M3_214	0.013	36.00
	E6M4_358	0.015	30.45
	E3M6_260	0.017	45.89
Protein	E6M2_640	0.005	81.18
	E4M1_382	0.010	88.90
	E4M4_217	0.011	84.09
	E4M1_353	0.013	88.22
	E6M3_190	0.013	88.33
Fibre	E3M6_260	4.42×10^{-4}	60.03
	E5M5_415	0.0013	39.55
	E3M2_145	0.0052	69.27
	E6M4_303	0.0074	54.31
	E3M8_125	0.0086	62.11

*Adjusted R^2, indicates the percentage of explained variation.

Genomewide AM is successful only if appropriate methods are implemented to control the effect of population structure. Inclusion of population structure would minimize type I errors due to the spurious associations between non-linked loci. The six models used in this study, accounting for Q (population structure) or for K (kinship estimates) or PCA (eigenvectors of PCA) primarily aimed at reducing the type I error. For all the traits under study, the models controlling relative kinship performed better than the model controlling population structure. Similarly, model controlling structure is better than PCA in explaining the phenotypic variation. In addition, MLM models identified 21 QTL for LP and seed quality traits. Among all, E3M6_260 was significant for LP, seed oil and fibre content, while E3M2_245 was associated with seed oil and fibre content. The number of QTLs also decreased drastically with high partial R^2 value when population structure was included in the MLM model. Although inclusion of PCA values did change R^2 values substantially, but Q+K MLM models recorded higher partial R^2 across all the traits.

During the recent years, molecular marker technology was successfully applied in cotton diversity studies, creating genetic linkage maps and identifying QTL for fibre traits using biparental cross derivatives or association mapping panel. Compared to other field crops, association mapping in cotton has not been explored to a great extent. A recent study by Kantartzi and Stewart (2008) identified 30 marker trait associations with 19 SSR markers in *G. arboreum* germplasm lines. The MLM models greatly reduced type I error and revealed true associations for fibre traits. However, measurement of the LD patterns for genomic regions and extent of LD among different populations of the target organisms is the start point to design and execute association mapping. A recent study reported significant LD between pair of SSR loci within 36–37 cM distance in the diverse upland cotton germplasm lines (Abdurakhmonov *et al.* 2008). Due to relatively less number of markers used in finding associations may result in low resolution.

Our results demonstrated the efficiency of MLM models in identifying true associations for seed quality traits. Adding more number of markers and expanding the mapping panel would result in greater precision and power. Looking at the complex pathways involved in the synthesis of oil and protein, the addition of more markers to catalogue multi-environment phenotypic variations would also improve the understanding of the genetic factors governing these traits.

Acknowledgements

We thank the Department of Agricultural Chemistry, LSU, Baton Rouge for seed quality trait analysis and all the RBTN coordinators for providing phenotypic data. Financial support from Cotton Incorporated is highly appreciated.

References

Abdurakhmonov I. Y., Kohel R. J., Yu J. Z., Pepper A. E., Abdullaev A. A., Kushanov F. N. *et al.* 2008 Molecular diversity and association mapping of fibre quality traits in exotic *G. hirsutum* L. germplasm. *Genomics* **92**(6), 478–487.

Achleitner A., Nicholas A., Tinker Zechner E. and Buerstmayr H. 2008 Genetic diversity among oat varieties of worldwide origin and associations of AFLP markers with quantitative traits. *Theor. Appl. Genet.* **117** (7), 1041–1053.

AOAC 1999 *Official methods of analysis.* 16th edition. Association of official analytical chemists, Washington, USA.

Badigannavar A. M., Myers G. O. and Jones D. C. 2012 Molecular diversity revealed by AFLP markers in upland cotton genotypes. *J. Crop Improv.* **26**, 627–640.

Bert P. F., Jouan I., Tourvieille de Labrouhe D., Serre F., Philippon J., Nicolas J. and Vear P. 2003 Comparative genetic analysis of quantitative traits in sunflower (*Helianthus annuus* L.). 2. Characterization of QTL involved in developmental and agronomic traits. *Theor. Appl. Genet.* **107**, 181–189.

Bradbury P. J., Zhang Z., Kroon D. E., Casstevens T. M., Ramdoss Y. and Buckler E. S. 2007 TASSEL: Soft ware for association mapping of complex traits in diverse samples. *Bioinform* **23**, 2633–2635.

Chen Z. F., Zhang Z. W. and Cheng H. L. 1986 The analysis of upland cotton quality. *Acta Agron. Sin.* **12**, 195–200.

Chung J., Babka H. L., Graef G. L., Staswick P. E., Lee D. J., Cregan P. B., Shoemaker R. C. and Specht J. E. 2003 The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Sci.* **43**, 1053–1067.

Dani R. G. and Kohel R. J. 1989 Maternal effects and genera tion mean analysis of seed-oil content in cotton (*Gossypium hirsutum*). *Theor. Appl. Genet.* **77**, 569–575.

Evanno G., Regnaut S. and Goudet J. 2005 Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* **14**, 2611–2620.

Excoffier L., Laval G. and Schneider S. 2005 Arlequin ver. 3.0: an integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **1**, 47–50.

Flint-Garcia S. A., Thuillet A. C., Yu J., Pressoir G., Romero S. M., Mitchell S. E. *et al.* 2005 Maize association population: a high resolution platform for quantitative trait locus dissection. *Plant J.* **44**, 1054–1064.

Gotmare V., Singh P., Mayee C. D., Deshpande V. and Bhagat C. 2004 Genetic variability for seed oil content and seed index in some wild species and perennial races of cotton. *Plant Breed.* **123**, 207–208.

Gupta P., Rustgi S. and Kulwal P. 2005 Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol. Biol.* **57**(4), 461–485.

Gupta V., Mukhopadhyay A., Arumugam N., Sodhi Y. S., Pental D. and Pradhan A. K. 2004 Molecular tagging of erucic acid trait in oilseed mustard (*Brassica juncea*) by QTL mapping and single nucleotide polymorphisms in FAE1 gene. *Theor. Appl. Genet.* **108**, 743–749.

Hardy O. J. and Vekemans X. 2002 SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2**, 618–620.

Heuzé V., Tran G., Bastianelli D., Hassoun P. and Lebas F. 2013 Cottonseed meal. Feedipedia.org. A programme by INRA, CIRAD, AFZ and FAO (http://www.feedipedia.org/node/550).

Hu X., Sullivan-Gilbert M., Gupta M. and Thompson S. A. 2006 Mapping of the loci controlling oleic and linolenic acid contents and development of *fad2* and *fad3* allele-specific markers in canola (*Brassica napus* L.) *Theor. Appl. Genet.* **113**, 497–507.

Kantartzi S. K. and Stewart J. M. 2008 Association analysis of fibre traits in *Gossypium arboreum* accessions. *Plant Breed.* **127**, 173–179.

Kianian S. F., Egli M. A., Phillips R. L., Rines H. W., Somers D. A., Gengenbach B. G. *et al.* 1999 Association of a major groat oil content QTL and an acetyl-CoA carboxylase gene in oat. *Theor. Appl. Genet.* **98**, 884–894.

Kohel R. J. 1978. Survey of *G. hirsutum* germplasm collections for seed oil percentage and seed characteristics. USDA-ARS Report. S-187.

Kohel R. J. and Cherry J. P. 1983 Variation of cottonseed quality with stratified harvests. *Crop Sci.* **23**, 1119–1124.

Kohel R. J., Glueck J. and Rooney L. W. 1985 Comparison of cotton germplasm collections for seed protein content. *Crop Sci.* **25**, 961–963.

Kumar S., Tamura K. and Nei M. 2004 MEGA4: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform.* **5**, 150–163.

Lacape J. M., Nguyen T. B., Thibivilliers S., Bojinov B., Courtois B., Cantrell R. G. *et al.* 2003 A combined RFLP–SSR–AFLP map of tetraploid cotton based on a *Gossypium hirsutum* × *Gossypium barbadense* backcross population. *Genome* **46**, 612–626.

Mansoor S. and Paterson A. H. 2012 Genomes for jeans: cotton genomics for engineering superior fibre. *Trends Biotech.* **30**, 521–527.

Mert M., Akiscan Y. and Gencer O. 2004 Inheritance of oil and protein in some cotton generations. *Asian J. Plant Sci.* **3**, 174–176.

Myers G. O., Baogong J., Akash M. W., Badigannavar A. M. and Saha S. 2009 Chromosomal assignment of AFLP markers in upland cotton (*Gossypium hirsutum* L.) *Euphytica* **165**, 391–399.

Nei M. and Li W. H. 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.

Panthee D. R., Pantalone V. R., West D. R., Saxton A. M. and Sams C. E. 2005 Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. *Crop Sci.* **45**, 2015–2022.

Peakall R. and Smouse P. E. 2012 GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. *Bioinformatics* **28**, 2537–2539.

Pritchard J. K., Stephens M. and Donnelly P. 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Rosenberg N., Pritchard J. K., Weber J. L., Cann H. and Kidd K. 2002 Genetic structure of human populations. *Science* **298**, 2381–2385.

SAS 2009 SAS Statistical Analysis Software for Windows 9.1.3. Cary, USA.

See D., Kanazin V., Kephart K. and Blake T. 2002 Mapping genes controlling variation in barley grain protein concentration. *Crop Sci.* **42**, 680–685.

Singh M., Singh T. H. and Chahal G. S. 1985 Genetic analysis of some seed quality characters in Upland cotton (*Gossypium hirsutum* L.) *Theor. Appl. Genet.* **71**, 126–128.

Sneath P. H. A. and Sokal R. R. 1973 *Numerical taxonomy: The principals and practice of numerical classification,* pp. 573. Freeman, San Francisco, USA.

Song X. and Tian-Zhen Zhang 2007 Identification of quantitative trait loci controlling seed physical and nutrient traits in cotton. *Seed Sci. Res.* **17**, 243–251.

Sun S. K., Chen J. H., Xian S. K. and Wei S. J. 1987 Study on the nutritional quality of cotton seeds. *Sci. Agric. Sin.* **5**, 12–16.

Tan Y. F., Sun M., Xing Y. Z., Hua J. P., Sun X. L., Zhang Q. F. and Corke H. 2001 Mapping quantitative trait loci for milling quality, protein content and color characteristics of rice using a recombinant inbred line population derived from an elite rice hybrid. *Theor. Appl. Genet.* **103**, 1037–1045.

Tar'an B., Warkentin T., Somers D. J., Miranda D., Vandenberg A., Blade and Bing D 2004 Identification of quantitative trait loci for grain yield, seed protein concentration and maturity in field pea (*Pisum sativum* L.) *Euphytica* **136**, 297–306.

Vos P., Hogers R., Bleeker M., Reijans M., Van de Lee T., Hornes M. *et al.* 1995 AFLP: A new technique for DNA fingerprinting. *Nucl. Acids Res.* **23**, 4407–4414.

Wallace T. P., Bowman D., Campbell B. T., Chee P., Gutierrez O. A., Kohel R. J. *et al.* 2009 Status of the USA cotton germplasm collection and crop vulnerability. *Genet. Resour. Crop Evol.* **56**, 507–532.

Weir B. S. 1996 *Genetic data analysis II*. Sinauer Associates, Sunderland, USA.

Wright S. 1951 The genetical structure of populations. *Ann. Eugen.* **15**, 323–354.

Wu J., Jenkins J. N., McCarty J. C. and Thaxton P. 2009 Seed trait associations with *Gossypium barbadense* L. chromosomes/arms in a *G. hirsutum* L. background. *Euphytica* **167**, 371–380.

Ye Z. H., Lu Z. Z. and Zhu J. 2003 Genetic analysis for developmental behavior of some seed quality traits in Upland cotton (*Gossypum hirsutum* L.). *Euphytica* **129**, 183–191.

Yu J., Pressoir G., Briggs W. H., Vroh B. I., Yamasaki M., Doebley J. F. *et al.* 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208.

Yu J., Zhang K., Li S., Yu S., Zhai H., Wu M. *et al.* 2012 Mapping quantitative trait loci for lint yield and fibre quality across environments in a Gossypium hirsutum × Gossypium barbadense backcross inbred line population. *Theor. Appl. Genet.* **126**, 275–287.

Zeng L., Meredith W. R. Jr, Gutiirrez O. A. and Boykin D. L. 2009 Identification of associations between SSR markers and fibre traits in an exotic germplasm derived from multiple crosses among Gossypium tetraploid species. *Theor. Appl. Genet.* **119**(1), 93–103.

Zhao J. Y., Becker H. C., Zhang D. Q., Zhang Y. F. and Ecke W. 2006 Conditional QTL mapping of oil content in rapeseed with respect to protein content and traits related to plant development and grain yield. *Theor. Appl. Genet.* **113**, 33–38.

Received 3 February 2014, in revised form 30 September 2014; accepted 17 October 2014
Unedited version published online: 20 October 2014
Final version published online: 12 March 2015