2010

Characterization of quantitative traits using association genetics tetraploid and genetic linkage mapping in diploid cotton (Gossypium spp.)

Ashok Badigannavar
Louisiana State University and Agricultural and Mechanical College, ashokmb1@gmail.com

CHARACTERIZATION OF QUANTITATIVE TRAITS USING ASSOCIATION GENETICS IN TETRAPLOID AND GENETIC LINKAGE MAPPING IN DIPLOID COTTON (*GOSSYPIUM* SPP.)

A Dissertation
Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The School of Plant, Environmental, and Soil Sciences

By
Ashok Badigannavar
BSc (Agri), University of Agricultural Sciences, Dharwad, India, 1997
MSc (Agri), University of Agricultural Sciences, Dharwad, India, 1999
May 2010

This dissertation is dedicated in memory of my late father…..

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my major professor Dr.Gerald O Myers for his valuable guidance and advice throughout my dissertation. The freedom he has provided during the course of study is commendable. My appreciation is also extended to the committee members; Dr.Rajasekharan, Dr.Collins Kimbeng, Dr. C. Knott and Dean nominee, Dr. Fangneng Huang. I am indebted to all my committee members for their incredible help and guidelines in the dissertation research, analysis and writing. I would like to thank Dr. Oard for his valuable suggestions in formulating the research proposal and Dept. Head and staff for their encouragement and support.

The financial support provided by the research sponsor, Cotton Incorporated is greatly appreciated. I would extend my sincere gratitude to Dr. A. H. Paterson for providing seeds of the diploid mapping population, Dr. F. Bourland, for providing the Arkansas genotypes, various public breeders and private seed companies allowing me to use their varieties, Dr. M. Ulloa for providing SSR markers, the LSU AgCenter, Cotton Fiber Lab for fiber sample analysis and to Dr. C.Kimbeng for their cooperation. The public breeders of the RBTN trials are acknowledged for providing the multistate cotton phenotypic data. I am extremely grateful to Computational Biology Unit, Microsoft High Performance Computing Facility, Cornell University for allowing me to use clusters for simulation studies.

I would like to thank my fellow colleagues, Jimmy Zumba, Tyson, Sreedhar, Suman, Samuel, Reddy, Rakesh, Charanjit and Marvellous for their help during the course of study. I am also thankful to Dr. N. Baisakh for his valuable suggestions during the course of study.

I would also like to thank my mother, brother, mama and niece for their love, continuous encouragement and being my motivational force. My heartfelt thanks are also extended to all my Indian and International friends without whom I would have missed the fun and wonderful life in Baton Rouge.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Adj. R^2	Adjusted Rsquare
AFLP	Amplified fragment length polymorphism
AIC	Akaike information content
AM	Association mapping
BIC	Bayesian information content
CAM	Cotton association mapping
CV	Predicted residual sum of squares with k-fold cross validation
DA	Discriminant analysis
ELO	Elongation percentage
EST	Expressed sequence tag
FL	Fiber strength
FS	Fiber strength
GLM	General linear model
HVI	High volume instrumentation
LD	Linkage disequilibrium`
LG	Linkage groups
LOD	Logarithm of odds
LY	Lint yield
MCMC	Markov chain monte carlo
MIC	Micronaire
MIM	Multiple interval mapping
MLM	Mixed linear model
MMR	Mixed multiple regression
PER	Protein efficiency ratio
PRESS	Predicted residual sum of squares
QTL/A	Quantitative trait loci/allele
RAPD	Random amplified polymorphic DNA
RBTN	Regional breeder’s trial network
RFLP	Restricted fragment length polymorphism
RIL	Recombinant inbred line
SBC	Schwarz Bayesian content
SCY	Seed cotton yield
SFI	Short fiber index
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
TRAP	Target region amplification polymorphism
UNI	Uniformity index

ABSTRACT

Cotton (*Gossypium spp.*) is the most extensively used natural fiber in the textile industry. Understanding the genetic diversity, population structure and marker trait associations are of great importance in marker assisted selection.

Microsatellite, AFLP and TRAP markers were used to construct a linkage map with 94 F_2 diploid individuals derived from a cross between *G. arboreum x G. herbaceum*. A total of 606 polymorphic markers gave rise to 37 linkage groups covering a total of 1109cM with an average distance of 7.92cM between each loci. Discriminant analysis identified three markers each for petal color and seed fuzziness, and four markers for petal spot. For quantitative traits, a total of 19 QTL's were identified and linked with five fiber traits using composite interval mapping. Markers such as qFL4-1, qFS4-2, qELO1-1 and qSI2-1 were found to be significantly linked with fiber length, strength, elongation and seed index respectively.

Association mapping principles were applied to upland cotton genotypes in order to examine population structure and marker trait associations. A set of 232 genotypes were genotyped using AFLP markers. The molecular diversity was in the range of 0.48-0.574 with molecular variance found to be 10% among the groups. Bayesian and MCMC based population structure analysis, there existed six subpopulations, in accordance with their geographical origin. The mixed and mixed-multiple regression (MMR) models identified significant markers for lint yield and fiber traits, showing low AICC, BIC and SBC values and high adj. R^2. Two way epistatic interaction analyses further confirmed their strong association.

In the similar study, a set of 75 upland cotton genotypes were analyzed for seed quality traits such as seed protein, oil and fiber content. Population structure based mixed models showed 32 significant markers, associated with these seed quality traits. MMR models identified several markers, notably E4M3_440, E4M3_200 and E5M7_195 for seed protein, oil and fiber content respectively.

Finally, 60 upland genotypes from RBTN program were screened with AFLP markers. The pairwise kinship estimates were ranging between 0.1-0.88 accounting for most of the shared ancestral alleles. The MMR models improved the efficiency of marker selection with 38 markers associated with eight traits.

CHAPTER 1 GENERAL INTRODUCTION

Cotton (*Gossypium* spp.) is the most extensively used natural fiber in the textile industry and is the sixth most abundantly grown oilseed crop. It is grown commercially in the tropical and subtropical regions of more than 50 countries. Worldwide, cotton production has been relatively stable for the last several years. In the United States however, planted acreage fell to 9.1 million in 2009 the lowest since 1983 and well below the 15.5 million acres planted in 2006. In Louisiana, producers planted 240,000 acres and are expected to harvest 420,000 bales, up 49 percent from last year's hurricane devastated crop (NASS, 2009). Due to the global economic downturn, world cotton consumption fell by 12% in 2008-2009 after a decade of uninterrupted growth. As the world economy gradually stabilizes, world cotton use is also expected to recover slowly. Increases in cotton consumption will mainly be driven by a rebound in Asia, in particular China (mainland), India and Pakistan.

Genetic improvements that enhance the economics of production and fiber processing characteristics will allow this natural renewable product to compete in favorably in the market place with petroleum derived synthetic fibers and enrich the livelihoods of millions of people worldwide. Therefore, over the years scientists have set a broad goal for genetic improvement of cotton through concerted application of traditional plant breeding, genetic engineering and molecular genetics tools. Traditionally, cotton being polyploid, has been considered as an excellent model system for studying plant genome size evolution, polyploidization and its fiber for single celled biological processes. Elucidating the cotton genomes will significantly contribute to our understanding of the functional and agronomic significance of polyploidy. The genus *Gossypium* consist of 45 diploid species divided in to 8 subgenomes (A-G and K) and five tetraploids (AD, Brubaker *et al.*, 1999). Of all the *Gossypium* species, two tetraploids (*G. hirsutum* and *G. barbadense*; 2n=4x=52) and two diploid species (*G. arboreum* and *G. herbaceum*; 2n=2x=26) are commercially grown for natural fiber. *G.*

hirsutum and *G. barbadense* being natural allopolyploids are derived from an interspecific hybridization of a African-Asian A-subgenome (*G. herbaceum* var. africanum) and an American D-subgenome(*G. raimondii*) species about 1-2 mya (Wendel and Cronn, 2003). The A genome species produce natural fiber, whereas the D genome does not. Significant impact of the D genome on fiber traits in the cultivated allotetraploids has been indicated by marker assisted QTL (Quantitative Trait Loci) localization (Jiang *et al*., 1998) and substitution line performance (Saha *et al*., 2006).

Efficient strategies for capturing the sequence diversity represented within the *Gossypium* genus are greatly influenced by large differences in genome size and organization across genus. As the cotton genome is relatively large at 2700 Mbp a highly saturated genetic map of cotton with 5000 cM long genome will require 3000 DNA probes to map at an average of 1cM density (Armuganathan and Earle, 1991). The architecture of the *Gossypium* genus and its subgenomes composition with 2C DNA content is illustrated in Fig 1.1 (Wendel and Cronn 2003);

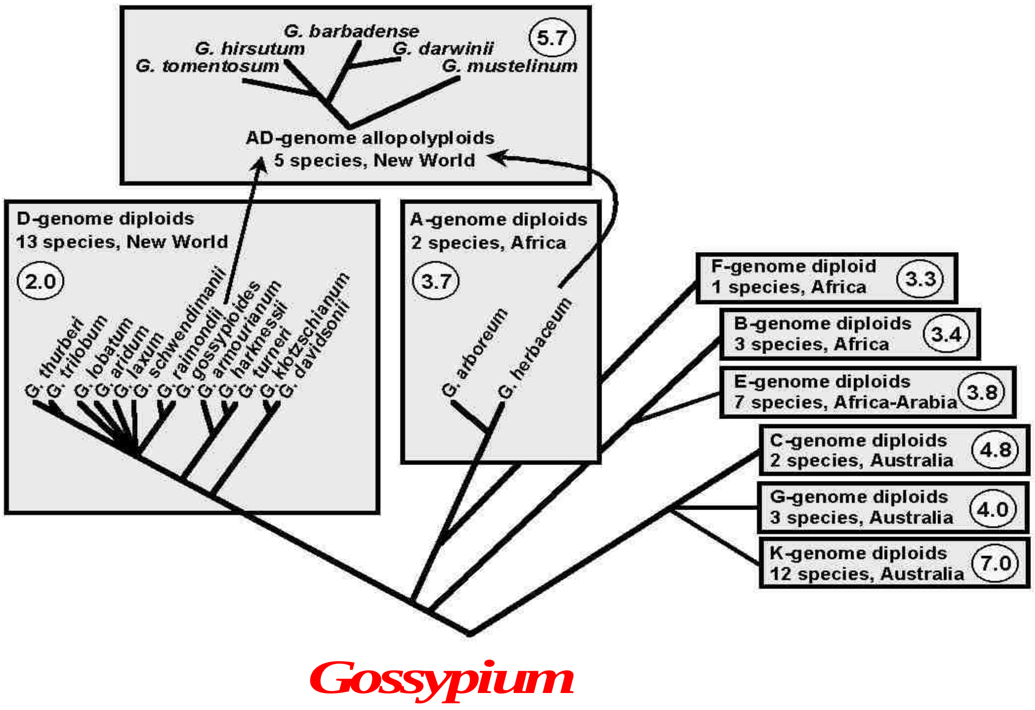


Fig 1.1 Evolutionary relationships among species of *Gossypium*. The 2C DNA content of each subgenomes is given in circle.

Traditional plant breeding procedures can be enhanced by using the linkage between markers and traits. An important step towards the establishment of such linkages is the development of genetic maps. Genetic mapping of traits comes down to finding linkages (associations) between mapped markers and phenotypic trait observations, mostly quantitative in nature. Finding such linkage can be done in several ways. Two commonly used approaches are; a) linkage analysis using a bi-parental mapping population segregating for the trait(s) of interest, or b) linkage Disequilibrium /association mapping using a well chosen (natural) population of lines, accessions, or genotypes.

1.1 Genetic Linkage Mapping and QTL Analysis of Fiber Traits in Diploid Cotton Using AFLP-SSR-TRAP Markers

Genetic linkage map construction has been recognized as an essential tool for plant molecular breeding using DNA markers because they are neutral, lack epistasis and are simply inherited in a Mendelian nature. Utilizing robust DNA markers that map to QTL's associated with fiber traits will be an important approach in fine mapping and marker assisted selection (MAS). The method of linkage analysis is well developed for bi-parental crosses between inbred lines. Estimation of recombination rates between loci allows the construction of a genetic linkage map. Associations between a trait and marker alleles identify the genomic regions in which the loci controlling the trait are located. In this way, QTL locations and effects are determined.

The A genome cottons occur naturally in Africa and Asia, while the D-genome species occurs only in the Americas. Meiotic pairing analysis has detected less bivalent formation between the tetraploid subgenomes than between the diploid A and D suggesting that the allotetraploid subgenomes are more divergent from one another than those of the descendants of their diploid progenitors (Endrizzi *et al*., 1962). Cytogenetic analysis has revealed that *G. herbaceum* (A_1) and *G. arboreum* (A_2) differ by a single translocation, while the A_t (A subgenome in tetraploid) differed from A, D and D_t (D subgenome in tetraploid) genomes by two reciprocal translocations (Endrizzi *et al.,* 1985).

Several types of molecular markers are available to dissect the complex genome of a crop such as cotton including random amplified polymorphic DNA (RAPD), Restricted fragment length polymorphism (RFLP), Amplified fragment length polymorphism (AFLP), Simple sequence repeats (SSR) and Expressed sequence tag (EST-SSR). To date several genetic maps of cotton genomes have been constructed using diverse molecular marker technologies and different mapping populations in tetraploid cottons (Reinisch *et al*., 1994; Ulloa *et al*., 2002; Rong *et al*., 2004; Mei *et al*., 2004; Nguyen *et al*., 2004 and Han *et al*., 2004). Few genetic maps have been developed in segregating populations involving diploid species. An RFLP linkage map was constructed for the diploid A genome with 275 loci using an F_2 interspecific *G. arboreum* × *G herbaceum* cross (Desai *et al.,* . 2006). The 13 chromosomes of the A genome were represented by 12 large linkage groups reflecting an expected inter-chromosomal translocation between the parents. Although the diploid mapping parents represent the closest living relatives of the allotetraploid A_t genome progenitor, two translocations and seven inversions were observed between the A and A_t genomes. The recombination rates are similar between them but the A_t genome shows a 93% increase in recombination relative to its diploid progenitors (Desai *et al*., 2006).

Genetic research on A genome cottons has declined with the decrease in their importance as a crop species during the first half of this century. Although tools to conduct molecular genetics research have been available for a long time, only limited research has been conducted on the Asiatic cotton species (Brubaker *et al*., 1999). Understanding the molecular genetics of the A genome cotton can be important for many reasons. For one, it provides a simple model system to study complex traits as it is commercially fiber producing species. Further, by knowing how fiber related QTLs are inherited in diploids, inferences on the mode of inheritance can be made to the existing maps of tetraploid cotton. Ideally, one can integrate all the QTLs associated with fiber and yield components in diploid and tetraploids with their inheritance pattern. Thus the present study aims to elucidate the inheritance, location and marker association with fiber traits in a simple diploid model system.

1.2 Association Mapping of Fiber Traits in Upland Cotton Using Molecular Markers

One of the limiting factors in genomic analysis of many plant species, including cotton, is that most genomic studies have been conducted in experimental populations developed from a bi-parental cross. Thus, while many QTLs have been reported, the effects of these QTLs often turn out to be unique to a specific genetic background, and there has been limited success in applying the results across breeding populations. Many researchers now consider that association analysis, whereby genes and QTL are detected in a random set of genotypes from a mixed genetic background, is a viable solution to this problem (Breseghello and Sorrells, 2006). The increased availability of molecular markers and the refinement of statistical tools have kindled renewed interest in this approach. Although association analysis shows great promise as an efficient and valuable tool for gene discovery, the analysis of marker-trait associations must account for the presence of population structure. Failure to do so can cause the detection of spurious associations between traits and unlinked markers.

Association mapping (AM) is based on the assumption that there is a set of markers available and either they represent actual genes (or alleles) or that of the markers are so close to the actual functional genes that they co-segregate and happen to be in linkage disequilibrium (LD). This implies that the LD mapping is done with a natural population in which association between traits and markers exists due to linkage disequilibrium. The degree of LD depends on the recombination events that have taken place in history (Nordborg *et al*., 2002). It is a result of the interaction between many factors, *e.g.* the mating system, recombination rate, selection, and population subdivision (Flint-Garcia *et al*., 2003). Not all LD occurring in a germplasm is due to linkage between loci. Linkage disequilibrium between unlinked loci can occur, attributable to population structure, admixture, outcrossing events and selection. Therefore, observed associations between markers and traits should be interpreted with care.

Two approaches are commonly applied in association mapping; (1) whole genome scans (Kraakman *et al.,* 2004) and (2) a candidate gene approach (Wilson *et al.*, 2004). Whole genome scans focus on identification of genomic regions on all chromosomes related to the trait of interest. Success and resolution of genome scans is dependent on the extent of LD. For example, increased LD decay, often represented by plotting LD versus genetic distance, requires a large number of closely linked markers, rendering the use of genome scans more laborious. Where a candidate gene for a trait has been identified, polymorphisms within the gene (SNPs) can be correlated with phenotypic variation (Thornsberry *et al.,* 2001) and are most useful when LD decays rapidly with increasing physical distance. The candidate gene approach has been effective at identifying single nucleotide polymorphisms in *Dwarf8* (Thornsberry *et al.,* 2001) and *Y1* (Palaisa *et al*., 2003) associated with phenotypic variation in flowering time and β-carotene accumulation, respectively, in maize.

The advantages of population-based association studies, utilizing a sample of individuals from germplasm collections or a natural population, over traditional QTL-mapping in biparental crosses are primarily due to; (1) availability of broader genetic variations with wider background for marker-trait correlations; (2) likelihood for a higher resolution mapping because of the utilization of recombination events from a large number of meiosis throughout the germplasm developmental history; (3) possibility of exploiting historically measured trait data for association, and (4) no need for the development of expensive and tedious biparental populations makes the approach time saving and cost-effective (Kraakman *et al*., 2004). The disadvantages of this approach are mainly Type I errors, associations could be caused by population structure and there would be a lack of linkage information among the markers identified for significant associations. All these can be attributed to population stratification caused by gene drift, founder effects or selection (Pritchard *et al*., 2000).

Several methods have been proposed for estimating population structure and modeling population structure in AM studies, including distance and model based methods (Pritchard *et al.*, . 2000; Peleg *et al.*, 2008). Distance based estimates of population structure are generally based on clustering of individuals with pair-wise genetic distance estimates between individuals (Nei 1972; Rogers 1972; Nei 1978). Although visually appealing, distance-based methods are not suitable for statistical inference. In contrast, model based methods assign individuals probabilistically to one or more sub-population. The most common model-based approach is Bayesian modeling where allele frequencies are used to estimate the likelihood of an individual belonging to a particular subpopulation. This approach allows assignment of individuals to respective populations that can be integrated into statistical models to account for population structure in AM studies. The software STRUCTURE (Pritchard *et al.*, 2000) has been developed to account for population structure and has been implemented in AM studies in a number of crop species.

Association or linkage disequilibrium (LD) mapping, based on pair-wise comparisons between observed and expected haplotype frequencies has been used extensively in human studies (Cardon and Abecasis, 2003) and in maize among polymorphic pairs of SNPs, notably insertions/deletions of individual candidate genes for maturity and plant height (Remington *et al.*, 2001; Thornsberry *et al.*, 2001). Cotton provides a good platform for using genome-wide association mapping to catalogue genes responsible for natural variation and identification of QTL's for economic traits. LD mapping involving 285 exotic *G. hirsutum* germplasm (including Uzbek, Mexican and African landrace stocks) was performed with 210 chromosome specific SSR's (Abdurakhomonov *et al*., 2008). The LD estimates were higher in exotic accessions than variety accessions. An exotic germplasm involving 260 *G. hirsutum* lines were used to associate polymorphic

SSR markers with fiber traits. A total of 314 polymorphic markers were able to divide the panel into six clusters and 59 markers were associated with fiber traits (Zeng *et al*., 2009). Fifty-six *G. arboreum* germplasm accessions introduced from nine regions of Africa, Asia and Europe were evaluated for major fiber traits using 98 SSR markers. The marker–trait associations based on single marker regression models for phenotypic traits were performed with correction for population structure. The study revealed 30 significant marker–trait associations with 19 SSR markers located on 11 chromosomes (Kantartzi and Stewart, 2008).

The numerous examples of association mapping studies performed in various germplasm resources, including the model plant *Arabidopsis,* demonstrates the enthusiasm with which LD-based association has met. The near-future completion of genome sequencing projects of crop species, powered with more cost-effective sequencing technologies, will certainly create a basis for application of whole genome-association studies accounting for rare and common copy number variants and epigenomics details of the trait of interest in plants (Abdurakhmonov *et al*., 2008).

1.3 Characterization and Marker Trait Associations of Seed Quality Traits in Upland Cotton (*Gossypium hirsutum* L.)

Cotton **(***G. hirsutum*) is primarily grown for fiber production; it is also the world's sixth largest source of vegetable oil**.** Cotton acreage has been cannibalized in recent years by corn and soybeans, a trend fueled in large part by the ethanol boom. Despite an anticipated 28 percent reduction in cotton production this year from the previous, the cottonseed crush will remain quite steady. This year’s estimated 4.71 million tons of cottonseed combines with ending stocks to set the stage for a crush of 2.7 million tons, compared to last year’s 2.76 million tons (NCPA report, 2008).

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. Commonly used in frying applications for snack foods and baked goods, cottonseed oil does not require hydrogenation (the process that produces artificial *trans* fatty acids) because of its inherent high stability. It is typically composed of about 26% palmitic acid (C16:0), 15% oleic acid (C18:1), and 58% linoleic acid (C18:2). The relatively high level of palmitic acid provides a degree of stability to the oil that makes it suitable for high-temperature frying applications, but is nutritionally undesirable due to the low-density lipoprotein cholesterol-raising properties of this saturated fatty acid (Cox *et al*., 1995). Cottonseed oil is one of only a few oils that are stable in the beta-prime crystal form, which is desirable in most solidified products because it promotes a smooth, workable consistency usually referred to as plasticity, which is important in baking applications. It also promotes relatively high levels of tocopherols (Vitamin E), a natural antioxidant; is cholesterol free; and satisfies kosher quality restrictions.

After crushing to remove the oil, cottonseed meal is used as a source of fodder protein in the livestock industry, but the sphere of its use in agriculture is limited. Constituting nearly half of a seed’s weight, the meal contains 23% high biological-value protein. Limiting its more widespread use is the presence of gossypol which binds with the proteins. The digestibility of the protein is diminished and consequently, is its assimilability in the animal. The fractionation of various protein components of the meal has shown that the amount of gossypol bound with the proteins depends on amino acid composition and structure. In view of this, the primary task in the technology of obtaining cottonseed proteins is the fraction of proteins containing different amounts of gossypol. For years, scientists have tried to breed cotton with gossypol levels safe for consumption. In the 1950s they succeeded, but because the toxin was missing from leaves as well as seeds, the plants proved defenseless against pests. With the help of a new technique called RNA interference, or RNAi, a gene-silencing mechanism succeeded in lowering the gossypol level in seeds only with minimum or no change in the rest of the plant (Ganesan *et al*., 2006).

Edible cottonseed has a high protein efficiency ratio (PER = 2.35) greater than that found in other vegetable proteins. It contains 64 g of protein per 100 g of edible cottonseed compared to 24 g of protein in beef. The protein in cottonseed is 100% assimilated by the body. It contains all nine essential amino acids, is extremely high in potassium, serves as a rich source of complex carbohydrates, and contains only polyunsaturated fatty acids. Its calcium-phosphorous ratio is considered ideal for building tissue for bone formation.

Whole cottonseed is high in protein, fat, fiber and energy. This combination of nutrients in one feedstuff is unusual. Whole cottonseed with the lint still attached is white and fuzzy in appearance. The typical cottonseed meal is composed of moisture (7%), ash (6.6%), protein (45.3%), fiber (6.3%), nitrogen-free extract (24.6%) and fat (10.2%). In order to balance the oil, protein and fiber content in the existing germplasm/cultivars, there is a need to survey the whole genome to identify genes/controlling elements responsible for these metabolic pathways.

1.4 Characterization of Upland Cotton Genotypes for Molecular Diversity and Marker Trait Associations

Plant breeders develop populations for variety development from crosses within regionally adapted germplasm. Understanding genetic diversity, population structure and marker trait associations with quantitative characters are of great importance in MAS. The narrow genetic base of upland cotton germplasm that is used in breeding programs is one of the factors recognizable for the lack of appreciable progress in improving yield and fiber traits over last two decades (Meredith 2000). Several studies have documented the decline in genetic diversity due to frequent use of only a few parents and the lack of contribution from the secondary gene pool (Bowman *et al*., 1996). The current cultivated upland cottons utilize an estimated 1% of the potential genetic variability available. Direct use of primitive accessions of cotton has been limited due to their photoperiodic sensitivity, negative linkages and poor fiber qualities. Care needs to be taken to intensively select in repeated backcrosses, keeping the desirable characteristics of the recurrent parent intact transferring a few desirable genes from wild species as possible.

The national collection of *Gossypium* species at Germplasm Research Unit TX, USA comprises of 9332 accessions representing 49 species from 74 countries assigned to three germplasm pools (Wallace *et al*., 2009). There is a need to screen the core germplasm with high density molecular map based PCR markers to fingerprint all accessions in order to minimize any sort of duplications. The development of a standard set of SSR markers that represents the diversity across the cotton genome is needed. Based on most of the previous studies in cotton on diversity, it is understood that genetic diversity exists in the primary gene pool. But there is much room for broadening the genetic base of the commercial germplasm. The Regional Breeder Testing Network (RBTN) has been developed as a mechanism for sharing particularly elite germplasm. This represents a valuable resource for research into genetic diversity and for the identification QTL's associated with fiber traits utilizing multi-location phenotypic and polymorphic molecular marker data in association mapping system.

In this context, the present study was undertaken to genetically dissect the cotton genome in order to identify associations between molecular markers and the developmental, fiber and seed quality traits. Surveying the genetic diversity in diploid (from A_1 and A_2 cross) and tetraploid cotton (representing US upland genotypes) may also provide a valuable insight into the interrelationships among the genotypes. The broad objectives of the investigation are listed as follows:

1) Genetic linkage mapping and QTL analysis of floral, seed and fiber traits in A genome diploid F_2 population using SSR, TRAP and AFLP markers.
2) Defining the cryptic population structure, genetic diversity and marker trait associations in US upland cottons.
3) Genetic diversity among upland cotton varieties and marker trait associations using the AFLP markers.
4) Molecular diversity and genetic association mapping of seed quality traits in upland cottons.

1.5 References

Abdurakhmonov IY, Kohel RJ, Yu JZ, Pepper AE, Abdullaev AA, Kushanov FN, Salakhutdinov IB, Buriev ZT, Saha S, Scheffler BE, Jenkins JN and Abdukarimov A 2008 Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm. Genomics 92(6), 478-487.

Arumuganathan K and Earle ED 1991 Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter, 9, 208–218.

Bowman DT, May OL and Calhoun DS 1996 Genetic base of upland cotton cultivars released between 1970 and 1990. Crop Sci 36:577–581.

Breseghello F and Sorrells ME 2006 Association mapping of kernel size and milling quality in Wheat (*Triticum aestivum* L.) cultivars, Genetics, 172, 1165-1177.

Brubaker CL, Paterson AH and Wendel JF 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42, 184-203.

Cardon LR and Abecasis GR 2003 Using haplotype blocks to map human complex trait loci. Trends in Genetics 19, 135–140.

Cox C, Mann J, Sutherland W, Chisholm A and Skeaff M 1995 Effects of coconut oil, and safflower oil on lipids and lipoproteins in persons with moderately elevated cholesterol levels. Journal of Lipid Research, 36, 1787–1795.

Desai A, Chee PW, Rong J, May OL and Paterson AH 2006 Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*. Genome. 49, 336-345.

Endrizzi JE 1962 The diploid-like cytological behavior of tetraploid cotton. Evolution 16, 325-329.

Endrizzi JE, Turcotte EL and Kohel RJ 1985 Genetics, cytology, and evolution of *Gossypium*. Advances in Genetics. 23, 271-375.

Flint-Garcia SA, Thornsberry JM and Buckler ES 2003 Structure of Linkage Disequilibrium in Plants. Annual Review of Plant Biology 54, 357-374.

Ganesan S, Campbell, LM, Puckhaber L, Stipanovic RD and Rathore KS 2006 Engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. Proceedings of National Academy of Sciences 103(48), 18054-18059.

Han ZG, Guo WZ, Song XL and Zhang TZ 2004 Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Molecular Genetics and Genomics 272, 308–327.

Jiang C-X, Wright RJ, Elzik KM and Paterson AH 1998 Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). Proceedings of National Academy of Sciences, 95: 4419-4424.

Kantartzi SK and Stewart J McD 2008 Association analysis of fiber traits in *Gossypium arboreum* accessions. Plant Breeding, 127, 173-179.

Kraakman AT, Niks W, Van den Berg RE, Stam PM and Van Eeuwijk FA 2004 Linkage Disequilibrium Mapping of yield and yield stability in modern spring barley cultivars. Genetics 168, 435-446.

Meredith WR Jr 2000 Cotton yield progress – why has it reached a plateau, Better Crops, 84: 6-9.

Mei M, Syed NH, Gao W, Thaxton PM, Smith CW, Stelly DM and Chen Z 2004 Genetic mapping and QTL analysis of fiber-related traits in cotton. Theoretical and Applied Genetics, 108, 280-291.

NAAS USDA, 2009 National Agricultural Statistical Survey. (www.nass.usda.gov)

NCPA Report, 2009 US cottonseed production, National Cottonseed Products Association, Inc., Cordova, TN

Nei M 1972 Genetic distance between populations. The American Naturalist 106, 283-292.

Nei, M 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89, 583-590.

Nguyen TB, Giband M, Brottier P, Risterucci AM, Lacape JM 2004 Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. Theoretical and Applied Genetics, 109:167–175.

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA and Weigel D 2002 The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nature Genetics, 30, 190-193.

Palaisa KA, Morgante M, Williams M and Rafalski A 2003 Contrasting Effects of Selection on sequence diversity and Linkage Disequilibrium at two phytoene synthase Loci. The Plant Cell 15, 1795-1806.

Peleg Z, Fahima T, Abbo S, Krugman T and Saranga Y 2008 Genetic structure of wild emmer wheat populations as reflected by transcribed versus anonymous SSR markers. Genome 51, 187-195.

Pritchard JK, Stephens M and Donnelly P 2000 Inference of population structure using multilocus genotype data. Genetics, 155, 945-959.

Reinisch AJ, Dong J, Brubaker CL, Stelly DM, Wendel JF, Paterson AH 1994 A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploidy genome. Genetics 138, 829–847.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM and Buckler ES 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. Theoretical and Applied Genetics ,98, 11479-11484.

Rogers JS 1972 Measures of genetic similarity and genetic distance. Studies in genetics. VII. Univ. Texas Publ. 7213, 145-153.

Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, Park C, Pierce GJ, Rainey KM, Rastogi VK, Schulze SR, Trolinder NL, Wendel JF, Wilkins TA, Dawn Williams-Coplin T, Wing RA, Wright RJ, Zhao X, Zhu L, Paterson AH 2004 A 3347-locus genetic recombination map of sequence tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics 166, 389–417.

Saha S, Jenkins JN, Wu J, Mc Carty JC, Gutierrez OA, Percy RG, Cantrell RG and Stelly DM 2006 Effects of chromosome specific introgression in upland cotton on fiber and agronomic traits. Genetics. 172, 1-12.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D and Buckler ES, 2001 *Dwarf8* polymorphisms associate with variation in flowering time. Nature Genetics 28, 286-289.

Ulloa M, Meredith WR, Shappley ZW, Kahler AL 2002 RFLP genetic linkage maps from four F2.3 populations and a joinmap of *Gossypium hirsutum* L. Theoretical and Applied Genetics, 104, 200-208.

Wallace T, Bowman D, Campbell BT, Chee P, Gutierrez OA, Kohel RJ, McCarty J, Myers GO, Percy R, Robinson F. Smith W, Stelly DM, Stewart JM, Thaxton P, Ulloa M and Weaver DB 2009 Status of the USA cotton germplasm collection and crop vulnerability. Genetic Resources for Crop Evolution, 56(4), 507-532.

Wendel JF and Cronn RC 2003 Polyploidy and the evolutionary history of cotton, Advances in Agronomy, 78, 139-186.

Wilson LM, Whitt SR, Ibáñez AM, Rocheford TR, Goodman MM and Buckler ES 2004 Dissection of maize kernel composition and starch production by candidate gene analysis. The Plant Cell 16, 2719-2733.

Zeng L, Meredith WR Jr, Gutierrez OA, Boykin DL 2009 Identification of associations between SSR markers and fiber traits in an exotic germplasm derived from multiple crosses among *Gossypium* tetraploid species. Theoretical and Applied Genetics 119(1), 93-103.

CHAPTER 2 GENETIC LINKAGE MAPPING AND QTL ANALYSIS IN DIPLOID COTTON USING AFLP-SSR-TRAP MARKERS

2.1 Introduction

The genus *Gossypium* consists of four cultivated cotton species. Among the diploid species (2n=2X=26), *G. arboreum* and *G. herbaceum* are generally cultivated on marginal and drought prone environments in Asia. They can be distinguished based on plant habit as well as leaf, bracteole and boll features (Fryxell, 1979). Long and narrow lobed leaves, bracteoles with fewer teeth and round tapering bolls are the characteristics of *G. arboreum*, while constricted leaf lobes, wide bracteole and round, less pitted bolls are the common features of *G. herbaceum*. Within the A genome, *G. herbaceum* and *G. arboreum* diverged relatively recently. Cytologically these species can be distinguished by a reciprocal translocation (Gerstel, 1953), while the A_t (A subgenome in tetraploid) differs from A, D and D_t (D subgenome in tetraploid) genomes by two reciprocal translocations. This suggests that *G. arboreum* arose as an incipient species with the origin through the fixation of the translocation (Endrizzi *et al*., 1985).

Potentially valuable genetic variability has been observed for developmental traits, yield and fiber characters in *G. arboreum* (Singh and Singh 1984) and *G. herbaceum* (Singh 1983). Old world Asiatic diploid cottons were economically important during early global expansion of commercial cotton production. In the 1950's, with the introduction of New world cotton, that had superior fiber quality and yield potential with desirable plant type, the area under diploid cotton cultivation drastically reduced. Diploid cotton, however, is a model system for studying the genetics of fiber development compared to the more complicated system in tetraploid New world cottons. Therefore an understanding of the genetic inheritance and genomic regions controlling the fiber genes of diploid cotton species is critical. In order to use the extant genetic diversity in the development of superior genotypes or transferring elite genes for biotic or abiotic stresses into cultivated tetraploids, molecular breeding techniques using molecular markers offers promising avenue compared to traditional breeding methods.

Molecular linkage maps provide essential tools for plant genetic research, facilitating quantitative trait locus (QTL) mapping, marker-assisted selection and map based cloning. The method of linkage analysis is well developed for bi-parental crosses between inbred lines. Estimation of recombination rates between loci allows for the construction of genetic linkage map. Besides, associations between a trait and marker alleles identify the genomic regions in which the loci controlling the trait are located. Several types of molecular markers are being employed to dissect the genome viz., RAPD (Random Amplification of Polymorphic DNA), RFLP (Restriction Fragment Length Polymorphism), AFLP (Amplified Fragment Length Polymorphism), SSR (Simple Sequence Repeat) and EST-SSR (Expressed Sequence Tag).

To date several genetic maps of cotton genomes have been constructed using diverse molecular markers and different mapping populations in tetraploid cottons (Reinisch *et al*., 1994; Ulloa *et al*., 2002; Rong *et al*., 2004; Mei *et al*., 2004; Nguyen *et al*., 2004; Han *et al*., 2004 and Zhang *et al*., 2009). Comparatively few genetic maps have been developed in segregating populations involving diploid species. Interspecific linkage maps of diploid cottons have been constructed for the A genome (*G. herbaceum* × *G. arboreum*), the D genome (*G. trilobum* × *G. raimondii*) (Brubaker *et al.,* 1999; Rong *et al.,* 2004; Desai *et al.,* 2006) and the G genome (*G. nelsonii* × *G. australe*) (Brubaker and Brown 2003) taxa. An RFLP linkage map using interspecific A genome diploid F_2 population mapped 275 loci (Desai *et al.,* 2006). The 13 chromosomes of the A genome were represented by 12 large linkage groups reflecting an expected inter-chromosomal translocation between the parents. Although the diploid mapping parents represent the closest living relatives of the allotetraploid A_t genome progenitor, two translocations and seven inversions were observed between the A and A_t genomes. The recombination rates are similar between them but the A_t genome shows a 93% increase in recombination relative to its diploid progenitors.

Among the different molecular marker technologies, the AFLP technique has also been frequently used in establishing the extent of genetic diversity and relatedness in cotton due to its high polymorphic nature. Evolutionary and genetic relationships of various germplasm resources including cultivars from subgenomes such as *G. raimondi, G. incanum, G. herbaceum* and *G. arboreum* were estimated using AFLPs (Iqbal *et al*., 1997). Genetic similarities revealed by AFLP analyses were in agreement with taxonomic relationships at the species level and this was also suggested by other groups using different marker systems (Abdalla *et al.,* 2001; Murtaza 2006).

The AFLP marker system has also been used extensively to develop genetic linkage maps and as a basis for map based QTL analysis. A map based on F_2 population developed from a cross between *G. hirsutum* acc. TM-1 x *G. barbadense* acc. 3–79 was constructed using RAPD and AFLP markers comprising 11 linkage groups that covered 521.7 cM (Altaf *et al*., 1997). In another study, 490 AFLP markers associated with agronomic traits were identified using an F_2 population developed from an interspecific cross (Reddy *et al.,* 1997). A backcross interspecific population was surveyed using 465 AFLP loci along with 229 SSRs, 192 RFLPs, and two morphological markers resulted in a map composed of 37 linkage groups and covered 4400 cM distance (Lacape *et al*., 2003). More than 50 AFLP markers have been surveyed on 92 recombinant inbred lines (RILs) of *G. hirsutum* grown in China and the USA, and identified AFLPs associated with fiber and agronomic traits. One to four markers were associated with 22–93% of the phenotypic variability of each of the seven traits which suggest that the selected markers could be used in MAS (Wu *et al.,* 2009).

Markers assigned to chromosomes are more useful than unlinked markers in MAS and map based cloning (Baogong, 2004). Out of 42 linkage groups developed using an interspecific F_2 population, 19 were assigned to 12 chromosomes using aneuploid interspecific hybrids and a set of

29 RFLP and SSR framework markers (Mei *et al*., 2004). Seven QTLs were also detected for six fiber-related traits; five of these were distributed among A-subgenome chromosomes (Mei *et al.*, 2004). To identify abundant polymorphisms for mapping, a trispecific F_2 mapping population was screened with AFLP and RAPD markers (Khan *et al*., 1998). A linkage map containing 51 linkage groups spanning about 6,663 cM was developed and suggested a higher level of recombination and polymorphism in the D genome than the A genome (Khan *et al.*, 1998). The possibility of identifying AFLPs as diagnostic markers for *G. hirsutum* and its closest relative *G. tomentosum* (endemic to the Hawaii) was explored in a study where 11 and 16 species-specific markers were identified for *G. tomentosum* and *G. hirsutum*, respectively (Hawkins *et al.*, 2005). These species-specific AFLP markers would be useful for detecting gene flow between *G. hirsutum* and *G. tomentosum* that had occurred in the past and might occur in the future. Thus AFLP system has proven to be valuable in evolutionary, molecular diversity and QTL or marker trait association studies in cotton.

Genetic research on A genome cottons declined with the decrease in their importance as crop species during the later half of the 20th century. Although tools to conduct molecular genetics research have been available for a long time, only limited research has been conducted on the Asiatic cotton species (Brubaker *et al*., 1999). Understanding the molecular genetics of A genome cotton can be important for many reasons. They can foremost serve as a simple model system to study complex quantitative traits, yet only a limited number of genetic maps and QTL studies have been conducted. There is a significant opportunity for further mining of the diploid genome with new marker systems to facilitate genetic mapping and MAS of fiber genes. In the present study, we used AFLP, SSR and TRAP markers to generate a framework genetic map of cultivated diploid cottons. We also describe herein the preliminary assessment of fiber QTL’s and detection of putative QTL’s using an interspecific F_2 population. Thus the broad objectives of the present study are;

1. Construction of an A genome diploid linkage map using AFLP, SSR and TRAP markers.
2. QTL analysis for qualitative and quantitative traits using both traditional linkage map based methods and robust General Linear Methods (GLM).

2.2 Materials and Methods

2.2.1 Plant Material and Phenotypic Analysis

An interspecific F_2 population was developed from a cross between *G. arboreum* (acc. SMA-4, PI529740) *x G. herbaceum* (acc. A-97, PI529670), (provided by Dr. A.H. Paterson, University of Georgia, Athens). The parents and 94 F_2 segregating plants were grown in the green house, at LSU AgCenter, Baton Rouge, LA. The phenotypic data on qualitative traits such as petal color (yellow or white), petal spot (absent or present) and seed hair (fuzzy or naked) was recorded for all 94 F_2 individuals and parents in the green house. The parent SMA-4 possesses yellow flowers with petal spot and naked seeds while A-97 has white flowers without petal spot and fuzzy seeds. The quantitative traits namely, fiber length (inches), fiber strength (g/tex), short fiber index (SFI), fiber elongation (%), seed index (g), and uniformity ratio were measured on an individual plant basis. The fiber analysis was done via HVI system at the LSU AgCenter Cotton Fiber Testing Laboratory. HVI measurements were repeated two times.

Micronaire is measured by relating airflow resistance to the specific surface of fibers. Fiber length is measured optically in a tapered fiber beard which is automatically prepared, carded, and brushed. Fiber strength is measured physically by clamping a fiber bundle between 2 pairs of clamps at known distance. The second pair of clamps pulls away from the first pair at a constant speed until the fiber bundle breaks. The distance it travels, extending the fiber bundle before breakage, is reported as elongation. Uniformity index is the ratio of mean length and upper half mean length expressed in percentage. Short fiber index is evaluated utilizing a prediction model to derive short fiber index from the HVI measurements of length and uniformity index.

Means of the phenotypic data from segregating individuals were used to test for normal distribution using PROC UNIVARIATE (SAS, 9.1.3, Cary, NC). Correlation analysis between pairs of traits was performed using PROC CORR in SAS. The correlation coefficients and a matrix plot were generated showing interrelationships among fiber traits.

2.2.2 DNA Isolation and Genotypic Analysis

The total genomic DNA from young leaves of the parents and F_2 plants was isolated using the cetyltrimethyl-ammonium bromide (CTAB) method as described previously (Zhang and Stewart 2000). Cotton leaves were frozen in liquid nitrogen after being collected and ground to a fine powder with a mortar and pestle. In a 50ml eppendorf tube, CTAB DNA extraction buffer (15ml) was added to each 1-1.5g finely ground sample. The supernatant was extracted twice with chloroform/isoamyl alcohol (24:1) after being incubated at 65^{o} C for 30 min. Then the supernatant was treated with ice cold isopropanol and RNase (Qiagen, Valencia, CA) in succession. The precipitated DNA was washed with 70% ethanol and dissolved in ddH_2O (200µl). DNA concentration was measured using a NanoDrop-1000 spectrophotometer (NanoDrop, Wilmington, DE) at an optical density ration of 260/280 nm. Samples yielding ratios between 1.8 and 2.0 were considered good quality DNA samples.

Sixty four primer combinations were used to generate AFLP data following the procedure given by Vos *et al*., (1995) with some modifications (Table: 2.1a). Individual plant DNA (20-50ng/µl) was digested with *EcoRI* (infrequent cutter with GAATTC recognition sequence) and *MseI* (frequent cutter with TTAA recognition sequence) restriction enzymes and oligonucleotide adapters specific to restriction sites were ligated to the resulting fragments through incubation (37^{o}C for 180 min) with DNA ligase. Pre-amplifications were done in an iCycler (BioRad Labs, Hercules, CA) using *EcoR*I+A and *Mse* I+C oligo primers. The amplification was carried out with 50ng/ul of oligo primers, 5mM dNTP's, 25mM $MgCl_2$, 10X buffer, Taq (5U/ul) and restrict- ligated template DNA in a total volume of 20 µl. The PCR was set up with initial denaturing for 94°C 2 min followed by 26 cycles at 94°C 1 min, 56°C 1 min., 72°C 1 min., and final extension at 72°C for 5min.

Table 2.1a Adapters and primers of AFLP marker system used for pre and selective amplification in diploid F_2 population

Primer/adapter	Nomenclature†	Sequence (5'-3')
ECORI primers:		
EcoRI linker 1	E-I	CTC GTA GAC TGC GTA CC
EcoRI linker 2	E-II	AAT TGG TAC GCA GTC TAC
EcoRI + A	E+A	GAC TGC GTA CCA ATT CA
E- AAC	E1	GACTGCGTACCAATTCAAC
E- AAG	E2	GACTGCGTACCAATTCAAG
E-ACA	E3	GACTGCGTACCAATTCACA
E-ACT	E4	GACTGCGTACCAATTCACT
E-ACC	E5	GACTGCGTACCAATTCACC
E-ACG	E6	GACTGCGTACCAATTCACG
E-AGG	E8	GACTGCGTACCAATTCAGG
E-AGA	E9	GACTGCGTACCAATTCAGA
MseI primers:		
MseI linker 1	M-I	GAC GAT GAG TCC TGA G
MseI linker 2	M-II	TAC TCA GGA CTC AT
MseI + C	M+C	GAT GAG TCC TGA GTA AC
M-CAA	M1	GATGAGTCCTGAGTAACAA
M-CAC	M2	GATGAGTCCTGAGTAACAC
M-CAG	M3	GATGAGTCCTGAGTAACAG
M-CAT	M4	GATGAGTCCTGAGTAACAT
M-CTA	M5	GATGAGTCCTGAGTAACTA
M-CTC	M6	GATGAGTCCTGAGTAACTC
M-CTG	M7	GATGAGTCCTGAGTAACTG
M-CTT	M8	GATGAGTCCTGAGTAACTT

†: Nomenclature is in accordance with the Lacape *et al*., 2003; Myers *et al*., 2009.

The pre amplified products were diluted with ddH_2O and selective amplification was done using two selective nucleotides. The EcoRI+ANN oligo primers were dye labeled with 700 and 800

IR dye. The PCR for selective amplification was carried out in a reaction volume of 10 μL consisting of 10X reaction buffer, 25 mM $MgCl_2$, 2.5 mM dNTPs, 1 μM each of EcoRI-ANN and MseI+CNN primers and 5U *Taq* polymerase (Promega, Madison, WI). The reactions were run on an *i*-Cycler (BioRad Labs, Hercules, CA). The PCR conditions for selective amplifications were as follows: 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 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. A total of 64 *EcoR* I - *Mse* I selective amplification primer combinations were used. The PCR amplified products were run on a LI-COR 4300 sequencer (LI-COR Inc., Lincoln, NE). The gels were saved onto a computer and scored manually. Presence of band was recorded as '1' and absence as '0', for a typical dominant marker system. Ambiguous data that could not be resolved were discarded. The nomenclature of AFLP loci was followed according to Lacape *et al.,* 2003 and Myers *et al*., 2009, which indicates the enzyme primer combinations with band size.

In addition, we used 44 SSR/EST-SSR markers (BNL, CIR and MUSS) which were selected from At subgenome of the previous tetraploid maps, potentially associated with fiber genes. The forward primer of these microsatellite markers were IR dye labeled (700 and 800) (MWG-Biotech, Germany). PCR amplification was performed in a total volume of 10μl containing 20-50ng of genomic DNA, each primer at 1μM, 5X buffer, 25mM $MgCl_2$, 5mM dNTP's and 1U of Taq polymerase (Promega, WI) with the following cycling profile: 1 cycle of 4min at 94°C, 35 cycles of 45s at 94°C, 45s at 55°C (varying for different SSR's depending on their T_m values), followed by 7 min at 72°C. The PCR was carried out using iCycler and the PCR products were separated using LICOR 4300 sequencer. The gels were saved onto a computer and scored manually as A (homozygous dominant), H (heterozygous) and B (homozygous recessive). Four combinations of TRAP markers, a two primer PCR technique (Hu and Vick, 2003) were also tried utilizing sequence information on sucrose synthase (SuSy) and sucrose phosphate synthase (SuPS) genes (Table: 2.1b). The forward IR dye labeled TRAP primers were combined with arbitrary reverse primers (kindly provided by the Sugarcane lab, SPESS, LSU). The above described PCR protocol was used to amplify genomic regions of 94 F_2 individuals. The PCR products were separated using (LICOR 4300) and the bands were scored similar to a dominant marker system.

Table 2.1b Forward and reverse primer sequences of TRAP markers used in diploid F_2 population.

Fixed/Forward primer	**Fixed primer sequence (5'-3')**	**GenBank number**
Sucrose Synthase (SuSy)	GGAGGAGCTGAGTGTTTC	AF263384
Sucrose Phosphate Synthase (SuPS)	CGACAACTGGATCAACAG	AB001338
Revere primer		
R1	GACTGCGTACGAATTAAT	IR Dye700
R2	GACTGCGTACGAATTTGA	IR Dye700

Allelic diversity at a given locus can be determined by Polymorphism Information Content (PIC) and it was calculated as '$PIC=1-\sum f_i^2$' where, f_i is the frequency of the i^{th} allele (Weir, 1996). PROC ALLELE was used to calculate the PIC values and frequency estimate was done using PROC Freq (SAS, 9.1.3, Cary, NC).

2.2.3 Linkage Map Construction

The segregation ratio for each marker was tested against expected Mendelian ratios using the Chi-square goodness of fit test. Only markers which are not significantly ($P \leq 0.05$) different from the expected 3:1 for dominant markers such as AFLP and TRAP and 1:2:1 for co-dominant markers such as SSR were utilized for map construction.

Linkage map construction was performed using JOINMAP 3.0 (Stam and Oojien 1995). The Kosambi map function was used (Kosambi, 1994) to convert recombination frequency to genetic map distance (centiMorgan, cM). All the linkage groups were determined at LOD (logarithm of odds) scores ≥3.0 and recombination frequency of 0.4 to provide evidence of linkage (Wu *et al*., 1992). The graphical representation of the LG was obtained using JOINMAP. Markers showing evidence of segregation distortion were marked specifically and used for mapping separately.

2.2.4 Data Analysis and QTL Mapping

a) Discriminant Analysis for Qualitative Traits

Qualitative traits such as petal color, petal spot and seed fuzziness were analyzed using Discriminant Analysis (DA). DA is used to classify cases into the values of a categorical dependent variable, usually a dichotomous (Fisher, 1936). Discriminant analysis has two steps: (1) an F test (Wilk's lambda) is used to test if the discriminant model as a whole is significant, and (2) if the F test shows significance, then the individual independent variables are assessed to see which differ significantly in mean (by group) and these are used to classify the dependent variable. The smaller the Wilk's lambda value for an independent variable, the more that variable contributes to the discriminant function.

The qualitative traits were divided into two groups based on yellow or white petal color, presence or absence of petal spot or seed fuzziness (present or absent). To identify the marker data that best differentiates training samples within each subpopulation, the parametric discriminant analysis (PROC STEPDISC of SAS 9.1.3) forward method was used in the first step. The non parametric method within the PROC DISCRIM procedure was then performed considering only the selected markers to construct and validate the class prediction function and to predict group membership. An error rate defined by 'percent correct classification' was calculated to measure ability of the markers to correctly assign individual lines to 5, 10 and 15% of the training samples. The CROSSVALIDATION option provides a better assessment of classification accuracy. This classification is also done for each observation; however, the discriminant function used in each case is constructed by taking that observation out of the data set. With high value of percent correct classification, an association between marker (s) and phenotype is inferred.

b) **QTL Analysis for Quantitative Traits**

A diploid segregation panel consisting of 94 individual cotton plants was evaluated for fiber length, strength, seed index, uniformity ratio, elongation percent and short fiber index. Using the linkage map and phenotypic information, QTL analysis was performed through interval and composite interval mapping via Windows QTL Cartographer 2.5 (Basten *et al*., 2001). Composite interval mapping (CIM) was carried out using the Zmapqtl component of Cartographer (Zeng and Weir, 1996). The analysis was performed with a maximum of five background markers based on the forward-backward regression method of selection. Zmapqtl provides estimates for the square of the partial correlation coefficient (R^2), the additive and the dominant effect. A LOD threshold of ≥2.5 (1000 permutations) was used to declare significant QTLs in the present investigation. A Chi-square test was performed to determine whether the allele frequency at each individual locus had normal segregation. The multiple interval mapping method (MIM) was employed whenever IM/CIM detected more than one QTL on the same linkage group to verify their significance.

QTL analysis was also performed using multiple regression employing PROC GLMSELECT in SAS. A variety of model selection methods are available, offering extensive capabilities for customizing the selection and stopping criteria. GLMSELECT compares most closely to PROC REG and PROC GLM. We used 52 types of GLMSLECT models. Stepwise selection method was used with all possible combinations of CHOOSE, SELECT and STOP. Different options used for these selection methods included, Bayesian Information Content (BIC), SBC (Schwarz Bayesian Information Criterion), Adjusted R^2, AICC (the Corrected Akaike Information Criterion), SL=0.15 (the significance level of the F statistic for entering or departing effects) and Cross validation (CV). Traits were considered as dependent variables and all markers were treated as independent variables. Each trait was analyzed separately and those independent variables with a calculated test statistic estimate less than the specified P value (0.05) were added to the model. To reduce Type I error, selected models were further tested with a validation step by using the ‘PRESS’ criterion in the ‘STOP’ option. The best model was then selected based on adjusted R^2 and least number of effects for a particular trait.

The QTL’s identified by the Cartographer and multiple regression methods were compared, with respect to significance of the marker, potential for explaining most of the phenotypic variability and their localization on the LG. The analysis were separately carried out and compared in order to identify common markers.

2.3 Results

2.3.1 Phenotypic Trait Analysis

The phenotypic data for fiber traits of the parents and the F_2 individuals are summarized in Table 2.2. The two parents differed significantly for most of the fiber traits except fiber strength.

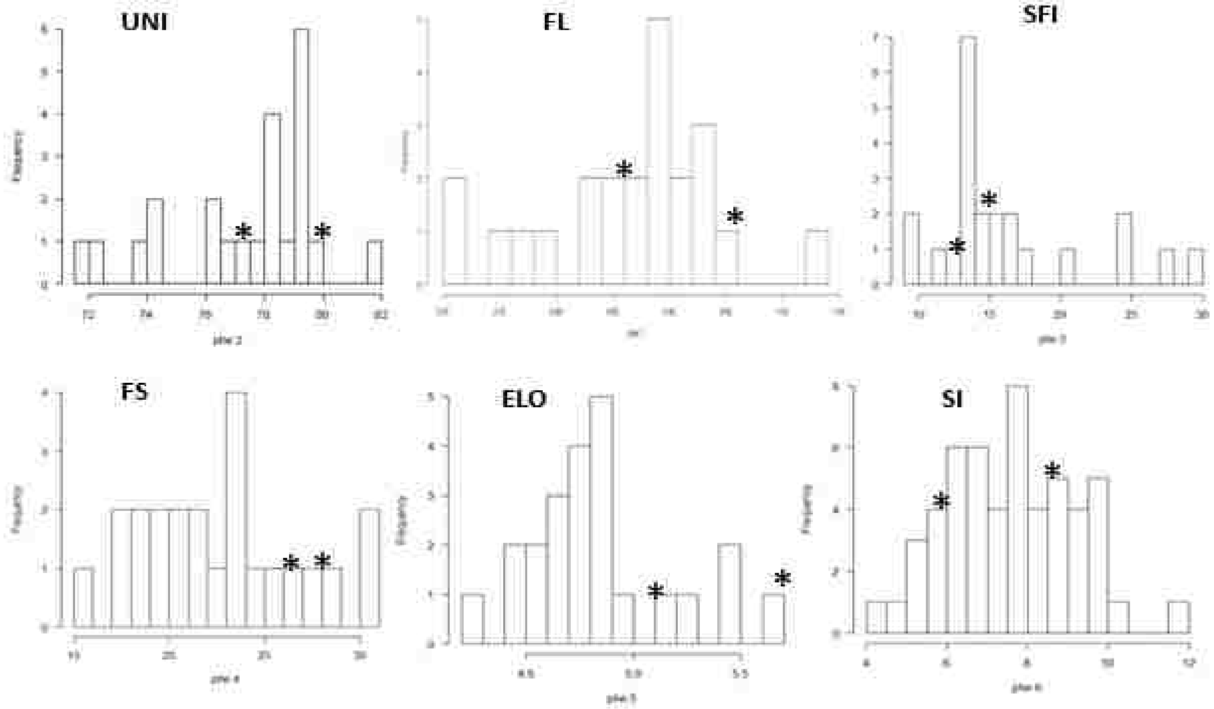
Table 2.2 Univariate analyses of FL, UNI, SFI, FS, ELO and SI characters in parental and diploid F_2 population

Parameters	FL (inch)	UNI	SFI	FS (g/tex)	ELO	SI (g)
Min	0.70	71.70	9.90	15.90	4.20	5.01
Max	1.02	81.70	29.80	30.90	5.70	11.90
Mean	0.86	77.42	16.40	22.71	4.89	7.79
SE	0.02	0.53	1.12	0.87	0.07	0.35
Var	0.01	6.45	29.07	17.58	0.13	2.82
SD	0.08	2.54	5.39	4.19	0.35	1.68
Skewness	-0.36	-0.74	1.09	0.35	0.58	0.50
Kurtosis	-0.48	-0.44	0.08	-0.99	-0.13	-0.27
Parents						
PI529740	1.00	80.6	12.1	26.50	5.90	5.90
PI529670	0.86	77.1	15.5	26.70	8.45	8.45

# FL= Fiber length, UNI= Uniformity index, SFI= Short fiber index, FS= Fiber strength, ELO= Elongation percentage, SI= Seed index, SD=Standard deviation, Var= Variance and SE= Standard error

The F_2 population showed transgressive segregants for all traits. Based on the wide range of values and high variance estimates, it is evident that moderate to high phenotypic diversity was present in the population. The wide ranges of values were evident for fiber traits, such as FL (0.7-1.02inch), UNI (71.7-81.7), SFI (9.9-29.8), FS (15.9-30.9g/tex) and SI (5-11.9g).

The frequency distribution for fiber traits among the F_2 individuals is graphically shown in Fig 2.1. Based on the amount of diversity present, it was concluded that the F_2 population possessed sufficient variation for QTL analysis. The correlation coefficients among the quantitative traits revealed that there was a significant positive relationship among the fiber traits (Table 2.3). The traits such as FL, UNI, FS and SFI were highly correlated with values ranging from 0.55-0.95. While, ELO had negative correlation with FL, UNI and FS. Seed index (SI) was positively correlated (not significant) with with FL, UNI and FS but negatively associated with SFI and ELO.



FL= Fiber length, UNI= Uniformity index, SFI= Short fiber index, FS= Fiber strength, ELO= Elongation percentage, SI= Seed index

Fig 2.1 Frequency distribution in a *G. arboreum x G. herbaceum* F_2 population for fiber quality and seed index. The mark (*) indicates parental values for each trait.

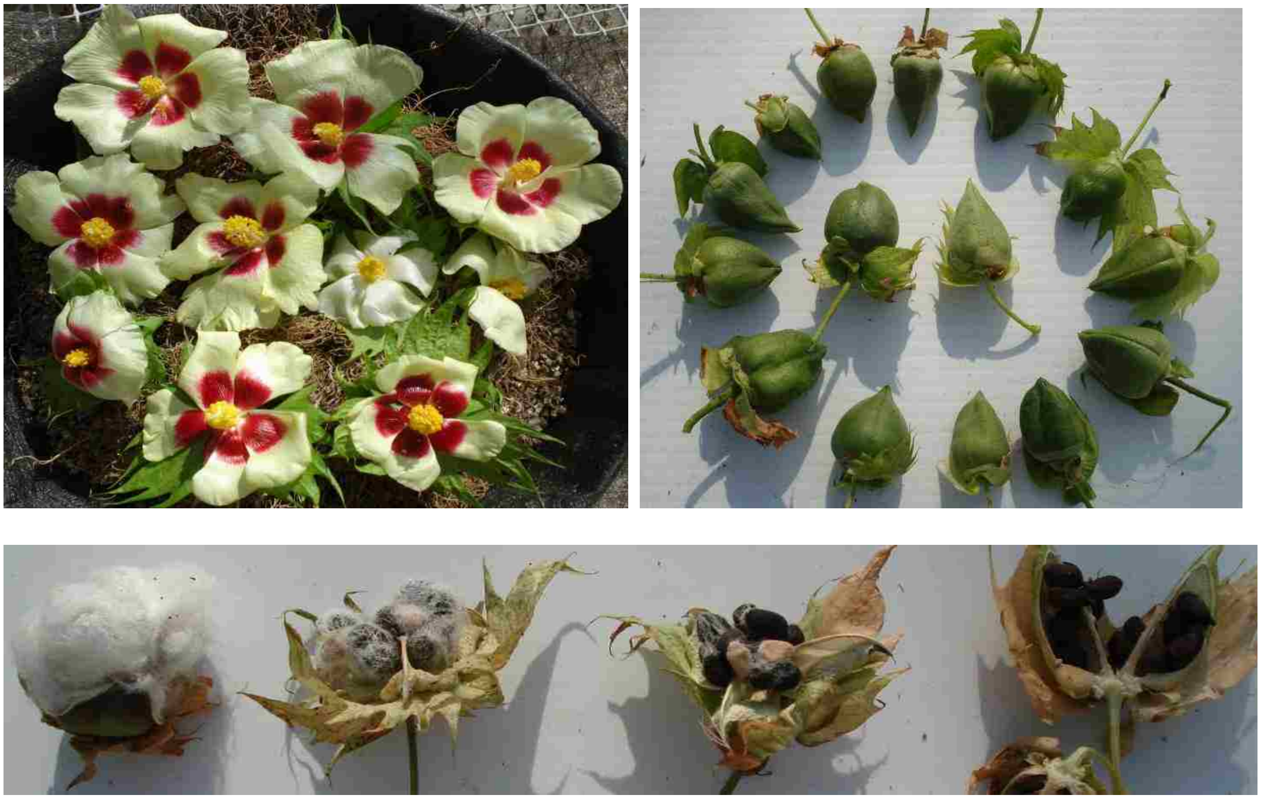


Fig 2.2 The phenotypic diversity present in the segregating F_2 population of A genome cottons, with respect to flower color and petal spot (center left: SMA-4 and center right: A-97 and F_2 are around), boll size and shape and seed fuzziness(extreme left: A-97 and extreme right: SMA-4 and center: F_2 segregants (clockwise from upper left).

The binaries of phenotypic diversity for flower color, boll size and shape and seed fuzziness is demonstrated in Fig 2.2. The parent SMA-4 possesses yellow flowers with petal spot and naked seeds while A-97 has white flowers without petal spot and fuzzy seeds. The segregation for the qualitative traits was recorded for each individual as categorical data.

Table 2.3 Pearson correlation coefficients among fiber traits of diploid F_2 population

	FL	UNI	SFI	FS	ELO	SI
FL	1					
UNI	0.94**	1				
SFI	0.86**	0.95**	1			
FS	0.55*	0.69**	-0.69**	1		
ELO	-0.52*	-0.52*	0.48*	-0.29	1	
SI	0.262	0.30	-0.51*	0.51*	-0.45*	1

*, ** Significant at $P \leq 0.05$, 0.01 respectively. FL= Fiber length, UNI= Uniformity index, SFI= Short fiber index, FS= Fiber strength, ELO= Elongation percentage, SI= Seed index

2.3.2 Molecular Analysis

Sixty four AFLP primer combinations were screened by selective amplification using diploid genomic DNA. A total of 539 polymorphic bands were generated with eight each ECORI and MseI primer combinations. In addition, SSR and TRAP markers generated 50 and 17 polymorphic loci respectively (Table 2.4). Among the different ECORI primers tried, E2-AAG and E6-ACG generated the highest number of polymorphic bands across all the MseI primers. The frequency of shared alleles among the F_2 population is presented in Fig 2.3. As expected, the F_2 segregants showed normal distribution with most of the individuals showing 60-80% similarity with a peak at 70%. Although the parents differed with respect to many distinguishable characters, the amount of genetic marker variability was moderate among the F_2 segregants.

Table 2.4 Summary of AFLP primers used and the number of polymorphic loci identified by each combination

Primer	M1†	M2	M3	M4	M5	M6	M7	M8	Total
E1	-	-	-	-	-	-	-	-	-
E2	-	29	19	14	13	9	5	9	98
E3	-	11	7	6	5	7	6	8	50
E4	-	15	-	9	10	16	11	11	72
E5	-	21	10	10	4	14	15	13	87
E6	-	15	33	13	5	12	9	11	98
E8	-	22	8	15	10	6	4	7	72
E9	-	22	18	3	9	4	4	3	63
SSR									50
TRAP									17

† = nomenclature of the AFLP markers is in accordance with Lacape *et al*., 2003 and Myers *et al*., 2009

The polymorphic information content (PIC) is commonly used in genetics as a measure of polymorphism and to estimate the informativeness of a marker locus used in linkage analysis. In the present study, PIC values varied from 0.087 to 0.37 with an average of 0.253 (Fig 2.4). The AFLP, TRAP and SSR markers produced moderate variability, mirroring to the narrow genetic base of the characters in the selected parents. Representative AFLP and SSR gel images showing typical marker segregation among the F_2 population is presented in Fig: 2.5.

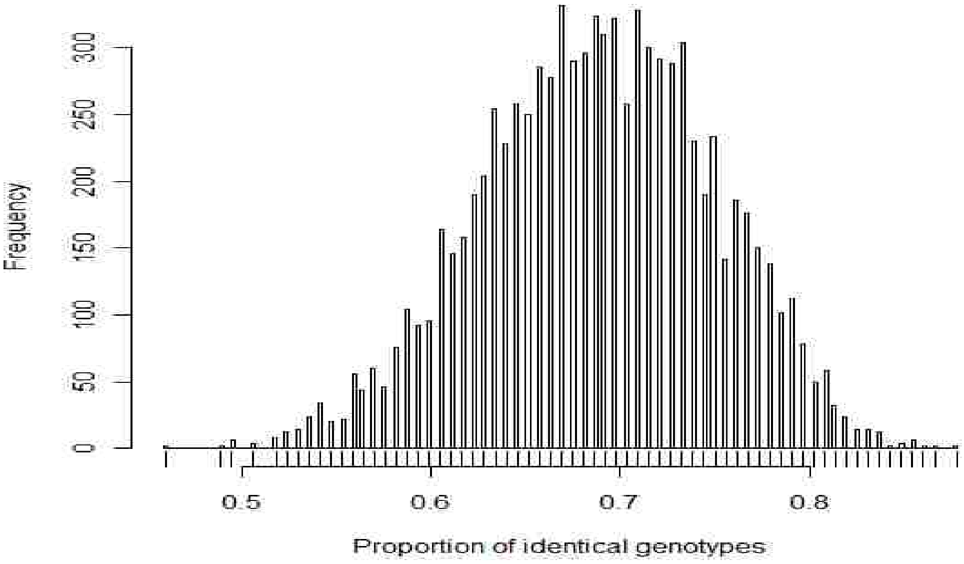


Fig 2.3 Frequency of shared alleles among the diploid F_2 segregating population. X axis: proportion of shared alleles; Y axis: frequency values

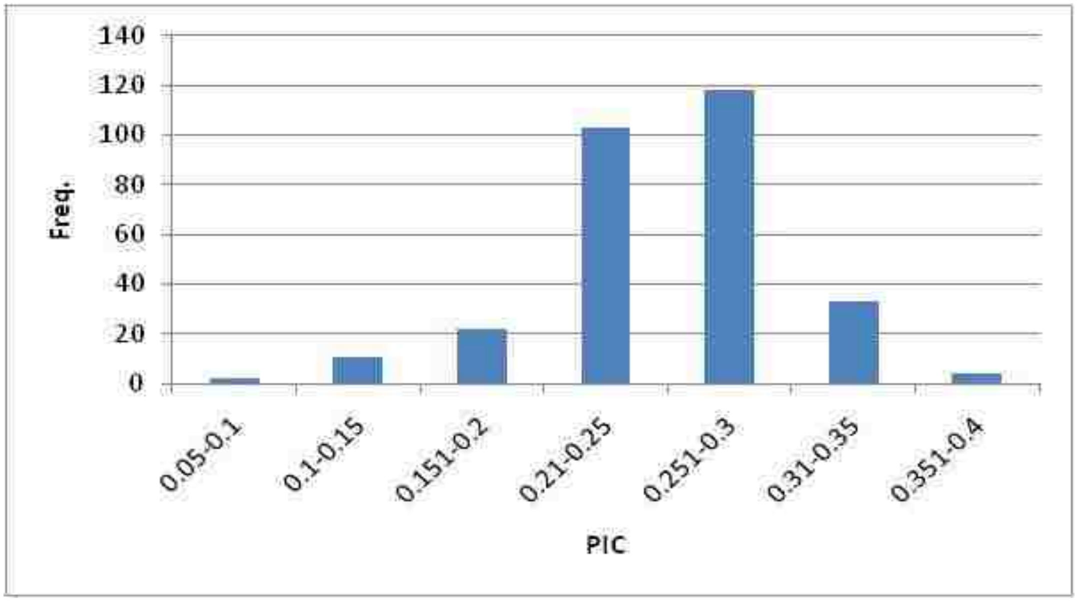
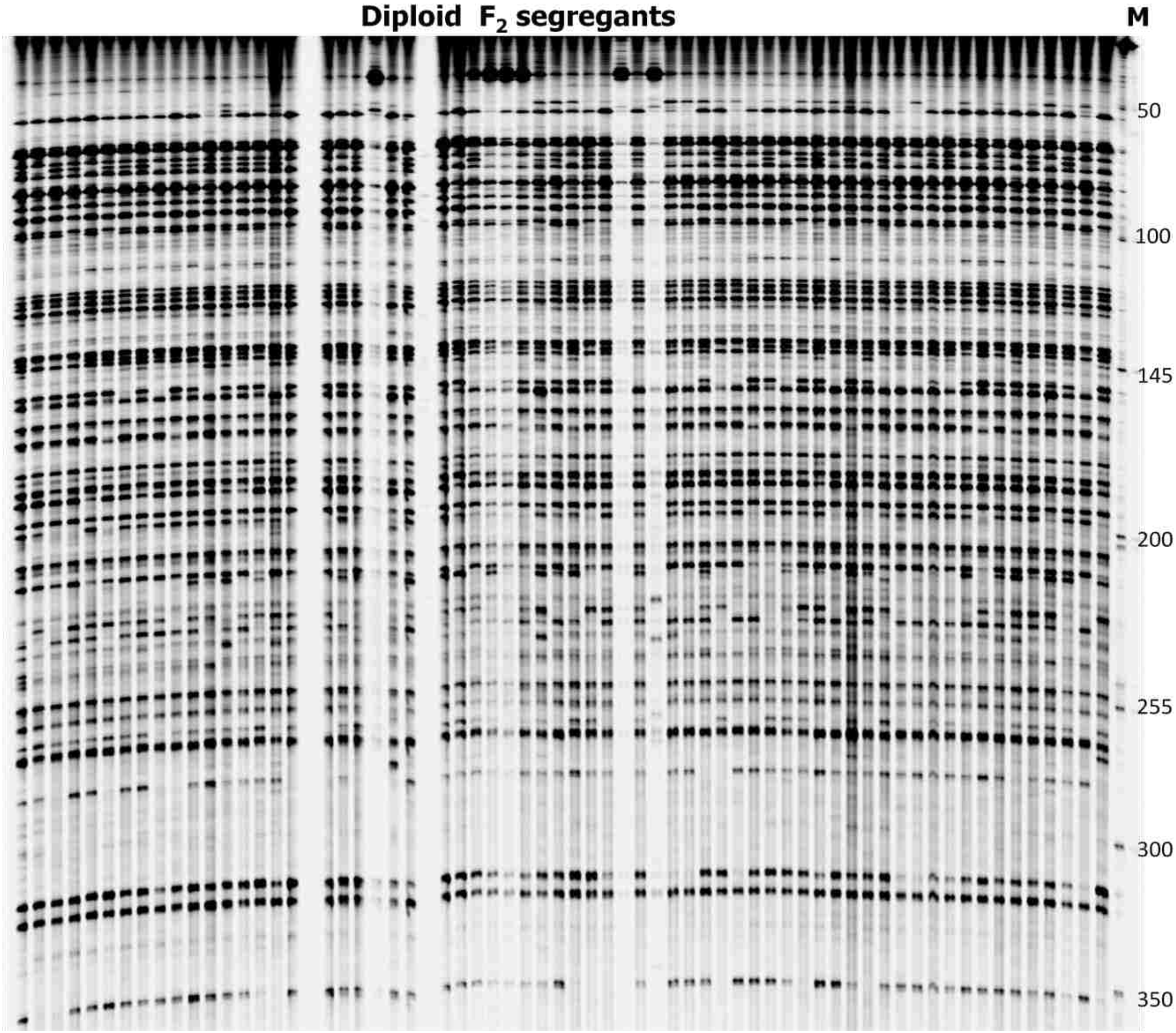


Fig 2.4 Frequency distribution of polymorphic information content (PIC) values in AFLP-SSR-TRAP markers in cotton association mapping. X axis: PIC estimates; Y axis: frequency values.

AFLP



SSR:

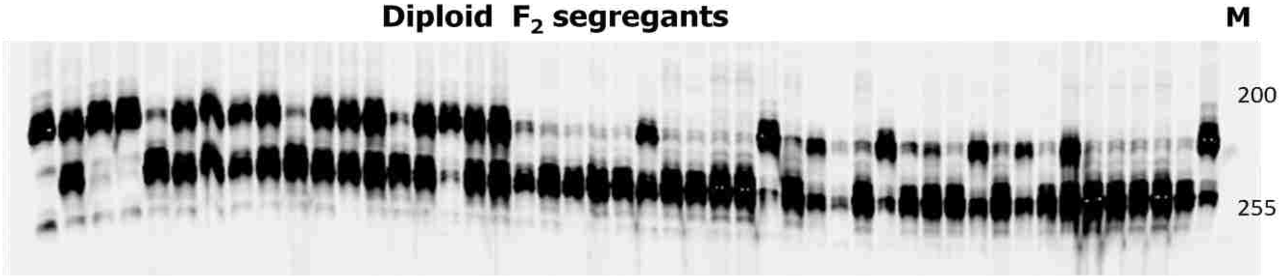


Fig 2.5 Representative gel pictures illustrating allele polymorphismof AFLP (E4M5) and SSR markers in diploid F_2 population from a cross between *G. arboreum x G. herbaceum*. M-molecular weight standard

2.3.3 Construction of Genetic Map

A total of 606 polymorphic markers were amplified from 94 F_2 individuals. Significant departures from the expected 3:1 (for AFLP and TRAP) and 1:2:1 (for SSR) segregation ratios were detected for 146 loci at $P \leq 0.05$, accounting for 24% of the polymorphic loci detected. A total of 140 markers were mapped on 37 linkage groups ranging from to 11 to 98cM in length (Table: 2.5). The remaining markers were ungrouped. The linkage groups were numbered from LG1-LG37 in descending order of length. The map covered a total of 1109 cM with an average distance of 7.92 cM between loci.

Nine linkage groups were considered as major ones (hosting more than 4 markers/LG) and the remaining were minor groups. The number of markers ranged from 2 to 21 per linkage group. One linkage group (LG7) consisted only of segregation distorted markers.

Few of the linkage groups had dense marker coverage. A majority of the linkage groups hosted evenly distributed markers. The diploid cotton has 26 chromosomes and the expected number of linkage groups is 13. Obviously, a greater number of polymorphic markers and a larger population size would aid in covering the number of linkage groups and fill in the remaining gaps.

2.3.4 QTL Analysis

a) Qualitative Traits

In cotton, the A genome diploids and the tetraploid species share a common morphology for various qualitative traits. In the present study, a survey of floral and seed morphology was done in the segregating F_2 population of a cross between the A_1 (SMA-4) and A_2 (A-97) genomes. A total of 606 markers including AFLP, SSR and TRAP were used to discriminate populations for two floral characters and seed fuzziness. The number of markers selected by the STEPDISC procedure applied after DA and the percent correct classification of F_2 individuals based on the selected markers is presented in Table 2.6.

Table 2.5 Number of genetic loci per LG, estimated LG length and average distance in the diploid linkage map

Linkage group	Number of loci	Estimated LG length (cM)	Average distance
LG1	21	98	4.7
LG2	7	96	13.7
LG3	3	25	8.3
LG4	8	77	9.6
LG5	8	55	6.9
LG6	7	43	6.1
LG7	6	17	2.8
LG8	3	38	12.7
LG9	3	33	11.0
LG10	3	24	8.0
LG11	2	13	6.5
LG12	2	19	9.5
LG13	2	17	8.5
LG14	2	25	12.5
LG15	2	11	5.5
LG16	2	20	10.0
LG17	2	20	10.0
LG18	17	60	3.5
LG19	4	45	11.3
LG20	4	45	11.3
LG21	3	26	8.7
LG22	3	30	10.0
LG23	3	23	7.7
LG24	3	45	15.0
LG25	2	23	11.5
LG26	2	31	15.5
LG27	2	23	11.5
LG28	2	22	11.0
LG29	2	3	1.5
LG30	2	18	9.0
LG31	2	12	6.0
LG32	2	20	10.0
LG33	2	27	13.5
LG34	2	25	12.5
LG35	2	26	13.0
LG36	2	22	11.0
LG37	2	22	11.0

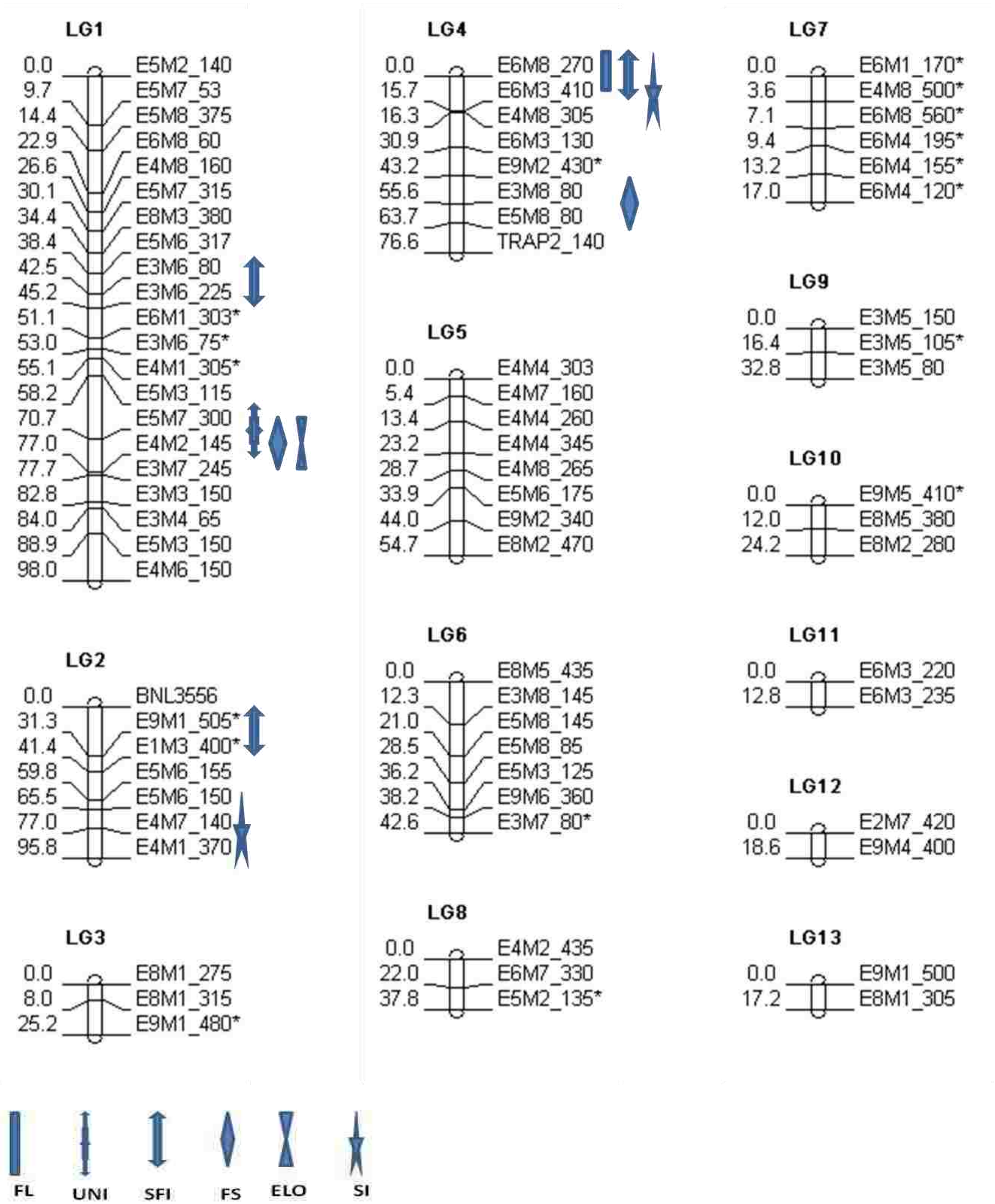
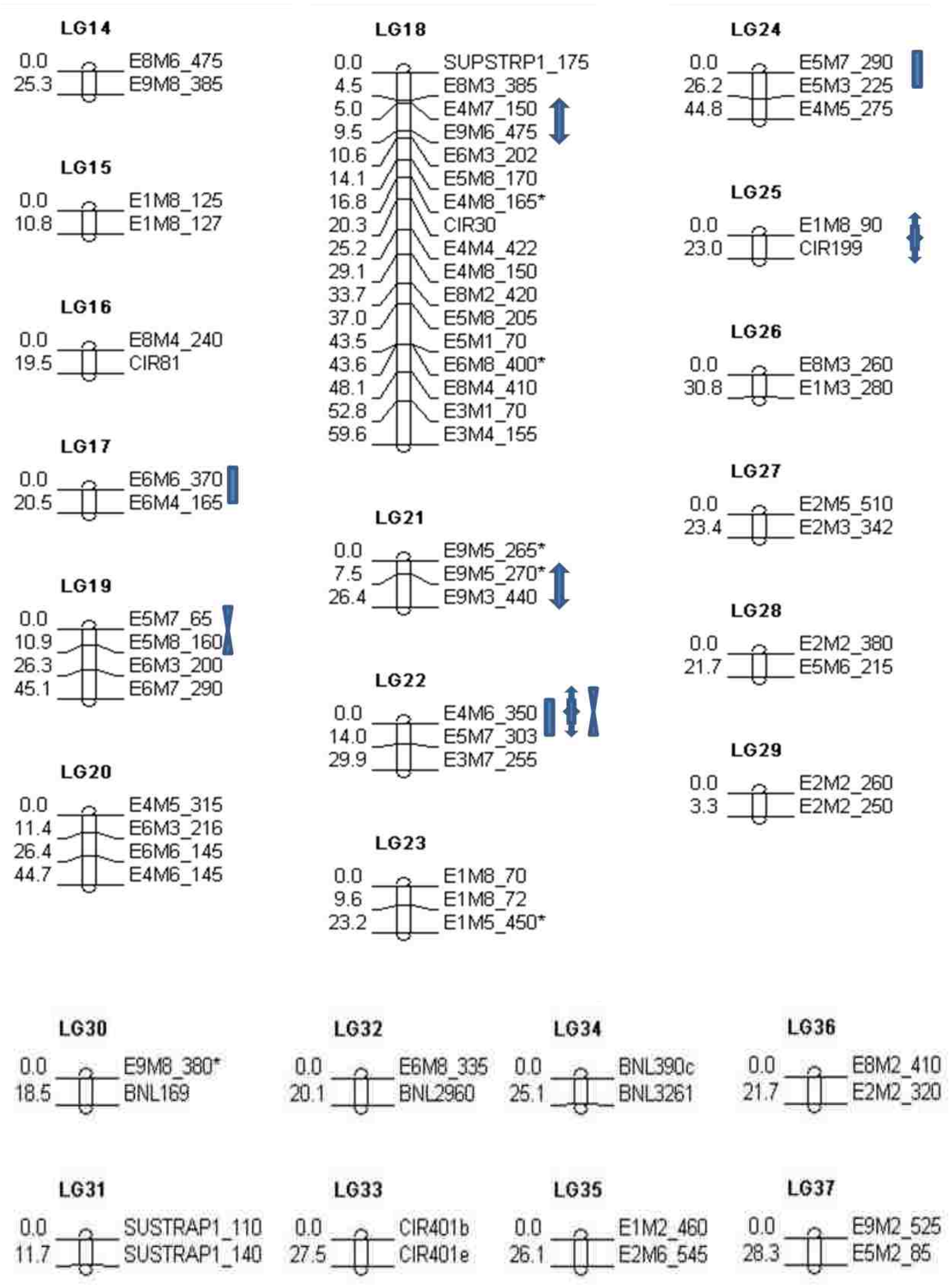


Fig 2.6 A genetic linkage map of the A genome diploid cotton based on the AFLP, SSR and TRAP markers. The map contains 37 linkage groups covering 1109cM with an average of 7.92 cM between loci. A total of 146 markers were identified as distorted ones, departing from the Mendelian segregation. They were represented with asterisk (*). QTL's for fiber traits and seed index are represented as boxes to the right side of each LG.

(Figure cont.)


LG14
0.0 E8M6_475
25.3 E9M8_385
LG15
0.0 E1M8_125
10.8 E1M8_127
LG16
0.0 E8M4_240
19.5 CIR81
LG17
0.0 E6M6_370
20.5 E6M4_165
LG19
0.0 E5M7_65
10.9 E5M8_160
26.3 E6M3_200
45.1 E6M7_290
LG20
0.0 E4M5_315
11.4 E6M3_216
26.4 E6M6_145
44.7 E4M6_145
LG18
0.0 SUPSTRP1_175
4.5 E8M3_385
5.0 E4M7_150
9.5 E9M6_475
10.6 E6M3_202
14.1 E5M8_170
16.8 E4M8_165*
20.3 CIR30
25.2 E4M4_422
29.1 E4M8_150
33.7 E8M2_420
37.0 E5M8_205
43.5 E5M1_70
43.6 E6M8_400*
48.1 E8M4_410
52.8 E3M1_70
59.6 E3M4_155
LG21
0.0 E9M5_265*
7.5 E9M5_270*
26.4 E9M3_440
LG22
0.0 E4M6_350
14.0 E5M7_303
29.9 E3M7_255
LG23
0.0 E1M8_70
9.6 E1M8_72
23.2 E1M5_450*
LG24
0.0 E5M7_290
26.2 E5M3_225
44.8 E4M5_275
LG25
0.0 E1M8_90
23.0 CIR199
LG26
0.0 E8M3_260
30.8 E1M3_280
LG27
0.0 E2M5_510
23.4 E2M3_342
LG28
0.0 E2M2_380
21.7 E5M6_215
LG29
0.0 E2M2_260
3.3 E2M2_250
LG30
0.0 E9M8_380*
18.5 BNL169
LG32
0.0 E6M8_335
20.1 BNL2960
LG34
0.0 BNL390c
25.1 BNL3261
LG36
0.0 E8M2_410
21.7 E2M2_320
LG31
0.0 SUSTRAP1_110
11.7 SUSTRAP1_140
LG33
0.0 CIR401b
27.5 CIR401e
LG35
0.0 E1M2_460
26.1 E2M6_545
LG37
0.0 E9M2_525
28.3 E5M2_85

DA identified three markers each for petal color and seed fuzziness and four markers for petal spot. The percent correct classification (obtained by cross-validation) was 100% with no error rate estimate. For petal color, DA selected AFLP markers E5M2_60, E5M6_205 and E9M1_560, which were able to discriminate the F_2 individuals with 100% correct classification in each training samples (5, 10 and 15%). Similar markers showed significant correlations with qualitative traits across different training samples selected for the study.

Table 2.6 Discriminant analyses selected markers for petal color, seed fuzziness and petal spot in a diploid F_2 population of cotton

Markers entered	Model R^2	Pr > F	Wilk's λ	Pr < λ†
Petal Color				
E5M2_60	0.39	0.0001	0.60	0.0001
E5M6_205	0.75	<0.0001	0.24	<0.0001
E9M1_560	1.00	<0.0001	0.00	<0.0001
Seed fuzziness				
E2M3_342	0.58	<0.0001	0.41	<0.0001
E8M8_510	0.84	<0.0001	0.15	<0.0001
E9M3_440	1.00	<0.0001	0.00	<0.0001
Petal Spot				
SUPSTRP1_175	0.32	0.0008	0.67	0.0008
E6M1_410	0.77	<0.0001	0.22	<0.0001
E9M2_520	0.87	<0.0001	0.12	<0.0001
E2M5_520	1.00	<0.0001	0.00	<0.0001

†: λ = Wilk's lambda used to test the significance of the discriminant model; Partial R^2 and Model R^2 were calculated from multiple regression (PROC REG, SAS Institute, ver. 9.1.2); % correct classification were calculated by leave one out validation within the training samples.

The 'Wilk's lambda' P values are significant for the Discriminant model as a whole and E9M1_560 was found to contribute more variation to the discriminant function. Similarly, DA identified E2M3_342, E8M8_510 and E9M3_440 as suitable marker for discriminating the population for seed fuzziness. The marker E2M3_342 is located on LG 27 and the marker

E9M3_440 on LG 21 where it is also associated with SFI. For the petal spot, DA identified SUSTRAP1_175 (LG 18), E6M1_410, E9M2_520 and E2M5_520 as significant markers for discriminating F_2 segregants with high adj. R^2.

b) Quantitative Traits

The locations, LG, LOD scores and additive and dominant effect estimates of major QTL's for all the fiber traits and seed index are given in Table 2.7. A total of 19 QTL's were identified and linked with five fiber traits or seed index by composite interval mapping. Of the 19 QTL's identified, LG4 and LG1 hosted markers linked with more than three traits.

A total of four QTL's were detected for the fiber length, which were located on linkage groups, 4, 17, 22 and 24 (Table 2.7). The qFL4-1 and qFL17-2 had high LOD values and explained 11.58 and 7.55% of the phenotypic variation, respectively. Three QTL's were detected for uniformity ratio in this population located on LG 1, 22 and 25 with R^2 values ranging from 5.6-9.5. The major QTL (LOD=9.48) was found in the interval of E1M8_90 - CIR-199. For SFI, five QTL's on LG 1, 2, 4, 18 and 21 were identified which had R^2 values ranging from 1.1-2.4. The major QTL (qSFI 2-2) was in the interval E9M1_505-E1M3_400, with an LOD value of 8.8 and dominant effect (17.7).

Two QTL's affecting fiber strength were identified, one, qFS4-2 on LG4 possessed an LOD value of 2.92 and both negative additive and dominant effects. The other QTL, qFS1-1, had an LOD value of 2.55 and explained about 9.5% of the phenotypic variation. A major QTL for uniformity index also mapped in the same region as one for fiber strength. A QTL located on E4M2_145 (qELO1-1) recorded a 2.95 LOD and explained 9.6% of phenotypic variation for elongation. Although there were two QTL's for seed index, only qSI2-1 located on LG2, possessed 4.17 LOD score, explaining 10.09% of the phenotypic variation. The significant QTLs identified showed additive and dominant effect across various fiber traits.

Table 2.7 Composite Interval Mapping for fiber traits using F_2 diploid cotton population from a cross between *G. arboreum* and *G. herbaceum*. QTLs are listed traitwise along with their position on LG, LOD, additive and dominant effects.

QTL†	LG	Position	LOD	Marker Interval	Additive*	Dominant	R^2
FL							
qFL4-1	4	4.0	5.1	E6M8_270-E6M3_410	-0.05	0.42	11.58
qFL17-2	17	14.0	5.1	E6M6_370-E6M4_165	-0.02	0.47	7.55
qFL22-3	22	2.0	2.7	E4M6_350-E5M7_303	0.14	-0.14	9.65
qFL24-4	24	20.0	4.3	E5M7_290-E5M3_225	0.03	0.78	4.80
UNI							
qUNI1-1	1	76.8	2.9	E5M7_300-E4M2_145	11.98	-14.4	9.56
qUNI22-2	22	2.0	2.5	E4M6_350-E5M7_303	12.40	-14.3	8.44
qUNI25-3	25	14.0	9.4	E1M8_90-CIR199	-1.31	25.6	5.60
SFI							
qSFI1-1	1	44.5	3.4	E3M6_80-E3M6_225	-0.36	18.44	1.41
qSFI2-2	2	33.1	8.8	E9M1_505-E1M3_400	0.37	17.17	1.90
qSFI4-3	4	12.0	7.5	E6M8_270-E6M3_410	-0.11	16.97	1.10
qSFI18-4	18	9.1	3.8	E8M3_385-E9M6_475	-1.50	19.74	2.36
qSFI21-5	21	23.5	3.4	E9M5_270-E9M3_440	0.30	-0.51	1.70
FS							
qFS1-1	1	77.0	2.5	E4M2_145	3.56	-5.31	9.50
qFS4-2	4	60.0	2.9	E5M3_115-E5M7_300	-1.35	-8.51	1.07
ELO							
qELO1-1	1	77.0	2.9	E4M2_145	0.76	-0.90	9.60
qELO19-2	19	10.0	2.5	E5M7_65-E5M8_160	0.86	-1.37	9.10
qELO22-3	22	2.0	2.5	E4M6_350-E5M7-303	0.78	-0.89	8.51
SI							
qSI2-1	2	91.0	4.1	E4M7_140-E4M1_370	1.61	0.45	10.09
qSI4-2	4	15.7	2.9	E6M8_270-E6M3_410	-1.32	7.60	6.80

† Nomenclature for the QTL was followed as per Shen *et al*., 2003.
*positive or negative additive effect leads to increased/decreased trait value with reference to SMA-4

Multiple regression is a statistical procedure that has been used to explore associations between molecular markers and quantitative traits. The assumption was made of a linear relationship between the markers and the quantitative trait of interest. In the present study, a total of 33 markers were identified to be associated with five fiber traits and seed index in diploid cotton (Table 2.8). The multiple regression method using 52 models with various selection options found to be robust enough to identify significant markers associated with fiber traits.

Table 2.8 Significant markers selected using PROC GLMSELECT for fiber traits in diploid F_2 mapping population.

Markers	Model R^2	Adj. R^2	AICC	BIC	SBC	PRESS	F value	Pr > F
FL[†]								
E9M2_380	0.170	0.161	-1.085	-200.132	-191.067	11.223	18.680	<.0001
BNL 3661	0.314	0.299	-1.251	-217.798	-204.200	9.497	18.830	<.0001
E4M2_145	0.406	0.386	-1.370	-231.138	-213.008	8.364	13.730	0.000
CIR 241	0.519	0.492	-1.533	-250.913	-223.717	7.042	10.530	0.002
E9M5_310	0.573	0.544	-1.627	-261.978	-230.250	6.309	10.870	0.001
E9M1_545	0.629	0.599	-1.741	-275.045	-238.785	5.600	12.820	0.001
E8M6_325	0.704	0.672	-1.911	-295.957	-250.631	4.745	11.270	0.001
E6M3_410	0.990	0.981	-3.45	-616.83	-403.80	2.369	6.27	0.015
FS								
E9M2_380	0.172	0.163	5.486	410.896	419.961	8004.312	18.870	<.0001
BNL 3661	0.316	0.301	5.317	393.051	406.649	6760.882	19.040	<.0001
E4M2_145	0.406	0.386	5.200	379.923	398.054	5965.966	13.490	0.000
E9M1_545	0.619	0.588	4.858	338.586	374.847	4105.780	12.040	0.001
SFI								
E9M2_380	0.151	0.141	4.947	360.757	369.822	4669.877	16.160	0.000
E1M5_485	0.264	0.247	4.828	347.504	361.102	4348.084	13.780	0.000
E9M8_395	0.363	0.342	4.706	333.953	352.083	3832.187	13.960	0.000
BNL3661	0.445	0.420	4.593	321.151	343.814	3484.333	12.990	0.001
E8M6_380	0.514	0.486	4.485	308.798	335.994	3009.731	12.360	0.001
ELO								
E9M2_380	0.172	0.163	2.378	121.835	130.900	357.780	18.880	<.0001
BNL3661	0.306	0.290	2.224	105.417	119.015	307.086	17.380	<.0001
E4M2_145	0.401	0.381	2.101	91.717	109.847	269.143	14.130	0.000
E8M8_300	0.617	0.585	1.757	50.247	86.508	194.137	12.880	0.001
SI								
E4M1_370	0.130	0.120	3.658	240.908	249.973	1292.379	13.560	0.0004
E9M3_200	0.303	0.287	3.460	220.304	233.901	1062.668	22.320	<.0001
E1M5_485	0.381	0.361	3.364	209.155	227.285	958.009	11.340	0.001
E6M3_500	0.466	0.442	3.241	195.427	218.090	878.357	14.000	0.000
E5M2_60	0.530	0.503	3.140	183.660	210.855	804.784	11.740	0.001
E9M7_350	0.729	0.696	2.724	132.420	182.278	507.689	11.180	0.001
UNI								
E9M2_380	0.173	0.164	7.897	635.104	644.170	89228.452	19.000	<.0001
BNL3661	0.311	0.296	7.737	618.043	631.641	76041.022	18.120	<.0001
E4M2_145	0.406	0.386	7.613	604.359	622.489	66688.242	14.110	0.000
CIR241	0.517	0.489	7.457	585.140	612.336	56425.459	9.880	0.002
E9M5_310	0.571	0.541	7.364	574.094	605.822	50569.328	10.850	0.001
E9M1_545	0.627	0.596	7.249	560.988	597.248	44872.756	12.860	0.001
E8M6_325	0.705	0.674	7.068	539.066	584.392	37547.020	11.710	0.001

[†] FL=fiber length, FS=fiber strength, SFI=short fiber index, ELO=elongation percentage, SI=seed index and UNI=uniformity ratio.

Fiber length was associated with seven significant markers. Markers such as E9M1_545, BNL3661 and E6M3_410 which possessed the highest Adj. R^2 values of 62.9, 31.4 and 99%, respectively, also had the lowest AICC, BIC and SBC values. The marker E9M3_410 was also detected by composite interval mapping with a LOD value of 5.1 and is located on LG4.

The F_2 population showed wide range of values for the FS (15.90-30.90g/tex). This phenotypic variation was efficiently captured (40-62%) by the markers E4M2_145 and E9M1_545. Low AICC, SBC and BIC values along with lowest P values ($P \leq 0.001$) for these markers showed the potential marker trait associations. The QTL qFS1-1 also mapped with the marker E4M2_145 on LG1 at 77cM with an LOD value of 2.55.

Markers such as BNL3661 and E8M6_380 were found to be strongly associated with SFI explaining 44.5 and 51.4% of phenotypic variation respectively. Both CIM and multiple regression methods identified several markers for SFI.

The marker E4M2_145 was found to be strongly associated with UNI and ELO using multiple regression methods. This marker was also found by the CIM method and mapped on LG1 (qUNI-1-1) and qELO1-1, for UNI and ELO traits respectively. Overall, this marker is associated with ELO, UNI, FL, and FS traits, found by CIM as well as multiple regression methods.

Seed index is a measure of seed weight (g/100 seed) and the F_2 segregants showed a wide range of values, 4.90-11.90g. Six markers were found to be significantly associated with seed index. Markers such as E4M1_370, E9M3_200 and E9M7_350 were able to capture 13, 30.3 and 72.9% of phenotypic variation respectively. Among these, E4M1_370 was also identified by the CIM method and located on LG2 at 91cM with an LOD of 4.17. For most of the traits, except SFI, composite interval mapping and multiple regression results are matching, due to the strong linkage of the marker and trait. Such a common markers were of high importance due to less probability of occurrence of false positives.

2.4 Discussion

2.4.1 Phenotypic and Molecular Diversity

Floral characters such as petal color, petal spot and seed related traits such as seed index and fuzziness showed a moderate degree of phenotypic diversity among the F_2 segregants. The *G. arboreum* petals are white without petal spot, while *G. herbaceum* has yellow flower with petal spot. These traits fit a monogenic inheritance models with the presence of petal spot as dominant over its absence and yellow petal color as dominant over white (Desai *et al*., 2006). The A genome cottons also exhibit a correlation between petal size and petal color with white color petals mostly associated with small flowers (Hutchinson, 1931).

Phenotypic diversity for fiber traits showed a wide range of values except for the ELO and UNI. The FL (0.7-1.02 inch), SFI (9.9-29.8), FS (15.90-30.90 g/tex) and SI (4.90-11.90g) showed a wide range of values in the segregants. The traits such as FL, UNI, SFI, FS and ELO were significantly correlated. Our results showed that approximately 75% of the F_2 individual plants were identical 60-70% at the molecular level for loci measured by markers. According to Kebede *et al*., (2007) microsatellite analysis revealed a low to moderate interspecific and intraspecific genetic diversity in *G. herbaceum* and *G arboreum* accessions. The study also indicated the extent of polymorphism up to 0.37 with an average of 0.253. The narrow genetic base of the parents could be due to the moderate diversity present in the contrasting characters under study. A similar study involving A_1 and A_2 genomes showed genetic similarities to the extent of 0.62 to 0.86 (Kebede *et al*., 2007).

2.4.2 Construction of Linkage Map

Among the 606 AFLP, SSR and TRAP markers used in this study, 460 markers were used to construct a diploid genetic map. Excluding 24% of the distorted markers, the map consists of 140 markers assembled on 37 linkage groups. The map covers 1109 cM with each loci at an average of

7.92 cM. Similar cross using 274 RFLP loci covered a map length of 1147 cM with an average distance of 4.2 cM between adjacent markers (Desai *et al*., 2006). Obviously more markers are needed to make this map more saturated.

Segregation distortion, the deviation of segregation ratios from expected Mendelian ratios has been reported in a wide range of plant species (Jenczewski *et al*., 1997). As many as 140 markers were considered to be segregation distorted accounting for 24% of the polymorphic markers scored. Segregation distortion may be due to the presence of lethal genes and/or overlapping fragments consisting of identically sized fragments (Hansen *et al*., 1999). It could also be related to different sizes of the parent genomes or to distorting factors, such as self-incompatibility alleles (Bert *et al*., 1999). Population size also influences the segregation distortion when the two markers are separated by more than 10cM (Hackett and Broadfoot, 2003).

2.4.3 QTL Analysis

Many genes are important to developmental, yield and fiber traits, but small population sizes, lack of marker saturation or over emphasis on tetraploid mapping without understanding the basic inheritance pattern of fiber genes in model diploid system has lead to meager success towards marker assisted selection (MAS) in cotton. Therefore, the present study attempts to map QTL's in a model A genome population.

Based upon discriminant analysis (DA) analysis, our study revealed that markers E5M2_60 and E5M6_205 were associated with petal color, while SUSTRAP1_175 and E6M1_410 were associated with petal spot. According to Desai *et al*., (2006), QTL's for floral characters showed a high correspondence among the At and Dt genomes and much lower correspondence among A and Dt. E2M3_342 was able to discriminate seed fuzziness and was found to be located on LG27. Seed fuzziness or naked seed was categorically discriminated by this marker and its parallel association with SFI indicated its role in suppressing seed fiber growth. In another *G. arboreum x G. herbaceum* segregating population, Rong *et al*., 2005 mapped the naked seed phenotype, *sma-4(fz)*, near the terminus of LG A.

Based on composite interval mapping, 19 QTL's were identified on 10 linkage groups for the five fiber and one seed related trait under study. The phenomenon of QTL clustering has been reported earlier in cotton (Shappely *et al*., 1998; Ulloa and Meredith, 2000; Qin *et al*., 2008). A total of four intervals were found to be involved in the control of more than two traits and located on LG 1, 2, 4 and 22. Not only are these fiber traits highly correlated, but they are also influenced by tight linkage, which was observed as linkage drag in breeding for these traits (Qin *et al*., 2008).

Since the parents, type of populations and marker system varied among different experiments reported in the literature and availability of few diploid linkage maps, detailed comparisons among different findings are difficult. With assignments of DNA markers or QTL's to specific chromosomes, such comparisons can be more valid.

For the significant QTL's identified, alleles associated with an increase in the trait value originated from both the parent. The potential QTL's having high LOD values, affecting UNI, FS on LG1 and 22 showed significant additive effects and explained 8.44-9.56% of the phenotypic variation. Similarly, Zhang *et al*., (2005) also observed additivity for UNI and FL on chromosome 5 of the upland cotton map, explaining 25% of trait variation. The marker E4M2_145 was found to be significantly associated with UNI, FS and ELO with high LOD values explaining up to 12% of the phenotypic variation. The significance of this marker was confirmed using composite interval mapping. In addition, multiple regression methods jointly confirmed the linkage of E4M2_145 with UNI, FS and ELO, E4M1_370 with SI, BNL 3661 and E8M6_380 with SFI and E9M3_410 with FL. Although the QTL's detected in this study have moderate genetic effects and their number is limited, the findings will help in validation and comparison to the tetraploid map.

2.5 Conclusion

The present study revealed moderate level of genetic diversity for the F_2 segregants. The study explored the diploid cotton genome as model system to map floral, fiber and seed traits. DA was effective in identifying potential markers which can differentiate among the floral traits. Both the composite interval mapping and multiple regression models confirmed the association of various QTL's to fiber and seed traits.

The construction of an A genome diploid map, combining AFLP, TRAP and SSR markers, can serve as a model for the advancement of cotton genetics, including the understanding of the inheritance of fiber genes. Adding additional markers to the existing map will assist in future map based cloning efforts and in gene discovery. However, the putative locations of the QTL do not necessarily represent physical distances (Shappley *et al*., 1998). Thus, a physical map is very much needed and would be of great value in cloning informative QTL's in cotton.

2.6 References

Abdalla AM, Reddy OUK, El-Zik KM and Pepper AE 2001 Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. Theoretical and Applied Genetics, 102, 222-229.

Altaf MK, Stewart J, Wajahatullah MK and Zhang J 1997 Molecular and morphological genetics of a trispecies F_2 population of cotton. Proceedings of Beltwide Cotton Conference, 448-452.

Baogong J 2004 Optimization of *Agrobacterium* mediated cotton transformation using shoot apices explants and quantitative trait loci analysis of yield and yield component traits in upland cotton (*Gossypium hirsutum*). PhD thesis. Louisiana State Univ.LA.

Basten CJ, Weir BS and Zeng ZB 2007 QTL Cartographer 2.5: A Reference manual and Tutorial for QTL Mapping. Department of Statistics, North Carolina State University, Raleigh, NC.

Bert PF, Charmet G, Sourdille P, Hayward MD and Balfourier 1999 A high-density molecular map for ryegrass (*Lolium perenne*) using AFLPmarkers. Theoretical and Applied Genetics 99, 445-452.

Brubaker CL and Brown AH 2003 The use of multiple alien chromosome addition aneuploids facilitates genetic linkage mapping of the *Gossypium* genome. Genome 46, 774-791.

Brubaker CL, Paterson AH and Wendel JF 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42, 184-203.

Desai A, Chee PW, Rong J, May OL and Paterson AH 2006 Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*. Genome. 49, 336.

Endrizzi JE, Turcotte EL and Kohel RJ 1985 Genetics, cytology, and evolution of *Gossypium*. Advances in Genetics, 23: 271-375.

Fisher RA 1936 The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7:179–188.

Fryxell PA 1979. The Natural History of the Cotton Tribe. Texas A&M University Press, Collage Station, Texas.

Gerstel, DU 1953 Chromosomal translocations in interspecific hybrids of the genus *Gossypium*. Evolution, **7,** 234–244.

Hackett CA and Broadfoot LB 2003 Effect of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity, 90: 33-38.

Han ZG, Guo WZ, Song XL and Zhang TZ 2004 Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Molecular Genetics and Genomics 272, 308-327.

Hansen M, Kraft T, Christiansson M, Nilsson NO 1999 Evaluation of AFLP in *Beta*. Theoretical and Applied Genetics, 98,845–852.

Hawkins JS, Pleasants J and Wendel, J.F. 2005 Identification of AFLP markers that discriminate between cultivated cotton and the Hawaiian Island endemic, *Gossypium tomentosum* Nuttall ex Seeman. Genetic Research on Crop Evolution, 52, 1069–1078.

Hu J, Vick BA 2003 TRAP target region amplified polymorphism, a novel marker technique for plant genotyping. Plant Molecular Biology Reporter, 21, 289–294.

Hutchinson JB 1931 Studies on the inheritance of corolla color and petal size in Asiatic cottons. Journal of Genetics, 24: 325-353.

Iqbal MJ, Aziz N, Saeed NA, Zafar Y 1997 Genetic diversity evaluation of some elite cotton varieties by RAPD analysis. Theoretical and Applied Genetics, 94,139-144.

Jenczewski EM, Gherardi I, Bonnin JM, Prosperi I, Oliveri 1997 Insight on segregation distortions in two intraspecific crosses between annual species of *Medicago* (Leguminosae). Theoretical and Applied Genetics, 94: 682–691.

Kebede H, Burow G, Dani RG and Allen RD 2007 A genome cotton as a source of genetic variability for upland cotton (*Gossypium hirsutum*). Genet Resources for Crop Evolution, 54, 885-895.

Khan MA, Zhang J, McD. Stewart J and Cantrell RJ 1998. Integrated molecular map based on a trispecific F_2 population of cotton. In: Beltwide Cotton Conference. 491-492.

Kosambi D 1944 The estimation of map distance from recombination values. Annals of Eugenics, 12, 172–175.

Lacape JM, Nguyen TB, Thibivilliers S, Bojinov B, Courtois B, Cantrell RG, Burr B and Hau B 2003 A combined RFLP–SSR–AFLP map of tetraploid cotton based on a *Gossypium hirsutum* ×*Gossypium barbadense* backcross population. Genome, 46, 612–626.

Mei M, Syed NH, Gao W, Thaxton PM, Smith CW, Stelly DM, and Chen ZJ 2004 Genetic mapping and QTL analysis of fiber-related traits in cotton (*Gossypium*). Theoretical and Applied Genetics, 108, 280-291.

Meredith WR Jr 2000 Cotton yield progress – why has it reached a plateau, Better Crops, 84: 6-9.

Murtaza N 2006 Cotton genetic diversity study by AFLP markers. Electronic Journal of Biotech. 9(4), 456-60.

Myers GO, Baogong J, Akash MW, Badigannavar AM, and Saha S 2009 Chromosomal assignment of AFLP markers in upland cotton (*Gossypium hirsutum* L.) Euphytica, 165(2), 391-399.

Nguyen TB, Giband M, Brottier P, Risterucci AM and Lacape JM 2004 Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. Theoretical and Applied Genetics, 109:167–175.

Qin H, Guo W, Zhang YM and Zhang T 2008 QTL mapping of yield and fiber traits based on a four way cross population in *Gossypium hirsutum.* Theoretical and Applied Genetics,117, 883-894.

Reddy A, Haisler RM, Yu J and Kohel RJ 1997 AFLP mapping in cotton. Plant Animal Genome Conference V, San Diego CA, USA.

Reinisch AJ, Dong J, Brubaker CL, Stelly DM, Wendel JF, Paterson AH 1994 A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploidy genome. Genetics 138, 829–847.

Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, Park C, Pierce GJ, Rainey KM, Rastogi VK, Schulze SR, Trolinder NL, Wendel JF, Wilkins TA, Dawn Williams Coplin T, Wing RA, Wright RJ, Zhao X, Zhu L and Paterson AH 2004 A 3347-locus genetic recombination map of sequence tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). Genetics 166, 389–417.

Rong J, Pierce GJ, Waghmare VN, Rogers CJ, Desai A, Chee PW, May OL, Gannaway JR, Wendel JF, Wilkins TA and Paterson AH 2005 Genetic mapping and comparative analysis of seven mutants related to seed fiber development in cotton. Theoretical and Applied Genetics, 111, 1137-1146.

SAS Institute, 2008 Documentation GLMSELECT. SAS Institute, Cary, NC.

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

Shappley ZW, Jenkins JN, Zhu J and McCarty JC 1998 Quantitative trait loci associated with agronomic and fiber traits of Upland cotton. The Journal of Cotton Science, 4, 153-163.

Singh P and Singh J 1984 Variability for some economic characters in the genetic stocks of *Gossypium arboreum* and *G. barbadense* cottons. Cotton Development, 14, 15–16.

Singh VV 1983 Range of variability in *Gossypium herbaceum* germplasm. Cotton Development, 14, 45.

Stam P, Van Oojien JW 1995 JoinMap version 2.0: software for the calculation of genetic linkage maps. CPRO-DLO, Wageningen, The Netherlands.

Ulloa M and Mereditch WR 2000 Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. Journal of Cotton Science, 4, 161-170.

Ulloa M, Meredith WR, Shappley ZW and Kahler AL 2002 RFLP genetic linkage maps from four $F_{2.3}$ populations and a joinmap of *Gossypium hirsutum* L. Theoretical and Applied Genetics, 104, 200–208.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M,Freijters A, Pot J, Peleman J, Kuiper M and Zabeau M 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23, 4407–4414.

Weir BS 1996 Genetic data analysis II, Sinauer, Sunderland, MA.

Wu J, Gutierrez OA, Jenkins JN, McCarty JC and Zhu J 2009 Quantitative analysis and QTL mapping for agronomic and fiber traits in an RI population of upland cotton, Euphytica, 165:231–245.

Wu KK, Burnquist W, Sorrells ME, Tew TL, Moore PH, and Tanksley SD 1992 The detection and estimation of linkage in polyploids using single-dose restriction fragments, Theoretical and Applied Genetics 83: 294-300.

Zeng ZB and Weir BS 1996 Statistical methods for mapping Quantitative Trait Loci. Acta Agronomica Sinica, 22, 535-549.

Zhang J. and J.McD Stewart, 2000 Economical and rapid method for extracting cotton genomic DNA. The Journal of Cotton Science, 4, 193-201.

Zhang ZS, Hu MC, Zhang J, Liu DJ, Zhang J, Zhang K, Wang W and Wan Q 2009 Construction of comphrensive PCR based marker linkage map and QTL mapping for fiber quality traits in upland cotton (*G. hirsutum*, L.). Euphytica, 24: 49-61.

Zhang ZS, Xiao YH, Luo M, Li XB, Luo XY, Hou L, Li DM and Pei Y 2005 Construction of a genetic linkage map and QTL analysis of fiber-related traits in upland cotton (*Gossypium hirsutum* L.). Euphytica, 144, 91-99.

CHAPTER 3 GENETIC ASSOCIATION MAPPING OF QUANTITATIVE TRAITS IN UPLAND COTTON

3.1 Introduction

Upland cotton (*Gossypium hirsutum* L., 2n = 52), one of four cultivated *Gossypium* species, is the world's leading fiber crop providing natural fiber for the textile industry (Endrizzi *et al*., 1985). Demands for enhancement of fiber quality traits such as fiber length and fiber strength have been increasing because of changes in spinning technology in the textile industry; however, most commercial cultivars, although high in yields, are lacking in desirable fiber quality. The primary breeding goal is how to simultaneously improve both yield and fiber quality.

Modern cotton cultivars show significant variation for agriculturally important traits (El-Zik and Thaxton 1989). For example, the longest, strongest and finest cotton fibers are produced by the *G. barbadense* cultivars of the Egyptian, Sea Island, and Pima groups. However, Upland *G. hirsutum* cultivars have earlier maturity, higher yields, and are adapted to a wider range of environments. An understanding of the genetic and genomic relationships of extant cotton species and cultivars is critical for the further utilization of cotton genetic diversity in the development of superior cultivars that combine the favorable qualities conditioned by this diverse germplasm.

A variety of molecular-marker technologies have been used to study the genetic diversity and relationships of crop species and their wild relatives. Studies using allozymes and RFLPs have been limited by low levels of polymorphism particularly at the intraspecific and even at the interspecific level. The amplified fragment length polymorphism (AFLP) method (Vos *et al.*, 1995) has been successfully used to analyze genetic diversity among a wide range of crop species and their wild relatives (Powell *et al.*,1996). AFLPs have been used to estimate genetic relationships in many studies including cotton (Pillay and Myers, 1999), lentil (Sharma *et al.*,1996), soybean (Maughan *et al.*, 1996), and barley (Becker *et al.*,1995). The major advantage of AFLP is its power to identify large numbers of potentially polymorphic loci. Evolutionary and genetic relationships of various germplasm resources including 43 cultivars of *G. raimondii, G. incanum, G. herbaceum and G. arboreum* were estimated using AFLP markers (Iqbal *et al.,* 2001). Molecular evidence for species distinctness from diversity analysis using AFLP markers in cultivated Indian diploid cotton indicated that *G. herbaceum* and *G. arboreum* formed two different clusters (Rana and Bhat 2004).

A bulked segregant analysis (BSA) approach combined with AFLP was used to identify additional molecular markers linked to the root knot nematode (*Meloidogyne incognita*) resistance genes *rkn1I* (Wang and Roberts 2006). AFLPs and SSRs were also used to search for novel markers linked to the *Xanthomonas compestris* (*Xcm)* resistance locus to facilitate introgression of this trait into *G. barbadense* through MAS. AFLP-RGA (Resistant Gene Analogs) was employed in cotton to search for polymorphisms in putative RGAs (Zhang *et al.,* 2007). The level of polymorphism detected with this technique was similar to that of AFLP. Approximately 300 polymorphic AFLP-RGA markers were identified, many of which were placed on an existing linkage map (Niu *et al.,* 2007).

The breeding process can be enhanced by using the linkage between markers and traits, which enables indirect selection via markers avoiding the phenotypic assessment of traits. This is especially important for traits whose expression is modified by the environment or for which conventional assays are difficult to do. An important step towards the establishment of such linkages is the development of genetic maps. Genetic mapping of traits comes down to establishing linkage between mapped markers and phenotypic trait observations, mostly quantitative in nature. Finding such linkage can be done in several ways. Two commonly used approaches are: a) linkage analysis using a bi-parental mapping population segregating for the trait(s) of interest, or b) linkage disequilibrium /association mapping using a well chosen (natural) population of lines, accessions, or cultivars.

Association mapping (AM) is based on the assumption that there is a set of markers available and either they represent actual genes (or alleles) or that of the markers are so close to the actual functional genes that they co-segregate and happen to be in linkage disequilibrium. This implies that the LD mapping is done with a natural population in which an association between traits and markers exists due to linkage disequilibrium. It has been used to study the genetics of complex traits in agricultural crops such as rice, maize, and barley (Iwata *et al.,* 2007). Association mapping studies make much broader use of available germplasm, thus ensuring a more comprehensive and precise mapping of QTLs. Association mapping identifies QTLs by examining the marker-trait associations, and enables researchers to use modern genetic technologies to exploit natural diversity and locate valuable genes in the genome (Zhu *et al.,* 2007).

The degree of LD in a germplasm depends on the recombination events that have taken place in history (Nordborg and Tavaré, *et al*., 2002). It is a result of the interaction between many factors, *e.g.* the mating system, recombination rate, selection, and population subdivision (Flint-Garcia *et al*., 2003). Not all LD occurring in a germplasm is due to linkage between loci. LD between unlinked loci can occur, attributable to population structure, admixture, outcrossing events and selection. Therefore, observed associations between markers and traits should be interpreted with care.

Two approaches are commonly applied in association mapping (1) whole genome mapping (Kraakman *et al.,* 2004) and (2) a candidate gene based approach (Wilson *et al.,* 2004). The candidate gene-association approaches rely on combining multiple lines of evidence to restrict the number of genes that are evaluated. Genome sequencing, comparative genomics, transcript profiling, low-resolution QTL analysis, and large scale knockouts provide opportunities to develop and refine candidate gene lists. These approaches are powerful at identifying candidate genes, but not at evaluating allelic affects. They can substantially reduce the amount of genotyping required, but most importantly, it can reduce the multiple issues created by testing thousands of sites across the genome.

Whole genome scans focus on the identification of genomic regions on all chromosomes related to the trait of interest. Success and resolution of this method depends on the extent of linkage disequilibrium (LD). The advantages of a population-based association study, which utilizes a sample of individuals from the germplasm collection or a natural population, over traditional QTL-mapping in biparental crosses are primarily due to; (1) availability of broader genetic variations with wider background for marker-trait correlations; (2) likelihood for higher resolution mapping because of the utilization of major recombination events from a large number of meiosis throughout the germplasm development history; (3) possibility of exploiting historically measured trait data for association; and (4) no need for the development of expensive and tedious biparental populations that makes the approach time saving and cost-effective (Kraakman *et al*., 2004). The disadvantages of this approach are mainly Type I errors; associations could be caused by population structure resulting in a lack of linkage information among the markers identified for significant associations. All these can be attributed to population stratification caused by gene drift, founder effects or selection (Pritchard *et al*., 2000).

Several methods have been proposed for estimating population structure and its modeling in AM studies (Pritchard *et al.*, 2000a and 2000b; Peleg *et al.,* 2008). Population structure is an important component in association mapping analyses because it can reduce both type I and type II errors between molecular markers and traits of interest in an autogamous species (Yu *et al.,* 2006). Distance based estimates of population structure are generally based on the clustering of individuals using pair-wise genetic distance estimates between individuals (Nei 1972; Rogers 1972; Nei 1978). In contrast, model based methods assign individuals probabilistically to one or more sub-populations. The most common model-based approach is Bayesian modeling where allele frequencies are used to estimate the likelihood of an individual belonging to a particular subpopulation. A mixed linear model (MLM) approach was found effective in removing the confounding effects of the population substructure in association mapping (Yu *et al.,* 2006) by using both the population structure information (Q-matrix) and pair-wise relatedness coefficients-'kinship' (K-matrix). The MLM or Q + K model works better than either the K model or Q model alone, as demonstrated in a highly structured *Arabidopsis* population (Yu *et al.,* 2006; Zhao *et al.,*2007). These approaches allow assignment of individuals to respective populations that can be integrated into statistical models to account for population structure in AM studies.

Various biometrical methods have been used in the past for estimating the mode of gene action controlling different agronomic and quality characters. In most of the genetic designs used, it is assumed that non-allelic interactions are absent, whereas the contrary is often true. Most methods also calculated a much larger standard error for the dominance component than for the additive component. Using modified QTL mapping one can identify two-way epistatic interactions by performing a complete pair-wise analysis of all the molecular markers. Fiber quality traits of cotton are inherited in a complex manner and tend to vary with the environment. Epistasis has been suggested to be the foundation of these complex traits. Adding epistasis to a model can increase the accuracy of prediction.

The numerous examples of association mapping studies performed in various plant germplasm resources including the model plant *Arabidopsis,* demonstrates the enthusiasm with which LD-based association mapping has met. Cotton provides a good platform for using genome-wide association mapping to catalogue genes responsible for natural variation and identification of QTL's for economic traits but relatively few studies have been done using this approach. LD mapping involving wild, Uzbekistan cotton varieties and exotic *G. hirsutum* germplasm lines was performed with 210 chromosome specific SSR's and detected higher linkage disequilibrium estimates in exotic accessions than varieties (Abdurakhomonov *et al*., 2008). In a study involving 260 *G. hirsutum* lines, 314 polymorphic SSR markers derived from exotic crosses were used to identify those associated with fiber traits. Structure analysis divided the panel into six clusters and 59 markers were associated with fiber traits (Zeng *et al*., 2009).

In light of the prospects of association mapping in other crops and as well as in cotton and the paucity of such studies in cotton, the present investigation was undertaken with following objectives;

1. Estimation of genetic diversity in a pool of genotypes representing US upland cottons.
2. Defining the cryptic population structure among US cotton genotypes.
3. Association of markers with yield and fiber quality parameters using various statistical models such as mixed and general linear models.

3.2 Materials and Methods

3. 2.1 Plant Material

A set of 220 upland cotton (*G. hirsutum* L.) and 12 genotypes from a standardized panel, representing subgenome donors, introgression breeding source, genetic standards and popular and or historical genotypes were considered for association mapping. The cotton association mapping panel (CAM) composition is given in Tables 3.1a and b. The entire CAM panel was divided into six groups based on the geographical origin of the breeding programs that developed them and or the region of their primary cultivation; Louisiana, Arkansas, South East (SE), Delta, Texas and Wild/Std. panel. A significant percentage of the genotypes were advanced breeding lines entered into the Regional Breeder's Trial Network (RBTN), a multistate cooperative testing program of public breeding programs.

Using the CTAB method, DNA was extracted from the young leaves of field grown plants (Zhang and Stewart 2000). The phenotypic data on yield and fiber traits for the RBTN entries in CAM panel was downloaded from the project website (www.cottonrbtn.com). The LA (kindly provided by Dr. G. Myers, Louisiana State University) and ARK (kindly provided by Dr. F. Bourland, University of Arkansas) genotypes phenotypic data, especially yield and fiber quality data was compiled from multiyear or multi location data (minimum of four environments and four replication).

Table 3.1a List of genotypes selected from five growing regions utilized for cotton association mapping

Louisiana
LA00405034
LA1110001
LA1110105
LA1110011
LA1110147
LA1110148
LA1110062
LA1110083
LA1110034
LA1110023
LA1110069
LA1110035
LA1110015
LA1110021
LA1110003
LA1110002
LA1110085
LA1110038
LA1110046
LA1110014
LA1110061
LA01407117
LA01407009
LA01407045
LA01407074
LA01407072
LA01407070
LA01407020
LA01407029
LA01407076
LA03404034
LA03404039
LA03404204
LA03404035
LA03404019
LA03404192
LA03404027
LA03404238
LA03404076
LA03404138
LA03404148
LA03404077
LA03404074
LA03404171
LA03404142
LA03404065
LA03404086
LA03404063
LA03404018
LA03404051
LA03404052
LA04307061
LA04307066
LA04307003
LA04307047
LA04307125
LA04308077
LA04308019
LA04307004
LA04308064
LA04308030
LA04308044
LA04307074
LA04307063
LA04308036
LA04307062
LA04307027
LA04307014

Arkansas
0101-10
0101-12
0101-24
0101-26
0101-34
0101-39
0101-41
0101-42
0101-46
0101-49
0101-55
0101-59
0102-11
0102-13
0102-48
0103-06
0103-45
0103-70
0104-03
0104-07
0104-10
0104-11
0104-20
0104-31
0104-36
0104-44
0104-47
0108-04
0108-20
0109-01
0109-11
0109-18
0110-16
0110-21
0110-38
0110-40
0107-03
0107-39
0111-20
0111-23
0111-24
0112-11
0112-25
0112-32
0112-34
0112-40
0114-03
0114-09
0114-11
0114-12
0114-20
0114-28
0114-46
0114-53
0022-11
0023-12
0023-13
0023-15
0023-16
0023-17
0034-15
0117-16
0120-21
0121-23
0105-15
0113-06
0113-15
0113-17
0113-19
0113-48
0113-49
0113-57
0001-01-03
0001-01-04
0001-01-09
0002-03-02
0002-19-04
0006-03-05
0006-11-05
0007-32-03
0008-22-10
0009-13-01
0011-11-03
0011-11-04
0012-03-08
0015-06-09
0015-06-11
0015-10-01
0015-11-04
0016-05-10
9704-13-05
9704-13-08
9706-36-05
9706-38-06
9706-39-10
9721-23-08
9801-36-03
9801-36-08
9801-37-04
9805-06-01
9811-15-07
9815-05-09
9803-17-04
9803-23-04
9803-23-08
9803-23-12
9823-05-04
0121-01
DELCOT277
0110-2NE
0141-15NE
99F-87

South Eastern
AU 1065
AU1107
AU1403
AU5210
AU6207
COKER100
GA2002212
GA2003118
GA2003156
GA3003131
PD03001
PD03011
PD2165
PD3025
PD99036
PD99041

Texas
FM800B2R
FM9060F
FM9063B2F
FM9068F
FM955LLB2
FM958
FM960B2R
FM960BR
FM965LLB2
FM991B2R
LANKART57
MCNAIR235
PM54

Delta
DPL393
DPL493
DPL491
DPL-458BR
DP393
DPL117B2RF
DP444BR
PHY485WRF
SG105
STV-4892BR

Table 3.1b Cotton Microsatellite Database (CMD) - a standardized Panel of cotton genotypes used to compare with association mapping genotypes (www.cottonmarker.org)

Code	Genotype	Characteristics
CMD01	TM-1	*G. hirsutum* (AD1)-genetic standard (BAC donor /RI parent)
CMD02	3-79	*G. barbadense* (AD2) -genetic standard (fiber QTLs /RI parent)
CMD03	Acala Maxxa	California Upland cotton (AD1) and BAC donor
CMD04	DPL 458BR	Upland cotton (AD1) with significant acreage
CMD05	Paymaster 1218BR	Upland cotton (AD1) with significant acreage
CMD06	Fibermax 832	Upland cotton (AD1) with significant acreage
CMD07	Stoneville 4892BR	Upland cotton (AD1) with significant acreage
CMD08	Pima S-6	Pima (AD2) germplasm breeding source
CMD09	*G. arboreum*	A subgenome representative
CMD10	*G. raimondii*	D subgenome representative
CMD11	*G. tomentosum*	Introgression breeding source
CMD12	*G. mustelinum*	Introgression breeding source

Environments were treated as replicates. The four replication data on lint yield, micronaire, fiber length, fiber strength, uniformity ratio, maturity coefficient and Short Fiber Index (SFI) were averaged across all testing locations to calculate variances. DP 393 was considered as check and all the comparisons were made in accordance with the performance of this cultivar. Especially for lint yield, DP 393 was taken as standard check and values of other CAM panel were adjusted to it. Fiber analysis data is derived from the High Volume Instrument (HVI) system. Correlation analysis for each trait was performed using PROC CORR in SAS.

3.2.2 Genotyping with Molecular Markers

Sixty four primer combinations were used to generate AFLP data (Table: 3.2) following the procedure given by Vos *et al*., (1995) with some modifications. Sample DNA was digested with *EcoRI* (infrequent cutter with GAATTC recognition sequence) and *MseI* (frequent cutter with TTAA recognition sequence) restriction enzymes and oligonucleotide adapters specific to restriction sites were ligated to the resulting fragments through incubation (37°C for 180 min) with DNA ligase via in an iCycler (BioRad Labs, Hercules, CA.)

Table 3.2 Adapters and primers of AFLP markers system used for pre and selective amplification in cotton association mapping.

Primer/adapter	Nomenclature*	Sequences(5'-3')
ECORI primers:		
EcoRI linker 1	E-I	CTC GTA GAC TGC GTA CC
EcoRI linker 2	E-II	AAT TGG TAC GCA GTC TAC
EcoRI + A	E+A	GAC TGC GTA CCA ATT CA
E- AAC	E1	GACTGCGTACCAATTCAAC
E- AAG	E2	GACTGCGTACCAATTCAAG
E-ACA	E3	GACTGCGTACCAATTCACA
E-ACT	E4	GACTGCGTACCAATTCACT
E-ACC	E5	GACTGCGTACCAATTCACC
E-ACG	E6	GACTGCGTACCAATTCACG
E-AGG	E8	GACTGCGTACCAATTCAGG
E-AGA	E9	GACTGCGTACCAATTCAGA
MseI primers:		
MseI linker 1	M-I	GAC GAT GAG TCC TGA G
MseI linker 2	M-II	TAC TCA GGA CTC AT
MseI + C	M+C	GAT GAG TCC TGA GTA AC
M-CAA	M1	GATGAGTCCTGAGTAACAA
M-CAC	M2	GATGAGTCCTGAGTAACAC
M-CAG	M3	GATGAGTCCTGAGTAACAG
M-CAT	M4	GATGAGTCCTGAGTAACAT
M-CTA	M5	GATGAGTCCTGAGTAACTA
M-CTC	M6	GATGAGTCCTGAGTAACTC
M-CTG	M7	GATGAGTCCTGAGTAACTG
M-CTT	M8	GATGAGTCCTGAGTAACTT

*Nomenclature is in accordance with the Lacape *et al*., 2003; Myers *et al*., 2009.

Pre-amplifications were done using *EcoR* I+A and *Mse* I+C oligo primers. The amplification was carried out with 50ng/ul of oligo primers, 5mM dNTP's, 25mM $MgCl_2$, 10X buffer, Taq polymerase (5U/µl), restrict ligated template DNA and ddH_2O in a total volume of 20ul. The PCR was set up with initial denaturing for 94^oC (2 min.) followed by 26 cycles at 94^oC (1 min), 56^oC (1 min), 72^oC (1 min) and final extension at 72^oC for 5min. The pre amplified products were diluted with ddH_2O and selective amplification was done using two selective nucleotides. The EcoRI+ANN oligo primers were dye labeled with 700 and 800 IR dye (MWG Biotech, Germany). The PCR for selective amplification was carried out in a reaction volume of 10 µL consisting of 10X reaction buffer, 25 mM $MgCl_2$, 2.5 mM dNTP 1µM each EcoRI-ANN ANN and MseI+CNN primers and 5U *Taq* polymerase (Promega, Madison, WI). The reactions were run on an *i*-Cycler (BioRad Labs, Hercules, CA, USA). The working PCR conditions for selective amplifications were standardized as follows: initial denaturing step at 94^oC for 2 min followed by 12 cycles at 94^oC for 30 s, 65^oC for 30 s (with 0.7^oC decrement every cycle) and 72^oC for 1 min, then followed by 23 cycles at 94^oC for 30 s, 56^oC for 30 s, and 72^oC for 1 min with a final extension step at 72^oC for 2 min. A total of 64 ***EcoR*** I - *Mse* I selective amplification primer combinations were used. The PCR amplified products were run on a LI-COR 4300 sequencer (LI-COR Inc., Lincoln, NE).

The gels were saved onto a computer and scored manually. Presence of band was recorded as '1' and absence as '0', a typical dominant marker system. Ambiguous data that could not be resolved were discarded. The nomenclature of AFLP loci was followed according to Lacape *et al.*, . (2003) and Myers *et al*., (2009), indicating the enzyme primer combination with band size.

3.2.3 Molecular Diversity and Population Structure

For each marker used, sub-populationwise diversity statistics including number of bands, unique bands, number of observed and effective alleles, Nei's genetic distances, expected heterozygosity and Shanon's information index were calculated using GenAlex 6.1 software (Peakall and Smouse, 2006). Allelic diversity at a given locus can be determined by Polymorphism Information Content (PIC) and was calculated as '$PIC=1-\sum f_i^2$ where, f_i is the frequency of the i^{th} allele (Weir, 1996). PROC ALLELE was used to calculate PIC values and frequency estimates were done using PROC FREQ (SAS 9.1.3, SAS Institute, Cary, NC).

Genetic differentiation among the subpopulation was estimated using hierarchical analysis of molecular variance (AMOVA; Excoffier *et al*., 2005) method in GenAlEx 6.1 (Peakall and Smouse, 2006). The pairwise F_{ST} values (Wright 1965) were estimated using the Bayesian model for dominant markers without prior knowledge of inbreeding coefficients. Wright's F-statistic is a hierarchical series of measures that indexes the fixation of different alleles in different populations. The pairwise F_{ST} values among the six predefined groups were calculated using AFLPSURV (Vekemans 2002). In order to know the possible structure in the set of CAM panel, various statistical analysis were performed on the basis of allelic frequencies. First, the Dice similarity coefficient was calculated using the formula $D = 2a/(2a + b + c)$, where a = 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). The genetic similarity coefficient matrix was then used to construct a tree with the neighbor joining procedure (Saitou and Nei, 1987) in MEGA software (Kumar *et al*., 2004). In addition, Principal Coordinate Analysis (PCoA) was performed using a genetic similarity matrix based on Nei – Li (1979) estimates to supplement the findings obtained from cluster analysis. All the above analyses were performed employing different modules of NTSYS-PC software, version 2.2 (Rohlf, 2000).

Correspondence analysis was also performed on the CAM panel using the marker matrix of band incidences (Greenacre, 1984). The multivariate nature of correspondence analysis can reveal relationships that would not be detected in a series of pair wise comparisons of variable. Another important feature is the graphical display of row and column points in biplots, which can help in detecting structural relationships among the variable categories and objects. The whole procedure was implemented in PAST software (Hammer *et al*., 2001) using AFLP marker data with predefined cultivar groups.

Bayesian model based clustering was performed using Structure software according to Pritchard *et al*., (2000a). The main criteria for this type of clustering is the allocation of individual genotypes into groups in such a way that Hardy-Weinberg equilibrium and linkage disequilibrium are valid within clusters, but absent between clusters. Gene flow between genetically distinct populations creates linkage disequilibrium (admixture linkage disequilibrium [ALD]) among all loci (linked and unlinked) that have different allele frequencies in the founding populations. Based on the prior information about the historical and popular cultivars included in the study along with standardized panel genotypes, we thought that sufficient exchange of favorable alleles among these genotypes can be accounted in the model. Therefore, the admixture model in Structure software was used and allele frequencies among populations were assumed to be correlated. To determine the optimum number of subpopulations, values for k ranging from 2-10 were performed with three independent runs for each value. Each run was carried out using 100,000 iterations with 100,000 burn-in iterations.

The optimum number of clusters (k) was determined based on the estimated logarithmic likelihood of the data (Yu *et al*., 2006). This value reaches a plateau when the minimum number of groups that best describes the population structure has been reached (Pritchard *et al*., 2000a; Evanno *et al*., 2005). Additionally, if there are separate populations the inferred value of alpha, which is defined as the ‘Dirichlet’ parameter for the degree of admixture, should remain constant (range ~0.2) while running the program. The mean alpha value for this data set was 0.0630 at k=6.

Another criterion for deciding the most appropriate value of k is the proportion of individuals belonging to the various populations should not be equal. If the population membership is symmetric (~1/K is 0.167) most of the individuals will be fairly admixed and one should infer that there is no real population structure. The membership of individuals in the populations determined by Structure for this data set was between 0.093–0.261. Therefore, based on the biological information on cultivar grouping and various statistics employed, it was evident that there exists at least six clusters. A graphical display of subpopulation composition from Structure software was generated using DISTRUCT (Rosenberg, 2002).

Nonrandom mating induces correlations in allelic states within and among loci, which can be used to understand the genetic structure of natural populations (Wright, 1965). For many species including cotton, it is important to quantify the contribution of two forms of nonrandom mating; inbreeding (mating among relatives) and population substructure (limited dispersal of gametes). To do this, 'INSTRUCT' model allowing for population structure and selfing rates was used (Gao *et al*., 2007). INSTRUCT implements a Markov Chain Monte Carlo (MCMC) algorithm for the generalized Bayesian clustering (extension of STRUCTURE) method to estimate the self fertilization rates or inbreeding coefficients and population-of-origin classification using multilocus marker data. The clustering of individuals into subpopulations is based on the genotypic data consisting of unlinked markers (Gao *et al*., 2007). The diploid model with 100000 burns, 200000 iterations, inferring populations structure with admixture specifications was run for 'k' ranging from 2-10. The data file was analyzed using the Computational Biology Application Suite for High Performance Computing, (Cornell University, Ithaca, NY). A graphical display of subpopulation composition from Instruct software was generated with DISTRUCT (Rosenberg, 2004). The deviance information criterion (DIC) was used to infer optimal k (Gao *et al.,* 2007). A common methodology to check the model convergence is by tracking the Gelman-Rubin convergence statistics (Brooks and Gelman, 1998). A Gelman–Rubin statistic under 1.2 indicates approximate convergence and it is used to assess when convergence occurs.

Pairwise kinship estimates were calculated using SPAGeDi software (Hardy and Vekemans, 2002). A kinship matrix consisting of coefficients along with Q-matrix obtained from

STRUCTURE, INSTRUCT and eigenvectors of PCoA, describing the assignment of each cultivar to a specific cluster, was used in the mixed model for association analysis.

3.2.4 Association Analysis: Statistical Models and Procedures

a) Mixed Models for Association Mapping

In association mapping, there is need to account for type I error or spurious associations/false positives. Incorporating the outcome of population structure and Principal coordinate analyses (PCA) increases the power to detect true marker trait associations. Eight statistical mixed models were tested (Table 3.3) for 568 AFLP markers and adjusted Rsquare values were computed for the fixed marker effects using TASSEL 2.1 beta version (Bradbury *et al*., 2007). Tests for significance were applied using the 'F' statistic associated with each marker. The model possessing the highest adj. R^2 was considered best at capturing the maximum variation. A cutoff P value (0.05) was used to determine whether a QTL was associated with a marker and adj. R^2 estimates were used to determine the magnitude of the QTL effect. Most of the marker trait associations were made based on a 215 genotypes subset that excluded the wild/std. panel for which phenotypic data was not available.

Table 3.3 Mixed models designed for association mapping in cotton using TASSEL software.

Code†	Model	Statistical equation
MT	Marker+Trait	$Y=A_{\text{ά}}+e$
MTS	Marker+Trait+Structure	$Y=A_{\text{ά}}+Q_{v}+e$
MTI	Marker+Trait+Instruct	$Y=A_{\text{ά}}+Q_{v}+e$
MTP	Marker+Trait+PCA	$Y=A_{\text{ά}}+Q_{v}+e$
MTK	Marker+Trait+Kinship	$Y=A_{\text{ά}}+Z_{u}+e$
MTSK	Marker+Trait++Structure+Kinship	$Y=X_{\text{ß}}+ A_{\text{ά}}+ Q_{v}+Z_{u}+e$
MTIK	Marker+Trait+Instruct+Kinship	$Y=X_{\text{ß}}+ A_{\text{ά}}+ Q_{v}+Z_{u}+e$
MTPK	Marker+Trait+PCA+Kinship	$Y=X_{\text{ß}}+ A_{\text{ά}}+ Q_{v}+Z_{u}+e$

† : Y = vector of phenotypic observations, ά= vector of allelic effects, e=vector of residual effects, ν=vector of population effects, ß=vectors of fixed effects other than allelic or population group effects, u=vector of polygenic background effects, Q=population membership assignment matrix, X, A and Z are incidence matrices of 1s and 0s relating y to ß, ά and u (Casa *et al*., 2008).

b) Mixed – Multiple Regression Models for Association Analysis

The GLMSELECT procedure in SAS performs effect selection in the framework of general linear models. A variety of model selection methods are available offering extensive capabilities for customizing and for using a wide variety of selection and stopping criteria. The GLMSELECT procedure compares most closely to PROC REG and PROC GLM. The PROC REG procedure supports a variety of model-selection methods but does not support a CLASS statement. The PROC GLM procedure supports a CLASS statement but does not include effect selection methods. The GLMSELECT procedure fills this gap. It focuses on the standard independently and identically distributed general linear model for univariate responses and offers great flexibility for, and insight into, the model selection algorithm.

In order to exploit the advantages of multiple regression procedures, all trait-wise significant markers selected by the mixed model procedures of TASSEL were screened for using fifty two GLMSLECT models. A stepwise selection method was used with all possible combinations of the CHOOSE, SELECT and STOP options. Options used included, Bayesian Information Content(BIC), SBC(Schwarz Bayesian Information Criterion), Adjusted R^2, AICC (Corrected Akaike Information Criterion), SL=0.15 (the significance level of the F statistic for entering or departing effects) and Cross validation (CV). Trait scores were considered as dependent variables and all markers were treated as the independent variables. Each trait was analyzed separately and those independent variables showing test statistic estimates of less than P= 0.05 were added into the model. To reduce Type I error, selected models were tested with a validation step by using the 'PRESS' criterion in the 'STOP' option. The best model was then selected based on adjusted R^2 and the least number of effects for a particular trait.

To estimate epistasis we calculated contrasts for two gene interactions in an additive x additive model in SAS (SAS, 9.1.3). The selected QTL's from mixed-multiple regression model were used to build an epistatic model. Those markers found significant at P<0.05 level were selected for each trait using PROC GLM.

3.3 Results and Discussion

3.3.1 Phenotypic Analysis

The cotton association mapping panel consisted of 232 genotypes representing five geographical regions of the USA along with standard panel (Table: 3.1a and 3.1b). The phenotypic data was collected from 215 genotypes (excluding standard panel) and statistically analyzed using PROC UNIVARIATE in SAS (Table: 3.4). The 215 genotypes had a mean lint yield of 95.78% in relation to the DP 393 check. The mean values observed for MIC (4.70), fiber length (1.14 inch), strength (31.10 g/tex), UI (84.34%), elongation (8.47) and SFC (4.78) were in accordance with national averages. The range and variances indicated a significant amount of phenotypic diversity present in the CAM panel. Similar variation was observed in a collection of 285 land race stocks from Africa and Mexico (Abdurakhmonov *et al*., 2008). Large genetic effects and relatively small genotype x environment (GXE) variance was observed in a 260 individual exotic mapping population (Zeng *et al*., 2009). The environmental factors influence in the development and modification of a trait.

Table 3.4 Phenotypic variation for yield and fiber traits in CAM panel

Parameters	LY†	LP	MIC	FL	FS	UI	ELO	SFC
N	215	215	215	215	215	215	215	215
Min	71.69	34.15	4.03	1.06	26.88	64.12	5.41	3.04
Max	147.85	45.93	5.91	1.26	36.74	87.19	11.39	8.01
Mean	95.78	40.11	4.70	1.14	31.10	84.34	8.47	4.78
SE	0.73	0.13	0.02	0.00	0.13	0.11	0.07	0.06
Variance	114.37	3.79	0.10	0.00	3.65	2.70	1.18	0.84
SD	10.69	1.95	0.32	0.03	1.91	1.64	1.08	0.92

†: LY = lint yield (standardized to DP 393); LP = Lint percentage; MIC = micronaire; FL = fiber length; FS = fiber strength; UI = uniformity index; ELO = elongation percentage; SFC = short fiber index.

The correlation studies among lint yield and fiber traits revealed significant trait relationships, for example, FS and UI, LY, MIC and FL; UI with MIC, FL (Table 3.5 and Fig: 3.1). The LY was positively correlated with MIC, FS, FL and ELO, while MIC was significantly negatively correlated with FL and ELO. Fiber length was positively correlated with FS, UI and negatively with ELO and SFC. A significant negative correlation was evident between SFC with other fiber traits.

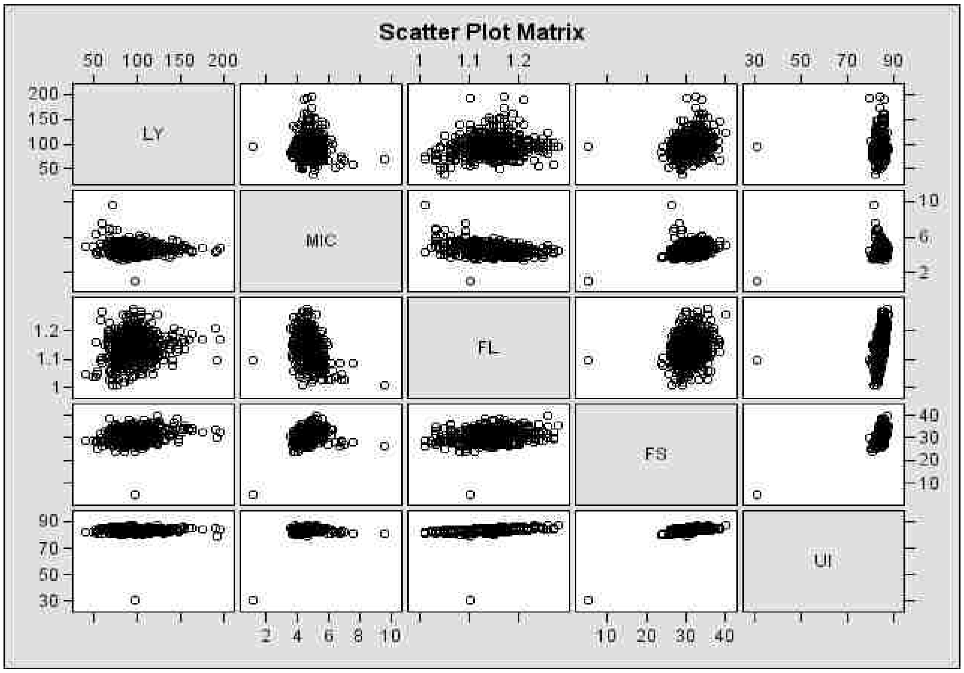


Fig 3.1 Pearson correlation matrix among lint yield and major fiber traits in cotton association mapping panel.

Table 3.5 Correlations between lint yield and major HVI fiber properties in upland cotton.

Traits	LY[†]	MIC	FL	FS	UI	ELO	SFC
LY	1						
MIC	0.0021	1					
FL	0.085	-0.374***	1				
FS	0.156***	0.341***	0.175***	1			
UI	0.0352	0.203***	0.31***	0.534***	1		
ELO	0.014	-0.0298	-0.122**	-0.0245	0.058	1	
SFC	-0.079	-0.115**	-0.238***	-0.436***	-0.326***	-0.443***	1

** P<0.01, *** P<0.001, † ; LY=Lint yield, MIC=Micronaire, FL=Fiber length, FS=Fiber strength, UI=Uniformity index, ELO=Elongation %, SFC=Short fiber content.

A similar positive correlation between FS and FL was also observed by earlier investigators using an exotic upland AM panel (Abdurakhmonov *et al.*, 2008). Some of the earlier studies also indicated negative correlations between MIC and FL, MIC and FS, FS and ELO in an upland and diploid association mapping panel (Zeng *et al.*, 2009; Kantartzi and Stewart, 2008).

3.3.2 Genetic Analysis with AFLP Markers

Based on geographical origin, the CAM panel was divided into six groups, viz., Louisiana, Arkansas, Southeast (SE), Texas, Delta and Wild/Std. panel. A total of 64 AFLP enzyme primer combinations were deployed to mine the cotton genome in order to estimate the extent of diversity present. Marker analysis of the panel resulted in 561 polymorphic loci. Heterozygosity was in the range of 0.33 to 0.39 among the groups (Table 3.6). Shannon-Weiner's Diversity Index (I), an index used in ecological studies to determine how diverse a population is, showed that diversity was moderate with values ranging from 0.485-0.574. The polymorphic Information Content measures the probability that two randomly chosen alleles from a population are distinguishable.

Table 3.6 Population genetic parameters for the inferred six clusters of cotton association mapping panel.

	N†	No. bands	Na	Ne	Mean He	I
LA	69	554	1.82	1.58	0.33	0.48
ARK	112	556	1.93	1.71	0.39	0.57
SE	16	558	1.84	1.7	0.38	0.54
Texas	14	562	1.84	1.68	0.37	0.53
Delta	12	560	1.81	1.68	0.37	0.52
W/SP	09	559	1.80	1.62	0.34	0.49

† ; N=No. of accessions, Na=No. of different alleles, Ne=No. of effective alleles, He=expected heterozygosity, I=Shannon's Information Index; LA= Louisiana, ARK=Arkansas, SE=South eastern, W/SP=wild or standard panel genotypes.

The PIC value for AFLP makers was in the range of 0.05-0.35 with an average of 0.254 (Fig 3.2). The frequency distribution of the PIC values demonstrated higher values for the range 0.15-0.35. The Nei's genetic diversity estimates revealed that all of the inferred groups were highly diverse with wild/Std. panel being the most diverse (0.118-0.197; Table 3.7) followed by the LA group with the SE and Delta groups.

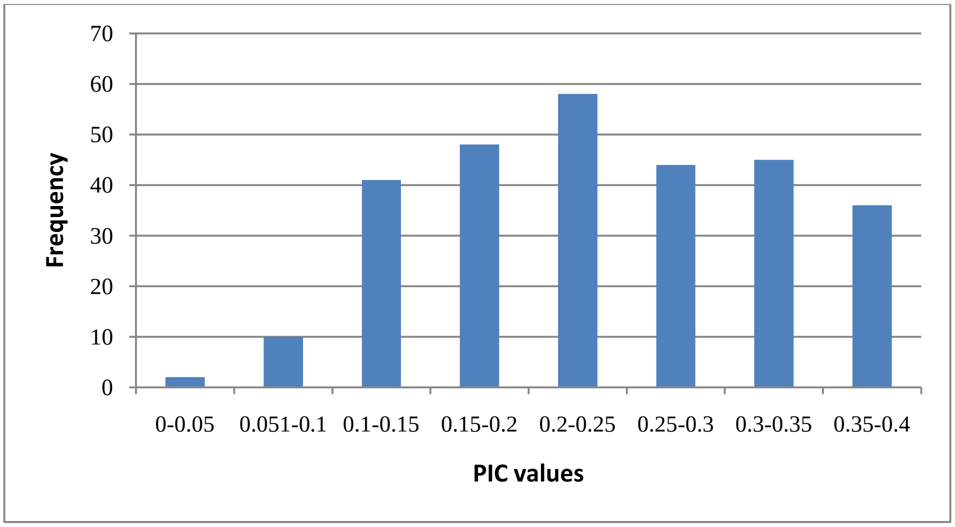


Fig 3.2 Frequency distribution for PIC values using AFLP markers in cotton association mapping panel. X axis: PIC values, Y axis: frequency estimates.

The allele frequency divergence among subpopulations as measured by nucleotide distances using Structure software (Pritchard *et al*., 2000a) revealed that allele frequency distances ranged from 0.148 (between SE and Std. panel) to 0.644 (between Arkansas and standard panel; Table 3.8). The Delta and std. panel seemed to be highly divergent from other groups.

Table 3.7 Nei's genetic diversity estimates for the inferred six clusters in cotton association mapping panel. X axis: PIC values, Y axis: frequency.

	LA†	**Ark**	**SE**	**T**	**Delta**	**Wild**
LA	0.000					
Ark	0.041	0.000				
SE	0.124	0.096	0.000			
T	0.133	0.100	0.060	0.000		
Delta	0.127	0.102	0.079	0.082	0.000	
Wild	0.197	0.148	0.132	0.118	0.125	0.000

†LA= Louisiana, ARK=Arkansas, SE=South eastern, T=Texas

A SSR based genetic diversity estimate of upland cultivars gave rise to 66 alleles with PIC values ranging from 0.18-0.62 (Candida *et al*., 2006); while other studies have reported PIC values of 0.08-0.89 (with an average of 0.55); (Lacape *et al*., 2007) and 0.05-0.82 (Liu *et al*., 2000). Previous association mapping studies reported a range of 0.007-0.380 PIC values in a 285 exotic upland panel and 0.006-0.50 for a panel of 334 Uzbekistan *G. hirsutum* accessions with average frequency of four SSR alleles per primer pair (Abdurakhmonov *et al*., 2008; 2009).

Table 3.8 Allele frequency divergence among inferred subpopulations in cotton association genotypes

	LA	ARK	SE	Texas	Delta	W/SP
LA	-					
ARK	0.595	-				
SE	0.498	0.587	-			
Texas	0.569	0.556	0.184	-		
Delta	0.576	0.555	0.282	0.305		
W/SP	0.564	0.644	0.148	0.239	0.324	-

SSR based allele frequency divergence estimates in an upland exotic panel resulted in values ranging from 0.11-0.27 (Zeng etl., 2009) and 0.00-0.66 in a diverse diploid panel (Kantartzi and Stewart, 2008). The average genetic distance within *G. hirsutum* accessions of specific ecotypes (Uzbekistan, Latin American and Australian) was very close and ranged from 0.12 to 0.14, while the highest GD among *G. hirsutum* varieties was observed within the Australian ecotype group (0.26) (Abdurakhmonov *et al*., 2009). These observations provide evidence for the existence of population substructure among cotton association mapping panels.

3.3.3 Analysis of Molecular Variance

The levels of genetic variation within and among the CAM groups identified by the cluster analysis were estimated from allelic frequencies using analysis of molecular variance, AMOVA (Weir and Cockreham 1984; Weir 1996). The within group genetic variation was 90percent while 10 percent of the variation was observed among the groups (Table 3.9). Wright’s (1965) F_{ST} (ø) statistic was used to evaluate the genetic differentiation between populations in the CAM panel (Table 3.10). The overall F_{ST} estimate was 0.0615.

Table 3.9 Analysis of Molecular Variance (AMOVA) among and within inferred groups

Source	df†	SS	MS	Estimated Variance	% variation
Among Populations	5	1600.123	320.025	8.076	10%
Within Populations	226	15932.795	70.499	70.499	90%
Total	231	17532.918		78.575	100%

† df=degrees of freedom, SS= sum of square, MS= mean sum of square

The pairwise F_{ST} values between the six groups indicated that genetic differentiation among clusters was highest between the LA and W/SP groups (0.158). Among groups, the Texas and SE had the lowest F_{ST} values (0.0001) indicating shared ancestry of these genotypes. The Texas genotypes under study seemed to support extensive utilization of putative ancestors from wild/Std. panel in their breeding program (lower F_{ST}=0.027). As 90% of the genetic variation was attributed to be within groups, highly significant variations were observed within predefined groups, the existence of population structure.

Table 3.10 Pairwise F_{ST} values between six inferred groups of cotton association genotypes.

	LA†	ARK	SE	Texas	Delta	W/SP
LA	-					
ARK	0.048	-				
SE	0.11	0.066	-			
Texas	0.111	0.063	0.0001	-		
Delta	0.101	0.0587	0.0074	0.007	-	
W/SP	0.158	0.0958	0.0382	0.027	0.0284	-

† LA= Louisiana, ARK=Arkansas, SE=South eastern, W/SP=Wild or Standard panel genotypes. Values were calculated as per the Wright (1965).

Prior results from a locus-by-locus AMOVA, employing only polymorphic AFLP markers among *G. tomentosum* and *G. hirsutum* accessions, demonstrated that there was little inter-population differentiation with only 13.2% of the variation occurred among populations and 86.8% of the variation residing within populations (Hawkins *et al.*, 2005). The within group component of genetic variance prevailed in an upland exotic association panel and accounted for 96.73% of the total variance. The 3.27% of the genetic variance observed among groups was significant with overall F_{ST} value of 0.032 (Abdurakhmonov *et al.*, 2008). A distribution of molecular genetic variation among (26.9%) and within (76.4%) six clusters of diploid accessions was reported by Kantartzi and Stewart (2008). In this study, the greatest proportion of genetic variance of cotton germplasm groups was attributed to within population groups, however the small variation observed among predefined groups was highly significant, suggesting the existence of population structure.

3.3.4 Kinship Estimates

Complex structures and familial relationships are common in inbred cultivated crops. In such crops, allele frequencies evolve between divergent structured populations via drift, mutation and selection. Differences in allele frequencies may be correlated with any morphological traits that differentiate two populations. A statistical correlation between a gene and a trait is not necessarily associated with causative relationship between a trait and gene, which can lead to false positives. The use of population structure and a matrix of kinship coefficients prove efficient in association studies (Yu *et al*., 2006). In the CAM panel, the pairwise kinship values varied from 0-0.69. Although 47% of the pairwise kinship estimates were close to zero, a significant percentage around 0.25 and 0.35 represented the relationships within families (Fig 3.3). About 16% of kinship pairs had a value of 0.25 and 22% had 0.35-0.49.

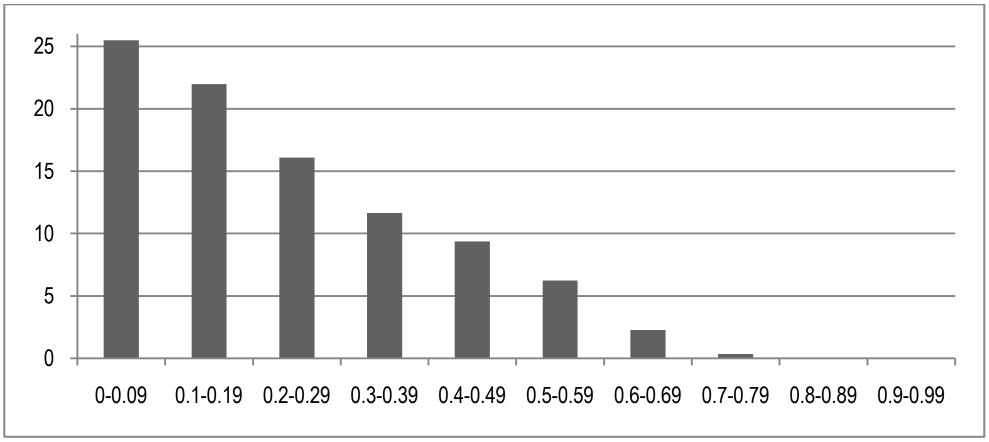


Fig 3.3 Relative frequency for kinship values estimated using allele frequency data in Cotton association mapping panel. X axis: range values for relative kinship estimates, Y axis: frequency values.

This indicates use of common ancestral genotypes in the history of most of the breeding programs due to their premium trait values. Abdurakhmonov *et al.*, (2008) observed that the majority of the pairs of cotton accessions (55%) had zero estimated kinship values, while the remaining pairs had a value of 0.05-0.25, suggesting involvement of some common parental genotypes in these germplasm groups. Kinship estimates can be used in mixed linear models, where in family structure is ignored. The inclusion of kinship improved model fit, as well as reducing the false positives and increasing the power to detect QTL.

3.3.5 Population Structure

Based on the neighbor joining analysis (NJ), the genetic distances among all the mapping genotypes is represented as a tree (Fig 3.4). The NJ tree consists of six clusters with a random spread of genotypes from the predefined CAM groups. The six broad clusters can be identified with LA and Arkansas genotypes spread out randomly. There is no distinct pattern observed and it is difficult to conclude the assignment of genotypes to their respective groups based on NJ analysis.

Correspondence analysis confirmed the population structure (Fig: 3.5). Genotypes representing LA and Arkansas regions grouped into a cluster on the left side, while most of the SE, Delta and Texas genotypes congregated in the center. The W/std. panel, owing to their high diversity, formed a small cluster in the top right side. From this analysis, a split could be proposed based upon geographical arguments and also based on molecular diversity. In addition, it also strongly supports the involvement of subgenome donors and Std. panel entries in the breeding programs of the Delta, Texas and SE regions. Thus it is concluded from this study that more ancestral sharing of alleles between LA and Arkansas and to a lesser extent between Arkansas and SE genotypes has occurred.

In order to gain additional insight into the genetic diversity of the CAM panel, Principal Coordinate analysis (PCoA) was performed using data from the genetic similarity matrix (Nei and

Li, 1979). Here, genetic relationships were most easily seen by plotting first three PCoA which explained 68% (35.51+18.56+14.25%) of the genetic variation (Fig 3.6). Three separate clusters were observed (LA, Ark, Std. panel) and are delineated based on their geographical origin.

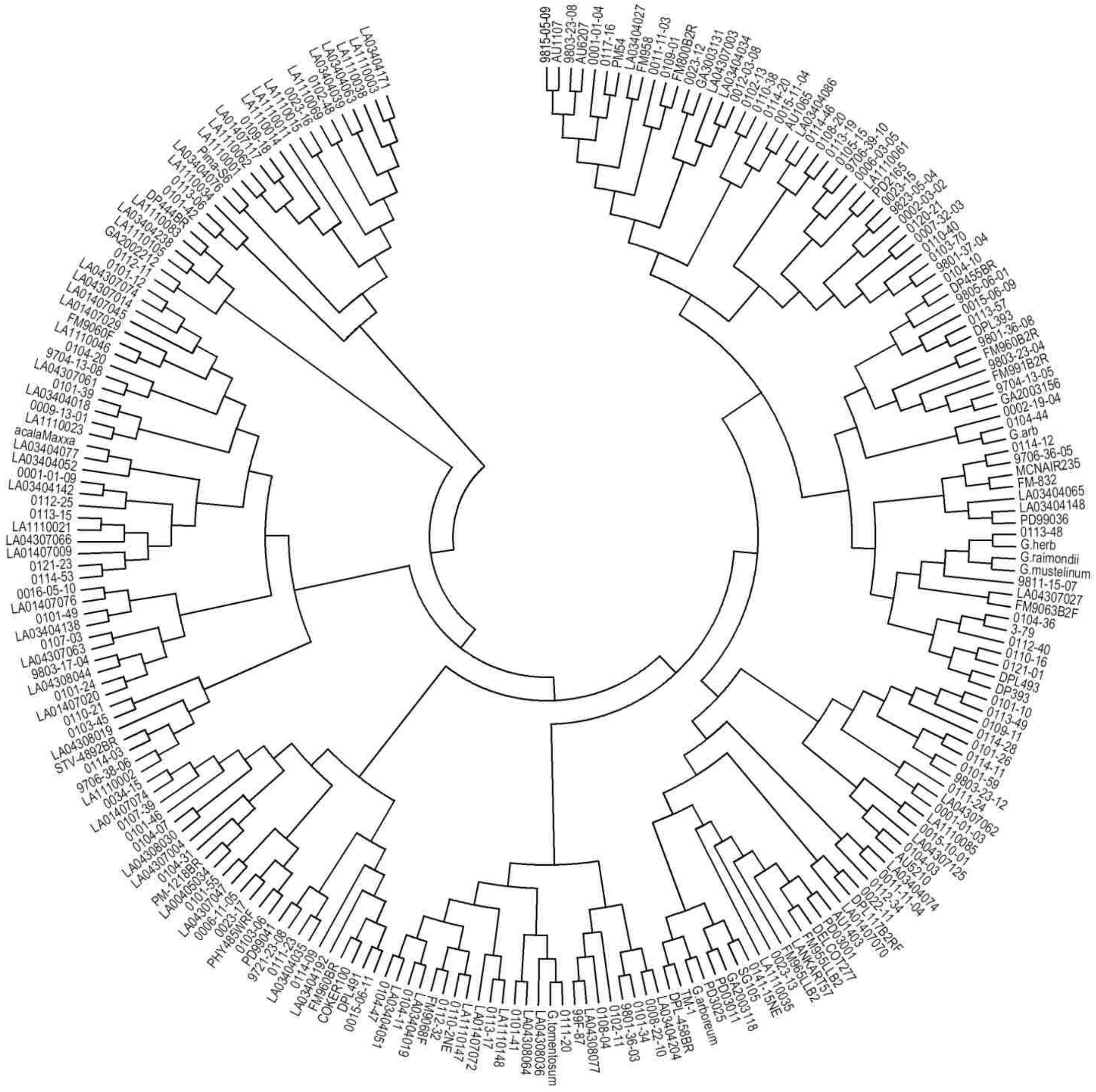
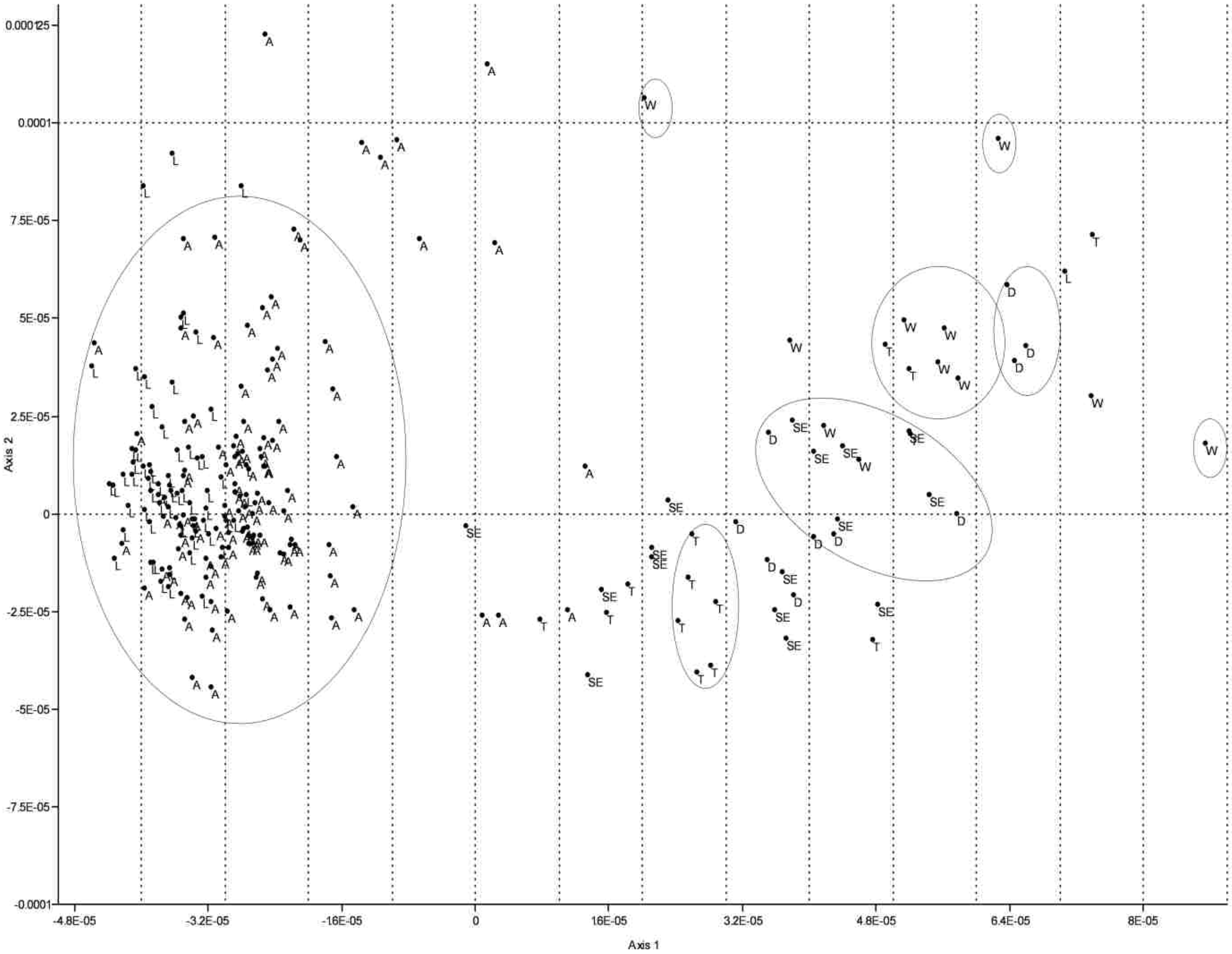


Fig 3.4 Neighbor-joining cluster analysis based on the pairwise Dice coefficient of association showing the genetic relationships among CAM panel. The DICE similarity coefficients were calculated in NTSYS software and tree diagram was constructed using MEGA software.

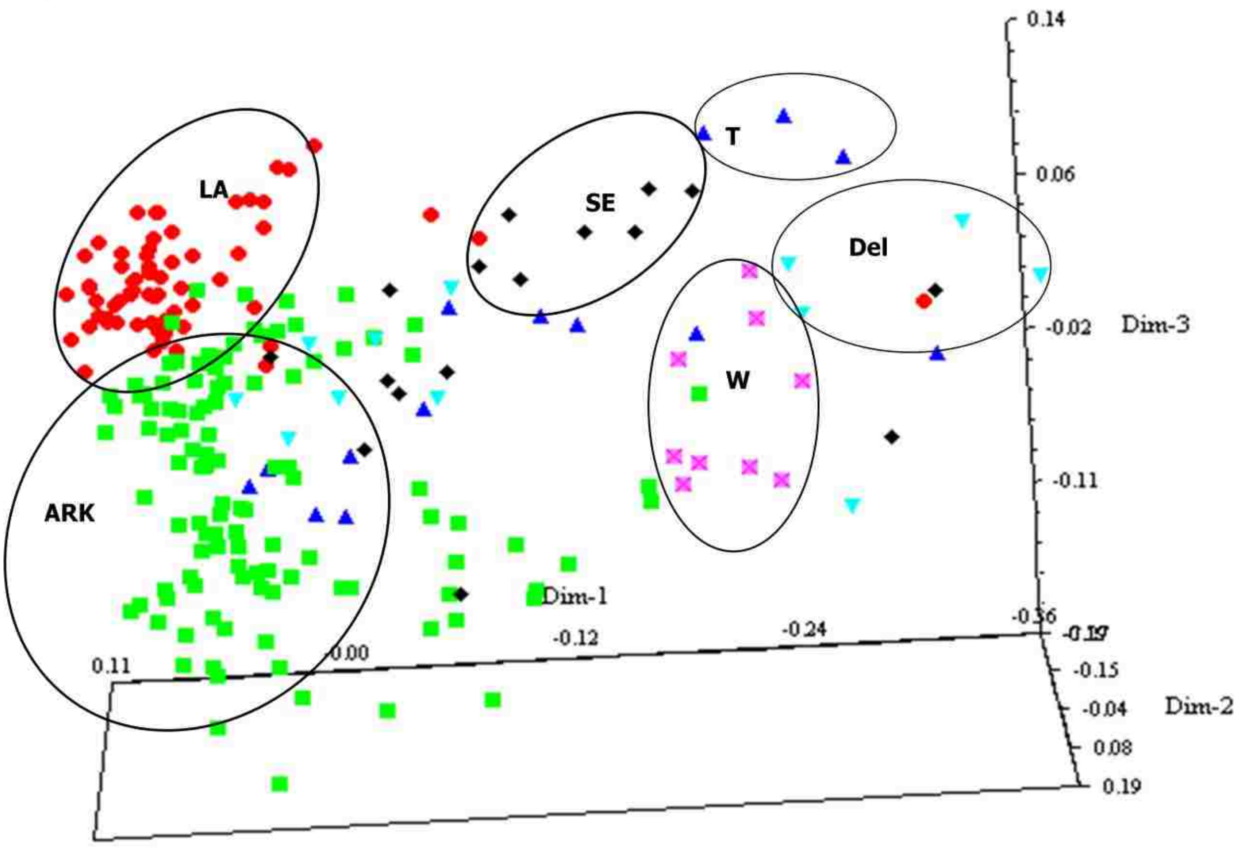
The LA and Arkansas groups seemed more genetically related, while std. panel was highly diverse, owing to the presence of wild or subgenomes contributors. Delta and SE genotypes are interspersed with each other with no definite pattern.



L=Louisiana; A=Arkansas, T=Teaxs, SE=South Eastern; D=Delta and W=Wild or Standard panel genotypes.

Fig 3.5 Correspondence analysis based on AFLP marker matrix. The marker matrix estimated across 232 cotton association genotypes. X and Y axis: coordinate 1 and 2 respectively.

In summary, most of the genotypes under study were fairly well grouped through correspondence and PCoA analysis, with few outliers. The information provided by the similar diversity analyses could help the breeders to plan their breeding programs.



LA=Louisiana, ARK=Arkansas, SE=South eastern, T=Texas, Del=Delta, W=Wild/standard panel

Fig 3.6 PCoA of cotton association mapping genotypes using AFLP marker matrix. The PCoA explained 68% of the genetic variation on three dimensional scales.

A Bayesian model based clustering method was used to infer population structure and assign individuals to discrete population based on AFLP markers. Multiple runs of Structure (ver. 2.2) were performed by setting k from 2 to 10. The posterior probability of the data (LnP(D)) showed an increasing trend, and from k=6 onwards, started getting constant(Fig: 3.7a). Due to the increasing trend even after the divergence at k=6 (which should otherwise plateau), the alpha (Dirichelt) parameter for the degree of admixture was estimated and it remained constant from k=6 onwards. A bar plot diagram showed that the splitting of Arkansas, Delta and Std. panel was not as expected (Fig: 3.9 left). The Arkansas group was the largest subgroup with 112 genotypes and showed two distinct sub groups with Structure, which is hard to explain. The primary composition of each of the ancestral blocks across 232 genotypes under study cannot be explained consistently using a Bayesian model.

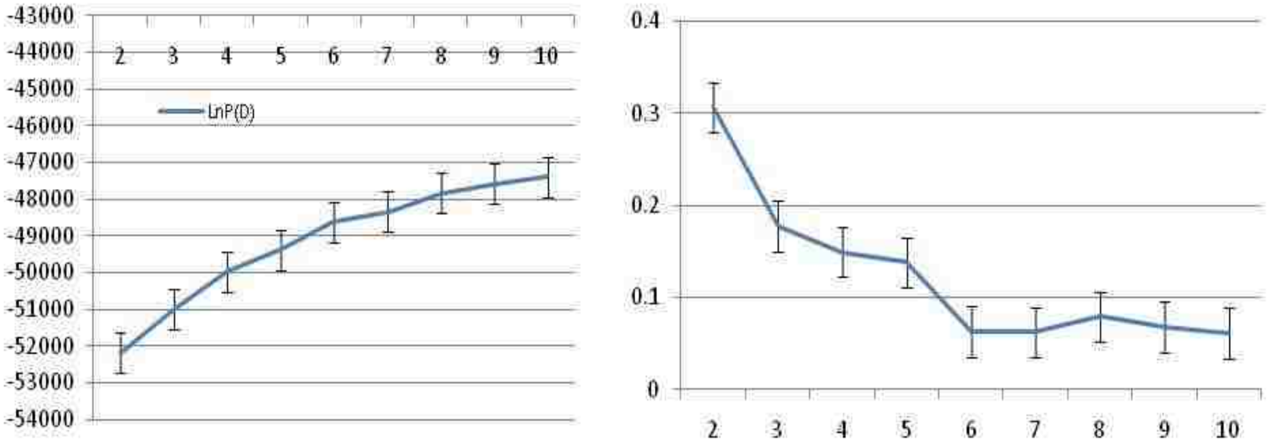


Fig 3.7 a) Posterior probabilities, LnP(D) as function of k, where k=2-10; b) Alpha (Dirichlet) values as function of k, where k=2-10. The LnP(D) and Alpha values are used to decide the ideal number of subpopulation existing. X axis: number of subpopulations assumed, Y axis: LnP(D) and Alpha values, respectively.

In a separate study upland cotton accessions were assigned to distinct clusters based on their geographical origin, viz., Uzbekistan, Australian and Latin America using Structure (Abdurakhmonov *et al.*, 2009). Similarly, analysis of genetic distance and population structure provided evidence of significant population structure amongst *G. arboreum* accessions and identified the highest likelihood at K = 6 (Kantartzi and Stewart, 2008).

To get more insight into CAM population structure, we used the MCMC algorithm for the generalized Bayesian clustering with Instruct software (Gao *et al.*, 2007). Cotton is basically a self pollinated crop with moderate chance of cross pollination (10-30%). Instruct revealed that posterior probabilities started increasing and become constant after k=6. The Deviance Information Criterion (DIC) also started stabilizing at k=6 (Fig: 3.8). The Gelman-Rubin convergence statistic was 0.999 at k=6 and supported model convergence. Visual comparison of Structure and Instruct bar plots revealed numerous differences with respect to grouping of genotypes in to subpopulations (Fig 3.9).

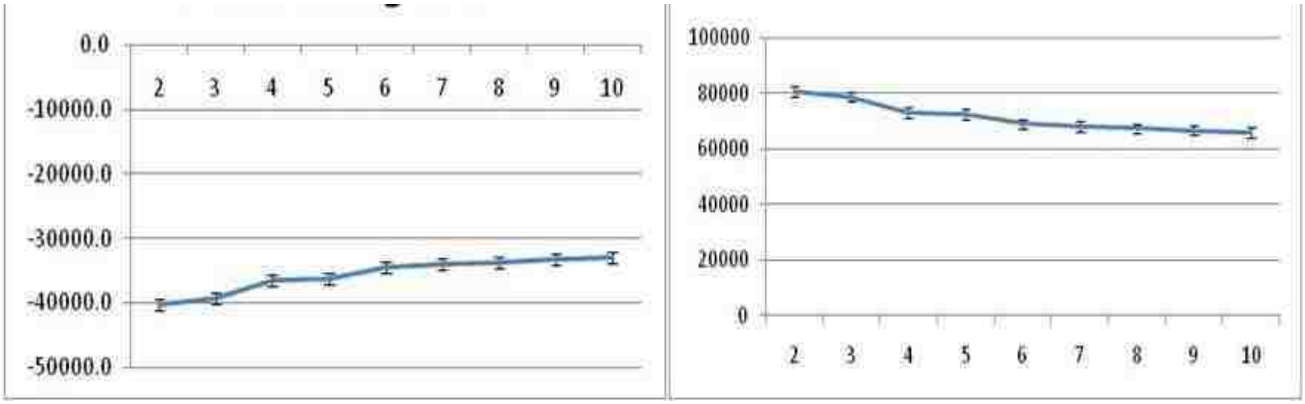


Fig 3.8 Posterior likelihood and deviance information content (DIC) statistics for CAM panel with k=2 to10 estimated using Instruct. X axis : Posterior log likelihood and DIC estimates respectively, Y axis: number of subpopulations assumed.

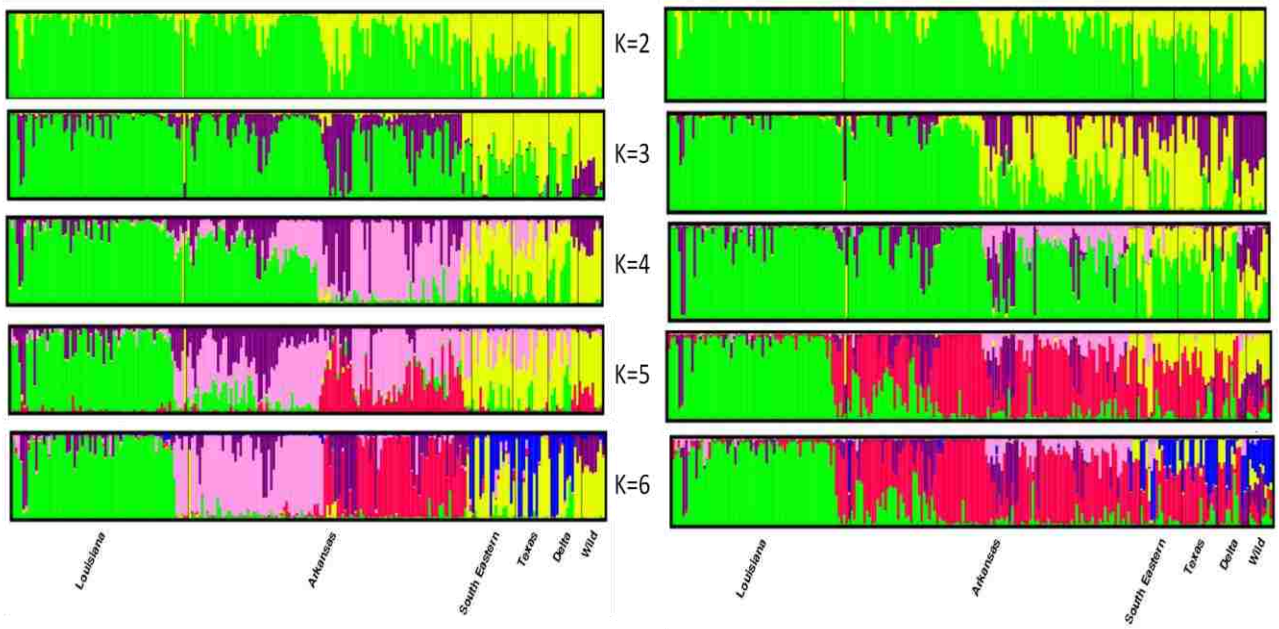


Fig 3.9 Bar plot of inferred population structure using Structure and InStruct softwares in CAM panel, with k = 2-6. Each individual is represented by a line partitioned in six colored segments that represent the individual's estimated membership fractions to each one of the six clusters.

Instruct seemed to more logically assign LA and Arkansas genotypes into distinct clusters although these clusters showed evidence of admixture. The number of Texas and Delta genotypes in the CAM panel was small in size, yet the MCMC algorithm fairly distinguished them and indicated that there was a considerable ancestral genomic exchange taken place during their development. The Louisiana genotypes were fairly intact with less admixture, while few had unexpected introgression from Arkansas. Overall, based on the biological significance and geographical adaptation, Instruct assigned CAM panel in to six sub clusters. This is also consistently supported by correspondence and PCoA analysis. Henceforth we considered six subpopulations as existing in the CAM panel for association analysis.

3.3.6 Association Analysis

a) Mixed Models Using TASSEL

Population structure and kinship among individuals does not only affect the amount and nature of diversity in a large inbred line collection, but can also lead to spurious associations (Gaut and Long, 2003). In this study, we tested the performance of eight models in minimizing type I error. We initially evaluated the naïve model (marker+trait) and then added population structure (either structure/Instruct) and eigenvectors of PCoA. These models were analyzed using the GLM procedures in TASSEL for all the eight traits under study (Fig 3.10).

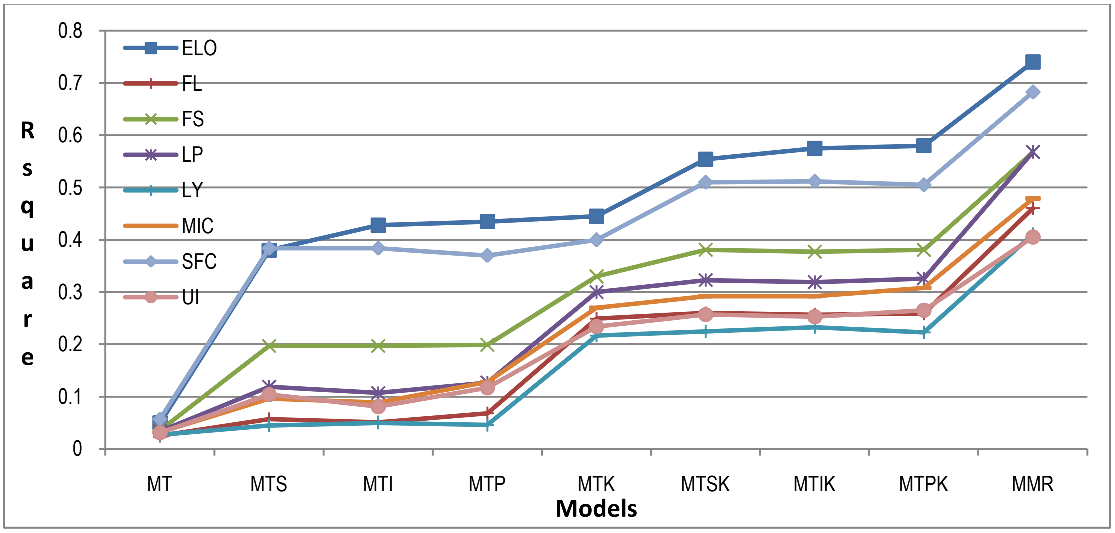


Fig 3.10 Genetic variations explained (adj. R^2) by different mixed and mixed-multiple regression models across yield and fiber traits. The mixed models were performed using TASSEL, while mixed multiple regression models using SAS. X axis: mixed and mixed-multiple regression models, Y axis: Adj. R^2.

M=marker; T= trait; I= Instruct; P=eigen values of PCA; K=kinship; S= structure; MMR=mixed multiple regression; ELO=Elongation percentage; FL=fiber length; FS=fiber strength; LP=lint percentage; LY=lint yield; MIC=micronaire; SFC=short fiber content; UI=uniformity index.

Utilizing 561 AFLP markers, the naïve model explained a negligible amount of the genetic variation with model R^2 ranging between 2.7-5.7% for the traits. Inclusion of Structure/PCoA resulted in an improved R^2 up to 43.5%. Mixed Linear Models (MLM), which consists of kinship, k along with population structure, or PCoA, were considered with k=6.

The MTIK model identified several markers associated with traits based on the cut off P value, 0.05 (Table 3.11). Fiber elongation had highest number of associated QTL (50) with R^2 value of 57.5%. The traits FL (24 QTL), LY (25 QTL) and MIC (29 QTL) had registered low model R^2 values of 25.66, 23.3 and 29.2%, respectively. Lint percent, being complex trait being influenced by many independent fiber traits, had 42 QTL's with an 31.9% of phenotypic variation being explained.

Thus MLM models incorporating information from Instruct or PCoA explained high degree of genetic variation; Instruct (57.5%) and PCoA (58%). Incorporating information about population structure from Structure software did not improve model efficiency. For most of the traits studied, the MTIK mixed model resulted in a high model R^2, except for the LY and MIC, which are highly influenced by environmental factors. Based on earlier results, where in Instruct assigned the CAM genotypes fairly well into six subpopulation, the MTIK method was selected as best among all models for association analysis. In MLM, the MTIK model was able to fit up to 60% for LY and SFC and between 53-57% for FL, FS, LP and MIC. Using the multiple QTLs graphs showing observed v/s predicted scores for fiber traits are given in Fig: 3.11. One of the initial association studies in cotton reported SSR marker associations using a small 56 accessions panel of diploid cottons. A total of 30 marker–trait associations were identified with 19 SSR markers located on 11 chromosomes (Kantartzi and Stewart 2008). Around 17 SSRs were associated with fiber quality traits such as, MIC, 23 with FL, 18 with UI, 19 with STR and 11 with ELO traits in the association mapping study of Abdurakhmonov *et al*., (2008).

Table 3.11 Quantitative trait alleles identified by the MTIK mixed model using TASSEL. Based on the high adj.R^2 and significant P value, the QTAs were identified for each trait in cotton association mapping panel.

Trait	Significant QTAs selected, given P<0.05
ELO	E6M4_297, E6M8_325, E6M3_520, E6M1_382, E4M1_348, E6M2_640, E6M2_375, E3M8_175, E3M8_305, E5M2_75, E5M2_110, E5M3_148, E3M4_50, E3M4_70, E5M4_450, E3M5_104, E3M5_250, E5M1_204, E5M1_395, E8M7_140, E9M8_370, E8M8_60, E8M8_330, E8M8_430, E9M4_280, ,E9M4_460, E8M4_385, E8M3_165, E8M6_55, ,E8M2_45, E9M5_50, E9M5_230, E8M5_70, E8M5_225, E9M3_160, E9M1_202, E8M1_195, E8M1_130, E2M8_315, E1M4_55, E6M7_55, E1M7_335, E6M5_310, E6M5_80,E2M5_295, E1M2_50, E4M6_60, E4M6_50, E1M6_130, E2M6_225
FL	E4M4_229, E4M4_177, E6M3_363, E6M2_375, E4M2_135, E3M4_364, E3M4_250, E5M1_204, E3M7_370,E8M8_605, E7M6_140,E8M2_75,E8M2_270, E2M3_60, E6M7_105, E1M8_97, E6M5_145,E1M5_200,E1M2_45, E1M2_210, E6M6_140, E4M6_175, E1M6_190, E1M6_270
FS	E6M8_362, E6M3_342,E6M1_218E4M1_382, E4M1_357, E4M1_348, E6M2_255, E4M2_206, E3M6_300, E5M3_345, E5M5_75, E5M7_325, E3M3_60, E9M7_350, E9M7_370, E8M3_200, E8M2_157, E9M1_58, E1M3_55, E1M3_150, E1M3_175, E4M7_180, E6M7_100, E6M7_125, E6M5_160, E1M5_60, E1M5_225, E2M5_295, E1M2_65, E4M5_70
LP	E6M4_249, E4M4_100, E4M3_219, E4M3_214, E4M3_220, E4M3_222, E4M1_348, E5M2_75, E3M6_70,E3M6_95, E3M6_300, E5M3_110, E3M4_70, E3M4_364, E3M5_355, E5M1_204, E5M7_70, E5M7_325,E3M3_90, E8M7_175, E8M7_295, E8M7_75 E8M4_420, E8M4_385, E7M6_140, E9M2_100, E9M1_140, E8M1_120, E6M7_180, E2M2_460,E1M8_85, E1M7_55, E1M7_112, E2M7_210, E2M5_295, E1M2_55, E1M2_215, E4M6_65, E4M6_220, E4M6_240, E1M6_210, E2M6_225
LY	E4M4_280, E4M1_357, E4M2_206, E5M8_260, E3M2_145, E3M6_70, E3M6_300, E3M4_60, E3M4_364, E5M4_450, E5M7_325, E8M7_75, E8M8_245, E8M4_385, E8M3_50, E9M1_54, E4M7_45, E4M7_195, E2M2_200, E4M8_320, E6M5_145,E6M5_150, E1M5_225, E6M6_75, E2M1_65
MIC	E6M4_325, E6M1_320, E6M1_196, E4M1_348, E4M2_265, E5M2_75, E5M6_145, E5M3_100, E5M4_225, E5M5_45, E5M5_415, E5M1_55, E5M7_325, E5M7_375, E9M8_330, E8M8_265, E9M4_460, E8M3_50, E8M3_255, E7M6_140, E9M2_60, E9M1_54, E9M1_140, E8M1_120, E1M4_200, E2M4_135, E1M2_210, E4M5_55, E2M6_180
SFC	E6M4_297, E6M4_270, E6M4_249, E4M4_280, E6M3_288, E6M1_218, E5M8_175, E5M8_260, E5M2_45, E5M6_40, E5M3_152, E3M5_98, E5M1_204, E5M1_395 , E5M7_325, E8M7_185, E8M7_280, E8M8_60, E7M6_65, E8M2_159, E8M2_185, E8M5_315, E8M1_130, E8M1_140, E2M8_155, E2M8_260, E2M2_280, E2M2_218, E1M5_225, E2M5_50, E2M5_190, E1M2_213, E2M6_185
UI	E4M4_217, E4M3_473, E4M1_348, , E6M2_364, E3M2_204, E5M2_75, E5M3_115, E5M3_175, E3M7_370, E3M3_195, E8M3_225, E9M2_208, E8M2_155, E8M2_159, E9M5_53, E8M5_315, E9M1_52, E1M3_220, E6M6_75, E6M6_140 ,E4M6_175, E1M6_60

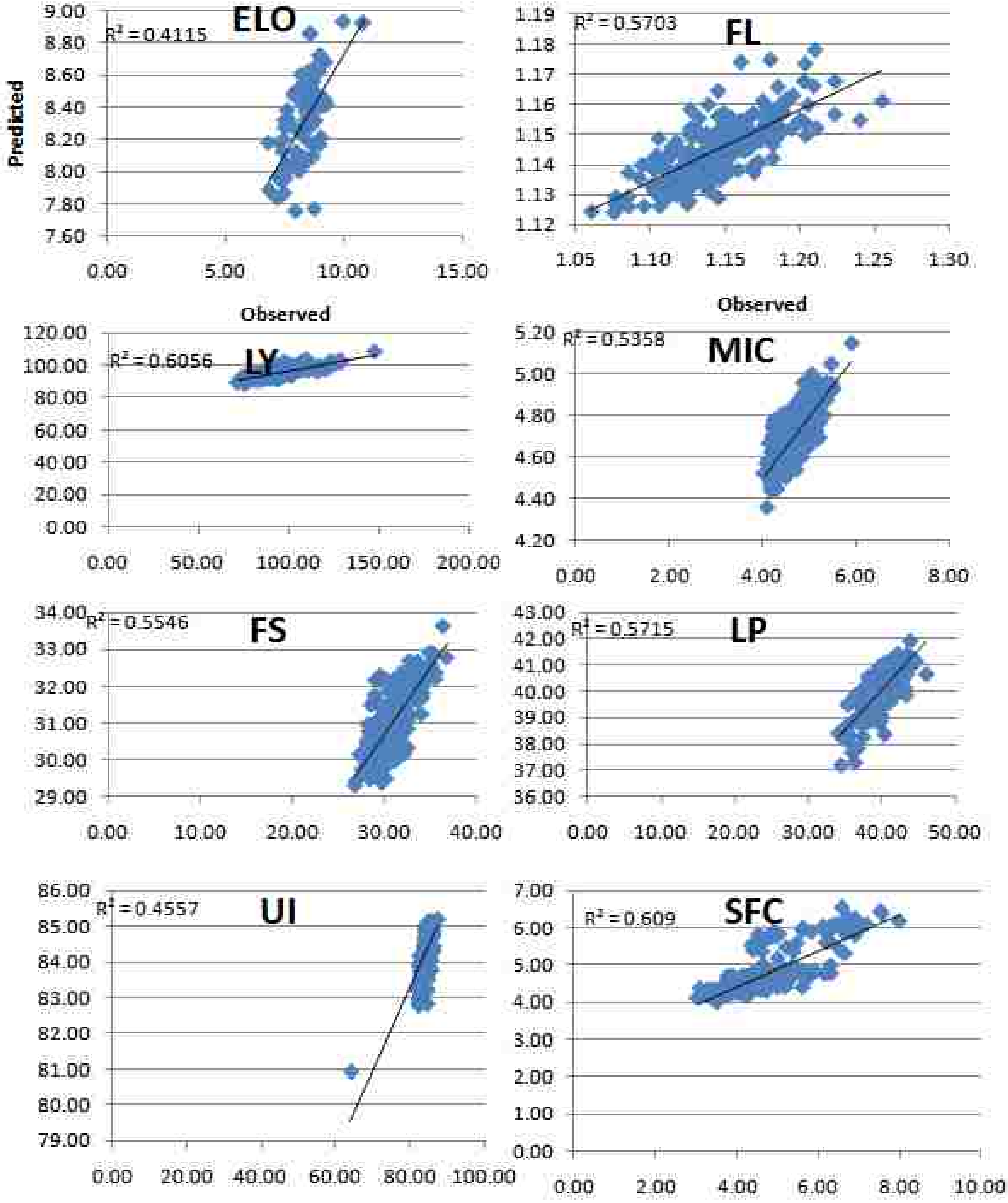


Fig 3.11 The observed v/s predicted scores of lint yield and fiber traits in CAM panel. Predicated values were based on the polymorphic AFLP-TRAP-SSR markers from mixed model analysis using TASSEL software.

In a mapping panel of 334 upland accessions, Mixed linear model (MLM), General linear model (GLM), and Structure analysis (SA) as implemented in TASSEL, identified 12-28 SSR's significantly associated with fiber traits from Uzbek and Mexican environments (Abdurakhmonov *et al*., 2009). Similar to the present study, Zeng *et al*., (2008) also noticed the power of the MLM method by inclusion of population structure and kinship data. As many as 12 of the 23 marker trait associations for yield components in a 260 mapping panel survived stringent correction and remained significant.

The success of association mapping in a polyploid species like cotton is getting improved by reducing the Type I errors, thereby setting up stringent threshold values for significance. The present study explored all possible mixed models in achieving true associations and reducing the false positives to identify QTL associated with lint yield and fiber quality traits in a diverse panel of genotypes with 6 distinct subpopulations.

b) Mixed-Multiple Regression Model

GLMSELECT, a general linear method for selecting models based on various statistical parameters is a new procedure, implemented in SAS. In order to consider all the markers simultaneously and perform stepwise multiple regression, 52 models were designed with different selection criteria and options. The significant QTL's identified by the MTIK mixed model for each trait were considered for validation using multiple regression. The mixed-multiple regression (MMR) model proved extremely powerful in improving the efficiency of the model by capturing 40.55-74% of the genetic variation for most of the traits under study (Table 3.12a- 3.12h). Among the 52 MLM-MMR models under study, the highest adj R^2 with minimum effective QTL's was selected from a model with the following options: CHOOSE=Adj.R^2, SELECT=AdjR^2 and STOP=Adj.R^2. All other models produced low R^2 values with high number of QTL's, which was seen as unreliable. These various statistical parameters were included to make the model more stringent, thus reducing the false positives efficiently.

Table 3.12a Significant QTL's selected from MLM-MMR based models for fiber elongation

QTL'S	Model R2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E6M4_297	0.179	0.175	-4.289	0.990	-5.410	2.452	<.0001
E6M1_382	0.375	0.366	-59.032	0.736	-61.278	-45.549	<.0001
E9M5_50	0.431	0.420	-77.213	0.652	-79.798	-60.359	<.0001
E9M3_160	0.483	0.471	-95.742	0.567	-98.387	-75.518	<.0001
E2M8_315	0.558	0.543	-125.392	0.430	-127.641	-98.427	<.0001
E1M7_335	0.524	0.510	-111.350	0.495	-113.893	-87.755	<.0001
E1M6_130	0.311	0.305	-40.053	0.824	-41.741	-29.941	<.0001
E6M3_520	0.614	0.595	-148.644	0.325	-149.968	-111.567	0.0028
E6M2_640	0.655	0.632	-166.441	0.246	-165.831	-119.252	0.0054
E3M5_104	0.682	0.656	-178.254	0.196	-175.147	-120.953	0.0265
E5M1_204	0.578	0.562	-133.458	0.394	-135.547	-103.122	0.0019
E8M7_140	0.665	0.642	-171.087	0.226	-169.677	-120.527	0.0130
E8M8_330	0.689	0.662	-181.012	0.186	-177.012	-120.340	0.0370
E9M5_230	0.696	0.668	-184.128	0.173	-179.058	-120.086	0.0310
E8M5_70	0.629	0.609	-155.309	0.295	-156.073	-114.861	0.0043
E9M1_202	0.674	0.649	-174.894	0.210	-172.665	-120.964	0.0206
E2M5_295	0.641	0.620	-160.138	0.274	-160.376	-116.320	0.0114
E4M6_60	0.703	0.674	-187.152	0.161	-180.907	-119.739	0.0330
E2M6_225	0.597	0.579	-141.225	0.359	-143.005	-107.519	0.0023

Fiber elongation is a property of fiber that is measured during the determination of bundle strength (Hertel, 1953). Increased fiber elongation is associated with improved yarn quality. The variability for fiber elongation values was from 5.41-11.39, with variance of 1.18 in the association panel. The MLM-MMR identified 19 significant markers out of the 50 from the mixed model alone (Table: 3.12a). Among all the markers selected, E2M8_315 and E1M7_335 proved to be significantly associated, explaining 54 and 51% phenotypic variation respectively. This is also supported by the low AICC, BIC, SBC statistics and highly significant P value.

Lint yield and lint percentage, are complex quantitative traits. In the present study, as many as 12 and 17 markers were associated with LY and LP respectively (Table 3.12b & c). The MLM-MMR identified E3M6_300 and E5M7_325 as common markers for both of these traits.

Table 3.12b Significant QTL's selected from MLM-MMR based models for Lint yield

QTL'S	Model R2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E5M7_325	0.086	0.082	1002.607	5.673	1003.354	1009.348	<.0001
E3M6_70	0.262	0.237	968.619	5.519	968.957	995.584	0.0160
E3M6_300	0.351	0.312	951.161	5.443	953.672	994.979	0.0329
E3M4_60	0.191	0.176	982.427	5.581	982.545	999.280	0.0072
E3M4_364	0.336	0.300	954.015	5.455	955.942	994.463	0.0203
E5M4_450	0.318	0.284	957.730	5.471	959.085	994.807	0.0200
E8M8_245	0.217	0.198	977.399	5.558	977.500	997.623	0.0090
E8M4_385	0.163	0.151	987.839	5.605	988.078	1001.322	0.0040
E4M7_45	0.300	0.269	961.445	5.487	962.346	995.152	0.0253
E4M7_195	0.129	0.121	994.298	5.635	994.744	1004.410	0.0014
E4M8_320	0.241	0.219	972.667	5.537	972.850	996.261	0.0108
E6M5_150	0.282	0.254	964.708	5.501	965.300	995.044	0.0175

Table 3.12c Significant QTL's selected from MLM-MMR based models for Lint percentage

QTL'S	Model R^2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E8M1_120	0.096	0.091	267.557	2.254	267.490	274.299	<.0001
E6M4_249	0.372	0.345	205.126	1.969	203.598	238.832	0.004
E5M2_75	0.254	0.236	234.255	2.101	232.416	254.479	0.002
E3M6_70	0.346	0.320	211.931	2.000	210.128	242.267	0.002
E3M6_95	0.481	0.442	176.217	1.843	178.212	230.147	0.017
E3M6_300	0.409	0.377	196.183	1.930	195.223	236.631	0.015
E5M3_110	0.502	0.459	171.350	1.824	174.857	232.022	0.041
E3M5_355	0.142	0.134	258.138	2.211	257.458	268.250	0.001
E5M1_204	0.491	0.450	173.937	1.835	176.614	231.238	0.047
E5M7_325	0.217	0.202	242.674	2.140	241.073	259.527	0.004
E9M2_100	0.315	0.292	219.876	2.036	217.879	246.841	0.002
E9M1_140	0.185	0.173	249.183	2.170	248.014	262.665	0.001
E1M7_55	0.391	0.361	200.514	1.949	199.224	237.591	0.012
E2M7_210	0.281	0.261	228.113	2.074	226.072	251.707	0.005
E4M6_65	0.426	0.391	192.021	1.912	191.491	235.839	0.016
E4M6_240	0.466	0.429	180.340	1.861	181.492	230.899	0.002
E2M6_225	0.441	0.405	188.258	1.896	188.231	235.447	0.020

Fiber strength is one of the most important fiber properties other than length contributing to cotton's use as a textile fiber. It translates directly into yarn strength and is related to spinnability. For fiber strength, MLM-MMR identified 17 markers significantly associated with high adj. R^2 ranging from 16.2-50.4%. Markers E4M1_382 and E4M2_206 registered low AIC, BIC and SBC values (Table 3.12d). Fiber length was associated with 17 markers compared to 24 by MLM alone (Table 3.12e). Significant markers were E4M2_135, E7M6_140, E1M2_45 and E1M6_270 all with high adj. R^2 values of 32, 38, 37 and 35.9%, respectively, and low AIC, BIC and SBC values.

Table 3.12d Significant QTL's selected from MLM-MMR based models for fiber strength

QTL'S	Model R^2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E4M2_206	0.238	0.231	224.995	2.057	224.642	235.107	<.0001
E4M1_382	0.166	0.162	242.448	2.138	242.487	249.189	<.0001
E6M3_342	0.278	0.268	215.432	2.013	214.711	228.914	0.0008
E6M1_218	0.499	0.466	157.072	1.751	158.675	204.261	0.0381
E4M1_348	0.349	0.333	197.312	1.930	196.303	217.536	0.0028
E6M2_255	0.488	0.457	159.684	1.762	160.770	203.502	0.0105
E3M6_260	0.423	0.400	177.397	1.839	176.767	207.733	0.0078
E5M3_345	0.532	0.494	148.187	1.715	152.103	205.488	0.0313
E9M7_370	0.471	0.442	164.677	1.784	165.086	205.125	0.0184
E8M3_200	0.441	0.417	172.316	1.817	171.987	206.023	0.0095
E4M7_180	0.543	0.504	145.087	1.702	150.022	205.758	0.0308
E6M7_100	0.511	0.477	153.713	1.737	156.035	204.272	0.0258
E6M7_125	0.521	0.485	151.233	1.727	154.256	205.163	0.0420
E6M5_160	0.456	0.429	168.582	1.801	168.558	205.659	0.0198
E1M5_60	0.320	0.307	204.557	1.963	203.678	221.410	0.0004
E1M5_225	0.382	0.364	188.025	1.887	187.143	211.620	0.0010
E2M5_295	0.402	0.382	182.807	1.864	181.981	209.772	0.0085

Fiber fineness or micronaire determines the spin limit and contributes to yarn strength and spinnability. Increased levels of fineness promote fiber to twist. The CAM panel had micronaire values in the range of 4.03-5.91. Low MIC can result from two major factors, immature fiber or genetically fine fiber. Maturity and fineness account for 90% of the variation in MIC reading.

Table 3.12e Significant QTL's selected from MLM-MMR based models for fiber length

QTL'S	Model R2	Adj. R2	AIC	AICC	BIC	SBC	Pr > F
E4M4_177	0.127	0.114	-1488.454	-5.912	-1489.059	-1474.97	0.0014
E6M3_363	0.083	0.075	-1480.064	-5.874	-1480.303	-1469.95	0.0032
E4M2_135	0.357	0.319	-1536.275	-6.126	-1535.660	-1492.45	0.0163
E3M4_364	0.171	0.155	-1497.599	-5.954	-1498.397	-1480.74	0.0010
E8M8_605	0.229	0.207	-1509.224	-6.007	-1510.257	-1485.63	0.0064
E7M6_140	0.432	0.380	-1550.889	-6.184	-1546.173	-1486.84	0.0435
E8M2_75	0.320	0.287	-1528.340	-6.092	-1528.654	-1491.26	0.0181
E8M2_270	0.301	0.271	-1524.435	-6.075	-1525.037	-1490.72	0.0129
E2M3_60	0.280	0.252	-1519.934	-6.055	-1520.761	-1489.59	0.0078
E6M7_105	0.338	0.302	-1532.116	-6.108	-1532.044	-1491.66	0.0197
E6M5_145	0.395	0.349	-1543.297	-6.154	-1541.103	-1489.36	0.0482
E1M5_200	0.045	0.040	-1473.215	-5.842	-1472.927	-1466.47	0.0018
E1M2_45	0.420	0.370	-1548.407	-6.175	-1544.625	-1487.73	0.0374
E6M6_140	0.371	0.331	-1539.179	-6.138	-1538.054	-1491.99	0.0325
E4M6_175	0.255	0.229	-1514.528	-6.031	-1515.523	-1487.56	0.0081
E1M6_190	0.201	0.182	-1503.525	-5.981	-1504.496	-1483.30	0.0056
E1M6_270	0.407	0.359	-1545.666	-6.164	-1542.758	-1488.36	0.0452

Table 3.12f Significant QTL's selected from MLM-MMR based models for Micronaire

QTL'S	Model R^2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E6M4_325	0.077	0.073	-505.143	-1.340	-504.783	-498.402	<.0001
E6M1_320	0.337	0.311	-562.215	-1.601	-561.926	-531.880	0.0213
E6M1_196	0.320	0.297	-558.666	-1.585	-558.567	-531.701	0.002
E4M2_320	0.183	0.171	-527.258	-1.442	-527.519	-513.776	0.0006
E5M7_375	0.222	0.207	-535.807	-1.481	-536.191	-518.954	0.0013
E9M8_330	0.368	0.337	-568.584	-1.628	-567.673	-531.507	0.0275
E8M3_50	0.396	0.360	-574.133	-1.651	-572.291	-530.315	0.0319
E8M3_255	0.409	0.371	-577.050	-1.663	-574.577	-529.861	0.0322
E7M6_140	0.382	0.348	-571.219	-1.639	-569.906	-530.772	0.0367
E9M1_54	0.136	0.127	-517.111	-1.395	-517.130	-506.999	0.0002
E9M1_140	0.255	0.237	-543.057	-1.514	-543.461	-522.833	0.0027
E8M1_120	0.353	0.325	-565.453	-1.615	-564.897	-531.746	0.0256
E2M4_135	0.288	0.267	-550.765	-1.549	-551.010	-527.171	0.0022
E1M2_210	0.425	0.385	-580.684	-1.679	-577.371	-530.125	0.0222

It is also being highly modified by the environmental factors and stress. The present study investigates 14 markers identified by MLM-MMR which were associated with MIC (Table: 3.12f).

The significant markers like E9M8_330, E8M3_50, E8M3_255 and E1M2_210 had high adj. R^2 (33-38.5%), supported by lower AIC, BIC and SBC values.

There were nine markers significantly associated with uniformity index (Table: 3.12g). The adj. R^2 was relatively low to moderate for this trait (10.6-33.4%). Markers such as E4M4_217, E5M3_115 and E3M7_370 were found significant with adj. R^2 values of 31.2, 34.6 and 33.5% respectively.

Table 3.12g Significant QTL's selected by MLM-MMR based models for uniformity index

QTL'S	Model R^2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E4M1_348	0.174	0.166	177.358	1.835	178.060	187.470	<.0001
E9M5_53	0.110	0.106	191.353	1.900	192.178	198.094	<.0001
E4M4_217	0.335	0.312	140.960	1.669	142.790	167.926	0.010
E5M3_115	0.374	0.346	131.968	1.629	134.982	165.674	0.0349
E5M3_175	0.228	0.217	164.869	1.778	165.649	178.352	0.0002
E3M7_370	0.360	0.335	134.647	1.641	137.156	164.983	0.0048
E8M2_155	0.261	0.247	157.594	1.744	158.475	174.447	0.0026
E1M3_220	0.313	0.293	145.873	1.691	147.282	169.467	0.0066
E4M6_175	0.288	0.271	151.512	1.717	152.596	171.736	0.0051

Table 3.12h Significant QTL's selected from MLM-MMR based models for SFI

QTL'S	Model R2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E4M4_280	0.306	0.303	-113.097	0.484	-113.397	-106.356	<.0001
E6M3_288	0.416	0.410	-148.127	0.321	-148.592	-138.015	<.0001
E5M8_175	0.457	0.449	-161.943	0.257	-162.687	-148.461	<.0001
E6M4_297	0.626	0.604	-224.253	-0.024	-222.480	-180.435	0.0177
E6M4_270	0.492	0.482	-174.270	0.201	-175.124	-157.417	0.0002
E5M6_40	0.616	0.595	-220.247	-0.007	-219.114	-179.799	0.0087
E5M3_152	0.635	0.612	-227.409	-0.037	-224.980	-180.220	0.0283
E5M1_204	0.559	0.544	-198.415	0.091	-199.042	-171.450	0.0015
E7M6_65	0.593	0.575	-211.866	0.030	-211.681	-178.159	0.0032
E8M2_159	0.602	0.583	-214.931	0.017	-214.443	-177.854	0.0285
E8M2_185	0.518	0.506	-183.463	0.159	-184.351	-163.240	0.0010
E2M8_260	0.536	0.523	-189.857	0.130	-190.754	-166.263	0.0044
E2M2_280	0.575	0.559	-204.723	0.062	-205.057	-174.387	0.0048
E2M2_218	0.643	0.618	-230.054	-0.048	-226.932	-179.495	0.0379

Short fiber content (SFC) is the percentage of fibers by weight with a length of less than 12.7 mm (Behery, 1993). The source for SFC comes from inherent nature of the genotype, the environment, or may be introduced by extensive mechanical handling of the cotton. Genetic factors such as those imparting fiber strength may also be involved in causing SFC. The association mapping study revealed 14 significant markers identified by MLM-MMR models that were associated with SFC (Table: 3.12h). Markers E4M4_280, E6M3_288, E5M8_175 and E2M2_218 were most responsible for the short fiber content.

Fiber traits associated with AFLP markers from this study were compared with earlier AFLP based mapping studies. It was hard to make any correlated Conclusion. Previous reports on associating AFLP markers with fiber traits using either general linear methods or combined MLM-MMR are few in cotton. Wu *et al*., (2007) reported E6M3_266 to have a strong association with LP. However, 1-4 markers were associated with 22-93% of the phenotypic variability of the fiber traits using GLM methods. The published association mapping studies do reveal the significance of MLM methods in reducing Type I errors (Abdurakhmonov *et al*., 2008; 2009; Zeng *et al*., 2008). The present study went further and explored multiple regression methods in order to validate the existence of QTL or trait associations.

The significant QTL's associated with fiber traits suggests that multiple linear regression models coupled with mixed model effect selection to be a promising approach for use in future cotton association based studies. The results provide strong evidence that through the application of multiple selection criteria such as R^2, BIC, AIC, AICC and SBC that it is possible to identify fewer markers that explain a greater proportion of the phenotypic variation, than the standard F tests commonly implemented in standard QTL mapping studies.

c) Epistasis for Fiber Quality Parameters

A total of 82 QTL's for fiber quality were identified by MLM-MMR model based QTL analysis. Although partial dominance and over dominance cannot be ruled out, additive genetic variance was predominant.

Common QTLs were detected in each trait found to be interacting with other significant QTL's. The QTL's identified through the additive epistatic model for major fiber traits are summarized in Table 3.13. With respect to lint yield, markers E4M7_45, E5M7_325 and E3M6_70 were found to be common and interacting with other markers. Our results indicate that additive gene action was the primary mechanism responsible for genetic variability in fiber quality traits.

Table 3.13 Significant QTL's identified interacting in additive epistatic manner for various fiber traits in cotton association mapping genotypes

ELO	MS	F Value	Pr > F	FS	MS	F Value	Pr > F
E1M6_130 x E2M6_525	2.528	8	0.0057	E6M2_255 x E6M7_100	9.23	4.97	0.0292
E9M5_50 x E4M6_60	1.859	5.89	0.0171	E6M7_125 x E2M5_295	8.03	4.32	0.0415
E2M8_315 x E1M7_335	1.456	4.61	0.0343	E6M2_255 x E4M7_180	7.46	4.01	0.0493
LP				**FL**			
				E8M2_270 x E1M2_45	0.006	10.47	0.0017
E6M4_249 x E4M6240	12.269	7.34	0.0079	E4M2_135 x E1M2_45	0.005	8.69	0.0041
E3M6_95 x E5M3_110	11.227	6.71	0.0109	E7M6_140 x E2M3_60	0.003	5.05	0.0273
E5M2_75 x E5M7_325	10.174	6.08	0.0153	E6M3_363 x E6M6_140	0.002	4.27	0.0419
E3M6_95 x E5M7_325	9.496	5.68	0.0190				
E3M6_70 x E5M3_110	9.185	5.49	0.0210	**MIC**			
E6M4_249 x E3M6_95	7.730	4.62	0.0339	E5M7_375 x E8M3_255	0.288	4.86	0.0292
				E4M2_265 x E8M3_50	0.235	3.97	0.0485
LY				E6M4_325 x E5M7_375	0.206	3.48	0.05
E4M7_45 x E4M8_320	768.94	11.7	0.0008	**SFC**			
E3M6_70 x E3M6_300	589.90	8.97	0.0032	E2M8_260 x E2M2_280	2.193	6.8	0.0101
E5M4_450 x E5M7_325	429.00	6.53	0.0116	E6M4_270 x E7M6_65	1.961	6.08	0.0149
E3M6_70 x E5M7_325	380.32	5.78	0.0173				
E8M8_245 x E4M8_320	319.52	4.86	0.029	**UI**			
E4M7_45 x E4M7_195	291.30	4.43	0.0369	E4M1_348 x E4M6_175	2.99	4.92	0.0278
E5M7_325 x E4M7_45	288.53	4.39	0.0378	E4M1_348 x E3M7_370	2.83	4.67	0.0321

3.4 Conclusion

While further validation is required, the markers showing strongest effects in this study provide ideal candidates for further study or future inclusion in strategies of marker assisted selection. The six groups identified in the CAM panel with high allelic divergence among the clusters and wide genetic distances proved to be efficient in capturing the enormous phenotypic variability present in the fiber traits. The insights provided by the in MLM-MMR approach reported herein, demonstrate the feasibility of this approach in reducing the false positives. Out of 568 AFLP markers used in this study, 255 markers were initially found to be significantly associated with eight traits using the traditional MLM approach. Inclusion of MMR improved the model, reducing the number of markers significantly associated with these traits to 111. The MMR based epistatic interactions revealed 49 QTLs responsible for eight fiber traits. Thus mixed MMR models were efficient in reducing the Type I error.

3.5 References

Abdurakhmonov IY, Saha S, Jenkins JN, Zabardast T, Burie Shukhrat E, Shermatov Scheffler BE, Pepper AE, Yu JZ, Kohel RJ and Abdukarimov A 2009 Linkage disequilibrium based association mapping of fiber quality traits in *G. hirsutum* L. variety germplasm. Genetica, 136, 401–417.

Abdurakhmonov IY, Kohel RJ, Yu JZ, Pepper AE, Abdullaev AA, Kushanov FN, Salakhutdinov IB, Buriev ZT, Saha S, Scheffler BE, Jenkins HN and Abdukarimov A 2008 Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm, Genomics 92(6), 478-87.

Becker J, Vos P, Kipier M, Salamini F and Heun M 1995 Combined mapping of AFLP and RFLP markers in barley. Molecular and General Genetics, 249(1), 65-73.

Behery HM 1993 Short fiber content and uniformity index in cotton. International Cotton Advisory Committee and center for Agricultural and Biosciences Review Article 4. Washington DC, DC, pp. 1-140.

Bradbury PJ, Zhang DE, Kroon TM, Casstevens, Ramdoss and Buckler ES 2007 TASSEL: Soft ware for association mapping of complex traits in diverse samples. Bioinformatics, 23, 2633-2635.

Brooks SP and Gelman A 1998 Alternative methods for monitoring convergence of iterative simulations, Journal of Computational and Graphical Statistic, **7**, 434–455.

Cândida HC, de Magalhães Bertini, Ivan Schuster, Tocio Sediyama, Everaldo Gonçalves de Barros and Maurílio Alves Moreira. 2006 Characterization and genetic diversity analysis of cotton cultivars using microsatellites. Genetics and Molecular Biology, 29(2), 321-329.

Casa AM, Pressoira G, Brown PJ, Mitchell SE, Rooney WL, Tuinstrac MR, Franks CD and Kresovicha S 2008 Community resources and strategies for association mapping in sorghum. Crop Science, 48, 30–40.

El-Zik KM, Thaxton PM 1989 Genetic improvement for resistance to pests and stresses in cotton. In: Frisbee RE, El-Zik KM, Wilson LT (eds) Integrated pest managment systems and cotton production. John Wiley and Sons, New York, pp 191–224.

Endrizzi JE, Turcotte EL, Kohel RJ 1985 Genetics, cytology and evolution of *Gossypium*. Advanced Genetics, 23, 271–375.

Evanno G, Regnaut S and Goudet J 2005 Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology, 14, 2611–2620.

Excoffier L, Laval G, and Schneider S, 2005 Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evolutionary Bioinformatics, Online 1, 47-50.

Flint-Garcia SA, Thornsberry JM and Buckler ES 2003 Structure of linkage disequilibrium in plants. Annual Review of Plant Biology, 54, 357–374.

Gao H, Williamson S and Bustamante CD 2007 A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176, 1635–1651.

Gaut BS and Long AD 2003 The slowdown on linkage disequilibrium. Plant Cell, 15(7), 1502-1506.

Greenacre MJ 1984 Theory and Applications of Correspondence Analysis. Academic Press, London.

Hammer Ø, Harper DAT and Ryan PD 2001 PAST: Paleontological statistics software package for education and data analysis, Palaeontologia Electronica, 4(1), art 4.

Hardy OJ and Vekemans X 2002 SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes **2:** 618-620.

Hawkins JS, Pleasants J and Wendel, J.F. 2005 Identification of AFLP markers that discriminate between cultivated cotton and the Hawaiian Island endemic, *Gossypium tomentosum* Nuttall ex Seeman. Genetic Research on Crop Evolution, 52, 1069–1078.

Hertel KL 1953 The Stelometer, measures fiber strength and elongation. Textile World, 103: 97-260.

Iqbal MJ, Reddy OUK, El-Zik KM and Pepper AE 2001 A genetic bottleneck in the ‘evolution under domestication’ of upland cotton *Gossypium hirsutum* L. examined using DNA fingerprinting. Theoretical and Applied Genetics, 103, 547–554.

Iwata, H., Y. Uga, Y. Yoshioka, K. Ebana, and T. Hayashi. 2007. Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasm. Theoretical and Applied Genetics, 114, 1437–1449.

Kantartzi SK and Stewart J McD 2008 Association analysis of fiber traits in *Gossypium* arboretum accessions. Plant Breeding, 127, 173-179.

Kraakman ATW, Martinez F, Mussiraliev B, Eeuwijk FA and Niks RE 2006 Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. Molecular Breeding, 17, 41–58.

Kumar S, Tamura K and Nei M 2004 MEGA4: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinformatics, 5(2), 150–163.

Lacape JM, Dessauw D, Rajab M, Noyer JL and Hau B 2007 Microsatellite diversity in tetraploid *Gossypium* germplasm: assembling a highly informative genotyping set of cotton SSRs. Molecular Breeding, 19, 45–58.

Lacape JM, Nguyen TB, Thibivilliers S, Bojinov B, Courtois B, Cantrell RG, Burr B and Hau B 2003 A combined RFLP–SSR–AFLP map of tetraploid cotton based on a *Gossypium hirsutum* ×*Gossypium barbadense* backcross population. Genome, 46, 612–626.

Liu S, Cantrell RG, McCarty JC Jr and Stewart JM 2000 Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Science, 40, 1459-1469.

Maughan PJ, Saghai-Maroof MA, Buss, GR and Huestis GM 1996 Amplified fragment length polymorphisms (AFLP) in soybean: species diversity, inheritance, and near- isogenic lines analysis. Theoretical and Applied Genetics, 93(3), 392- 401.

Myers GO, Baogong J, Akash MW, Badigannavar AM and Saha S 2009 Chromosomal assignment of AFLP markers in upland cotton (*Gossypium hirsutum* L.) Euphytica, 165(2), 391-399.

Nei M 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89, 583–590.

Nei M 1972 Genetic distances between populations. American Naturalist, 106, 283–292.

Nei M, and Li WH 1979 Mathematical model for studying genetic variations in terms of restriction endonucleases. Proceedings of National Academy of Sciences, USA 76, 5369–5373.

Niu C, Doug J, Hinchliffe, Yingzhi Lu, Cantrell RG, Wang C, Roberts P and Zhang J 2007 Identification of molecular markers linked to root-knot nematode resistance in cotton (*Gossypium hirsutum* L.). Crop Science, 47, 951-960.

Nordborg M and Tavare S 2002 Linkage disequilibrium: What history has to tell us. Trends in Genetics, 18, 83–90.

Peleg Z, Fahima T, Abbo S, Krugman T and Saranga Y 2008 Genetic structure of wild emmer wheat populations as reflected by transcribed versus anonymous SSR markers. Genome 51, 187-195.

Peakall R and Smouse PE 2006 GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes, 6, 288-295.

Pillay M and Myers GO 1999 Genetic diversity in cotton assessed by variation in ribosomal RNA genes and AFLP markers. Crop Science, 39(6), 1881-1886.

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S and Rafalski A 1996 The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding 2, 225-238.

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

Pritchard JK, Stephens M, Rosenberg NA and Donnelly P 2000 Association mapping in structured populations. American Journal of Human Genetics, 67, 170–181.

Rana MK and Bhat KV 2004 A comparison of AFLP and RAPD markers for genetic diversity and cultivar identification in cotton. Journal of Plant Biochemistry and Biotechnology, 13, 19–24.

Rogers JS 1972 Measures of genetic similarity and genetic distance. Studies in genetics. VII. Univ. Texas Publ. 7213,145-153.

Rohlf FJ 1990. NTSYS-pc. Numerical taxonomy and multivariate analysis system, Version 2.02. Exeter Soft ware, New York.

Rosenberg NA 2004 *Distruct*: a program for the graphical display of population structure. Molecular Ecology Notes, 4, 137-138.

Rosenberg N, Pritchard JK, Weber JL, Cann H and Kidd H 2002 Genetic structure of human populations. Science, 298, 2381–2385.

Saitou N and Nei M 1987 The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, 4, 406-425.

SAS Institute, 2008 Documentation GLMSELECT. SAS Institute, Cary, NC.

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

Sharma SK, Knox MR, and Ellis THN 1996 AFLP analysis of the diversity and phylogeny of *Lens* and its comparison with RAPD analysis. Theoretical and Applied Genetics, 93(5-6), 751-758.

Sneath PHA and Sokal RR 1973 Numerical taxonomy The principals and practice of numerical classification. W.H. Freeman and Co., San Francisco, California. Pp 573.

TASSEL 2009 User manual, trait analysis by association, evolution and linkage. www.maizegenetics.net/tassel

Vekemans X, Beauwens T, Lemaire M and Roldan-Ruiz I 2002 Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Molecular Ecology, 11, 139-151.

Vos P, Hogers R, Bleeker M, Reijans M, Van T, Hornes M, Frijters A, Pot J, Peleman J and Kuiper K 1995 AFLP: A new technique for DNA fingerprinting. Nucleic Acids Research, 23, 4407-4414.

Wang C and Roberts PA 2006 Development of AFLP and CAPS markers for root knot nematode resistance in cotton. Euphytica, 152(2), 185-196.

Weir BS 1996 Genetic Data Analysis II, 2nd Ed. Sinauer Associates, Inc., Sunderland, MA.

Weir BS and Cockerham CC 1984 Estimating F statistics for the analysis of population structure. Evolution: International Journal of Organic Evolution, 38, 1358–1370.

Wilson LM, Whitt SR, Ibáñez AM, Rocheford TR, Goodman MM and Buckler ES 2004 Dissection of Maize kernel composition and starch production by candidate gene analysis. The Plant Cell 16, 2719-2733.

Wright S 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution, 19, 395-420.

Wu J, Jenkins JN, McCarty JC, Zhong M and Swindle M 2007 AFLP marker associations with agronomic and fiber traits in cotton. Euphytica. 153, 153-163.

Yu J, Pressoir G, Briggs WH, Vroh I, Yamasaki BM, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S and Buckler ES 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics, 38, 203–208.

Zeng, L.,William R. Meredith Jr. Osman A. Gutie´rrez DL and Boykin 2009 Identification of associations between SSR markers and fiber traits in an exotic germplasm derived from multiple crosses among tetraploid species. Theoretical and Applied Genetics, 119, 93–103.

Zhao K, MJ Aranzana S Kim C Lister C, Shindo C, Tang C, Toomajian H, Zheng C, Dean P Marjoram and M Nordborg 2007 An *Arabidopsis* example of association mapping in structured samples. PLoS Genetics, 3:e4.

Zhang J, Yuan Y, Niu C, Hinchliffe DJ, Lu Y, Yu S, Percy RG, Ulloa M and Cantrell RG 2007 AFLP-RGA markers in comparison with RGA and AFLP in cultivated tetraploid cotton, Crop Science, 47:180-187.

Zhang J and Stewart J. McD 2000 Economical and rapid method for extracting cotton genomic DNA. Journal of Cotton Science, 4: 193-201.

Zhu Q, Zheng X, Luo J, Gaut BS and Ge S 2007 Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: Severe bottleneck during domestication of Rice. Molecular Biology and Evolution, 24, 875-888.

CHAPTER 4 CHARACTERIZATION OF UPLAND COTTON GENOTYPES FOR MOLECULAR DIVERSITY AND MARKER TRAIT ASSOCIATIONS

4.1 Introduction

Cotton (primarily *Gossypium hirsutum* L. and *G. barbadense* L.) is the most extensively used natural fiber in the textile industry and is the sixth most abundantly grown oilseed crop. It is grown commercially in the tropical and subtropical regions of more than 50 countries. Worldwide, cotton production has been relatively stable for the last several years. In the United States however, planted acreage fell to 9.1 million in 2009 the lowest since 1983 and well below the 15.5 million acres planted in 2006 (NASS, 2009). In Louisiana, producers planted 240,000 acres and expected to harvest 420,000 bales, up 49 percent from last year's hurricane devastated crop (NASS, 2009). Due to the global economic downturn, world cotton consumption fell by 12% in 2008/09 after a decade of uninterrupted growth. As the economy gradually stabilizes, world cotton use is expected to recover slowly. Increases in cotton consumption will mainly be driven by a rebound in Asia, in particular China (mainland), India and Pakistan.

Fiber quality has become an increasingly important consideration in marketing cotton and choosing varieties. Modern spinning technologies demand cotton with the most consistent and highest-quality fiber properties. Conventional breeding has played an important role in yield and fiber quality improvement of upland cotton. The advent of molecular markers may make it possible for plant breeders to even more rapidly and precisely improve crop economic and agronomic traits (Tanksley and Hewitt 1988).

Molecular marker technology can also be a valuable tool for exploring the genetic diversity in cotton. A variety of molecular-marker technologies have been used to study the genetic diversity and relationships between cultivated cottons and their wild relatives. Of these methods, random amplified polymorphic DNAs (RAPDs) have been most widely used (Multani and Lyon 1995;

Tatineni *et al.,*1996; Iqbal *et al.,* 1997). RFLPs (Wendal and Brubaker 1993) have numerous advantages over RAPD's (reproducibility), but have been limited in their use due to their technical complexity. Recently, the amplified fragment length polymorphism (AFLP) method (Zabeau and Vos 1993; Vos *et al.,* 1995) has also been successfully used to analyze genetic diversity among a wide range of crop species and their wild relatives (Hill *et al.,*1995; Maughan *et al.,* 1996; Powell *et al.,* . 1996). AFLP's have higher repeatability than RAPD's and are technically easier than RFLP's. Their highly polymorphic nature is also an advantage, especially in *Gossypium* genus, where intraspecific polymorphism is low. At least in cultivated cottons, recent studies using molecular markers suggest a fairly high degree of genetic uniformity and similarity. Van Becelaere *et al.,* (2005) and Lu and Myers (2002) reported very high levels of genetic similarity ranging from 0.91 to 0.97 and 0.93 to 0.98, respectively.

The narrow genetic base of upland cotton germplasm that is used in breeding programs is one of the factors in failing to achieve appreciable amount of progress in improving yield and fiber traits over last two decades (Meredith 2000). Some studies have postulated that decline in genetic diversity is due to frequent use of few parents and lack of contribution from the secondary gene pool (Bowman *et al.,* 1996). Thus there is need to improve the genetic base of the existing genotypes by tapping the secondary and tertiary gene pools. Several breeding programs have been initiated over the past few years to breed superior genotypes through the co-ordinated efforts of several breeders across the US.

The National collection of *Gossypium* species at Germplasm Research Unit TX, USA comprises of 9332 accessions representing 49 species from 74 countries assigned to three germplasm pools (Wallace *et al.,* 2009). There is a need to screen the core germplasm with high density molecular map based PCR markers to fingerprint all accessions, in order to minimize any sort of duplications. The development of a standard set of SSR markers that represents the diversity across the cotton genome is needed. Based on most of the previous studies in cotton on diversity, it is understood that genetic diversity exists in the primary gene pool. But there is much room for broadening the genetic base of the commercial germplasm. The National regional breeder trial network (RBTN) has been the mainstay in developing new upland varieties incorporating various traits to combat biotic and abiotic stresses apart from improving fiber traits. The newly developed Louisiana and other upland cotton genotypes suitable for cultivation in wide agro climatic conditions has to be screened for their inherent genetic diversity and for the presence of novel QTL's associated with fiber traits utilizing multi-location phenotypic and polymorphic molecular marker data in association mapping system.

Hence the present study was planned to determine the efficiency of AFLP for estimating genetic diversity among a collection of 60 accessions of upland cotton and also for the identification of potential marker trait associations for major fiber traits.

4.2 Materials and Methods

4.2.1 Plant Material and Phenotypic Analysis

A set of 60 upland cotton genotypes from Louisiana, Regional Breeding testing Network and a set of newly developed heat tolerant genotypes were included in the study (Table 4.1). The Regional Breeder's Trial Network (RBTN) is a multistate testing program of public breeding lines. The genotypes were segregated into categories based upon region of origin.

Plants were field grown in 2008 as per LA Cooperative Extension Service guidelines at the Dean Lee Research Station in Alexandria, LA. Leaf samples from representative plants were collected and bulked for DNA extraction. Phenotypic data on yield and fiber traits was obtained from the RBTN trial website (www.cottonrbtn.com). The four replication data on lint yield, micronaire, fiber length, strength, uniformity ratio, maturity coefficient and Short Fiber index (SFI) was averaged to calculate mean and variances using SAS 9.1.3 (SAS Institute, Cary, NC).

Deltapine DP 393 (Bridge and Gowan 2005; US patent 6930228) and Phytogen 72, Acala (US PVP 200100115) were considered as check and all the comparisons were made in relation with the performance of these genotypes.

Table 4.1 List of Upland cotton genotypes selected for the study with their description

Code	Cultivar	Description†	Code	Cultivar	Description
LA-1	AU-5491	SE region	LA-31	LA 05307113	Louisiana region
LA-2	NM-03012	SW region	LA-32	LA 05307095	Louisiana region
LA-3	GA-2004230	SE region	LA-33	LA 05307027	Louisiana region
LA-4	04PST-250	Delta	LA-34	LA 05307087	Louisiana region
LA-5	0020-31ne	Arkansas region	LA-35	LA 05307107	Louisiana region
LA-6	PD-04012	SE region	LA-36	LA04308035	Louisiana region
LA-7	0028-16ne	Arkansas region	LA-37	LA 05307094	Louisiana region
LA-8	04PST 246	Delta	LA-38	AGC 208	SW region
LA-9	ARK 0015-06-11	Arkansas region	LA-39	8824	Delta
LA-10	0147-22ne	Arkansas region	LA-40	PX03201-38-5	Heat tolerant
LA-11	ACALA 1517-99	SW region	LA-41	PX03202-83-3	Heat tolerant
LA-12	TAM B 182-34	Texas	LA-42	PX03202-9-1	Heat tolerant
LA-13	AU-6103	SE region	LA-43	PX03203-25-2	Heat tolerant
LA-14	AU-5367	SE region	LA-44	PX03203-65-3	Heat tolerant
LA-15	8921-2-2-14-13-11	Arkansas region	LA-45	PX03204-21-1	Heat tolerant
LA-16	04-PST-275	Arkansas region	LA-46	PX03201-66-7	Heat tolerant
LA-17	0149-17ne	Arkansas region	LA-47	PX03201-19-3	Heat tolerant
LA-18	GA-2004089	SE region	LA-48	PX03201-38-5	Heat tolerant
LA-19	GA-2004303	SE region	LA-49	PX03203-65-3	Heat tolerant
LA-20	LA 05307083	Louisiana region	LA-50	PX03202-9-1	Heat tolerant
LA-21	LA 05307029	Louisiana region	LA-51	PX03201-19-2	Heat tolerant
LA-22	LA 0530761	Louisiana region	LA-52	PX03201-66-1	Heat tolerant
LA-23	LA 05307025	Louisiana region	LA-53	PHYTOGEN 72	California Acala
LA-24	LA 05307119	Louisiana region	LA-54	SG 747	Delta
LA-25	LA 05307073	Louisiana region	LA-55	PX03201-19-4	Heat tolerant
LA-26	LA 05307042	Louisiana region	LA-56	PX03203-25-2	Heat tolerant
LA-27	LA 05307062	Louisiana region	LA-57	PX03204-21-1	Heat tolerant
LA-28	LA 05307028	Louisiana region	LA-58	PX03202-83-3	Heat tolerant
LA-29	LA 05307057	Louisiana region	LA-59	PX03202-65-1	Heat tolerant
LA-30	LA 05307088	Louisiana region	LA-60	PX03201-66-8	Heat tolerant

† : SE=South eastern; LA=Louisiana; SW=South west;

Seed cotton yield and lint yield were standardized by setting the yield of DP 393 as equal to 100%. Fiber analysis was conducted by using High Volume Instrument (HVI) system. The phenotypic data was subjected to ANOVA to determine replication and genotypic differences. Correlation analysis for pair of traits was performed using PROC CORR in SAS.

4.2.2 Genotyping with AFLP Markers

Sixty four primer combinations were used to generate AFLP data (Table 4.2) following procedure given by Vos *et al.,* (1995) with minor modifications. Sample DNA was digested with *EcoRI* (infrequent cutter with GAATTC recognition sequence) and *MseI* (frequent cutter with TTAA recognition sequence) restriction enzymes and oligonucleotide adapters specific to restriction sites were ligated to the resulting fragments through incubation (180 min, 37 °C) with DNA ligase using a iCycler (BioRad Labs, Hercules, CA).

Table 4.2 Adapters and primers of AFLP marker system used for pre and selective amplification in upland cottons.

Primer/adapter	Nomenclature*	Sequences(5'-3')
ECORI primers:		
EcoRI linker 1	E-I	CTC GTA GAC TGC GTA CC
EcoRI linker 2	E-II	AAT TGG TAC GCA GTC TAC
EcoRI + A	E+A	GAC TGC GTA CCA ATT CA
E- AAC	E1	GACTGCGTACCAATTCAAC
E- AAG	E2	GACTGCGTACCAATTCAAG
E-ACA	E3	GACTGCGTACCAATTCACA
E-ACT	E4	GACTGCGTACCAATTCACT
E-ACC	E5	GACTGCGTACCAATTCACC
E-ACG	E6	GACTGCGTACCAATTCACG
E-AGG	E8	GACTGCGTACCAATTCAGG
E-AGA	E9	GACTGCGTACCAATTCAGA
MseI primers:		
MseI linker 1	M-I	GAC GAT GAG TCC TGA G
MseI linker 2	M-II	TAC TCA GGA CTC AT
MseI + C	M+C	GAT GAG TCC TGA GTA AC
M-CAA	M1	GATGAGTCCTGAGTAACAA
M-CAC	M2	GATGAGTCCTGAGTAACAC
M-CAG	M3	GATGAGTCCTGAGTAACAG
M-CAT	M4	GATGAGTCCTGAGTAACAT
M-CTA	M5	GATGAGTCCTGAGTAACTA
M-CTC	M6	GATGAGTCCTGAGTAACTC
M-CTG	M7	GATGAGTCCTGAGTAACTG
M-CTT	M8	GATGAGTCCTGAGTAACTT

*Nomenclature is in accordance with the Lacape *et al*., 2003; Myers *et al*., 2009.

Pre-amplifications were done using *EcoR* I+A and *Mse* I+C oligo primers. The amplification was carried out with 50ng/ul of oligo primers, 5mM dNTP's, 25mM $MgCl_2$, 10X buffer, Taq polymerase(5U/µl) and restrict ligated template DNA making total volume of 20µl. The PCR was set up with initial denaturing for 94^{o}C for 2 min followed by 26 cycles at 94^{o}C for 1 min, 56^{o}C for 1 min., 72^{o}C for 1 min., and final extension at 72^{o}C for 5min. The pre amplified products were diluted with ddH_2O. Selective amplification was done using two selective nucleotides. The EcoRI+ANN oligo primers were dye labeled with 700 and 800 IR dye (MWG Biotech, Germany). The PCR for selective amplification was carried out in a reaction volume of 10 µL consisting of 10X reaction buffer, 25 mM $MgCl_2$, 2.5 mM dNTPs, 1 µM each of EcoRI-ANN and MseI+CNN primers and 5U *Taq* polymerase (Promega, Madison, WI). The reactions were run on an *i*-Cycler (BioRad Labs, Hercules, CA). Touchdown PCR was used for selective amplifications using the following profile: initial denaturing step at 94^{o}C for 2 min followed by initial 12 cycles at 94^{o}C for 30 s, 65^{o}C for 30 s (with 0.7^{o}C decrement every cycle) and 72^{o}C for 1 min, then followed by 23 cycles at 94^{o}C for 30 s, 56^{o}C for 30 s, and 72^{o}C for 1 min with a final extension step at 72^{o}C for 2 min. A total of 64 *EcoR* I - *Mse* I selective amplification primer combinations were used. The PCR amplified products were run on a LI-COR 4300 sequencer (LI-COR Inc., Lincoln, NE). Gels images were saved onto a computer, printed and scored manually. Presence of a band was recorded as '1' and absence as '0', as per a typical dominant marker system. Ambiguous data that could not be resolved was discarded. The nomenclature of AFLP loci was followed according to Lacape *et al.*, (2003); and Myers *et al*., (2009), indicating the enzyme primer combinations with band size.

4.2.3 Molecular Analysis

For each marker used, sub-populationwise diversity statistics including number of bands and Nei's genetic distances were calculated using GenAlex 6.1 software (Peakall and Smouse 2006). Allelic diversity at a given locus can be determined by Polymorphism Information Content (PIC) and it was calculated as ‘PIC=1-$\sum f_i^2$ where, f_i is the frequency of the i^{th} allele (Weir, 1996). PROC ALLELE was used to calculate PIC values and frequency estimate was done using PROC Freq (SAS, 9.1.3).

In order to know the possible structure in the set of core panel, various statistical analyses were performed on the basis of allelic frequencies. First, the Dice similarity coefficient was calculated using the formula D = 2a/(2a + b + c), where *a* = 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 data were calculated for each pair of genotypes (distance =1- similarity) and used for UPGMA clustering in MEGA 4.0 (Kumar *et al.,* 2004). In addition, Principal Coordinate Analysis (PCoA) was also performed using a genetic similarity matrix based on the Nei – Li (1979) to supplement the findings obtained from cluster analysis. All the above analyses were performed employing PAST software (Hammer *et al*., 2001).

Correspondence analysis was performed on core panel by marker matrix of band incidences (Greenacre 1984). The multivariate nature of correspondence analysis can reveal relationships that would not be detected in a series of pair wise comparisons of variable. Another important feature is the graphical display of row and column points in biplots, which can help in detecting structural relationships among the variable categories and objects. The whole procedure was implemented in PAST software (Hammer *et al*., 2001) using AFLP marker data with predefined cultivar groups.

4.2.4 Association Analysis; Statistical Models and Procedures

a) Mixed Models for Marker-Trait Association

For a successful marker trait association, one has to account for type I error or spurious associations/false positives. Incorporating the outcome of population structure and PCA increases the power to detect true marker trait associations. In view of this, we tested four statistical mixed models for 254 AFLP markers and adjusted R^2 values were computed for the fixed marker effects using TASSEL 2.1 beta version (Bradbury *et al*., 2007). Tests for significance were applied using F statistic associated with the marker. The model possessing highest adjusted R^2 was considered best among all, capturing maximum variation explained by the model. The cutoff P value (0.05) determines whether a QTL is associated with the marker and R^2 estimates magnitude of the QTL effects. Most of the marker trait associations were made based on 60 genotypes.

b) Mixed – Multiple Regression Models for Association Analysis

In order to exploit the advantages of multiple regression procedures, we used all those traitwise significant markers selected by mixed model procedures using TASSEL and screened for 52 PROC GLMSLECT models. Stepwise selection method was used with all possible combinations of CHOOSE, SELECT and STOP. Different options were used for these selection methods such as, Bayesian Information Content (BIC), SBC (Schwarz Bayesian Information Criterion), Adj. R^2, AICC (the Corrected Akaike Information Criterion), SL=0.15 (the significance level of the F statistic for entering or departing effects) and Cross validation (CV). Traits were considered as dependent variable and all the markers were treated as independent variables. Each trait was analyzed separately and those independent variables showing test statistic estimate less than the P value (0.05) were added in the model. To reduce the Type I error, selected models were tested with validation step by using 'PRESS' criterion in 'STOP' option. The best model was then selected based on adjusted R^2 and less number of effects for a particular trait.

4.3. Results

4.3.1 Phenotypic Analysis

The 60 genotypes were evaluated in Louisiana to obtain estimates of agronomic performance and fiber quality. The phenotypic data for seed cotton yield (SCY) and fiber traits was obtained from the RBTN coordinators (summarized in Table: 4.3). The Louisiana cultivar, LA05307042 recorded 16% more SCY than the check variety, while among the heat tolerant lines, PX3201-38-5 yielded 24% more SCY compared to the check.

Table 4.3 Phenotypic variability for LY and fiber quality traits among upland cotton genotypes

Variable	SCY†	LY	LP	MIC	FL	FS	UI	ELO	SFC
N	60	60	60	60	60	60	60	60	60
Mean	97.35	89.71	44.47	4.80	1.16	31.69	84.30	7.85	5.49
SD	12.12	15.99	8.60	0.35	0.07	2.23	1.61	1.66	1.37
Min.	64.06	35.25	32.50	3.91	1.06	26.45	74.19	4.00	3.33
Max.	143.01	141.49	64.71	5.86	1.42	37.50	87.44	10.82	8.60

† SCY=Seed cotton yield (Standardized); LY=Lint yield (Standardized); LP=Lint percentage (%); MIC=Micronaire; FL=Fiber length (inches); FS=Fiber strength (g/tex); UI=Uniformity index; ELO=Elongation percentages (%); SFC=Short fiber index.

The cultivar 8824-1-2-25-192-8 from the Delta region registered the highest SCY (43%) improvement over the check variety. With respect to fiber quality parameters, fine (3.91 MIC) and extra long (1.42 inch) fibers were observed in TAMB182-34, while PX3203-65-3 had strong fibers with an estimated value of 37.50 g/tex. The uniformity index showed a moderate range value of 74-87%, with a mean of 84%. Short fiber content (SFC) describes the amount of short fibers within a sample that are below half an inch in fiber length. Irrespective of the genotypes, SFC values ranged from 3.33-8.60 with a mean of 5.49.

Most of the Louisiana genotypes possessed coarse fiber (MIC 4.6-5.2), high uniformity (84%), medium staple (1.07-1.11 inch) and very strong fibers (30.22-32.9 g/tex). The heat tolerant genotypes bred and screened in the SW region showed much variability for the micronaire (4.15-5.23), fiber length (1.08-1.28 inch) and strength (31.46-37.50 g/tex).

Pearson correlation analysis identified significant positive and negative relationships among the phenotypic traits measured in this study (Fig 4.1 and Table 4.4). Significant negative correlation was observed between LY and FL, FS; between MIC with FL; and between ELO with FL and FS.

Significant positive correlations were detected between MIC and ELO and between FL and FS.

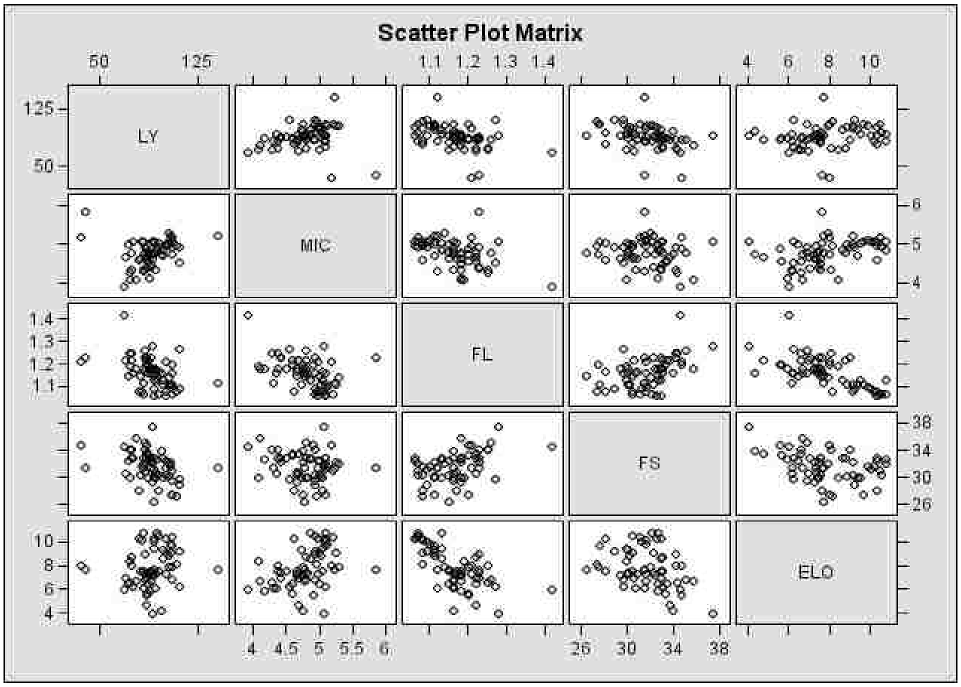


Fig 4.1 Scatter plot showing pair wise Pearson correlation coefficients among fiber traits

Significant negative correlations were found between SFC and both MIC and ELO.As expected, SFC had negative correlation with most of the fiber traits except for FS and FL.

Table 4.4 Phenotypic correlations (Pearson) for lint yield and fiber traits in upland cotton

	LY†	LP	MIC	FL	FS	UI	ELO	SFC
LY	1							
LP	0.308	1						
MIC	0.142	-0.00	1					
FL	-0.427**	0.08	-0.450**	1				
FS	-0.372*	-0.417**	-0.159	0.451*	1			
UI	-0.024	0.287	-0.127	0.410	0.105	1		
ELO	0.182	-0.049	0.382*	-0.663**	-0.374*	-0.151	1	
SFC	-0.128	-0.106	-0.418**	0.293	0.278	-0.101	-0.710**	1

† LY=Lint yield (standardized); LP=Lint percentage; MIC=Micronaire; FL=Fiber length (inches); FS=Fiber strength (g/tex); UI=Uniformity index; ELO=Elongation (%); SFC=Short fiber index.

4.3.2 Molecular Diversity Analysis

Information on the memberships of individuals in specific clusters and the relatedness of individuals are important in the characterization of a diverse group of genotypes. The pairwise kinship values were estimated based on 254 polymorphic AFLP markers using TASSEL software. The pairwise kinship values varied between 0.1-0.88 with average of 0.55 (Fig 4.2). Many of the genotypes under study shared common ancestral genotypes and 59% of the pairwise estimates were in the range of 0.6-0.88 indicating significant relatedness.

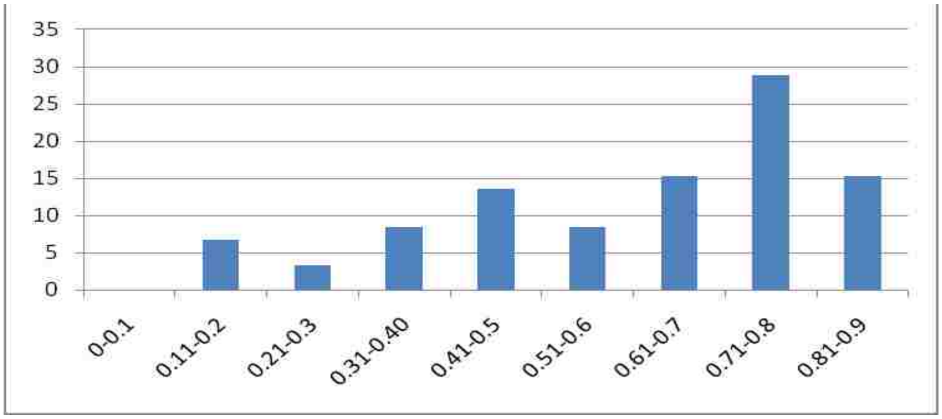


Fig 4.2 Percent kinship values among set of 60 Upland cotton genotypes as assessed by AFLP markers (X axis: percent kinship estimates; Y-axis: frequency).

The expected heterozygosity under Hardy-Weinberg genotypic proportions, also known as Nei's genetic diversity index, was 0.27 for the AFLP markers analyzed. The estimates of genetic diversity were in the range of 0.1–0.340 with an average of 0.23 (Fig 4.3).

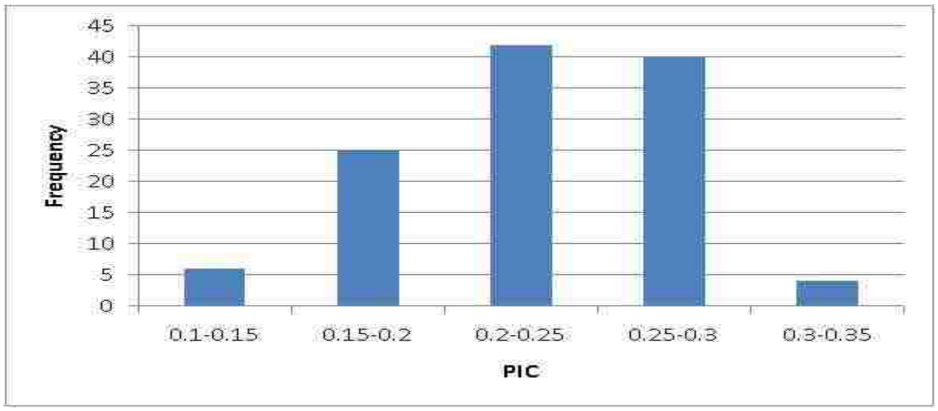


Fig 4.3 Polymorphic information content values for AFLP markers in a set of 60 upland cotton genotypes (X axis: polymorphic information content; Y-axis: frequency)

Around 80% of the AFLP markers showed a PIC range of 0.15-0.3. Earlier studies have reported polymorphic information content values in cotton of 0.05-0.82 with average of 0.31 (Liu *et* *al*., 2000) and between 0.08-0.89 with an average of 0.55 (Lacape *et al*., 2007). In the present study, we observed less diversity for the AFLP markers used. The narrow genetic composition of the genotypes in this study explains the lower mean observed here and yet is indicative of the efficiency of AFLP marker technology in capturing allelic diversity.

Genotypes can be grouped into clusters based on genetic similarity/dissimilarity matrices. Various graphical or tree based algorithms utilizing marker information can partition the genetic variability into single or multidimensional scales. Correspondence analysis is one such descriptive technique for investigating the association between markers and graphically displays the patterns in the data. In the present study, most of the genotypes formed a single cluster (Fig 4.4).

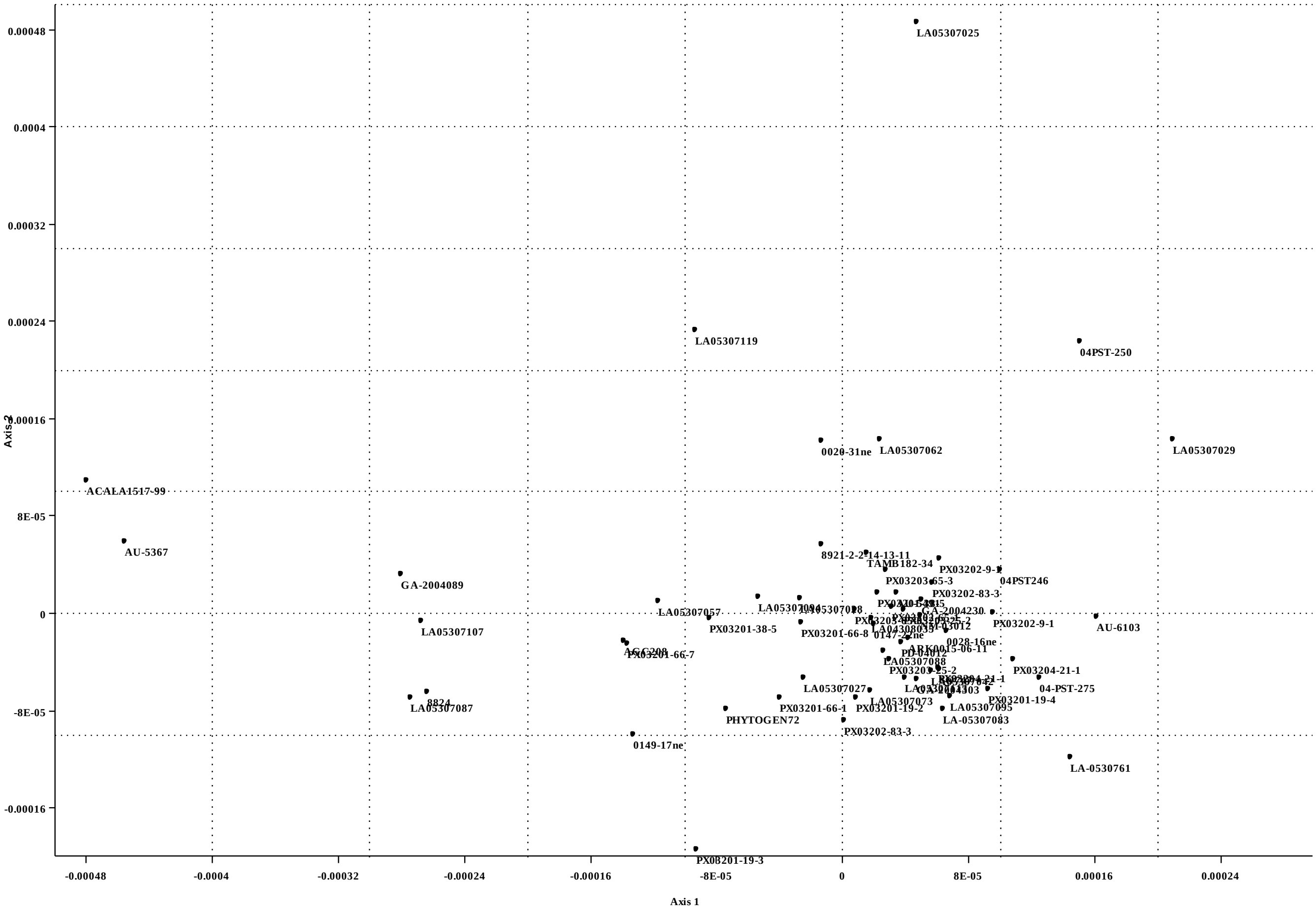


Fig 4.4 Correspondence analysis showing upland cotton genotypes using AFLP marker matrix. The plot was generated using PAST software using marker matrix (X-axis: dimension 1 and Y-axis: dimension 2).

One small cluster consisted of 8824, GA2004089, LA05307087 and LA 05307107. Some of the genotypes such as ACALA 1517-99, AU5367 and LA 05307025 were distinct outliers and more diverse from the other genotypes. In order to visualize the genetic relationships within the upland genotypes, Principal Coordinate Analysis (PCoA), based on genetic similarity matrices (Nei and Li, 1979) was used. The first two eigenvectors accounted for 63% of the variation observed. PCA (Fig. 4.5) again placed most of the genotypes into one cluster. The plot illustrated results very similar to the correspondence analysis. No obvious clustering was observed with respect to geographical origin of the genotypes under study.

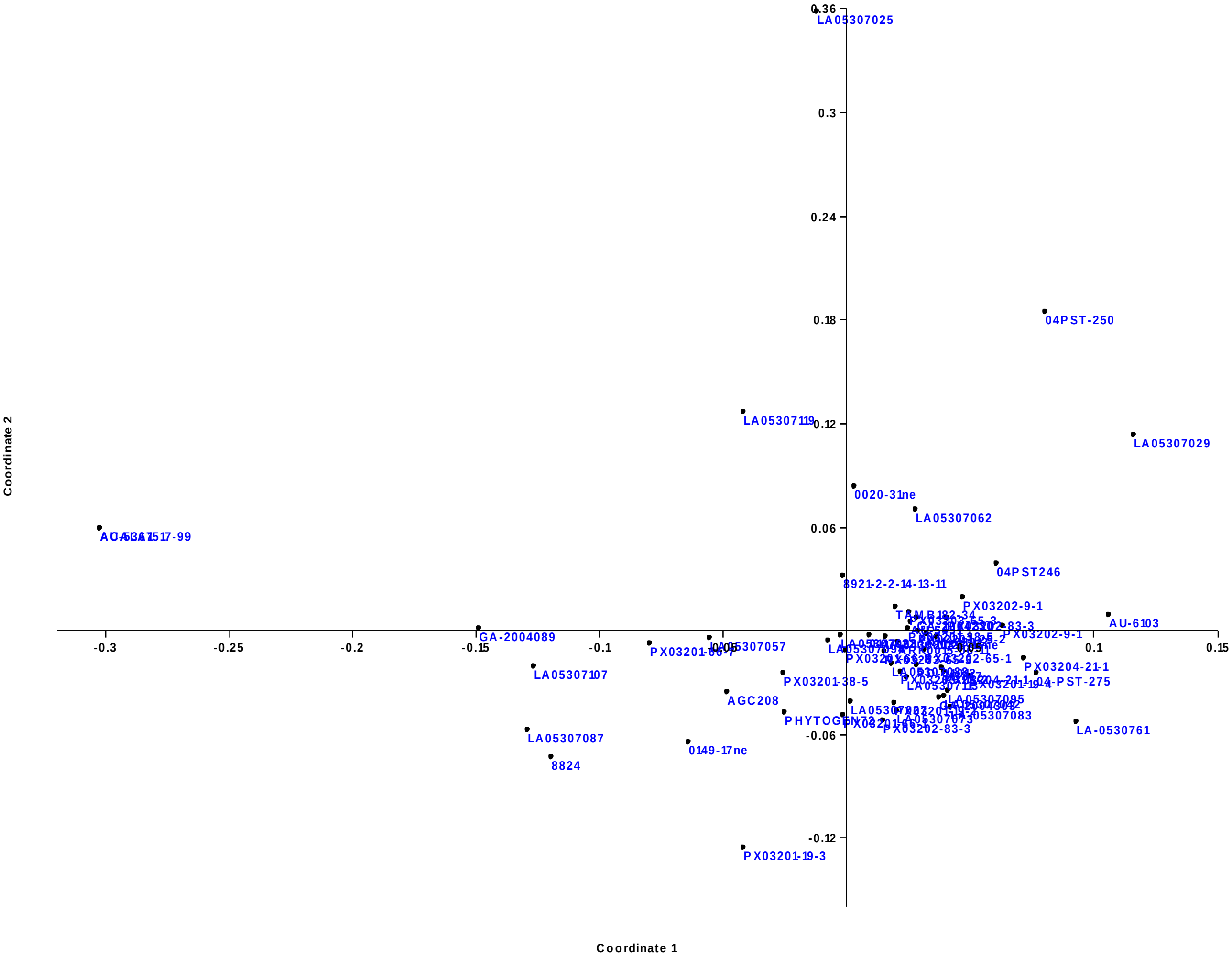


Fig 4.5 PCoA analysis of upland cotton genotypes assessed using DICE similarity coefficients in NTSYS software. X and Y axis describes coordinate 1 and 2 respectively.

An Unweighted Pair Group Method with Arithmetic mean (UPGMA) dendrogram of the 60 upland genotypes was constructed (Fig 4.6) based on the dissimilarity matrix using 234 AFLP markers. Ark-15-6-11 and 0147-22ne were highly similar, while AU 5367, ACALA 5367 and LA 05307025 were highly diverse from rest of the population. Most of the other genotypes were found to be genetically similar. The diverse Louisiana genotypes were LA5307025, LA5307029 and LA5307119, while among the heat tolerant group; PX3201-66-7, PX3201-19-3 and PX 3201-66-8 were the most diverse. Some of the publicly bred genotypes such as AU-5367, 04PST 250, 0020-31ne and GA2004089 were found to be dissimilar to rest of the genotypes. Nevertheless, the DICE distance estimates among all the 60 genotypes reached a maximum of less than 0.15, indicating the relative lack of genetic diversity in this group of genotypes as a whole.

4.3.3 Marker Trait Associations

a) Mixed Models for Marker Trait Associations

Association mapping using AFLP markers for LY and fiber traits was done using GLM and mixed models implemented in TASSEL (Bradbury *et al*., 2007). Initial analysis to detect population structure (Pritchard *et al*., 2000) did not find any clusters. The population was genetically related and formed only one group. Subsequently considered was PCA and kinship data in a mixed model. Initially the naïve model, comprising marker scores and trait data resulted in a low adj. R^2 (average 10.06%), while marker and PCA showed 16.9% adj. R^2. Including kinship data in both the mixed model and the simple model (marker+kinship) increased the adj. R^2 value to 29%. Inclusion of PCA data improved model R^2 value even further (41.7%). Using a MTPK (Marker+Trait+PCA+Kinship) model found significant associations between markers and quantitative traits (Table 4.5). As many as 112 markers were found to be significantly associated with the eight traits under study ($P<0.05$). Among all the traits, SFC had the highest number of associated markers (26), while UI was associated with the least number of markers (10).

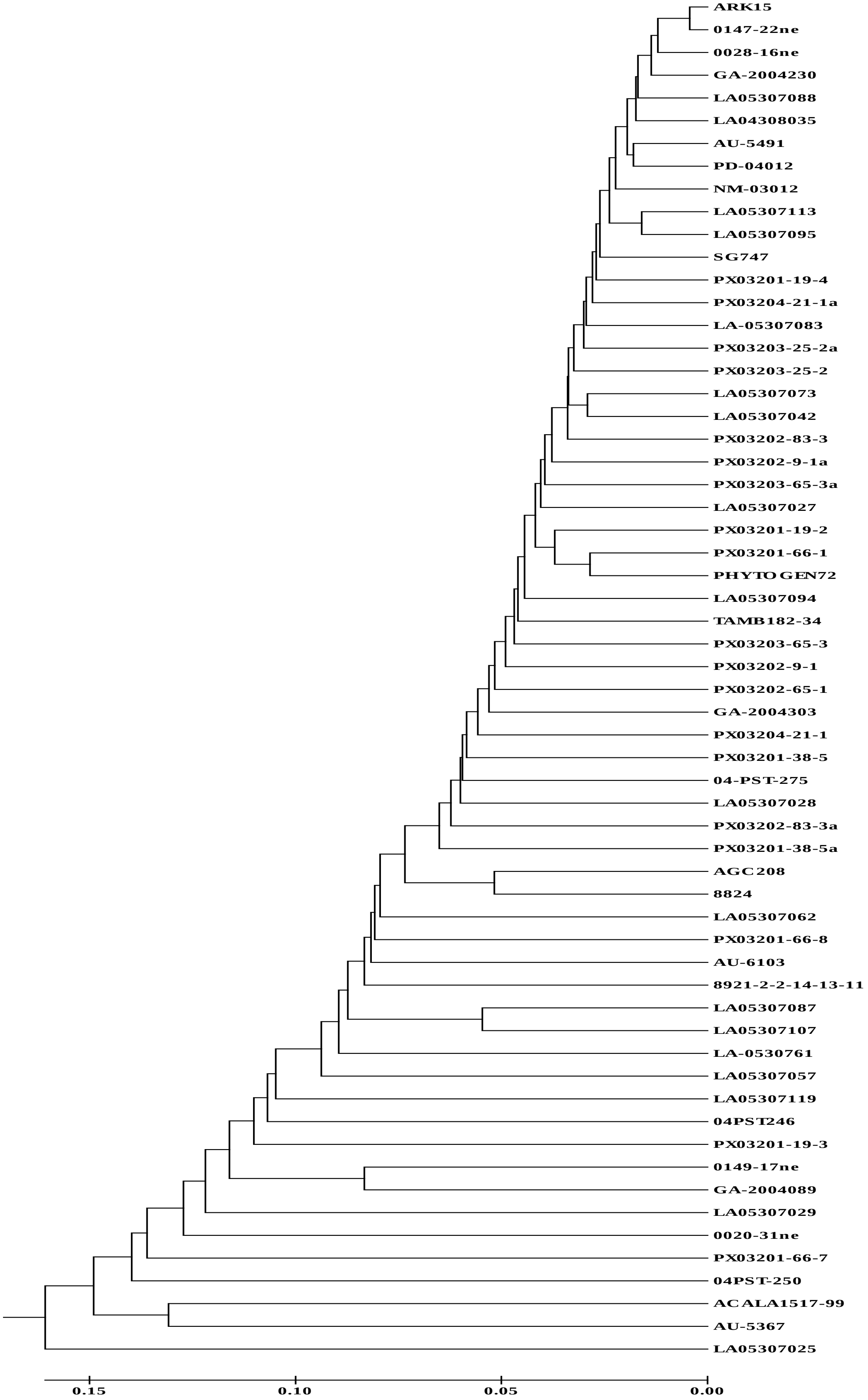


Fig 4.6 UPGMA dendrogram of 60 upland cotton genotypes based on DICE distance estimates calculated using AFLP markers.

Table 4.5 Significant markers selected from mixed model using AFLP markers in upland cotton. The models were evaluated in TASSEL software.

Trait	Significant QTLs selected based on MTPK mixed model
ELO†	E4M2_55, E4M2_70, E6M8_280, E9M5_70, E5M5_170,E5M4_365, E9M2_170, E1M5_65, E2M4_60, E9M3_185, E8M3_200
FL	E6M8_140, E9M5_265, E3M5_150, E5M5_70, E5M5_170, E5M4_365, E9M2_170, E9M7_135, E1M5_175, E2M8_265, E2M4_60, E9M3_185, E8M3_200
FS	E8M5_90,E3M5_70, E3M5_150, E5M5_145, E5M3_100, E5M3_160, E5M3_204, E5M2_70, E5M2_260, E8M2_65, E8M4_65, E1M7_170, E2M7_180, E9M7_135, E1M6_140
LP	E6M5_140, E8M5_90, E5M1_60, E5M5_70, E5M5_145, E5M3_100, E5M4_120, E9M2_150, E9M2_155, E8M2_85, E1M7_75, E1M7_115, E8M1_155, E8M3_60
LY	E6M5_140, E8M5_90, E5M1_60, E5M5_70, E5M5_145, E5M3_100, E5M4_120, E9M2_150, E9M2_155, E8M2_85, E1M7_75, E1M7_115, E8M1_155,E8M3_60
MIC	E6M2_255, E4M2_55, E6M8_140, E5M1_115, E5M1_170, E3M8_170, E5M8_120, E5M5_70, E3M3_60, E5M4_365, E3M2_150, E6M6_80, E6M6_100, E1M2_60, E9M7_135
SFC	E6M8_140, E6M1_130, E5M4_365, E4M7_140, E6M6_150, E8M2_85, E1M1_53, E9M3_300, E4M2_55, E4M8_350, E5M1_300, E5M5_170, E5M4_365, E9M2_180, E8M2_65, E2M4_60
UI	E6M2_370, E3M5_150, E5M4_270, E1M7_75, E1M7_185, E1M7_200, E1M8_55, E2M4_60, E6M3_65, E6M3_110

† LY=Lint yield; LP=Lint percentage; MIC=Micronaire; FL=Fiber length; FS=Fiber strength; UI=Uniformity index; ELO=Elongation ratio; SFC=Short fiber index.

Most of these markers stayed significant as progressed the analysis from the naïve to the MTPK model. Among the highly significant markers selected, E6M8_140 was associated with FL, MIC, LY and SFC. Other common markers were E3M5_150 for FL and FS, E5M3_204 for FS and LY, E8M2_85 for LP and LY and E2M4_60 for ELO, FL, FC and UI. As many as five markers were common between ELO and FL, viz., E9M2_170, E1M5_65, E2M4_60, E9M3_185, E8M3_200 and E5M4_365. The correlation between the lint yield and fiber properties could be the reason as the same set of markers were influencing the different traits.

b) Mixed – Multiple Regression Models for Association Analysis

Yu *et al*., (2006) commented on the efficiency of mixed models as well as on their ability to reduce the incidence of false positives. In order to further reduce the number of potential false positives and increase the efficiency of marker trait association models, we performed mixed-multiple regression (MMR) analysis. In this type of statistical analysis, all significant markers from the MTPK mixed model (from TASSEL) are validated under stringent statistical parameters using general linear models. The GLMSELECT (SAS) procedure was used as a MMR model selection procedure or a set of candidate models. We used 52 different general linear models with an array of CHOOSE, SELECT and STOP options and different model selection criteria : SBC, Adj. R^2, AIC, AICC, BIC and PRESS. The MMR method proved highly efficient in capturing most of the genetic variation with 38 significant markers (Table 4.6) for eight traits under the study. A total of 297 markers were identified by GLM and 108 by mixed models. After accounting for shared markers across yield and fiber traits, a total of 254 unique polymorphic markers were found and used in subsequent analyses. As most of the fiber traits are interrelated, we noticed several set of markers found common governing more than one fiber trait. The sequential validation of markers is an improved method for reducing false positives and identifying truly significant associations.

Table 4.6 Composition of the number of markers selected for yield and fiber traits by alternate marker-trait association models with range values for R^2

Traits	GLM	MTPK	MTPK-GLM
ELO	43(99%)	11(16-46%)	6 (17-42%)
LY	19 (99%)	14(86-93%)	5 (19-50%)
LP	21(99%)	14(42-56)	4 (30-56%)
FL	48(99%)	13(10-36%)	6 (18-56%)
FS	36(99%)	15(14-23%)	4(15-38%)
MIC	52(99%)	15(35-57%)	3 (12-21%)
UI	35(99%)	10(14-21%)	4 (12-36%)
SFC	43(99%)	16(26-53%)	6(15-57%)
Total	297	108	38

Mixed multiple regression models improved the efficiency of selection of significant markers associated with the fiber traits studied herein. Most of the markers had high R^2 values, to the extent of 12.4-57.4% (Table 4.7). Lint yield and lint percentage (LP), being the most complex dependent variables, were associated with five and four QTL's, respectively. The most significant QTL were E6M8_140 and E5M4_365 for LY and E5M5_145 and E1M7_75 for LP.

Micronaire, of the fiber traits under study, is the most affected by environmental factors. Three significant QTL's, viz., E5M8_120, E5M4_365 and E1M2_60 were associated with this character. Fiber length and ELO were associated with the common QTL's, E9M3_185 and E5M4_365. The QTL E5M5_170 was associated with both ELO and SFC traits. There were four markers associated with UI and the most significant one was E2M4_60.

Table 4.7 Marker trait associations in upland cotton using Mixed-Multiple regression models.

LY	**MODELR^2**	**ADJ.R^2**	**AIC**	**AICC**	**BIC**	**SBC**	**PRESS**	**Pr > F**
E6M8_140	0.20	0.19	294.69	5.95	295.26	298.87	8000.63	0.0001
E5M4_365	0.29	0.27	289.32	5.86	289.91	295.60	7636.79	0.008
E5M7_210	0.37	0.34	284.01	5.78	285.03	292.38	6982.78	0.009
E8M2_85	0.45	0.41	278.06	5.69	280.13	288.53	6433.68	0.007
E8M5_110	0.53	0.47	272.85	5.62	277.22	287.51	6077.18	0.02
LP								
E5M5_145	0.30	0.29	237.93	5.00	238.09	242.12	3209.48	<0.0001
E1M7_75	0.56	0.53	215.55	4.65	217.41	226.02	2192.21	0.002
E5M3_100	0.39	0.37	231.15	4.89	231.24	237.43	2844.80	0.004
E5M1_60	0.47	0.45	224.59	4.79	225.09	232.97	2531.83	0.005
MIC								
E5M8_120	0.31	0.27	-149.66	-1.44	-146.71	-141.29	4.947	0.006
E5M4_365	0.12	0.10	-138.69	-1.27	-137.22	-134.51	5.855	0.007
E1M2_60	0.21	0.18	-143.59	-1.34	-141.70	-137.31	5.380	0.011
FL								
E5M5_70	0.18	0.17	-333.22	-4.51	-333.20	-329.03	0.236	0.001
E5M4_365	0.31	0.28	-341.45	-4.64	-341.43	-335.16	0.202	0.00
E9M3_185	0.39	0.36	-347.55	-4.74	-347.23	-339.17	0.180	0.00
E3M5_150	0.47	0.43	-353.41	-4.83	-352.30	-342.93	0.166	0.00
E6M8_140	0.53	0.49	-358.82	-4.91	-356.44	-346.26	0.155	0.01
E9M5_265	0.57	0.52	-361.60	-4.94	-357.98	-346.94	0.150	0.01
FS								
E8M4_65	0.15	0.13	89.36	2.53	90.10	93.55	261.93	0.00
E5M3_100	0.26	0.24	82.76	2.42	83.72	89.04	234.41	0.00
E1M7_170	0.38	0.33	76.31	2.33	78.25	86.78	209.78	0.02
E5M3_204	0.32	0.28	79.78	2.38	81.03	88.16	222.45	0.03

Contd…..

UI	**MODELR^2**	**ADJ.R^2**	**AIC**	**AICC**	**BIC**	**SBC**	**PRESS**	**Pr > F**
E2M4_60	0.12	0.10	51.95	1.90	52.99	56.14	173.83	0.00
E3M5_150	0.20	0.17	48.33	1.85	49.43	54.58	169.38	0.02
E1M7_75	0.30	0.26	42.22	1.75	44.27	50.80	156.24	0.00
E1M8_55	0.35	0.31	39.28	1.71	41.91	49.67	150.87	0.03
SFC								
E5M5_170	0.17	0.16	28.42	1.51	29.25	32.61	93.86	0.00
E5M4_365	0.27	0.25	22.46	1.42	23.49	28.74	84.83	0.00
E4M8_350	0.33	0.29	19.66	1.38	20.98	28.03	80.80	0.03
E5M1_300	0.38	0.33	16.97	1.34	18.88	27.44	74.91	0.03
E4M2_55	0.43	0.37	14.31	1.30	17.16	26.88	70.89	0.04
ELO								
E5M4_365	0.15	0.13	54.01	1.94	54.07	58.20	146.36	0.00
E5M5_170	0.26	0.24	47.22	1.83	47.18	53.50	130.81	0.00
E1M5_65	0.34	0.31	42.43	1.75	42.52	50.81	119.37	0.01
E9M2_170	0.43	0.39	35.60	1.65	36.55	46.07	107.88	0.00
E4M2_70	0.48	0.43	32.18	1.60	33.97	44.75	99.83	0.02
E9M3_185	0.57	0.50	26.81	1.55	32.10	45.66	95.80	0.04

4.4 Discussion

The present study explores the efficiency of AFLP markers in capturing phenotypic variability using association mapping principles. The high genetic similarity between the genotypes included in this study is attributed to the use of common ancestral genotypes in the breeding programs. Narrow genetic diversity has also been observed in other cotton association mapping studies (Abdurakhmonov *et al*., 2008; 2009).

Progress in using breeding approaches to improve fiber quality traits is dependent upon exploiting genetic variability. Genetic diversity studies on *G. hirsutum* germplasm collections from Africa, Uzbekistan and Mexico regions identified diversity for fiber traits within the germplasm. Cluster analysis also suggested that diversity remains in the PeeDee germplasm collection following 50 years of breeding (Campbell *et al*., 2009).

Genetic diversity studies conducted previously in *Gossypium* species, inferred from isozyme, random amplification of polymorphic DNA (RAPDs), restricted fragment length polymorphism (RFLPs), amplified fragment length polymorphism (AFLPs), and SSRs data have reported a low level of molecular diversity within *G. hirsutum* cotton germplasms (Abdurakhmonov, 2007). Our results obtained from genetic distance analysis confirmed the narrow genetic base among elite *G. hirsutum* cotton genotypes. Zeng *et al*., (2009) attributed moderate allele frequency divergence (0.11-0.27) among six groups of upland genotypes to be due to natural selection for fitness among exotic genes in the local environment. The range of genetic similarity in present study is much higher than the previous reports (Guiterez *et al*., 2002; Rahaman *et al*., 2002; Zhang, 2005). Based on PCA, correspondence and UPGMA analysis it is evident that distantly related primary gene pool members or secondary gene pool of the cotton have been utilized in the development of the 60 upland genotypes studied here.

According to Abdalla *et al*., (2001), one possible explanation for low genetic diversity is that selfing (following hybridization) will result in the decrease in the number of loci that are polymorphic in subsequent generations by 50%. In addition to creating a set of closely related descendent genotypes, various markers would have independently become fixed to one or the other parental allele. Thus, high levels of similarity within upland cluster could be due to the fact that these genotypes have been subjected to a greater degree of inter-cultivar gene flow (Kellogg *et al*., . 1996; Wendel and Doyle 1998).

Refining the MLM approach of Yu *et al*., (2006), we considered the use of PCA and kinship estimates to eliminate spurious associations. This approach identified a number of AFLP markers significantly associated with yield and fiber traits. Improvement upon the MLM approach in our study came from multiple regression based GLM studies. The MLM-MR approach reduced the number of significant markers. The general linear method has been used before in cotton, with molecular markers, where it reduced the number of significant markers by 6-21%. This study for the first time explored the MLM-MMR statistical approaches using AFLP markers in cotton.

In line with the present study, Wu *et al*., (2007) observed a large number of AFLP markers strongly associated with yield, boll weight and lint percentage. Here only a few AFLP markers were selected using linear regression models. Out of an original set of 297 significant markers for eight cotton traits, the addition of PCA and kinship data reduced this number to 108. Using MLM-MMR approach, the number of significant markers was reduced even further to 38. Zhang *et al*., (2009) also reported that the specification of additional criteria can reduce the number of significant QTLs identified. This is the first report of such model based selection criteria being applied to AFLP data in cotton.

4.5 Conclusion

The narrow genetic base of upland cotton germplasm that is used in breeding programs is one of the factors in failing to achieve appreciable amount of progress in improving yield and fiber traits. The present investigation attempts to determine efficiency of AFLP markers in estimating genetic diversity in 60 Upland accessions of Louisiana. Genetic distance analysis confirmed the narrow genetic base among *G. hirsutum* genotypes. The PCA and kinship estimates in MLM approach identified number of significant AFLP markers associated with yield and fiber traits. The MLM-MMR approach using AFLP markers found to be useful in reducing the false positives and improving reliability of the data.

4.6 References

Abdalla AM, Reddy OUK, El-Zik KM and Pepper AE 2001 Genetic diversity and relationships of diploid and tetraploid cottons revealed using AFLP. Theoretical and Applied Genetics, 102, 222– 229.

Abdurakhmonov IY, Kohel RJ, Yu JZ, Pepper AE, Abdullaev AA, Kushanov FN, Salakhutdinov IB, Buriev ZT, Saha S, Scheffler BE, Jenkins HN and Abdukarimov A 2008 Molecular diversity and association mapping of fiber quality traits in exotic *G. hirsutum* L. germplasm, Genomics 92(6), 478-487.

Abdurakhmonov IY, Saha S, Jenkins JN, Zabardast T, Burie Shukhrat, Shermatov, Scheffler BE, Pepper AE, Yu JZ, Kohel RJ and Abdukarimov A 2009 Linkage disequilibrium based association mapping of fiber quality traits in *G. hirsutum* L. variety germplasm. Genetica 136, 401–417.

Bowman DT, May OL and Calhoun DS 1996 Genetic base of upland cotton cultivars released between 1970 and 1990. Crop Science, 36:577–581.

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

Bridge JM and Mc Gowan R. 2005 Cotton cultivar DP 393, US patent 6930228, Date issued 16th October 2005.

Campbell BT, Williams VE and Park W 2009 Using molecular markers and field performance data to characterize the Pee Dee cotton germplasm resources. Euphytica, 169, 285-301.

Greenacre MJ 1984 Theory and Applications of Correspondence Analysis. Academic Press, London.

Gutierrez OA, Basu S, Saha S, Jenkins JN, Shoemaker DB, Cheatham CL and McCarty JC Jr 2002 Genetic distance among selected cotton genotypes and its relationship with F2 performance. Crop Science, 42,1841–1847.

Hammer, Harper DAT, Ryan PD 2001 PAST: Paleontological statistics software package for education and data analysis Palaeontologia Electronica, vol 4, issue 1, art 4.

Hill M, Witsenboer H, Zabeau M, Vos P, Kesseli R and Michelmore R 1995 PCR-based fingerprinting using AFLPs as a tool for studying genetic relationships in *Lactuca* spp. Theoretical and Applied Genetics, 93: 1202–1210.

Iqbal MJ, Aziz N, Saeed NA and Zafar Y 1997 Genetic diversity evaluation of some elite cotton varieties by RAPD analysis. Theoretical and Applied Genetics, 94,139-144.

Kellogg EA, Appels R and Mason-Gamer RJ 1996 When genes tell different stories: the diploid genera of Titicaceae (Gramineae). Systematic Botany, 21, 321–347.

Kumar S, Tamura K and Nei M 2004 MEGA4: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinformatics, 5(2), 150–163.

Lacape JM, Dessauw D, Rajab M, Noyer JL and Hau B 2007 Microsatellite diversity in tetraploid *Gossypium* germplasm: assembling a highly informative genotyping set of cotton SSRs. Molecular Breeding, 19, 45–58.

Lacape JM, Nguyen TB, Thibivilliers S, Bojinov B, Courtois B, Cantrell RG, Burr B and B. Hau 2003 A combined RFLP–SSR–AFLP map of tetraploid cotton based on a *Gossypium hirsutum* ×*Gossypium barbadense* backcross population, Genome, 46, 612-626.

Liu S, Cantrell RG, McCarty JC Jr and Stewart JM 2000 Simple sequence repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Science, 40, 1459-1469.

Lu HJ and Myers GO 2002 Genetic relationships and discrimination of ten inXuential Upland cotton varieties using RAPD markers. Theoretical and Applied Genetics,105:325–331.

Maughan PJ, Saghai Maroof MA, Buss GR and Huestis GM 1996 Amplified fragment length polymorphism (AFLP) in soybean: species diversity, inheritance, and near-isogenic line analysis. Theoretical and Applied Genetics ,93, 392–401.

Meredith WR Jr 2000 Cotton yield progress – why has it reached a plateau, Better Crops, 84: 6-9.

Multani DS and Lyon BR 1995 Genetic fingerprinting of Australian cotton cultivars with RAPD markers. Genome 38, 1005-1008.

Myers GO, Baogong J, Akash MW, Badigannavar AM and Saha S 2009 Chromosomal assignment of AFLP markers in upland cotton (*Gossypium hirsutum* L.) Euphytica, 165(2), 391-399.

NAAS USDA, 2009 National Agricultural Statistical Survey. (http://www.nass.usda.gov)

Nei M and Li WH 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of National Academy of Sciences, USA 76, 5269–5273.

Peakall R and Smouse PE 2006 GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 6, 288-295.

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S and Rafalski A 1996 The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers forgermplasm analysis. Molecular Breeding 2, 225-238.

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

Rahman M, Hussain D and Zafar Y 2002 Estimation of genetic divergence among elite cotton cultivars-genotypes by DNA Wngerprinting technology. Crop Science 42, 2137-2144.

SAS Institute, 2008 Documentation GLMSELECT. SAS Institute, Cary, NC.

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

Tanksley SD and Hewitt J 1988 Use of molecular markers in breeding for soluble solids content in tomato - a re-examination. Theoretical and Applied Genetics 75, 811 - 823.

TASSEL 2009 User manual, trait analysis by association, evolution and linkage. www.maizegenetics.net/tassel.

Tatineni V, Cantrell RG and Davis DD 1996 Genetic diversity in elite cotton germplasm determined by morphological characteristics and RAPD. Crop Science, 36, 186–192.

Van Becelaere G, Lubbers EL, Paterson AH and Chee PW 2005 Pedigree- vs DNA marker-based genetic similarity estimates in cotton. Crop Science 45, 2281–2287.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M,Freijters A, Pot J, Peleman J, Kuiper M and Zabeau M 1995 AFLP : a new technique for DNA fingerprinting. Nucleic Acids Research, 23, 4407–4414.

Wallace T, Bowman D, Campbell BT, Chee P, Gutierrez OA, Kohel RJ, McCarty J, Myers GO, Percy R, Robinson F. Smith W, Stelly DM, Stewart JM, Thaxton P, Ulloa M and Weaver DB 2009 Status of the USA cotton germplasm collection and crop vulnerability. Genetic Resources and Crop Evolution, 56(4), 507-532.

Wendel JF and Brubaker CL 1993 RFLP diversity in *Gossypium hirsutum* L. and new insights into the domestication of cotton. American Journal of Botany, 80:71.

Wendel JF and Doyle JJ 1998 Phylogenetic incongruence: window into genome history and molecular evolution. In: Soltis DE, Soltis PS, Doyle JJ (eds) Molecular systematics of plants. II.DNA Sequencing. Kluwer Academic, Boston Dordrecht London, pp 265-296.

Wu J, Jenkins JN, McCarty JC, Zhong M and Swindle M 2007 AFLP marker associations with agronomic and fiber traits in cotton. Euphytica. 153, 153–163.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S and Buckler ES 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics, 38, 203–208.

Zabeau M and Vos P 1993 Selective restriction fragment amplification: a general method for DNA fingerprinting. European Patent Application number: 92402629.7, Publication Number EP 0534858.

Zeng L, Meredith WR Jr, Gutiirrez OA and Boykin DL 2009 Identification of associations between SSR markers and fiber traits in an exotic germplasm derived from multiple crosses among *Gossypium* tetraploid species. Theoretical and Applied Genetics 119(1), 93-103.

Zhang ZS, Hu MC, Zhang J, Liu DJ, Zhang J, Zhang K, Wang W and Wan Q 2009 Construction of comphrensive PCR based marker linkage map and QTL mapping for fiber quality traits in upland cotton (*G. hirsutum*, L.). Euphytica, 24: 49-61.

Zhang J, Lu Y, Cantrell RG and Hughs E 2005 Molecular marker diversity and field performance in commercial cotton cultivars evaluated in the Southwestern USA. Crop Science 45,1483–1490.

CHAPTER 5 CHARACTERIZATION AND MARKER TRAIT ASSOCIATIONS OF SEED QUALITY TRAITS IN UPLAND COTTON (*Gossypium hirsutum*)

5.1 Introduction

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 relatively high level of palmitic acid provides a degree of stability to the oil that makes it suitable for high-temperature frying applications, but is nutritionally undesirable due to the low-density lipoprotein cholesterol-raising properties of this saturated fatty acid (Cox *et al*., 1995). Cottonseed oil is one of a few oils that are stable in the beta-prime crystal form, which is desirable in most solidified products as it promotes a smooth, workable consistency usually called plasticity.

Cottonseed meal is left after oil extraction and used as a source of fodder protein in the livestock industry, but the sphere of its use in agriculture is limited. Constituting nearly half of a seed's weight, the meal contains 23% of high biological-value protein. The presence of bound gossypol in proteins considerably changes their properties, including their biological value. The gossypol in cottonseed feed products could be toxic to some animals in certain situations. Some of the classical signs of chronic gossypol toxicity are loss of appetite, weakness, emaciation, weight loss, decreased egg size and hatchability in poultry. These symptoms have been observed consistently in non-ruminants and occasionally in young ruminants or in mature ruminants with very high free gossypol intakes. The ability of ruminants to tolerate higher oral doses of gossypol than non ruminants is due to the binding of free gossypol by soluble ruminant proteins (Hudson *et al*., 1988).

The fractionation of various protein components of the meal has shown that the amount of gossypol bound with the proteins depends on their amino acid composition and structure. Therefore, the primary task in the technology of obtaining cottonseed proteins is the fraction of proteins containing different amounts of gossypol. For years, scientists have tried to breed cotton with gossypol levels safe for consumption. In the 1950s they succeeded, but because the toxin was missing from leaves as well as seeds, the plants were more susceptible to damage from pests. With the help of a new technique called RNA interference, or RNAi, a gene-silencing mechanism has been developed that lowered the gossypol level in seeds while sparing the rest of the plant (Ganesan *et al*., 2006).

Edible cottonseed has a higher protein efficiency ratio (PER = 2.35) than other vegetable proteins. It contains 64 g of protein per 100 g of edible cottonseed compared to 24 g of protein in beef. It contains all nine essential amino acids, is extremely high in potassium, is a rich source of complex carbohydrates, and contains only polyunsaturated oil. Its calcium-phosphorous ratio is considered ideal for building tissue for bone formation. Whole cottonseed is high in protein, fat, fiber and energy. This combination of nutrients in one feedstuff is unusual. Whole cottonseed with the lint still attached is white and fuzzy in appearance. The typical cottonseed meal is composed of moisture (7%), ash (6.6%), protein (45.3%), fiber (6.3%), nitrogen-free extract (24.6%) and fat (10.2%). In order to balance the oil, protein and fiber content in the existing germplasm/cultivars, there is a need to survey the genome to identify genes/controlling elements responsible for these metabolic pathways.

Protein and oil concentration, kernel index and kernel 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; Singh *et al*., 1985; Ye *et al.,* . 2003). Seed traits may be simultaneously controlled by seed nuclear genes, cytoplasmic genes and maternal nuclear genes (Ye *et al*., 2003). Previous studies showed 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 these traits in conventional cotton breeding programs. Genetic mapping provides a useful tool to understand the genetic architecture of quantitative traits at the molecular level. DNA markers linked to quantitative trait loci (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; Pe´rez-Vich *et al*., 2004), oilseed mustard (Gupta *et al*., 2004) and canola (Hu *et al*., 2006). In cotton, 11 single QTL's were found associated with oil and protein content (Song and Zhang 2007). Amino acid specific epistatic QTL's were also detected, which explained 4.43-9.55% of the phenotypic variation. A recent study using chromosome substitution lines identified chromosome 4 of the 3-79 in a *G.barbadense,* introgressed TM-1 background, to be associated with seed oil, protein and fiber percentage (Wu *et al*., 2009). None of the studies in cotton, to date, have explored the possibility of screening a broad array of germplasm for molecular marker associations with these traits using association/LD principles.

The present study was planned to identify and map genomic regions associated with seed protein, seed oil and fiber content in a diverse collection of upland cotton cultivars. The study also explores the extent of genetic variability present in upland cultivars to facilitate selection of these in traits in introgression breeding.

5.2 Materials and Methods

5.2.1 Plant Material

A set of 75 *G. hirsutum* upland cotton genotypes and 2 diploid genotypes were selected for analyzing seed quality traits. (Table 5.1). The entire upland mapping panel was divided into five groups based on their geographical origin viz., Louisiana (25), Arkansas(17), SE (22), Delta (4), Texas/SW(6). In addition two diploid subgenomes representatives' *G. arboreum* (A_1) and *G. herbaceum* (A_2) we also considered for comparison.

Table 5.1 List of genotypes used for analyzing seed quality traits in upland cotton

Cultivar	Region	Cultivar	Region
LA1110001	Louisiana	AU-5491	South eastern
LA1110147	Louisiana	AU1065	South eastern
LA1110148	Louisiana	AU1107	South eastern
LA03404204	Louisiana	AU1403	South eastern
LA01407117	Louisiana	AU5210	South eastern
LA01407009	Louisiana	AU6207	South eastern
LA1110023	Louisiana	AU-6103	South eastern
LA1110035	Louisiana	AU-5367	South eastern
LA03404148	Louisiana	GA2002212	South eastern
LA03404171	Louisiana	GA2003118	South eastern
LA03404065	Louisiana	GA2003156	South eastern
LA1110061	Louisiana	GA3003131	South eastern
LA01407074	Louisiana	GA-2004089	South eastern
LA01407072	Louisiana	GA-2004303	South eastern
LA04307004	Louisiana	GA-2004230	South eastern
LA04307074	Louisiana	PD03001	South eastern
LA04307063	Louisiana	PD03011	South eastern
LA1110014	Louisiana	PD3025	South eastern
LA03404051	Louisiana	PD99036	South eastern
LA04308044	Louisiana	PD99041	South eastern
LA04307027	Louisiana	PD-04012	South eastern
LA-05307083	Louisiana	COKER100	South eastern
LA05307029	Louisiana	DPL393	Delta
LA-0530761	Louisiana	DP393	Delta
LA05307094	Louisiana	SG105	Delta
9801-37-04	Arkansas	SG747	Delta
9811-15-07	Arkansas	ACALA1517-99	South west
9815-05-09	Arkansas	FM958	Texas
9803-17-04	Arkansas	NM-03012	South west
9803-23-04	Arkansas	TAMB182-34	Texas
9801-37-04	Arkansas	TM-1	Texas
0015-06-11	Arkansas	MCNAIR235	Texas
0147-22ne	Arkansas	PX03203-25-2	South west
0110-2ne	Arkansas	*G. arboreum*	Diploid
0141-15ne	Arkansas	*G. herbaceum*	Diploid
0020-31ne	Arkansas		
0028-16ne	Arkansas		
0149-17ne	Arkansas		
8921-2-2-14-13-11	Arkansas		
04-PST-275	Arkansas		
04PST-250	Arkansas		
04PST-246	Arkansas		

Most of the genotypes, except the historical ones were selected from advanced breeding lines tested in the Regional Breeder's Trial Network (RBTN), a multistate testing program of public breeding lines covering cotton producing regions (www.cottonrbtn.com). Plants were field grown in 2008 as per LA cooperative extension service guidelines at the Dean Lee Research Station in Alexandria, LA. Leaf samples from representative plants were collected and bulked for DNA extraction. Phenotypic data on yield was obtained from the RBTN trial website (www.cottonrbtn.com). The four replication data on seed cotton yield and lint percentage was averaged to calculate variances using SAS (SAS 9.1.3, SAS Institute, Cary, NC). Deltapine, DP 393 was considered as the check variety and all the comparisons were made in relation with the performance of this cultivar. For lint yield, the values of other CAM panel were adjusted based on the relative performance of the check variety, DP393.

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

5.2.2 Genotyping with AFLP Markers

Sixty four primer combinations were used to generate Amplified Fragment Length Polymorphism (AFLP) data (Table: 5.2a) following the procedure given by Vos *et al.,* (1995) with minor modifications. Sample DNA was digested with *EcoRI* (infrequent cutter with GAATTC recognition sequence) and *MseI* (frequent cutter with TTAA recognition sequence) restriction enzymes and oligonucleotide adapters specific to restriction sites were ligated to the resulting fragments through incubation (180 min, 37 °C) with DNA ligase using a iCycler (BioRad Labs, Hercules, CA.).

Table 5.2a Adapters and primers of AFLP marker system used for pre and selective amplification in upland cottons.

Primer/adapter	Nomenclature†	Sequence (5'-3')
ECORI primers:		
EcoRI linker 1	E-I	CTC GTA GAC TGC GTA CC
EcoRI linker 2	E-II	AAT TGG TAC GCA GTC TAC
EcoRI + A	E+A	GAC TGC GTA CCA ATT CA
E- AAC	E1	GACTGCGTACCAATTCAAC
E- AAG	E2	GACTGCGTACCAATTCAAG
E-ACA	E3	GACTGCGTACCAATTCACA
E-ACT	E4	GACTGCGTACCAATTCACT
E-ACC	E5	GACTGCGTACCAATTCACC
E-ACG	E6	GACTGCGTACCAATTCACG
E-AGG	E8	GACTGCGTACCAATTCAGG
E-AGA	E9	GACTGCGTACCAATTCAGA
MseI primers:		
MseI linker 1	M-I	GAC GAT GAG TCC TGA G
MseI linker 2	M-II	TAC TCA GGA CTC AT
MseI + C	M+C	GAT GAG TCC TGA GTA AC
M-CAA	M1	GATGAGTCCTGAGTAACAA
M-CAC	M2	GATGAGTCCTGAGTAACAC
M-CAG	M3	GATGAGTCCTGAGTAACAG
M-CAT	M4	GATGAGTCCTGAGTAACAT
M-CTA	M5	GATGAGTCCTGAGTAACTA
M-CTC	M6	GATGAGTCCTGAGTAACTC
M-CTG	M7	GATGAGTCCTGAGTAACTG
M-CTT	M8	GATGAGTCCTGAGTAACTT

† : Nomenclature is in accordance with the Lacape *et al.*, 2003; Myers *et al.*, 2009.

Pre-amplifications were done using *EcoR* I+A and *Mse* I+C oligo primers. The amplification was carried out with 50ng/ul of oligo primers, 5mM dNTP's, 25mM $MgCl_2$, 10X buffer, Taq polymerase(5U/µl) and restrict ligated template DNA making total volume of 20µl. The PCR was set up with initial denaturing for 94^{o}C for 2 min followed by 26 cycles at 94^{o}C for 1 min, 56^{o}C for 1 min., 72^{o}C for 1 min., and final extension at 72^{o}C for 5min. The pre amplified products were diluted with ddH_2O. Selective amplification was done using two selective nucleotides. The EcoRI+ANN oligo primers were dye labeled with 700 and 800 IR dye (MWG Biotech, Germany). The PCR for selective amplification was carried out in a reaction volume of 10 µL consisting of 10X reaction buffer, 25 mM $MgCl_2$, 2.5 mM dNTPs, 1 µM each of EcoRI-ANN and MseI+CNN primers and 5U *Taq* polymerase (Promega, Madison, WI). The reactions were run on an *i*-Cycler (BioRad Labs, Hercules, CA). Touchdown PCR was used for selective amplifications using the following profile: initial denaturing step at 94^oC for 2 min followed by initial 12 cycles at 94^oC for 30 s, 65^oC for 30 s (with 0.7^oC decrement every cycle) and 72^oC for 1 min, then followed by 23 cycles at 94^oC for 30 s, 56^oC for 30 s, and 72^oC for 1 min with a final extension step at 72^oC for 2 min. A total of 64 *EcoR* I - *Mse* I selective amplification primer combinations were used. The PCR amplified products were run on a LI-COR 4300 sequencer (LI-COR Inc., Lincoln, NE). Gels images were saved onto a computer, printed and scored manually. Presence of a band was recorded as '1' and absence as '0', as per a typical dominant marker system. Ambiguous data that could not be resolved was discarded. The nomenclature of AFLP loci was followed according to Lacape *et al.,* (2003); and Myers *et al*., (2009), indicating the enzyme primer combinations with band size.

5.2.3 Molecular Diversity Analysis:

For each marker used, sub-populationwise diversity statistics including the number of observed and effective alleles, Nei's genetic distances, expected heterozygosity and Shannon's information index were calculated using GenAlEx 6.2 software (Peakall and Smouse, 2006). Allelic diversity at a given locus can be determined by Polymorphism Information Content (PIC) and it was calculated as '$PIC=1-\sum f_i^2$ where, f_i is the frequency of the i^{th} 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, NC).

Genetic differentiation among the subpopulation was estimated using hierarchial analysis of molecular variance (AMOVA; Excoffier *et al*., 2005) via GenAlEx 6.2. In order to identify possible structure, various statistical analyses were performed on the basis of allelic frequencies. First, the Dice similarity coefficient was calculated using the formula D = 2a/(2a + b + c), where *a* = 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 data were calculated for each pair of genotypes (distance =1- similarity) and used for UPGMA clustering in MEGA 4.0 (Kumar *et al.,* . 2004). In addition, Principal Coordinate Analysis (PCoA) was also performed using a genetic similarity matrix based on the formula of Nei and Li (1979) to supplement the findings obtained from cluster analysis. All the above analyses were performed employing Paleontological Statistics (PAST) software (Hammer *et al*., 2001).

A Bayesian model based clustering was performed using the software program Structure according to Pritchard *et al*., (2000). The main criteria for this type of clustering is the allocation of individual genotypes to groups in such a way that Hardy-Weinberg equilibrium and linkage disequilibrium are valid within clusters but absent between clusters. The admixture model was selected in the software and allele frequencies among populations were assumed to be correlated. Each run was carried out using 100,000 iterations with 100,000 burn-in iterations. The optimum number of cluster (k) was determined based on the estimated logarithmic likelihood of the data (Yu *et al*., 2006). This value reaches a plateau when the minimum number of groups that best describes the population structure has been reached (Pritchard *et al*., 2000; Evanno *et al*., 2005). In addition, alpha values were also monitored to assess the minimum number of subpopulation. The alpha value becomes lowest and starts to plateau. The minimum number of subpopulation at this stage would be the ideal k value. A graphical display of subpopulation composition from Structure software was generated using DISTRUCT (Rosenberg, 2002).

5.2.4 Association Analysis

a) Mixed Models for Association Mapping

Six statistical mixed models (Table: 5.2b) and their adjusted R^2 values were computed for fixed marker effects using TASSEL 2.1, beta version (Bradbury *et al.*, 2007). Several models incorporated the outcome of population structure and PCA analysis in an effort to increase the power to detect true marker trait associations. Tests for significance were calculated using the F statistic associated with the marker. The model possessing the highest adjusted R^2 was considered the best since it captured the maximum variation. A cutoff P value of 0.05 was used to determine whether a QTL was associated with a marker. R^2 estimates were used to calculate the magnitude of the QTL effects. Most of the marker trait associations were made based on 77 genotypes.

Table 5.2b Mixed models designed for association mapping of seed quality traits in upland cottons using TASSEL software.

Code	Model	Statistical equation†
MT	Marker+Trait	$Y=A_{\acute{\alpha}}+e$
MTS	Marker+Trait+Structure	$Y=A_{\acute{\alpha}}+Q_{v}+e$
MTP	Marker+Trait+PCA	$Y=A_{\acute{\alpha}}+Q_{v}+e$
MTK	Marker+Trait+Kinship	$Y=A_{\acute{\alpha}}+Z_{u}+e$
MTSK	Marker+Trait++Structure+Kinship	$Y=X_{ß}+ A_{\acute{\alpha}}+ Q_{v}+Z_{u}+e$
MTPK	Marker+Trait+PCA+Kinship	$Y=X_{ß}+ A_{\acute{\alpha}}+ Q_{v}+Z_{u}+e$

†: Y = vector of phenotypic observations, ά= vector of allelic effects, e=vector of residual effects, ν=vector of population effects, ß=vectors of fixed effects other than allelic or population group effects, u=vector of polygenic background effects, Q=population membership assignment matrix, X, A and Z are incidence matrices of 1s and 0s relating to y to ß, ά and u(Casa *et al*., 2008).

b) Mixed – Multiple Regression Models for Association Analysis

The GLMSELECT in SAS performs effect selection in the framework of general linear models. A variety of model selection methods are available, offering extensive capabilities for customizing the wide variety of selection and stopping criteria. The GLMSELECT compares most closely to PROC REG and PROC GLM. The PROC REG procedure supports a variety of model-selection methods but does not support a CLASS statement. The GLM procedure supports a CLASS statement but does not include effect selection methods. The GLMSELECT procedure fills this gap. It focuses on the standard independently and identically distributed general linear model for univariate responses and offers great flexibility for and insight into the model selection algorithm.

In order to exploit the advantages of multiple regression procedures, we used all those traitwise significant markers selected by mixed model procedures using TASSEL and screened for 52 PROC GLMSLECT models. Stepwise selection method was used with all possible combinations of CHOOSE, SELECT and STOP. Different options were used for these selection methods i.e., Bayesian Information Content (BIC), SBC (Schwarz Bayesian Information Criterion), Adjusted R^2, AICC (the Corrected Akaike Information Criterion), SL=0.15 (the significance level of the F statistic for entering or departing effects) and Cross validation (CV). Traits were considered as dependent variables and all the markers were treated as independent variables. Each trait was analyzed separately and those independent variables showing test statistic estimates of less than P=0.05 were added in the model. To reduce the Type I error, selected models were further tested with validation step by using ‘PRESS’ criterion in ‘STOP’ option. The best model was then selected based on adjusted R^2 and the fewest number of effects for a particular trait.

Following simple GLM and MLM in TASSEL, Mixed multiple regressions in GLMSELECT enormously improved the efficiency of statistical model selection in order to cull out false positives and increasing the power to detect QTL.

5.3 Results

5.3.1 Genetic Analyses

A total of 64 ECoRI-MseI primer combinations were screened across 77 cotton genotypes and 234 polymorphic fragments were scored. Based on the prior knowledge and the confirmation of 5 subgroups via Structure analysis, several genetic diversity parameters were calculated. The

Shannon Index, a measurement used to compare diversity between two or more subpopulations, ranged between 0.45-0.61 (Table: 5.3). The number of effective alleles was highest for Arkansas (1.7) while lowest for Delta genotypes (1.5). The heterozygosity for the AFLP markers ranged from 0.318 (DELTA) to 0.43 (ARK). ARK and SE showed the highest heterozygosity among all the subgroups studied. The pairwise Nei genetic similarity between upland genotypes ranged from 0.822 between DELTA and LA to 0.948, between SE and ARK subgroups (Table: 5.4). Across different subpopulations, we observed moderate to low genetic diversity.

Table 5.3 The genetic diversity parameters for five subgroups in upland cotton genotypes.

Pop	Na	Ne	I	He	UHe
LA	1.885	1.613	0.507	0.347	0.354
ARK	1.966	1.795	0.615	0.431	0.444
SE	1.966	1.788	0.610	0.427	0.437
DELTA	1.709	1.589	0.453	0.318	0.353
SW/T	1.748	1.637	0.484	0.341	0.364

Na=No. of different alleles, Ne=No. of effective alleles, I=Shannon's index, He=Expected heterozygosity, UHe=Unbiased expected heterozygosity, LA=Louisiana, ARK=Arkansas, SE=South Eastern, SW/T=South West/Texas

The frequency distribution values for relative kinship revealed that the relatedness ranged from 0-0.9 (Fig: 5.1). 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 often share common genotypes in their breeding development programs. The polymorphic Information Content (PIC) measures how different populations are distinguished based on probability of randomly chosen alleles. The frequency distribution for PIC using AFLP markers ranged from 0-0.40 with more than 90% of them falling between 0.15-0.40 (Fig: 5.1).

Table 5.4 Pairwise Population Matrix of Nei Genetic identity among upland cotton genotypes

	LA†	ARK	SE	DELTA	SW/T
LA	1.000				
ARK	0.928	1.000			
SE	0.908	0.948	1.000		
DELTA	0.822	0.863	0.881	1.000	
SW/T	0.875	0.898	0.907	0.864	1.000

†: LA=Louisiana; ARK=Arkansas; SE=South eastern; SW/T= South western or Texas

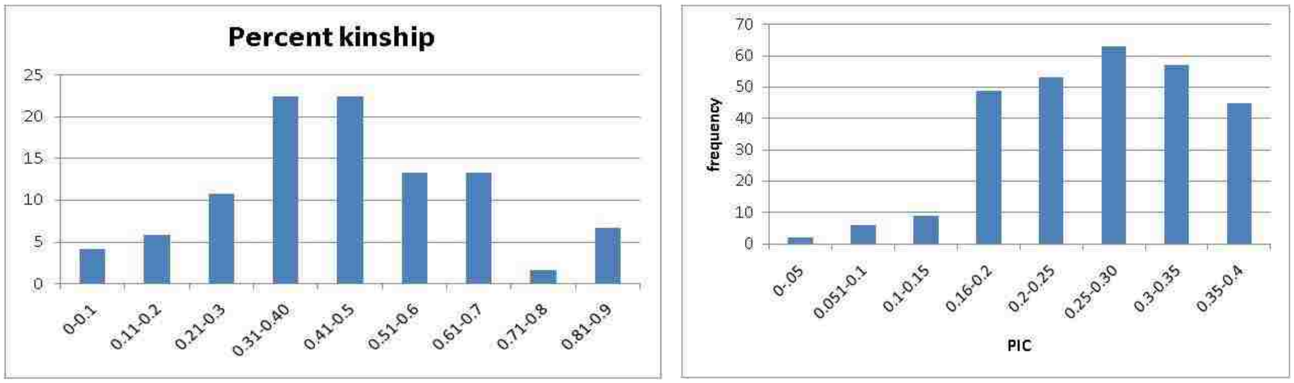


Fig 5.1 Frequency distribution for percent kinship and PIC estimates for AFLP markers in upland cotton genotypes. X axis: percent kinship and PIC estimates respectively; Y axis: frequency values.

5.3.2 Phenotypic Analyses

The cotton upland genotypes considered for this study was comprised of 75 upland and 2 diploid elite germplasm lines developed by breeding programs covering five relatively distinct geographical regions. Data on the yield components SCY and LP were collected from RBTN coordinators and averaged across four replications. The mean values were used to perform univariate analysis. The standardized seed cotton yields ranged from 64% (GA-2004089) to 139.16% (LA1110001) with a mean of 97.11 (Table: 5.5). Lint percentage varied from 35.67% to 57.35% with average of 42.97%. These two traits showed considerable genetic variance among the upland cottons. Looking at seed traits, the seed protein content ranged from 18.05% to 28.45% with an average of 23.4%, oil content ranged from 6.47% to 25.16% with an average of 17.86%, while fiber content varied between 15.88% to 37.12% with an average of 20.23%.

Table 5.5 Univariate analysis of yield and seed quality traits in upland cotton genotypes

Traits	N†	Min.	Max.	Mean	SE	Variance	SD	Median
Protein	77	18.05	28.45	23.4	0.47	16.73	4.09	24.1
Oil	77	6.47	25.16	17.86	0.36	10.07	3.17	17.96
Fiber	77	15.88	37.12	20.23	0.36	10.12	3.18	19.54
SCY	77	64.52	139.16	97.11	1.79	226.75	15.06	97.34
LP	77	35.67	57.35	42.97	0.57	22.67	4.76	41.53

†: N= Number of genotypes, SE=Std. error, SD=Std. deviation; LP=lint percentage; SCY=seed cotton yield

The frequency distribution graphs for lint percentage and quality traits were presented in Fig: 5.2. A majority of the germplasm lines and genotypes showed a LP ranging between 37.6-45%.

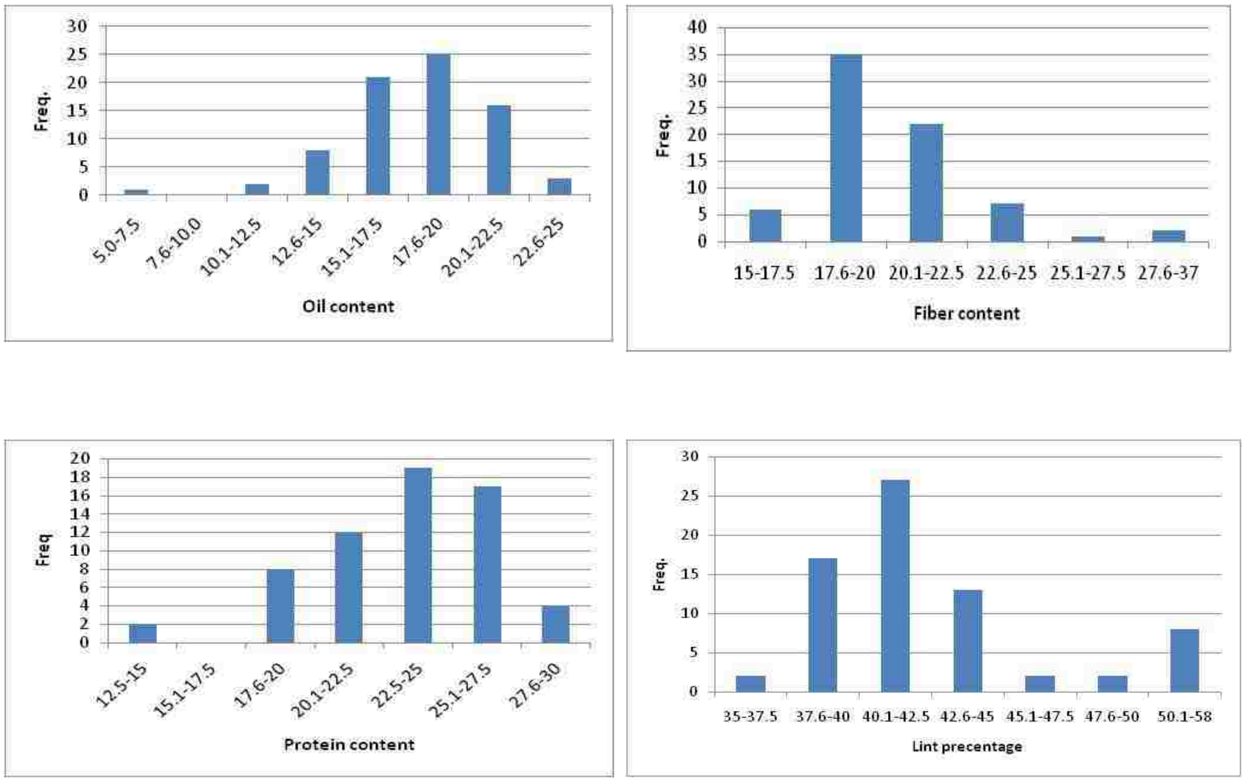


Fig 5.2 Frequency distribution for lint percentage and seed quality traits in upland cotton genotypes. X axis: oil content (%); fiber content (%); protein content (%) and lint percentage (%) respectively. Y axis: frequency values.

Generally, high lint percentages favor more seed cotton/boll and smaller seed (seed index) than does low lint percentage. The major frequency classes for oil content were between 12.6-22.5%, for fiber content were between 17.6-25% and for oil content were between 17.6-27.5% across all the genotypes studied. Only a few extreme peaks were observed.

The correlations among the yield and quality traits are graphically represented in Table: 5.6 and Fig: 5.3. There were a significant negative correlations between fiber content with oil and protein percentage. While not significant, Protein and oil percentages were negatively correlated, leading to the fact that both cannot be balanced in a single cultivar. All other correlations, particularly those between SCY and LP with seed quality traits were not significant. Of these,

Table 5.6 Correlation coefficients among yield and seed quality traits of upland cottons

Traits	**SCY†**	**LP**	**Protein**	**Oil**	**Fiber**
SCY	1				
LP	-0.074	1			
Protein	0.045	-0.240	1		
Oil	-0.040	0.027	-0.224	1	
Fiber	0.268	0.033	-0.340*	-0.61**	1

* significant at P≤0.05; ** significant at P≤0.01, †: SCY=seed cotton yield; LP=lint percentage

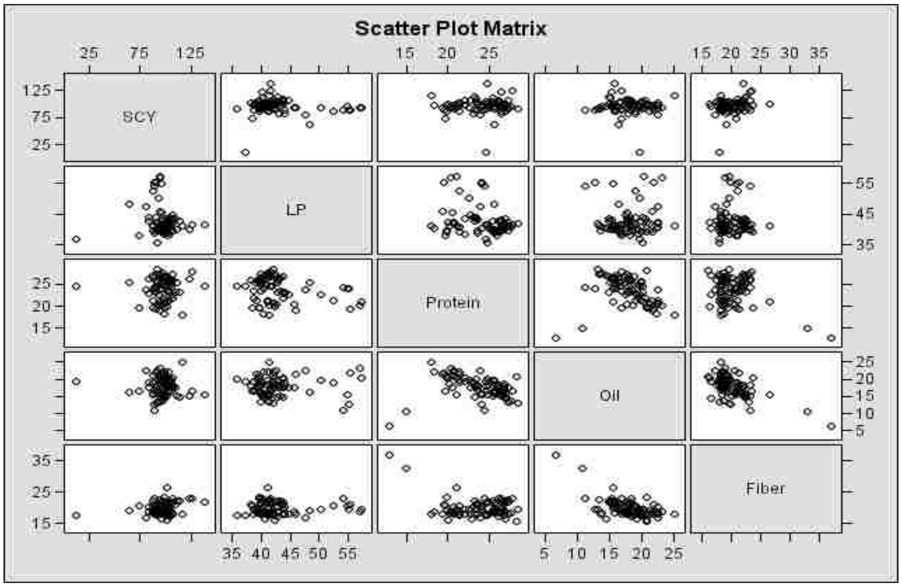


Fig 5.3 Scatter plot showing correlations among yield and seed quality traits in upland cotton

however, two were relatively large; SCY with fiber (0.268) and LP with Protein (-0.240). Typically high yielding cotton has a high LP which is most easily achieved by decreasing seed size (G. Myers, personal communication). In this study, fiber content was determined from hulled seeds. The hull is expected to be higher in fiber than the embryo, so as seed size decreases (SCY increases) there is a positive correlation with precent fiber. Similarily, since a majority of seed protein is in the embryo, as lint percentage increases (smaller seed), it is expected that protein percentages would decrease.

5.3.3 AMOVA and Cluster Analysis

In order to estimate genetic diversity within and among the predefined subpopulations, we calculated Wright's F_{ST} index (Table: 5.7). In addition, an estimate molecular variance present in the upland genotypes using 234 AFLP markers using AMOVA test (Table: 5.8) was done. Based on the pairwise F_{ST} estimates, SE and SW/T (South Western/Texas) was very closely related (0.0095), while Delta and LA was highly diverse (0.141). The average estimate of F_{ST} was 0.0529 indicating a low level of genetic differentiation among groups. The AMOVA also revealed that although most of the genetic diversity was attributable to differences within populations (94%), there was still some variation among groups (6%). The DICE distances among individuals were plotted in a two-dimensional graph using PCoA analysis (Fig: 5.4).

Table 5.7 Pairwise F_{ST} values estimated based on Weir and Cockerham (1984) approach for five subgroups of upland cottons.

F_{ST}	LA†	ARK	SE	DELTA	SW/T
LA	0				
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	0

†: F_{ST}=Wright's fixation index; LA=Louisiana; ARK=Arkansas; SE=South eastern; SW/T=South west-Texas

Table 5.8 Analysis of Molecular Variance (AMOVA) among and within subgroups of upland genotypes

Source	df†	SS	MS	Estimated Variance	%variance
Among Pops	4	270.901	67.725	2.280	6%
Within Pops	72	2495.775	34.664	34.664	94%
Total	76	2766.675		36.944	100%

†: df=degrees of freedom; SS=Sum of square; MS=mean sum of square

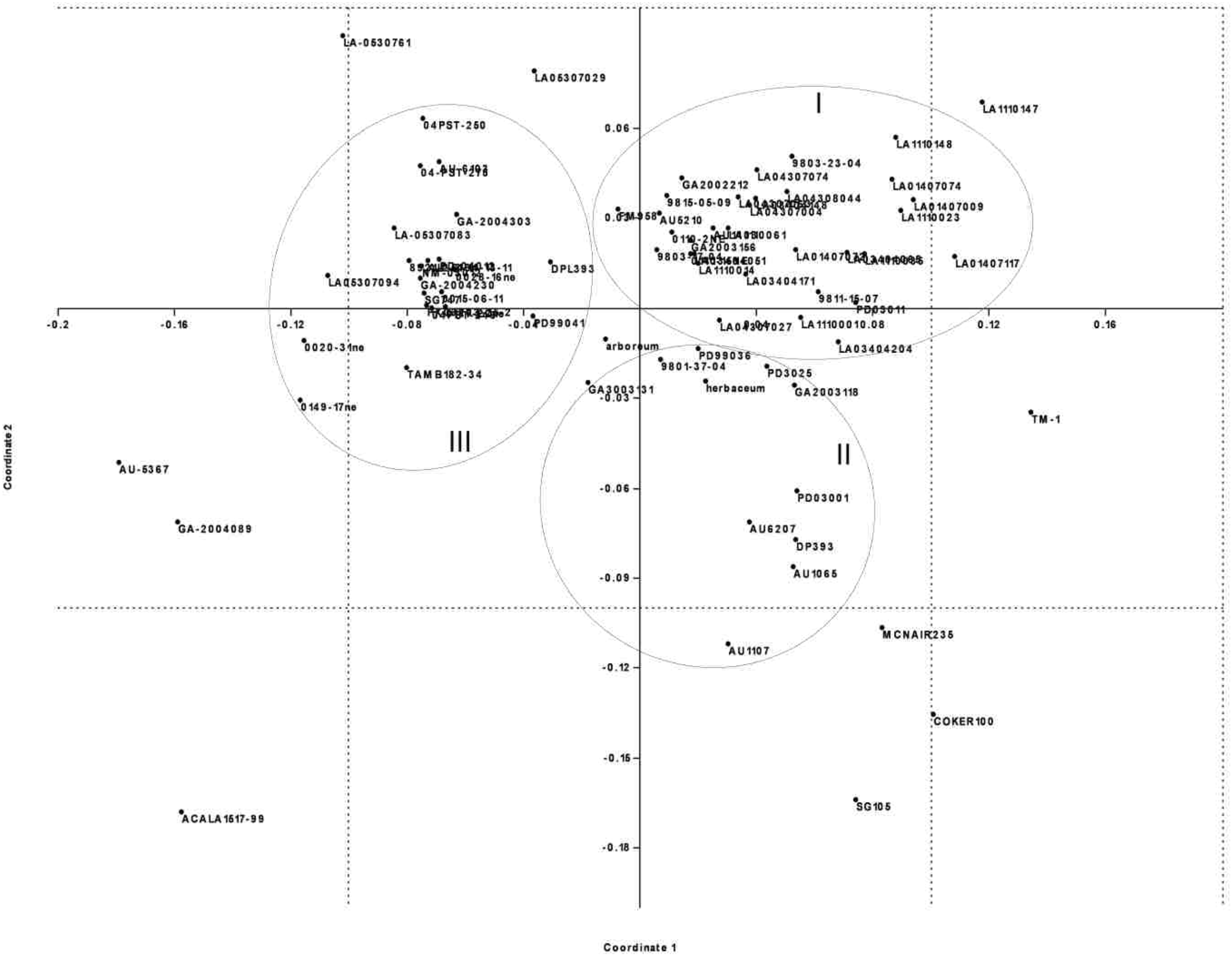


Fig 5.4 PCoA based on DICE similarity coefficients using AFLP markers in upland cotton genotypes. The PCoA was constructed using PAST software, which formed distinct three clusters. X and Y axis specify co-ordinate 1 and 2 respectively.

The first two co-ordinates explained 49% of the genetic variation. The general grouping did not clearly establish the separation of samples according to the geographical origin of each population. Most of the LA genotypes grouped in top right side (I) with some Arkansas genotypes too. The second cluster consisted of SE genotypes, except GA-2004089. The third cluster on left side (III) comprised of Arkansas and few representatives from LA, SE, Delta and SW groups. Some diverse genotypes like LA111047, LA0530761, Acala1517-99 moved away from any of the designed clusters. The diploid species, *G. herbaceum* was located in second cluster, while *G. arboreum* was seen in the middle of the three clusters. Overall the PCoA did not give clear separation of genotypes.

Results of the DICE genetic distances and cluster analysis are presented in the form of a dendrogram in Fig: 5.5. On the basis of the DICE coefficients, the 77 genotypes can be classified into two major groups, one comprised of most of the LA, SE, ARK genotypes, while the other groups consisted of diverse LA, SE and SW/Texas genotypes along with the outgroup, diploid species. Genotypes such as GA-2004089, LA0530761 and 04-PST-275 were highly diverse compared to rest of the upland genotypes.

5.3.4 Population Structure

In order to assess the levels of genetic structure within the five identified clusters, the estimate of posterior distribution of pairwise Wright's F_{ST} (Wright, 1951), a measure of the genetic variance among populations was also calculated using 100,000 iterations (Fig: 5.6). F_{ST} values between all groups were significant ($P<0.001$) and ranged from 0.2 to 0.53, supporting the existence of genetic structure.

For the AFLP data, the clustering of genotypes using STRUCTURE did produce a clear discrimination of the genotypes into predefined groups with some exceptions. The population structure analysis revealed that LnP(D) estimates increased with increase of k up to k=4 and then suddenly dropped and continued to increase again leading to plateau at k=6 (Fig: 5.6). There could be a possibility of either k=4 or 5 in this population. Going with prior information of k=5 based on

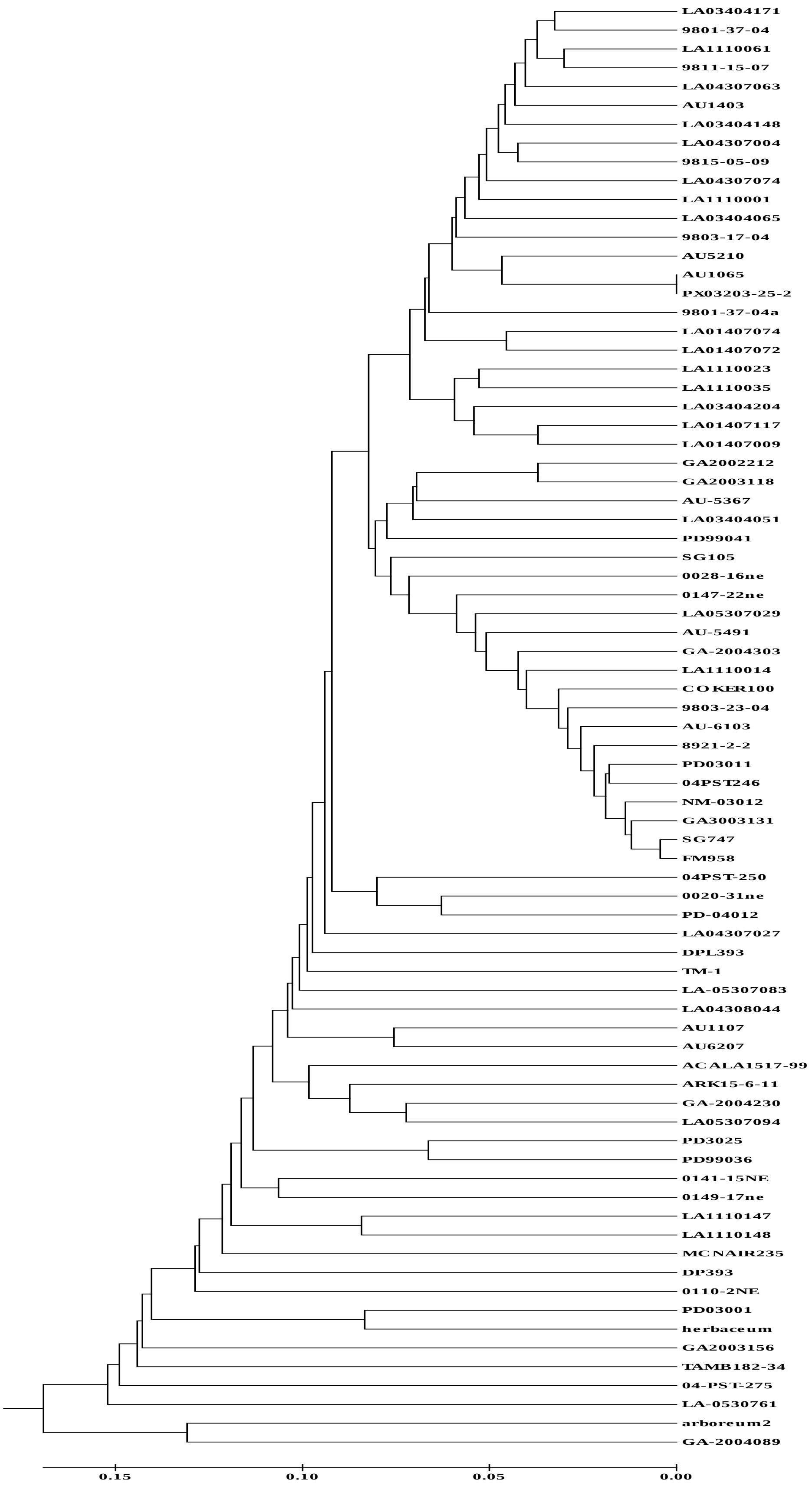


Fig 5.5 UPGMA based dendrogram of 77 upland cotton genotypes estimated using DICE distances

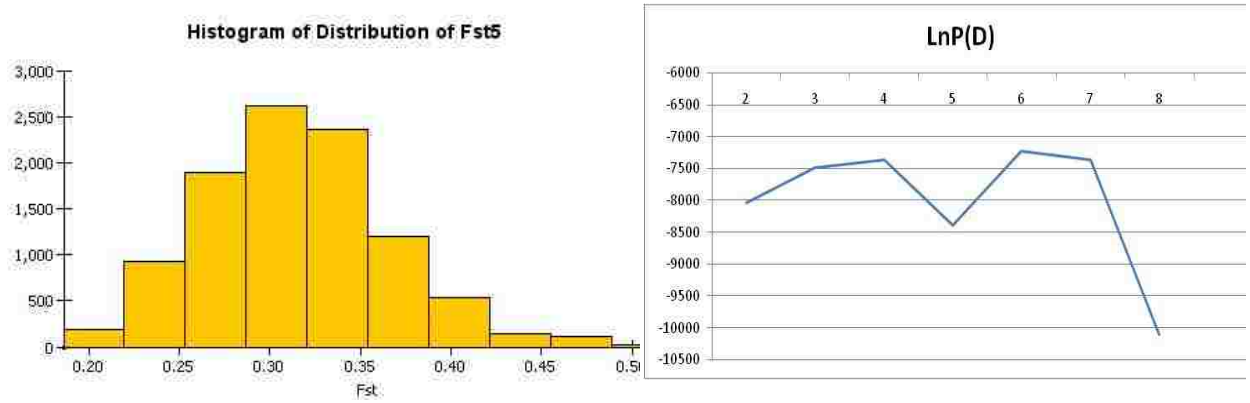


Fig 5.6 Distribution of pairwise F_{ST} values (k=5) and Posterior probability, lnP(D) of the data as function of the number of subpopulations(k), where k ranged from 2-8.

geographical grouping, we decided to set k=5 for all our future analysis. The bar plot diagram on population structure is presented in Fig: 5.7.

The barplot indicated LA genotypes showing uniformity with less admixture, mainly from Delta, SW/T and Ark ancestral genes. Similar is the case with ARK, and SE clusters too. Although the bars indicate that some genotypes have a genetic background with a large fraction from one of the five predefined subpopulations, substantial intermixing between the groups was evident.

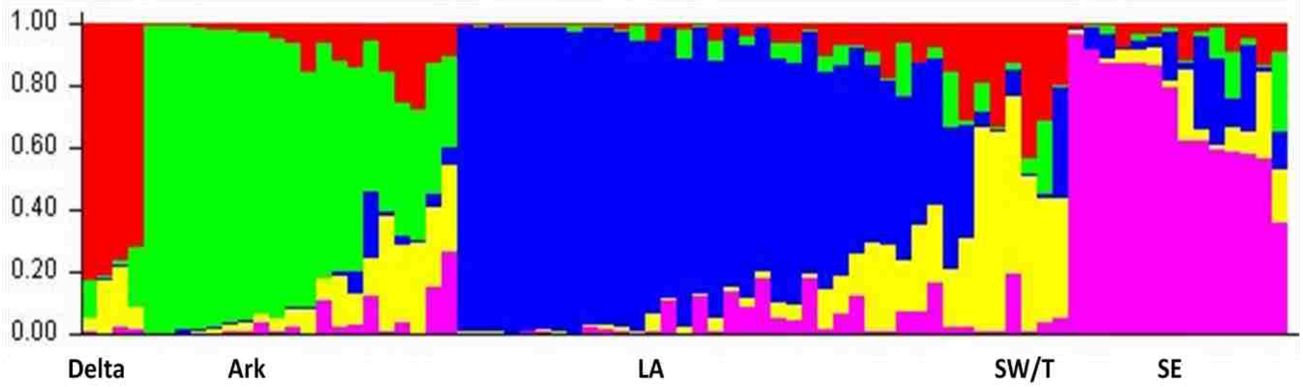


Fig 5.7 Bar plot representing population structure of five subgroups of upland cotton. Each individual genotype is represented by a line partitioned in five colored segments that represent estimated membership fractions too each one of the five subgroups. The bar plot was generated using Structure (Pritchard *et al*., 2000) software following admixture model.

5.3.5 Association Analyses

The population structure and kinship analysis is important to check spurious associations and minimize Type I error in association mapping. We tested the performance of five different association models in controlling for false positives or spurious associations (Fig: 5.8). The models studied were: 1) model that did not control population structure or relatedness; naïve (MT), 2) model that accounted for either PCA (MTP) or 3) population structure (MTS), 4) a naïve mixed model(MTK), 5) a model with kinship and or population structure (MTSK) and 6) a kinship model with PCA (MTPK). The pairwise relatedness of each indivuals based on allelic information was obtained through TASSEL software. The relative performance of each model was evaluated based on the extent of genetic variation explained by them (model R^2). The MTSK model was found efficient among all those studied in explaining the highest genetic variation in phenotypic trait values. The model was able to describe 30-50% of variation for the seed quality traits and 60% for the lint percentage. Over all, MTSK model gave the best fit with the fewest effects (markers) and high model R^2. Therefore, we selected MTSK as the mixed model for determining marker trait associations.

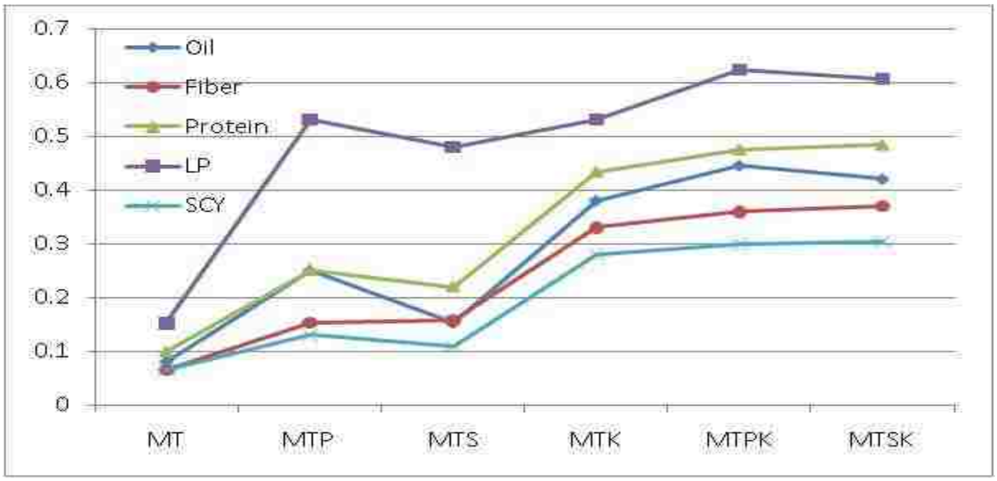


Fig 5.8 Performance of the mixed models based on the proportion of genetic variation explained (model R^2) in upland cotton. The mixed models were designed in SAS using PROC GLMSELECT statistics. X axis: Models selected; Y axis: Model R^2. M=marker; T=trait; S=structure; P=eigenvalues of PCA; K=kinship estimates

Association analysis identified marker trait associations ($P<0.05$) from MTSK mixed model for all the seed quality traits evaluated, viz., SCY, LP, protein, oil and fiber content (Table: 5.9). The MTSK model identified 45 significant markers ($P<0.05$) for five traits. Traits such as SCY and

LP were associated with five and eight markers respectively, while seed quality traits, oil, protein and fiber content were found significantly associated with 12, 15 and 5 markers respectively. Markers such as E3M5_255, E6M1_218 and E3M6_260, E4M1_365 were strongly associated with SCY and LP respectively with high adj. R2. The seed quality traits oil and protein content were governed by common markers; E4M3_255, E4M3_218, E6M2_595 and E3M3_60. Fiber content was governed by 5 markers, with E4M4_177 and E5M7_195 having highly influential.

Table 5.9 Significant QTLs (P<0.05) for yield and seed quality traits in upland cotton. The QTLs were identified using mixed model (MTSK) of TASSEL software.

Traits	Significant QTL's identified					
SCY	E3M5_255,	E6M1_218	E6M3_473,	E5M1_55,	E5M7_158	
Lint %	E3M6_260, E6M4_249,	E4M1_365, E5M3_65	E4M4_242,	E4M2_440	E3M3_255,	E3M8_305,
Oil	E4M3_255, E3M3_130,	E6M4_341, E4M2_206,	E3M7_370, E3M3_60,	E4M3_200, E3M2_145,	E6M2_595, E6M2_364,	E4M3_218, E5M7_70
Protein	E4M1_382, E5M2_642, E5M7_180,	E5M3_230, E4M3_255, E4M3_218,	E6M2_320, E6M3_285, E6M1_196	E6M2_595, E3M7_210,	E4M3_440, E5M4_170,	E3M3_60, E4M3_245,
Fiber	E4M4_177,	E5M7_195,	E5M6_170	E5M1_395,	E5M6_130	

In order to further validate markers selected from mixed models, multiple regression using 52 mixed-multiple regression models was performed. Mixed multiple regression supposedly reduces false positives by simultaneously comparing all the markers in stepwise regression. Among the 52 MLM-MMR models under study, high Adj. R^2 with minimum effective QTL's were selected from a model with CHOOSE=Adj.R^2, SELECT=AdjR^2 and STOP=Adj.R^2. The other models produced low R^2 values with high number QTL's, which was unreliable.

As many as 14 significant markers were identified for five traits using Mixed-MMR approach (Table: 5.10). For SCY, E5M1_55 and E6M1_218 were significantly associated, while LP was governed by E4M4_242, E4M1_365 and E6M3_260. The seed quality trait protein was associated with E4M3_440, E6M2_595 and E6M1_196, while oil content was associated with

E4M3_200, E6M2_364 and fiber content with E5M7_195. These significant markers recorded high adj. R^2, lower P values and the lowest AIC, BIC and SBC estimates.

Table 5.10 Significant QTLs identified for yield and seed quality traits using Mixed-Multiple regression models in upland cotton. The QTLs were identified using PROC GLMSELECT of SAS.

Protein	Model R^2	Adj. R^2	AIC	AICC	BIC	SBC	Pr > F
E4M3_440	0.2125	0.202	158.4213	3.0877	159.3483	163.1089	<.0001
E6M2_595	0.3033	0.2845	150.9882	2.9941	152.106	158.0196	0.0027
E6M1_196	0.3839	0.3585	143.5283	2.901	145.302	152.9036	0.0028
E5M3_130	0.4263	0.3944	140.0381	2.8602	142.4716	151.7571	0.0239
Oil							
E4M3_200	0.1455	0.1341	168.7024	3.2212	169.9945	173.3901	0.0006
E6M2_364	0.217	0.1958	163.9766	3.1628	165.4378	171.0081	0.0113
E3M2_145	0.2799	0.2503	159.5324	3.1088	161.4715	168.9076	0.0138
E4M3_255	0.3259	0.2885	156.4413	3.0733	159.0364	168.1603	0.0297
Fiber							
E5M7_195	0.3166	0.3075	151.9285	3.0033	154.1251	156.6162	<.0001
Seed cotton yield							
E5M1_55	0.0835	0.0705	386.3652	6.3989	388.1783	390.9185	0.0138
E6M1_218	0.1484	0.1237	383.0788	6.3566	385.2394	389.9088	0.0249
Lint percentage							
E4M4_242	0.4688	0.4612	181.8918	3.559	182.6868	186.4451	<.0001
E4M1_365	0.5484	0.5353	172.1962	3.4277	173.3405	179.0262	0.0009
E6M3_260	0.6028	0.5853	164.9529	3.3314	166.8073	174.0595	0.0032

5.4 Discussion

The achievement and progress of conventional breeding in improving the complex genetic base for cotton seed quality traits is limited. There has been no exclusive breeding work initiated in improving seed quality traits. Recently molecular markers have provided a useful base for understanding and manipulating the genetic factors governing seed quality traits. In the present study, association mapping was successfully employed for the identification of AFLP markers associated with the seed quality traits oil, protein and fiber content in addition to yield parameters.

The 64 primer combinations of AFLP markers generated heterozygosity in the range of 0.347(LA) to 0.431(ARK). The frequency of kinship revealed that genetic relatedness is prominent among the genotypes under study. High levels of similarity within upland cluster are due to the fact that these genotypes have been subjected to a great degree of inter-cultivar gene flow (Kellogg *et al.*, . 1996; Wendel and Doyle 1998). However, the PIC values ranged from 0.15-0.40 for most of the genotypes, thus AFLP markers are useful to distinguish genotypes based on allelic frequencies.

The phenotypic data on seed quality traits suggested wide variability for protein (18-28.45%), oil (6-25.6) and fiber content (15.88-37.2%). Kohel (1978) and Song and Zhang (2007) also suggested that there was wide variability for seed oil, weight and seed oil index in the *G. hirsutum* germplasm collection. A wide range of variability has been observed for seed oil content in the wild species and perennial races of *G. arboreum*. The highest seed oil content (22.89%) was observed in the wild species *G. lobatum* and the lowest (10.26%) was recorded in *G. stocksii* (Gotmare *et al.,* 2004). In the present study diploids had the lowest values for oil content (6.47% for *G. herbaceum* and 10.79% for *G. arboreum*) in comparison to the tetraploid accessions (25.16% in ARK-9811-15-07). Mert *et al*., (2004) reported oil content varying between 19.1-25.2%, while protein percentage ranged from 22.9-26.2% across two locations in upland cotton, whereas the present study showed better range of values for the oil (6.47-25.16%) and protein (18.05-28.45%) content in the upland cotton. Interspecific cross derivatives offer an even wider variability for the quality traits. A TM-1 x Hai7124 generated BC_1S_1 population recorded 28.97-40% kernel oil and 32-47% of protein content (Song and Zhang, 2007). Based on these results, association mapping can be a good choice in order to identify significant markers associated with seed traits utilizing upland and diploid cotton accessions (historical/wild) or traditional QTL mapping using interspecific segregating populations.

Seed quality traits are directly influenced by the lint percentage, seed cotton yield, seed number, seed index or weight, seed coat content, moisture level and external environmental factors. In the present study, SCY and LP were negatively but not significantly correlated (-0.074). We observed positive correlations between SCY and protein and between LP and oil, while negative correlations between SCY and oil and between LP and protein content were observed. To increase oil and protein in a given seed size would require the increase to take place at the expense of other residual constituents, e.g., by reducing the seed coat. Most cultivated upland cotton lines show a decrease in seed coat thickness compared to their primitive ancestors, but the seed coat is required as protective cover during development of the embryo. The thin seed coat lines are prone to break during ginning and fiber processing leading to embryo damage (Kohel *et al*., 1985).

Simultaneous improvement of oil and protein is complicated, owing to their negative correlation (-0.224 in the present study) has been reported. Several reports in the past have also noticed such a pattern in upland and interspecific crosses. 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 in seed percentages also decrease with harvest date, but the greatest change is in the amount of oil (Kohel and Cherry, 1983). Here, fiber content was negatively correlated with protein and oil content, with non-significant positive correlations with SCY and LP. Ye *et al*., (2003) revealed significant phenotypic correlation between oil, protein and lysine index at various developmental stages. Looking to the complex pathways involved in the synthesis of oil and protein, the addition of more markers to catalogue multi environment phenotypic variation would improve the understanding of genetic factors governing these traits.

Several studies have been conducted to understand the inheritance pattern and gene action governing quality traits. Seed index was found to be predominantly under the control of genes acting additively thus this trait could easily be manipulated through selection for the production of pure line variety. The 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 fiber 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 epistatic manner controlling these traits. Analysis of reciprocal backcrosses suggests the existence of maternal effects (Dani and Kohel 1989). Recently, there is only one report documenting linkage based QTL mapping of seed quality traits in upland cotton (Song and Zhang, 2007). The present study explored the possibility of identifying QTL's responsible for oil, protein and fiber content using association mapping approaches based on extensive statistical models to explain phenotypic variation.

Based on the MLM approach (Yu *et al*., 2006), we considered population structure, principal component analysis and kinship to eliminate spurious associations. The present study identified a number of AFLP markers significantly associated with fiber traits. Initially using mixed models, we identified 45 significant markers associated with seed cotton yield (SCY), lint percentage (LP) and quality traits. The potential mixed model, utilizing population structure data, identified common markers (E4M3_255, E4M3_218, E6M2_595 and E3M3_60) governing seed oil and protein content. The adj. R^2, which measures the quantity of explainable genetic variation ranged from 30-60%. Similarly a significant QTL (qPP-D9-1) for total protein percentage was identified in a BC_1

population involving *G. hirsutum* and *G. barbadense* parents but it did not reflect large variations in protein components (Song and Zhang 2007).

More support and validity of the MLM approach in our study came from multiple regression based GLM studies. As many as 14 significant markers were associated with the five traits. Markers such as E4M3_440 and E6M2_595 for seed protein, E4M3_200, E6M2_364 for seed oil and E5M7_195 for fiber content showed high adj. R2 and low AIC, BIC and SBC statistics. The consistency of these markers was confirmed both in mixed model and mixed-multiple regression models. Most of the markers selected through mixed-MMR models provide a good insight into deciding the most robust model. The robustness and high efficiency of the models in explaining the phenotypic variation provide a tool for their further use in fine mapping and MAS.

5.5 Conclusion and Future Work

In cotton, yield and fiber quality are the main crop features, whereas cotton seed components are typically seen as by-products and are not major breeding objectives. There has been little systematic effort in improving the nutritional quality of cotton meal beyond efforts to remove or eliminate gossypol. This is partly attributable to our lack of understanding of the genetics and complexity of the traits involved and is responsible for achieving marginal success in improving these traits. With the help of high throughput genomic tools, efforts have been initiated to dissect the pathways underlying these traits. For example, a genetically modified fatty acid composition of cottonseed oil using the hairpin RNA-mediated gene silencing technique was developed and demonstrated successfully (Liu *et al*., 2002). Similarly, Ganesan *et al*., (2006) successfully used RNAi to disrupt gossypol biosynthesis in cottonseed tissue by interfering with the expression of the *delta-cadinene synthase* gene during seed development. These results illustrate that targeted genetic modification provides a mechanism to improve this important source of nutrition. Marker assisted introgression and transfer of specific alleles would also undoubtedly increase the efficiency of seed quality focused breeding programs in the future.

5.6 References

AOAC 1999 Official Methods of Analysis of the Association of official Analytical Chemists, pp. 1-12.

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

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y and Buckler ES 2007 TASSEL: Soft ware for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

Chen ZF, Zhang ZW and Cheng HL 1986 The analysis of Upland cotton quality. Acta Agronomica Sinica 12, 195–200.

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC and Specht JE 2003 The seed protein, oil, and yield QTL on soybean linkage group I. Crop Science 43, 1053-1067.

Casa AM, Pressoira G, Brown PJ, Mitchell SE, Rooney WL, Tuinstrac MR, Franks CD and Kresovicha S 2008 Community resources and strategies for association mapping in sorghum. Crop Science, 48, 30–40.

Cox C, Mann J, Sutherland W, Chisholm A and Skeaff M 1995 Effects of coconut oil, and safflower oil on lipids and lipoproteins in persons with moderately elevated cholesterol levels. Journal of Lipid Research, 36, 1787–1795.

Dani RG and Kohel RJ 1989 Maternal effects and generation mean analysis of seed-oil content in cotton (*Gossypium hirsutum*). Theoretical and Applied Genetics, 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. Molecular Ecology 14,2611–2620.

Excoffier L, Laval G and Schneider S 2005 Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online, 1, 47-50.

Ganesan S, Campbell, LM, Puckhaber L, Stipanovic RD and Rathore KS 2006 Engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. Proceedings of National Academy of Sciences, 103(48), 18054-18059.

Gotmare V, Singh P, Mayee CD, 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 Breeding 123, 207-208.

Gupta V, Mukhopadhyay A, Arumugam N, Sodhi YS, Pental D and Pradhan AK 2004 Molecular tagging of erucic acid trait in oilseed mustard (*Brassica juncea*) by QTL mapping and single nucleotide polymorphisms in *FAE1* gene. Theoretical and Applied Genetics 108, 743–749.

Hammer, Harper DAT, Ryan PD 2001 PAST: Paleontological statistics software package for education and data analysis, Palaeontologia Electronica, vol 4, issue 1, art 4.

Hu X, Sullivan-Gilbert M, Gupta M and Thompson SA 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.). Theoretical and Applied Genetics 113, 497–507.

Hudson LM, Kerr LA and Maslin WR 1988 Gossypol toxicosis in a herd of beef calves. Journal of American Veterinary Medical Association 192(9), 1303.

Kellogg EA, Appels R and Mason-Gamer RJ 1996 When genes tell different stories: the diploid genera of Titicaceae (*Gramineae*). Systematic Botany, 21, 21–347.

Kianian SF, Egli MA, Phillips RL, Rines HW, Somers DA, Gengenbach BG, Webster FH, Livingston SM, Groh S, O'Donoughue LS, Sorrells ME, Wesenberg DM, Stuthman DD and Fulcher RG 1999 Association of a major groat oil content QTL and an acetyl-CoA carboxylase gene in oat. Theoretical and Applied Genetics 98, 884–894.

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

Kohel RJ and Cherry JP 1983 Variation of Cottonseed Quality with stratified harvests. Crop Science, 23:1119-1124.

Kohel RJ, Glueck J and Rooney LW 1985 Comparison of cotton germplasm collections for seed protein content. Crop Science. 25(6), 961-963.

Kumar S, Tamura K, Nei M 2004 MEGA4: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinformatics 5(2), 150-163.

Lacape JM, Nguyen TB, Thibivilliers S, Bojinov B, Courtois B, Cantrell RG, Burr B and B Hau 2003 A combined RFLP–SSR–AFLP map of tetraploid cotton based on a *Gossypium hirsutum* ×*Gossypium barbadense* backcross population. Genome, 46, 612–626.

Liu Q, Singh SP and Green AG 2002 High-Stearic and high-Oleic Cottonseed oils Produced by Hairpin RNA-Mediated Post-Transcriptional Gene Silencing. *Plant Physiology,* 129, 1732–1743.

Mert M, Akiscan Y and Gencer O 2004 Inheritance of oil and protein in some cotton generations. Asian Journal of Plant Sciences, 3(2), 174-176.

Myers GO, Baogong J, Akash MW, Badigannavar AM and Saha S 2009 Chromosomal assignment of AFLP markers in upland cotton (*Gossypium hirsutum* L.) Euphytica, 165(2), 391-399.

Nei M and Li WH 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of National Academy of. Sciences, USA 76, 5269–5273.

Panthee DR, Pantalone VR, West DR, Saxton AM and Sams CE 2005 Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Science, 45, 2015–2022.

Pe´rez-Vich B, Knapp SJ, Leon AJ, Ferna´ndez-Martı´- nez JM and Berry ST 2004 Mapping minor QTL for increased stearic acid content in sunflower seed oil. Molecular Breeding, 13, 313–322.

Peakall R and Smouse PE 2006 GENAlEx 6.1: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 6, 288-295.

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

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

SAS INSTITUTE 2008 Documentation GLMSELECT. SAS Institute, Cary, NC.

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

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

Singh M, Singh TH and Chahal GS 1985 Genetic analysis of some seed quality characters in Upland cotton (*Gossypium hirsutum* L.). Theoretical and Applied Genetics, 71, 126–128.

Sneath PHA and Sokal RR 1973 Numerical taxonomy: The principals and practice of numerical classification. W.H. Freeman and Co., San Francisco, California. pp 573.

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

Sun SK, Chen JH, Xian SK and Wei SJ 1987 Study on the nutritional quality of cotton seeds. Scientia Agricultura Sinica 5, 12–16.

Tan YF, Sun M, Xing YZ, Hua JP, Sun XL, Zhang QF 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. Theoretical and Applied Genetics,103, 1037–1045.

Tar'an B, Warkentin T, Somers DJ, 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.

TASSEL 2009 User manual, trait analysis by association, evolution and linkage. www.maizegenetics.net/tassel

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M,Freijters A, Pot J, Peleman J, Kuiper M and Zabeau M 1995 AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research, 23, 4407-4414.

Wendel JF and Doyle JJ 1998 Phylogenetic incongruence: window into genome history and molecular evolution. In: Soltis DE, Soltis PS, Doyle JJ (eds) Molecular systematics of plants. II.DNA Sequencing. Kluwer Academic, Boston Dordrecht London, pp 265–296.

Ye ZH, Lu ZZ 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 WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S and Buckler ES 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38,203-208.

Zhao JY, Becker HC, Zhang DQ, Zhang YF 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 Theoretical and Applied Genetics 113, 33–38.

Weir BS and Cockerham CC 1984 Estimating F statistics for the analysis of population structure. Evolution: International Journal of Organic Evolution, 38, 1358–1370.

Wright S 1951 The genetical structure of populations. Annals of Eugenics 15, 323-54.

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

CHAPTER 6 SUMMARY AND CONCLUSIONS

Cotton (*Gossypium spp.*) is the most extensively used natural fiber in the textile industry. Understanding the genetic diversity, population structure and marker trait associations are of great importance in marker assisted selection. The present study was undertaken to genetically dissect the cotton genome in order to identify associations between molecular markers and the developmental, fiber and seed quality traits. Surveying the genetic diversity in diploid (involving A_1 and A_2 subgenome cross derivatives) and tetraploid cottons (representing US upland genotypes) may also provide a valuable insight into the interrelationships among the genotypes.

The diploid species, *G. arboreum* and *G. herbaceum* are generally cultivated on marginal and drought prone environments in Asia. Microsatellite, AFLP and TRAP markers were used to construct a linkage map with 94 F_2 diploid individuals derived from a cross between *G. arboreum x G. herbaceum*. A total of 606 polymorphic markers gave rise to 37 linkage groups covering a total of 1109cM with an average distance of 7.92cM between each loci. Discriminant analysis identified three markers each for petal color and seed fuzziness, and four markers for petal spot. For quantitative traits, a total of 19 QTL's were identified and linked with five fiber traits using composite interval mapping. Markers e.g., qFL4-1, qFS4-2, qELO1-1 and qSI2-1 were found to be significantly linked with fiber length, strength, elongation and seed index respectively. The construction of an A genome diploid map, combining AFLP, TRAP and SSR markers, can serve as a model for the advancement of cotton genetics, including the understanding of the inheritance of fiber genes. Adding additional markers to the existing map will assist in future map based cloning efforts and in gene discovery.

Association mapping principles were applied to upland cotton genotypes in order to examine population structure and marker trait associations. A set of 232 genotypes were genotyped using AFLP markers. Based on 568 polymorphic markers, molecular diversity was found to be in the range of 0.48-0.574 with a variance around 10% among the groups. Based upon Bayesian and MCMC, population structure analysis, there existed six subpopulations, in agreement with their geographical origin. The mixed and mixed-multiple regression (MMR) models identified significant markers for lint yield and fiber traits, showing low AICC, BIC and SBC values and high adj. R^2. Out of 568 AFLP markers used in this study, 255 markers were initially found to be significantly associated with eight traits using the traditional MLM approach. Inclusion of MMR improved the model, reducing the number of markers significantly associated with these traits to 111. The MMR based epistatic interactions revealed 49 QTLs responsible for eight fiber traits. Thus mixed MMR models were efficient in reducing the Type I error. This sequential validation of marker is an improved method for reducing false positives and identifying truly significant associations.

The narrow genetic base of upland cotton germplasm that is used in breeding programs is one of the factors in failing to achieve appreciable amount of progress in improving yield and fiber traits over last two decades. Hence the present study was planned to determine the efficiency of AFLP for estimating genetic diversity among a collection of 60 accessions of upland cotton and also for the identification of potential marker trait associations for major fiber traits. The pairwise kinship estimates were ranging between 0.1-0.88 accounting for most of the shared ancestral alleles. Genetic distance analysis confirmed the narrow genetic base among *G. hirsutum* genotypes. The PCA and kinship estimates in MLM approach identified a number of significant AFLP markers associated with yield and fiber traits. The MMR identified 38 markers associated with eight traits. These models improved the efficiency of marker trait association by reducing the false positives.

The achievement and progress of conventional breeding in improving the complex genetic base for cotton seed quality traits is limited. There has been very limited breeding work exclusively devoted to improving the seed quality traits. Therefore, the present study was planned to identify AFLP markers associated with the yield and seed quality traits using association mapping principles. A set of 75 upland cotton genotypes were analyzed for seed quality traits such as seed protein, oil and fiber content. Population structure based mixed models showed 32 significant markers associated with these seed quality traits. MMR models identified several markers, notably E4M3_440, E4M3_200 and E5M7_195 for seed protein, oil and fiber content respectively. Marker assisted introgression and transfer of specific alleles would also undoubtedly increase the efficiency of seed quality focused breeding programs in the future.

VITA

Ashok Badigannavar, was born in North Karnataka province of southern India. He earned a Bachelor of Science (Agri.) and Master of Science (Agri.) degrees from the University of Agricultural Sciences, Dharwad, India, specializing in genetics and plant breeding. He moved on to work as ‘Research Associate’ in ‘Hybrid Cotton Project’ at Agricultural Research Station, Dharwad. In 2004, he joined a private company and served as ‘Cotton Breeder’. In 2006, he joined the School of Plant, Environmental and Soil Sciences, Louisiana State University for the doctoral program and will graduate in Spring 2010.