

**QUANTITATIVE TRAIT LOCI MAPPING FOR  
AGRONOMIC AND FIBER QUALITY TRAITS IN UPLAND COTTON  
(*GOSSYPIUM HIRSUTUM* L.) USING MOLECULAR MARKERS**

**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 Department of Agronomy**

**by**

**Muhanad Walid Akash**

**B.S., The University of Jordan, Jordan, 1995**

**M.S., The University of Jordan, Jordan, 1997**

**M.S., Louisiana State University, U. S. A., 2003**

**December, 2003**

## **ACKNOWLEDGEMENTS**

I would like to express my sincere appreciation to my major professor, Dr. Gerald Myers, for his encouragement, support, friendship, and insight throughout the course of my studies. I would also like to express gratitude to my minor advisor in genetics, Dr. Manjit Kang, for providing significant input on much of my graduate work. I also thank Drs. Brad Venuto, Michael Stine, Don LaBonte, and Jeff Hoy for serving on my advisory committee and for their time and suggestions in the preparation of the dissertation.

My special thanks goes to Dr. Altaf Khan, Mary Bowen for the support and help in molecular genetic lab. A special acknowledgement to my major professor in The Department of Experimental Statistics, Dr. Barry Moser. His invaluable advice and constructive criticism undoubtedly helped me to develop the skills required of a statistician.

Most of all, I am grateful to my parents, sisters, and brothers (Khaled, Mohammed, Mu'tasem, and Haitham) for their love, support and encouragement rendered in the course of my graduate studies at Louisiana State University.

## TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS.....	ii
LIST OF TABLES.....	vi
LIST OF FIGURES.....	xi
ABSTRACT.....	xii
CHAPTER 1 INTRODUCTION.....	1
1.1 Upland Cotton.....	1
1.2 Origin.....	1
1.3 Cotton Breeding and Genetics.....	3
1.4 Objectives.....	5
1.5 References.....	6
CHAPTER 2 LITERATURE REVIEW.....	9
2.1 Molecular Markers.....	9
2.1.1 Biochemical Markers.....	9
2.1.2 DNA-Based Markers.....	10
2.1.2.1. Hybridization-Based (non-PCR) Techniques.....	10
2.1.2.2. Arbitrarily Primed PCR and Other PCR-Based Multi-Locus Profiling Techniques.....	11
2.1.2.3. Sequence Targeted and Single Locus PCR.....	12
2.2 DNA-Markers Used in Cotton.....	12
2.2.1 Restriction Fragment Length Polymorphism .....	13
2.2.2 Amplified Fragment Length Polymorphism .....	14
2.2.3 Randomly Amplified Polymorphic DNA.....	16
2.2.4 Simple Sequence Repeat .....	17
2.2.5 Inter Simple Sequence Repeat .....	20
2.3 Fingerprinting and Diversity Studies.....	21
2.4 Linkage Maps.....	21
2.5 Mapping Quantitative Trait Loci.....	23
2.6 Methods of QTL Mapping.....	25
2.7 QTL X Environment Interaction .....	25
2.8 Marker-Assisted Selection.....	26
2.9 References.....	28
CHAPTER 3 THE DEVELOPMENT OF A GENETIC MAP FOR UPLAND COTTON ( <i>GOSSYPIUM HIRSUTUM</i> L.) COMPRISED OF AMPLIFIED FRAGMENT LENGTH POLYMORPHISMS.....	37
3.1 Introduction .....	37
3.2 Materials and Methods.....	39
3.2.1 Mapping Population.....	39
3.2.2 DNA Preparation.....	41

3.2.3 DNA Quantification.....	42
3.2.4 Amplified Fragment Length Polymorphism Analysis.....	42
3.2.5 Gel Analysis.....	46
3.2.6 Marker Naming.....	49
3.2.7 Map Construction.....	49
3.3 Result and Discussion.....	49
3.4 References.....	55
CHAPTER 4 MOLECULAR QTL MAPPING FOR AGRONOMIC TRAITS IN UPLAND COTTON ( <i>GOSSYPIUM HIRSUTUM</i> L.).....	59
4.1 Introduction.....	59
4.2 Materials and Methods.....	63
4.2.1 Plant Material.....	63
4.2.2 Phenotypic Measurement.....	63
4.2.3 Linkage Analysis.....	64
4.2.4 QTL Analysis.....	64
4.2.5 QTL X Environment Interaction.....	65
4.3 Results and Discussion.....	65
4.3.1 Quantitative Traits.....	66
4.3.2 QTL for Lint Weight Per Boll (LY).....	68
4.3.3 QTL for Seedcotton Weight Per Boll (BW).....	70
4.3.4 QTL for Boll Number Per Plant (B/P).....	72
4.3.5 QTL for Lint Percentage (LP).....	73
4.3.6 QTL X Environment Interaction.....	75
4.4 References.....	79
CHAPTER 5 MOLECULAR QTL MAPPING FOR FIBER QUALITY TRAITS IN UPLAND COTTON ( <i>GOSSYPIUM HIRSUTUM</i> L.).....	83
5.1 Introduction.....	83
5.2 Materials and Methods.....	86
5.2.1 Plant Material.....	86
5.2.2 Phenotypic Measurement.....	87
5.2.3 Linkage Analysis.....	87
5.2.4 QTL Analysis.....	88
5.2.5 QTL X Environment Interaction.....	89
5.3 Results and Discussion.....	89
5.3.1 Quantitative Traits.....	89
5.3.2 QTL for Fiber Elongation (E).....	92
5.3.3 QTL for Fiber Length (L).....	93
5.3.4 QTL for Fiber Uniformity (U).....	95
5.3.5 QTL for Fiber Strength (S).....	98
5.3.6 QTL for Fiber Micronaire (M).....	100
5.3.7 QTL X Environment Interaction.....	102
5.4 References.....	107

CHAPTER 6 MULTIPLE IMPUTATION FOR MISSING DATA IN MOLECULAR PLANT BREEDING STUDIES.....	111
6.1 Introduction.....	111
6.1.1 Patterns of Missing Data.....	112
6.1.2 Types of Missing Data.....	113
6.1.2.1 Missing at Random (MAR).....	113
6.1.2.2 Missing not at Random (MNAR).....	113
6.1.2.3 Missing Completely at Random (MCAR).....	113
6.1.3 Methods for Handling Missing Data.....	114
6.1.3.1 Case Deletion.....	114
6.1.3.2 Single Imputation.....	115
6.1.3.3 Multiple Imputation (MI).....	117
6.1.3.3.1 The MI Procedure.....	117
6.1.3.3.2 MI Efficiency.....	118
6.1.3.3.3 The MIANALYZE Procedure.....	119
6.2 Materials and Methods.....	121
6.2.1 Data Preparation.....	121
6.2.2 MI Methods.....	121
6.2.3 Data Analysis.....	122
6.3 Results and Discussion.....	122
6.3.1 Complete Data Analysis.....	122
6.3.2 Propensity Score and Regression Methods.....	124
6.3.3 MCMC Monotone-Data and MCMC Full-Data Imputation methods.....	125
6.3.4 Logistic Regression and Discriminant Function Methods...	125
6.3.5 Missing not at Random (MNAR) .....	127
6.4 References.....	129
CHAPTER 7. SUMMARY AND CONCLUSIONS.....	131
7.1 References.....	133
VITA. ....	134

## LIST OF TABLES

		Page
2.1	Comparison of different DNA-marker systems.....	12
2.2	Example of correlations between genome size, enzyme combination (EC), pre-amplification (PA) and amplification strategy (Amp) in various organisms. Genome size is indicated in megabases. Restriction enzymes include <i>EcoRI</i> (E), <i>MseI</i> (M), <i>PstI</i> (P) and <i>TaqI</i> (T). The number of selective bases for AFLP (pre)amplification is given in the last two columns (Vos et al., 1995).....	15
3.1	Reported molecular marker based linkage maps in cotton for (a) intraspecific and (b) interspecific crosses.....	39
3.2	Mean performance for agronomic and fiber quality traits of Paymaster 54 and Pee Dee 2165 at Alexandria and Baton Rouge, LA, in 2002.....	40
3.3	Adapters and primers used for pre-amplification and selective amplification of AFLP procedure.....	43
3.4	Protocol components for digestion and ligation of genomic DNA .....	44
3.5	Reagents used in the Preamplification step (a) and selective amplification step (b).....	46
3.6	Proc Freq code to test for marker segregation distortion (SAS Version 9).....	48
3.7	Number of monomorphic and polymorphic (total) and number of polymorphic amplified fragment length polymorphism (AFLP) primer combinations between two lines (Pee Dee 2165 and Paymaster 54) of Upland cotton. ....	50
3.8	Marker distributions among the linkage groups of Upland cotton .....	54
4.1	Normality tests for Upland cotton agronomic traits: lint weight per boll (LY), seedcotton weight per plant (BW), boll number per plant (B/P), and lint percentage (LP) at Baton Rouge (B) and Alexandria (A).....	66
4.2	The correlation among Upland cotton agronomic trait. Data combined for two locations (Alexandria and Baton Rouge). The number on the top is the correlation coefficient and the number below is its correspondent P value.....	68
4.3	Putative QTL and their interval position influencing Upland cotton lint weight per boll trait. IM and CIM were used under MapMaker/QTL and	

	QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A) .....	69
4.4	AFLP markers that were associated with putative QTL influencing Upland cotton lint weight per boll trait using simple and logistic regression at Baton Rouge (B).....	70
4.5	Putative QTL and their interval position influencing Upland cotton seedcotton weight per plant trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).....	71
4.6	AFLP markers that were associated with putative QTL influencing Upland cotton seedcotton weight per plant trait using simple and logistic regression at Baton Rouge (B).....	71
4.7	Putative QTL and their interval position influencing Upland cotton boll number per plant trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).....	72
4.8	AFLP markers that were associated with putative QTL influencing Upland cotton bolls numberper plant trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A) .....	73
4.9	Putative QTL and their interval position influencing Upland cotton lint percentage trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).....	74
4.10	AFLP markers associated with putative QTL influencing Upland cotton lint percentage trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A).....	74
4.11	QTL X environment interaction LOD using Module Jzmapqtl of QTL-CARTOGRAPHER in Upland cotton.....	75
4.12	The QTL summary for Upland cotton agronomic traits.....	77
5.1	Normality tests for Upland cotton fiber quality traits: elongation (E), length (L), uniformity (U), strength (S), and micronaire (M) at Baton Rouge (B) and Alexandria (A).....	91
5.2	The correlation among Upland cotton fiber quality traits. Data combined for two locations (Baton Rouge and Alexandria). The	

	number on the top is the correlation coefficient and the number below is its correspondent P value.....	91
5.3	Putative QTL and their interval position influencing Upland cotton fiber elongation trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).....	92
5.4	AFLP markers that were associated with putative QTL influencing Upland cotton fiber elongation trait using simple and logistic regression at Baton Rouge (B).....	93
5.5	Putative QTL and their interval position influencing Upland cotton fiber length trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).....	94
5.6	AFLP markers that were associated with putative QTL influencing Upland cotton fiber length trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A).....	95
5.7	Putative QTL and their interval position influencing Upland cotton fiber uniformity trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).....	96
5.8	AFLP markers that were associated with putative QTL influencing Upland cotton fiber uniformity trait using simple and logistic regression at Baton Rouge (B).....	97
5.9	Putative QTL and their interval position influencing Upland cotton fiber strength trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).....	99
5.10	AFLP markers that were associated with putative QTL influencing Upland cotton fiber strength trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A).....	100
5.11	Putative QTL and their interval position influencing Upland cotton fiber micronaire trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A). ....	101
5.12	AFLP markers that were associated with putative QTL influencing Upland cotton fiber micronaire trait using simple and logistic regression	

	at Baton Rouge (B) and Alexandria (A).....	102
5.13	QTL X environment interaction LOD using Module Jzmapqtl of QTL-CARTOGRAPHER in Upland cotton.....	103
5.14	The QTL summary for Upland cotton fiber quality traits.....	106
6.1	Pattern of missing data. A: arbitrary pattern and B: monotone pattern. Here, an "X" means that the variable is observed and a "." means that the variable is missing.....	112
6.2	Imputation Methods in SAS PROC MI (SAS_V9 On line Doc.).....	118
6.3	Percent efficiency of MI estimation by number of imputation m and percentage of missing data $\lambda$ .....	119
6.4	Complete data analysis showing Pearson correlation coefficients and corresponding P-values for lint weight per boll (LY), lint percentage (LP), seedcotton weight per boll (BW), boll number per plant (B/P).....	123
6.5	Complete data analysis showing logistic regression parameter estimates and their associated confidence limits and P-values.....	123
6.6	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Regression Method.....	124
6.7	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Propensity Score.....	125
6.8	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Monotone-Data Imputation.....	126
6.9	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Full-Data Imputation...	126
6.10	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Discriminant Function.....	127
6.11	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Logistic Regression.....	127
6.12	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Full-Data Imputation for MNAR type of missingness.....	128

6.13	The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Monotone-Data Imputation for MNAR type of missingness.....	128
------	--	-----

## LIST OF FIGURES

		Page
1.1	U. S. Cotton Productivity in Kg ha <sup>-1</sup> from 1930 through 2002 (USDA, 2002).....	05
3.1	High quality undigested Upland cotton genomic DNA.....	44
3.2	Preamplification products (10 µL/lane) create a visible smear in the 100 to 1000 bp range.....	45
3.3	AFLP segregating pattern in F <sub>2</sub> population from the cross of Paymaster 54 and Pee Dee 2165. The amplification was made using 8 different primer combinations.....	47
3.4a	Major Genetic linkage groups of Upland cotton ( <i>Gossypium hirsutum</i> L.) comprised of 102 amplified fragment length polymorphism markers.	52
3.4b	Minor Genetic linkage groups of Upland cotton ( <i>Gossypium hirsutum</i> L.) comprised of 41 amplified fragment length polymorphism markers...	53
4.1	Frequency distribution for each Upland cotton agronomic trait in the F <sub>2:3</sub> population at Baton Rouge (B) and Alexandria (A). The data shown for boll number per plant at Baton Rouge (BP_B) and lint weight per boll at Alexandria (LY_A) were transformed (square root and log transformation, respectively) .....	67
4.2	A comparison of QTL positions for Upland cotton lint weight per boll (LY), seed cotton weight per boll (BW), bolls number per plant (B/P), and lint percentage (LP) using composite interval mapping (CIM) and interval mapping (IM).....	78
5.1	Frequency distribution for each cotton fiber quality trait in the F <sub>2:3</sub> population at Baton Rouge (B) and Alexandria (A). The data shown for micronaire at Baton Rouge (M_B) and strength at Alexandria (S_A) were transformed (square root and log transformation, respectively)....	90
5.2	A comparison of QTL positions for Upland cotton micronaire (M), strength (S), uniformity (U), length (L), and elongation (L) using composite interval mapping (CIM) and interval mapping (IM).....	105

## ABSTRACT

The breeding of Upland cotton (*Gossypium hirsutum* L.) cultivars that combine high yield and fiber quality is a major challenge to the breeder. The understanding of the quantitative trait loci (QTL) contributing to agronomic and fiber quality traits offers an excellent route to solve this problem. A QTL analysis was carried out after an  $F_{2:3}$  population composed of 138 lines, derived from the intraspecific cross between Paymaster 54 and PeeDee 2165, was developed and a linkage map including 143 AFLP markers was constructed. The  $F_{2:3}$  population was grown in two locations, Alexandria and Baton Rouge in LA. The 143 linked markers were assigned to 13 major and 15 minor linkage groups, the 28 linkage groups cover a genetic distance of 1773.2 cM. This gives coverage of 37.7% of the cotton genome (4700 cM). Single-marker analysis, including simple and logistic regression, and interval marker analysis, including interval mapping (IM) and composite interval mapping (CIM), was used. Interval mapping was used to study QTL interaction effects with the environment.

For the agronomic traits, the same five QTL were detected, using a significant threshold of 2 LOD, in both IM and CIM. These include two for lint weight per boll, two for seedcotton weight per plant, and one for lint percentage, which collectively, based on IM analysis, explained 32.5%, 28.6%, and 4.4% of the phenotypic variation, respectively. In total, seven and nine different QTL were detected by IM and CIM, respectively. For the fiber quality traits, the same nine QTL were detected in both IM and CIM. These

include one for fiber elongation, one for length, two for uniformity, three for strength, and two for micronaire, which collectively, based on IM analysis, explained 50.9%, 18.7%, 69%, 49.6%, and 25.3% of the phenotypic variation, respectively. In total, nine and 19 different QTL were detected in IM and CIM, respectively. Eleven QTL were found to have significant interaction effects with the two locations.

Future efforts in QTL mapping should focus on developing more saturated maps, using larger population sizes, and more powerful statistical algorithms and theories for identifying QTL and elucidating QTL X environment interactions.

## CHAPTER 1 INTRODUCTION

### 1.1 Upland Cotton

Cotton is the most important textile fiber crop and the world's second-most important oil-seed crop after soybean (*Glycine max* L.) (Poehlman and Sleper, 1995). It is grown commercially in the temperate and tropical regions of more than 50 countries, including the United States, India, China, Central and South America, the Middle East, and Australia (Fryxell, 1979; Smith, 1999).

Until recently, the United States was the world's leading cotton producer. In the 1990s, China became the leading cotton-producing country followed by the United States and the republics of the former Soviet Union. In the United States, Upland cotton was grown on 13.5 and 12.2 million acres in 2001 and 2002, respectively (NASS, 2003). In addition to being the world's major natural source of textile fiber and an important oil-seed crop, both cotton seed and its meal are used in food and feed products. Cotton feed products are widely used in Central American countries and India where cotton is considered a low cost, high quality protein ingredient (Ensminger et al., 1990).

Typically, cotton is harvested as 'seedcotton' that is then ginned to separate the seed and lint. The ginned seed is covered in short, fuzzy fibers that must be removed before the seed can be used for planting or crushed for oil. Fiber is further processed by spinning to produce yarn that is knitted or woven into fabrics.

### 1.2 Origin

Radiation of the genus *Gossypium* was accompanied by substantial evolution of chromosome structure and size. Results from using 16 nuclear and chloroplast

genes revealed that the cotton genome groups radiated in rapid succession following the formation of the cotton genus (Cronn et al., 2002).

The *Gossypium* genus comprises about 50 species with a basic chromosome number of 13. New species continue to be discovered. Of the known species, 45 are diploid with 26 chromosomes and there are at least five allotetraploid species with 52 chromosomes (Fryxell, 1992; Brubaker et al., 1999).

Based on chromosomal similarities, Poehlman and Sleper (1995) showed that these 50 species are commonly grouped into eight genome groups designated A through G and K (Edwards and Mirza, 1979; Endrizzi et al., 1985). Each genome represents a group of morphologically similar species that can rarely form hybrids with species from other genomic groups. Diploid species with A, B, E, or F genomes are African or Asian in origin and referred to as Old World species. Diploid species with the C or G genomes are Australian in origin. Diploid species containing the D genome originated in the Western hemisphere. The 5 allotetraploid species containing the AADD genome combination are referred to as New World species (Endrizzi et al., 1985).

The world's cotton fiber was produced from four of the 50 species, *G. arboreum* L. (n = 13, A genome), *G. herbaceum* L. (n = 13, A genome), *G. barbadense* L. (n = 26, AD genome), and *G. hirsutum* L. (n = 26, AD genome). The tetraploid species, *G. barbadense* and *G. hirsutum*, dominate world cotton production with a large number of improved varieties having been developed (Zhao et al., 1998).

*G. hirsutum* is the principal cultivated cotton and accounts for about 90% of the world's cotton production. In nature, *G. hirsutum* is a perennial shrub

approximately 1.5 m in height. However, *G. hirsutum* is grown as an annual crop. The Sea Island form of *G. barbadense* was introduced into the Nile Valley of Egypt where it became known as Egyptian cotton and was prized for its fine, long, and strong fibers. Egyptian cotton was subsequently introduced to Arizona, where it is known as Pima cotton. *G. barbadense* accounts for about 9% of the world's cotton production (Poehlman and Sleper, 1995).

### **1.3 Cotton Breeding and Genetics**

Yield in cotton, as in many other crops, is determined by the interaction among numerous yield components. Since yield is a complex trait, there is a need to break it down into manageable, manipulable units. The major ones in cotton are lint weight per boll (LY), boll number per plant (B/P), seedcotton weight per plant (BW), and lint percentage (LP). Culp and Harrell (1975) suggested that the breeder might select for medium to small bolls with the greatest possible number of small seed per boll to maintain a high lint percentage. Seed index (weight of 100 seeds), and lint index (lint weight on 100 seeds) are other yield components that indirectly affect the total yield. Fiber quality is evaluated by a combination of traits: micronaire (fiber maturity), length (longer fiber can be spun into finer yarn), strength, elongation (elasticity), and uniformity.

The determination of the locations of quantitative trait loci (QTL) for agriculturally important characteristics promises increased efficiency in selection through the use of marker-assisted selection (MAS) and opens the door for their future genetic manipulation and possible transfer between different plant species. The basic theory and tools for QTL detection were all in place by 1923 when Sax (1923) reported the association of seed size in beans (a quantitatively inherited

character) with seed-coat pigmentation (a discrete monogenic trait). The underlying assumption of using marker loci to detect polygenes is that of linkage disequilibrium, defined as the non-random association of alleles at different loci in a population, which exists between alleles at the marker locus and alleles of the linked polygene(s) (Tanksley, 1993). The idea of using single-gene markers to systematically characterize and map individual polygenes controlling quantitative traits was simple (Thoday, 1961). If the segregation of a single-gene marker could be used to detect and estimate the effect of a linked polygene and if these single-gene markers were scattered throughout the genome of an organism, it should be possible to map and characterize all of the polygenes affecting a character (Tanksley, 1993). In practice, the first linkage maps of QTL in Upland cotton were provided by Shappley et al. (1994) and Reinisch et al. (1994). Recently, several cotton QTL have been identified. For example, QTL for agronomic and fiber traits using restriction fragment length polymorphism (RFLP) markers have been identified (Shappley et al., 1998); RFLP markers were used to identify QTL for leaf morphology (Jiang et al., 2000), and QTL for stomatal conductance were discovered using randomly amplified polymorphic DNA (RAPD) and simple sequence repeat (SSR) markers (Ulloa et al., 2000). Other researchers have identified QTL for agronomic traits using RAPD and amplified fragment length polymorphism (AFLP) markers (Khan et al., 1998), for density of leaf and stem trichomes using RFLP markers (Wright et al., 1999), and for cotton productivity, physiological and fiber quality traits using RFLP markers (Saranga et al., 2001). However, no QTL involved in the expression of agronomic or fiber quality traits have been detected by means of AFLP markers.

## 1.4 Objectives

While numerous applied genetic, agronomic and conventional plant breeding and protection advances have had significant, positive impacts on cotton crop productivity, concerns have been raised about recent yield plateaus (Figure 1.1), crop production profitability, and fiber quality (Report of the American Cotton Producers, 1999) . In an effort to overcome these concerns, cotton researchers are increasingly making use of modern biotechnological tools particularly through the identification of quantitative trait loci and their allelic association with molecular markers such as randomly amplified polymorphic DNA (RAPD), simple sequence repeat (SSR) markers and amplified fragment length polymorphism (AFLP).

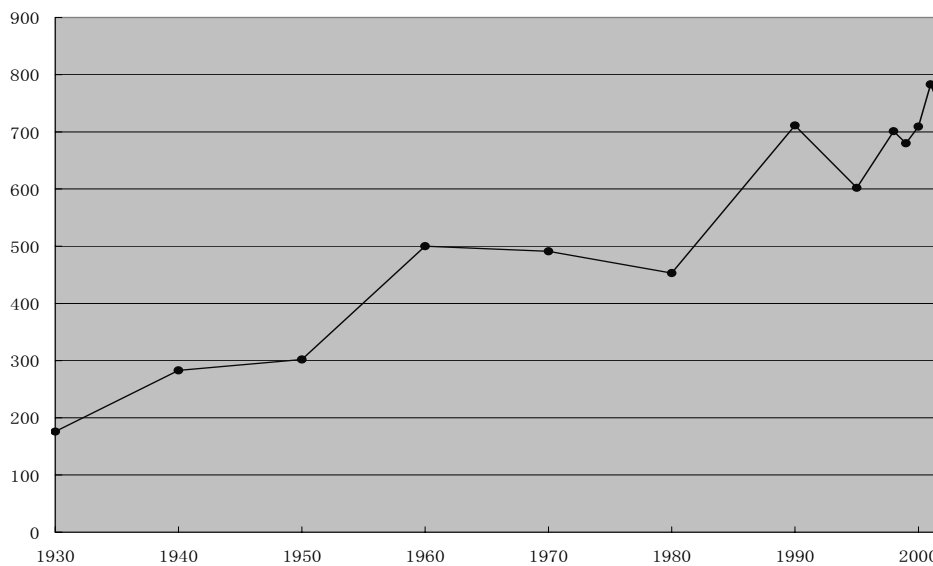


Figure 1.1 U. S. Cotton Productivity in Kg ha<sup>-1</sup> from 1930 through 2002 (USDA, 2002)

The objectives of this study were: (1) establish an AFLP-based linkage map in cotton, (2) identify and map QTL controlling agronomic and fiber quality traits in Upland cotton through their association with AFLP markers, (3) identify DNA markers

for use in cotton improvement that might provide a beginning basis for cloning specific genes that influence yield components in cotton and (4) investigate multiple imputation for missing data in molecular plant breeding studies.

This dissertation includes four separate manuscripts corresponding with these objectives. The first manuscript (chapter three) describes the details about the development of F<sub>2:3</sub> population, marker analysis, and map construction. Explanation of mapping methods including single-marker analysis (SMA) (simple and logistic regression) and interval-marker analysis (IMA) (interval mapping (IM) and composite interval mapping (CIM)) is presented in the next two manuscripts (chapters four and five). Chapter four includes QTL mapping for cotton agronomic traits (lint weight per boll (LY), boll number per plant (B/P), bolls weight (BW), lint percentage (LP)) while chapter five deals with QTL mapping for cotton fiber quality traits (fiber elongation (E), length (L), strength (S), uniformity (U), and micronaire (M)). In the last manuscript (chapter six), multiple imputation for missing data in molecular plant breeding studies is discussed. Chapter two provides a literature review of QTL mapping in general and in cotton. In specific, the final chapter (chapter seven) includes summary and conclusions.

## 1.5 References

- Cronn, R. C., R. L. Small, T. Haselkorn and J. F. Wendel. 2002. Rapid diversification of the cotton genus (*Gossypium: Malvaceae*) revealed by analysis of sixteen nuclear and chloroplast genes. *American Journal of Botany*. 89: 707-725.
- Culp, T. W., and D. C. Harrell. 1975. Influence of lint percentage, boll size, and seed size on lint yield of Upland cotton with high fiber strength. *Crop Sci*. 15: 741-746.
- Brubaker, C. L., F. M. Bourland and J. F. Wendel. 1999. The origin and domestication of cotton. In C. W. Smith and J. T. Cothren [eds]. *Cotton: Origin,*

- history, technology, and production. p. 3-31. John Wiley & Sons, New York, NY.
- Edwards, G. A., and M. A. Mirza. 1979. Genomes of the Australian wild species of Cotton. II. The designation of a new G Genome for *Gossypium bickii*. *Can. J. Genet. Cytol.* 21: 367-372.
- Endrizzi, J. E., E. L. Turcotte and R. Kohel. 1985. Genetics, cytology and evolution of *Gossypium*. *Advances in Genetics.* 23: 271-273.
- Ensminger, M. E., J. E., Oldfield, and W. W. Heinemann. 1990. Excerpts with reference to cottonseed components. In: Ensminger Publishing Company, USA.
- Fryxell, P. A. 1979. The Natural History of the Cotton Tribe. Texas A&M University Press, Collage Station, Texas.
- Jiang, C-X., R. J. Wright, S. S. Woo, T. A. Del Monte, and A. Paterson. 2000. QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). *Theor. Appl. Genet.* 100: 409-418.
- Khan, M. A., J. Zhang, J. McD. Stewart, and R. G. Cantrell. 1998. Integrated molecular map based on a trisppecific F<sub>2</sub> population of cotton. In: 1998 Proc. Beltwide Cotton Conference. 491-492. National Cotton Council Am., Memphis, TN.
- New Mexico Agricultural Statistics Service. 2003. Special survey results acreage and yield review. [www.nass.usda.gov/nm](http://www.nass.usda.gov/nm)
- Poehlman, J. M., and D. A. Sleper. 1995. Breeding field crops. Fourth Edition. Iowa State University Press, USA 494 p.
- Reinisch, M.J., J. Dong, C.L. Brubaker, D.M. Stelly, J.F. Wendel, and A.H. Paterson. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: Chromosome organization and evolution in a disomic polyploid genome. *Genetics.* 138:829-847.
- Report of the American Cotton Producers. 1999. Blue Ribbon Yield Committee National Cotton Council Am., Memphis, TN.
- Sax, K. 1923. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics.* 8: 552-560.
- Saranga, Y., M. Menz, C-X. Jiang, R. L. Wright, D. Yakir, and A. H. Paterson. 2001. Genomic dissection of genotype x environment interactions conferring adaptation of cotton to arid conditions. [www.genome.org](http://www.genome.org).

- Shappley, Z.W. 1994. RFLPs in cotton (*Gossypium hirsutum* L.): Feasibility of use, diversity among plants within a line, and establishment of molecular markers and linkage groups among two F<sub>2</sub> populations. M.S. thesis. Mississippi State Univ., Mississippi State.
- Shappley, Z. W., J. N. Jenkins, J. Zhu, and J. C. McCarty, Jr. 1998. Quantitative trait loci associated with agronomic and fiber traits of Upland cotton. *The Journal of Cotton Science* 4: 153-163.
- Smith, W. C. 1999. Production Statistics. In Smith, W. C. and J. T. Cothern (eds) . *Cotton: Origin, history, technology, and production*. John Wiley and Sons, Inc. New York.
- Tanksley, S. D. 1993. Mapping polygenes. *Annu. Rev. Genet.* 27: 205-233.
- Thoday, J. M. 1961. Location of polygenes. *Nature.* 191: 368-370.
- Ulloa, M., R. G. Cantrell, R. G. Percy, E. Zeiger, and Z. Lu. 2000. QTL analysis of stomatal conductance and relationship to lint yield in an interspecific cotton. *The Journal of Cotton Science.* 4: 10-18.
- Wright, R. J., P.M. Thaxton, K.M. El-Zik, and A. H. Paterson. 1999. Molecular mapping of genes affecting pubescence of cotton. *The American Genetic association* 90:215-219.
- Zhao, X., Y. Ji, X. Ding, D. M. Stelly, and A. H. Paterson. 1998. Macromolecular organization and genetic mapping of a rapidly evolving chromosome-specific tandem repeat family (B77) in cotton (*Gossypium*). *Plant Molecular Biology.* 38: 1031-1042.

## CHAPTER 2 LITERATURE REVIEW

### 2.1 Molecular Markers

Until recent advances in molecular genetics, breeders have been improving both qualitative and quantitative inherited traits by conventional breeding methods based on phenotypic evaluation and selection, which are resource-consuming. Currently, two main types of molecular markers, biochemical markers and DNA-based markers, are available for genetic studies. Tanksley (1983) listed five properties that distinguish molecular markers from morphological markers. These properties are: (1) genotypes can be determined at the whole plant, tissue and/or cellular level, (2) a relatively large number of naturally occurring alleles exist at many loci, (3) phenotypic neutrality; deleterious effects are not usually associated with different alleles, (4) alleles at many loci are codominant, thus all possible genotypes can be distinguished, and (5) few epistatic or pleiotropic effects are observed.

#### 2.1.1 Biochemical Markers

Markert and Moller (1959) were first to describe the differing forms of bands that they were able to visualize with specific enzyme stains. They were the first to introduce the term biochemical polymorphisms often referred to as allozyme or isozyme markers. By the early 1980s, biochemical markers had been employed as a general tool for mapping QTL (Weller et al., 1988). In cotton, *Pgm<sub>7</sub>*, which encodes a monomeric phosphoglucomutase isozyme, was the first biochemical locus to be mapped and recently has been localized to the long arm of chromosome 12 (Saha and Stelly, 1994).

Isozymes are functionally similar forms of enzymes (Murphy et al., 1990). Allozymes are different forms of the same enzyme resulting from allelic variation

(Crozier, 1993), which display differential mobility with electrophoretic techniques and can be detected by staining for enzyme activity (Conkle et al., 1982). The net charge of the protein influences its movement in an electrical field (Hartl, 1988); other important factors influencing protein migration are its size and shape (Murphy et al., 1990). Biochemical studies met with considerably more success than previous studies using morphological markers. However, the number of genetic markers provided by isozyme assays was insufficient in many plant breeding applications. (Tanksley, 1983; Tanksley, 1993).

### 2.1.2 DNA-Based Markers

The introduction of the polymerase chain reaction (PCR) has enabled the development of powerful genetic markers. However, most recent DNA-based markers fall into one of three basic categories depending upon the techniques that are used:

#### 2.1.2.1 Hybridization-Based (non-PCR) Techniques

This technique is exemplified by restriction fragment length polymorphism (RFLP) analysis. Herein, probes are hybridized to filters containing DNA that has been digested with restriction enzymes. The resultant fragments are separated by gel electrophoresis and transferred onto filters by southern blotting. Hybridization can also be carried out with probes for minisatellite or microsatellite sequences to yield a variable number of tandem repeats (VNTR) and allow oligonucleotide fingerprinting (Karp and Edwards, 1997).

### 2.1.2.2 Arbitrarily Primed PCR and Other PCR-Based Multi-Locus Profiling Techniques

A common feature of these techniques is the lack of requirement for sequence information from the genome under investigation. However, they differ in the length and sequence of the primers used, the stringency of the PCR conditions, and the method of fragment separation and detection. In RAPD analysis (an arbitrarily-primed PCR technique), the amplification products are separated on agarose gels in the presence of ethidium bromide and visualized under ultraviolet light. In the other PCR-based multi-locus profiling techniques, the primers used are semi-arbitrary in that they are based on restriction enzyme sites, e. g. AFLP where DNA is digested with two restriction enzymes, adaptors are ligated, and then PCR is carried out with generic primers that comprise a common part corresponding to the adaptors and restriction site and a unique part corresponding to the selective bases or sequences that are interspersed in the genome, such as repetitive elements, transposable elements and microsatellites (Karp and Edwards, 1997).

Microsatellite (repeat)-primed PCR (MP-PCR) includes: (a) unanchored single SSR primer amplification reactions (SPAR) in which the variation is not SSR-based, (b) inter SSR (ISSR) in which the variation is between SSRs rather than at SSR; in this technique, SSR primer anchored at the 5' or 3' end, (c) randomly amplified microsatellite polymorphism (RAMP) performed between a 5' anchored mono, di, or a tri-repeat and an arbitrary decamer primer; this technique does reflect the variation in SSR, and (d) selective amplification of microsatellite polymorphic loci (SAMPL) in which one labeled SSR primer (anchored) and one unlabeled adaptor primer are used (Karp and Edwards, 1997)

### 2.1.2.3 Sequence Targeted and Single Locus PCR

In general, a limitation of arbitrarily amplified DNA is the lack of allelic information for both dominance and assignment of alleles to loci. These problems are overcome with PCR directed to specific single-locus targets, for which a prerequisite is knowledge of the sequence of the target or flanking target regions. If SSR loci are cloned and sequenced, primers to the flanking region can be designed to produce a sequence-tagged microsatellite site (STMS), or SSR markers as they are often called. The SSRs are highly attractive markers because each primer pair typically identifies a single locus that, because of the high mutability of SSR loci, may have many alleles (Karp and Edwards, 1997).

## **2.2 DNA-Markers Used in Cotton**

A comparison of DNA-markers used in cotton improvement is shown in Table 2.1.

Table 2.1 Comparison of Different DNA-Marker Systems.

	RFLP	RAPD	SSR	AFLP	ISSR
Definition	Southern blotting of restricted fragments	PCR of random primers	PCR of microsatellite	Detection of DNA restriction fragments by PCR	PCR of inter simple sequence repeats
Abundance	High	High	High	High	Medium-high
Level of polymorphism	Medium	Medium	High	Medium	Medium
Codominance of alleles	Yes	No	Yes	No	No
Loci number	1-2	3-15	1	40-120	3-12
Locus specificity	Yes	No	Yes	No	No

(Table cont'd)

	RFLP	RAPD	SSR	AFLP	ISSR
Reproducibility	High	Low	High	High	Medium-high
Labor intensity	High	Low	Low	Medium	Low
Prior sequence information	Required	Not required	Required	Not required	Not required
Amount of DNA required	2-10 µg	10-20 ng	20-50 ng	20-500 ng	10-20 ng
Development cost	High	Low	High	Medium	Low
Assay cost	Medium	Low	Medium	Medium	Low
Technical demands	High	Low	Low-medium	Medium	Low
Amenability to automation	No	Yes	Yes	Yes	Yes
Reproducibility	High	Fair	High	High	Medium-high

This information suggests that RFLP, SSR and AFLP are the most effective and reliable methods for detecting polymorphism. However, given the large amount of DNA required for RFLP detection and the difficulties in automating RFLP analysis, AFLP and SSR are the most promising methods. In cotton, most work has been done using RFLPs (Shapple et al., 1998b).

### 2.2.1 Restriction Fragment Length Polymorphism

Restriction fragment length polymorphism (RFLP) was the first DNA marker technology to be utilized; the first true RFLP map in a crop plant (tomato) was constructed in 1986 with 57 loci (Bernatzky and Tanksley, 1986). It refers to the variation among individuals in the lengths of DNA fragments produced by restriction

enzymes that cut DNA at specific sites. RFLP analysis, in its original form, consists of DNA digestion with a restriction enzyme, separation of the restriction fragments by agarose gel electrophoresis, transfer of the separated restriction fragments to a filter by Southern blotting and detection by autoradiography. The differing sizes of the DNA fragments may result from base substitutions, additions, deletions, or sequence rearrangements within restriction enzymes recognition sequences (Avisé, 1994).

In a study of heterosis and varietal origins, Meredith (1992) reported the first RFLP evaluations in Upland cotton. Also, a detailed RFLP map was used to map genes affecting density of leaf and stem trichomes (Wright et al., 1999). Detailed RFLP maps of cotton with 41, 5, 31, 24 and 17 linkage groups were developed by Reinisch et al. (1994), Shappley et al. (1996;1998a;1998b) and Ulloa and Meredith (2000), respectively.

### 2.2.2 Amplified Fragment Length Polymorphism

One of the newest and most promising methods for mapping is the amplified fragment length polymorphism (AFLP) technique, previously known as selective restriction fragment amplification (SRFA) (Zabeau and Vos, 1993). It was originally conceived to allow the construction of very high density DNA marker maps.

The AFLP technique is now one of the most frequently used molecular marker technologies in modern crop improvement research. It is based on the detection of DNA restriction fragments by PCR amplification (Vos et al., 1995). For AFLP analysis, only a small amount of purified genomic DNA is needed (20-500 ng, Table 2.1); this is digested with two different restriction enzymes, generally a rare-cutter (e.g., *EcoRI*) and a frequent cutter (e.g., *MseI* or *TaqI*). Adapters are designed such that the initial restriction site is not restored after ligation, which allows simultaneous

restriction and ligation. Depending on genome size, restriction-ligation generates thousands of adapted fragments. For visualization after electrophoresis, only a subset of these fragments is amplified. To achieve selective amplification of a subset of these fragments, primers are extended into the unknown chromosomal restriction fragments. An extension of one selective nucleotide amplifies 1 of 4 of ligated fragments, two selective nucleotides in both primers amplify 1 of 256 of the fragments, whereas three selective nucleotides in both primers amplify 1 of 4096 of the fragments. To minimize artifacts for large genome size protocols (Table 2.2), most protocols incorporate two amplifications. The first is performed with a 1-bp extension, followed by a 3-bp extension.

Table 2.2 Example of correlations between genome size, enzyme combination (EC), pre-amplification (PA) and amplification strategy (Amp) in various organisms. Genome size is indicated in megabases. Restriction enzymes include *EcoRI* (E), *MseI* (M), *PstI* (P) and *TaqI* (T). The number of selective bases for AFLP (pre)amplification is given in the last two columns (Vos et al., 1995).

Organisms	Genome size	EC	PA	Amp
1. Cosmids, BACs†, PACs‡	0.01		E/M	- 0/0
2. YACs§, microorganisms	0.1-1	E/M	-	0/1
3. Microorganisms	1-5	E/M	-	1/1
4. Microorganisms	5-20	E/M	-	1/2
5. Fungi	20-100		E/M	- 2/2
6. Plants, invertebrates	100-500	E/M	0/1	2/3
7. Plants	500-5000	E/M	1/1	3/3
8. Plants	>5000	P/M	1/1	3/3
9. Mammals, vertebrates	ca. 3000	E/T	1/1	3/3

†Bacterial artificial chromosome

‡P1-derived artificial chromosome

§Yeast artificial chromosomes

Polyacrylamide gel electrophoresis is then used for DNA separation. Several detection methods could be used, ranging from simple agarose electrophoresis, based on one enzyme with single adapter (Gibson et al., 1998) to automated genotyping using a DNA sequencer. The numbers of fragments that can be analyzed

in one reaction is typically 40-120 restriction fragments. However, the presence of polymorphic fragments is due to the insertions or deletions within the amplified fragments, mutations in the restriction sites or in the selective primer extension sites. Altaf et al. (1997) developed a map of 11 linkage groups that covered 521.7 cM of the cotton genome (4700 cM) with a mean distance of 16.8 cM between markers, using both RAPD and AFLP.

The AFLP technique has been used for the construction of linkage maps in several crops, such as barley (Costa et al., 2001; Vaz Patta et al., 2003) and rice (Xu et al., 2000); for marker saturation in barley (Lahaye et al., 1998), rice (Maheswaran et al., 1997) and potato (Bendahmane et al., 1997); for the analysis of genetic diversity in *Arabidopsis* (Erschadi et al., 2000; Breyne et al., 1999); for molecular phylogeny in potato (Kardolus et al., 1998); and for cultivar identification in potato (McGregor et al., 2002).

### 2.2.3 Randomly Amplified Polymorphic DNA

While no one marker system can be considered ideal for all molecular marker applications, the randomly RAPD method provides a valuable tool in the repertoire of a molecular geneticist and has many advantages: non-radioactive detection, no prior DNA sequence information for a genome is required, universal primers work in any genome, very small amount of genomic DNA is needed, experimental simplicity, and no need for expensive equipment beyond a thermocycler and a transilluminator (Rafalski, 1997). However, it is often criticized for its lack of reproducibility (Jones et al., 1997). Several factors may influence reproducibility of RAPD profiles within and between laboratories including DNA concentration, reproducibility of thermocycler

profiles, primer quality and concentration, choice of DNA polymerase, and pipetting accuracy (Rafalski, 1997).

RAPD markers (Williams et al., 1990; Welsh and McClelland, 1990) have provided a significant advance in the construction and saturation of genetic maps in tomato (*Lycopersicon esculentum* L.) (Martin et al., 1991), rice (*Oryza sativa* L.) (Monna et al., 1994), and wheat (*Triticum monococcum* L. and *T. boeoticum* L.) (Kojima et al., 1998). RAPDs have greatly facilitated linkage mapping in cotton (Khan et al., 1999; Khan et al., 1998; Zhang et al., 2002). Zhang et al. (2002) used both RAPDs and SSR to construct a map containing 43 linkage groups to investigate the homeologous chromosomal regions of the A and D sub-genomes in the allotetraploid cotton genome.

#### 2.2.4 Simple Sequence Repeat

The applicability of microsatellite markers in genome analysis primarily depends on three inherent circumstances: abundance, hypervariability and in most cases, stable Mendelian inheritance (Ellegren, 1993). Simple sequence repeats or microsatellites are short, tandemly repeated DNA sequence motifs that consist of two to six nucleotide core units, and were initially described in humans (Litt and Luty, 1989). They are highly abundant in eukaryotic genomes but also occur in prokaryotes at lower frequencies. They seldom include more than 70 repeat units and are interspersed throughout the genome. These small repetitive DNA sequences provide the basis for a PCR-based, multi-allelic, co-dominant genetic marker system. The high incidence of detectable polymorphisms through changes in repeat numbers is caused by an intramolecular mutation mechanism called DNA slippage. However,

the most common mutations involve the gain or loss of a single repeat unit (Schlotterer and Tautz, 1992).

The regions flanking the microsatellite are generally conserved among genotypes of the same species. PCR primers relative to the flanking regions are used to amplify SSR-containing DNA fragments. The length of the amplified fragment will vary according to the number of repeated residues and this can simply be measured by electrophoresis of amplified products (Ellegren, 1993). The ability of these hypervariable regions to reveal high allelic diversity is particularly useful in distinguishing between closely related genotypes. SSRs are now considered as the marker of choice for self-pollinated crops with little intraspecific polymorphism (Roder et al., 1998). Furthermore, the reproducibility of SSRs is such that they can be efficiently used by different laboratories to produce consensus data, which makes them useful for genome mapping projects and results in their successful isolation and application within many plant species (Dietrich, 1996; Dib et al., 1996; Schmidt and Heslop, 1998).

Multiplex PCR bins of SSR primers and semi-automated detection of the amplified products are the key factors for high-throughput genotyping and improving the efficiency of genetic mapping and marker-assisted programs utilizing SSR markers. Multiplex PCR is based on the simultaneous amplification of several microsatellite loci in a single PCR tube. The most critical step for the establishment of multiplex PCRs is to choosing the correct PCR condition, primer combinations, and annealing temperature. This step can be avoided by amplifying the microsatellite loci separately and subsequently pooling the PCR product. Analyzing the pooled microsatellite PCRs on a single gel still provides considerable time savings

(Schlotterer, 1998). For example, 13 multiplex PCR bins were optimized to contain, on average, four cotton SSR primer pairs per bin (Liu et al., 2000).

Current microsatellite analysis relies on size determination of the entire PCR product consisting of the microsatellite stretch and flanking regions. The number of repeats can be calculated by subtraction of the flanking nucleotides and dividing the remaining base pairs by the size of the repeat unit. As a rule of thumb, the separation capacity of the gel should be at least half the size of the repeat unit. Therefore, sizing of PCR products on agarose gels for most microsatellites is not appropriate as they provide too little resolution. The most commonly used gel type is a 6% denaturing polyacrylamide gel as heteroduplex molecules generated during the late PCR cycles of heterozygous individuals will result in a third band (sometimes also a fourth), which may cause an incorrect assignment of alleles (Schlotterer, 1998).

The original and most sensitive approach for microsatellite detection is based on radioactivity with two different methods for labeling the PCR product: incorporation of labeled nucleotides and end-labeling one of the PCR primers. However, silver staining, blotting hybridization and fluorescent dyes on automated sequencers can be used as non-radioactive detection methods. The use of fluorescent dyes on automated sequencers is a relatively new detection method where the amplified PCR products are labeled with a fluorescent dye (either by incorporation during PCR or by using an end-labeled PCR primer). When activated by laser light, this dye emits a signal that can be detected and by comparing the migration of the PCR product with a length marker, accurate sizing is possible (Schlotterer, 1998). Switching to this method eliminates the problems of dealing with

radioactivity-based systems, such as exposure of investigators to radioactivity, high cost of radioactive waste disposal, time-consuming paper work involved in radioisotope use, and short half-lives of radioisotopes. Fluorescent-labeled primers remain stable for years if properly stored (David and Menotti-Raymond, 1998).

Genome maps based, at least in part, on microsatellite markers now exist for a number of plant species, such as *Arabidopsis* (Bell and Ecker, 1994), tomato (Broun and Tanksley, 1996), rice (Cho et al., 2000), wheat (Borner et al., 2000), tetraploid potato (Bradshaw et al., 1998) and other animal and plant species (Powell et al., 1996; Gyapay et al., 1994; Sverdlov et al., 1998).

In cotton, SSRs represent a new class of genetic markers. Liu et al. (2000) used 65 SSR primer pairs to amplify 70 marker loci localized to a specific cotton chromosome or genome. The SSR markers identified in this study provide a framework that can be used with further conventional linkage mapping to other DNA markers to expand the genome-wide coverage of the cotton genetic map. In fact, a linkage map was recently produced with 199 RAPD and SSR DNA markers to assist in selection for cotton stomatal conductance; two putative QTL for this difficult-to-measure physiological trait were identified on two cotton linkage groups (Ulloa et al., 2000)

#### 2.2.5 Inter Simple Sequence Repeat

In contrast to the SSR marker technique that amplifies with primers located on the flanking single-copy DNA, microsatellite anchored primers that anneal to an SSR region can amplify regions between adjacent SSRs. The inter-simple sequence repeat (ISSR) technique uses primers that are complementary to a single SSR and anchored at either the 5' or 3' end with a one- to three-base extension (Zietkiewicz

et al., 1994). This anchor ensures that the primer binds only to one end of a complementary SSR locus. Amplicons generated consist of the region between neighboring and inverted SSRs. As a result, the highly complex banding pattern obtained will often differ greatly between genotypes of the same species. ISSRs have been used for linkage map construction in wheat (Kojima et al., 1998; Nagaoka and Ogihara, 1997), potato (*Solanum tuberosum* L.) (Prevost and Wilkinson, 1999), citrus (*Poncirus trifoliata* L.) (Sankar and Moore, 2001), watermelon (*Citrullus lanatus* L.) (Hashizume et al., 2003) and chickpea (*Ascochyta rabiei* L.) (Flandez-Galvez et al., 2003). However, no ISSR base linkage map has been reported in cotton.

### **2.3 Fingerprinting and Diversity Studies**

Plant breeders desire their new varieties to be distinct, uniform and stable (D, U and S criteria) (Cooke, 1995). In the past, the ability to discriminate between varieties was heavily dependent on morphological traits. Lately, DNA markers have been employed as a promising method of fingerprinting. For example, Lu and Myers (2002) evaluated the level of genetic diversity of 10 influential cotton varieties using RAPD markers. They were able to individually identify all tested varieties by specific markers in genetic fingerprinting. Similar to other crops, an understanding of the evolutionary and genomic relationships of cotton species and cultivars is critical for further utilization of extant genetic diversity in the development of superior cultivars (El-Zik and Thaxton, 1989).

### **2.4 Linkage Maps**

The molecular information of a crop genome is usually presented in the framework of a genetic linkage maps that are useful to locate or tag genes of interest, to facilitate MAS, and map based cloning. With the introduction of molecular markers

in quantitative genetics, the problem of finding suitable genetic markers can be considered solved. It is now clear that a genetic map nearly saturated with polymorphic molecular markers can be generated for almost any species. In fact, such maps have already been produced for many species of economic or scientific interest.

Genetic maps are also essential to locate the genes that are involved in the expression of traits. This can easily be done for simple heritable traits based on one gene, but is also possible for complex traits that are based on more genes (QTL). In the latter case, large segregating populations ( $n > 100$ ) are required to unravel the number of loci involved in the trait (Jeuken et al., 2001).

In cotton, the contribution of new markers to generate a more saturated Upland cotton linkage map will enhance our understanding of its genetics and improve cotton breeding efficiency, especially when quantitative traits are implicated. Shappley (1994) and Reinisch et al. (1994) separately reported the first linkage map constructed with RFLP markers. Shappley (1994) constructed five linkage groups in a cross of Upland cotton, while 41 linkage groups were constructed in a cross between *G. hirsutum* and *G. barbadense*. Since then, Wright et al. (1999), Saranga et al. (2001), Shappley et al. (1996), Shappley et al. (1998a;1998b), Yu et al. (1998), Kohel et al. (2001), Jiang et al. (2000), and Ulloa and Meredith (2000) have developed detailed RFLP maps of cotton. Also, other types of markers have been used in linkage map construction. This includes, RAPDs in a study conducted by Kohel (2001), Zhang et al. (2002) who used both RAPDs and SSRs to construct 43 linkage groups. Altaf et al. (1997) constructed 11 linkage groups using both RAPDs

and AFLPs. To the author's knowledge, there are no refereed publications using AFLP in constructing linkage groups in Upland cotton.

## **2.5 Mapping Quantitative Trait Loci**

Most agronomically important characteristics of crops are inherited quantitatively and are under the influence of both the environment and the genetic factors determined by QTL (Gelderman, 1975). Since it is not practical to infer an individual's genotype from its phenotype, it is difficult task to identify and characterize the QTL.

Mendel (1866) wrote that complex variation in the color of flowers might be due to the independent action of several genetic factors. It was not until 1923, when Sax demonstrated an association between seed weight and seed-coat color in beans. This association was proposed to be due to linkage between genes controlling seed-color and one or more genes controlling seed-size (Sax, 1923). Thoday (1961) was the first to use multiple genetic markers to map individual polygenes controlling a quantitative trait.

Most studies involving the identification of QTL begin with two inbred lines, which differ in the trait(s) of interest. Crossing these two parental lines gives the first filial ( $F_1$ ) generation, which is the start for the construction of backcross,  $F_{2:3}$ , and recombinant inbred populations (derived by inbreeding  $F_2$  progeny until they become virtually homozygous lines by selfing or sibbing) (Tanksley, 1993). In the  $F_2$  population, each individual will receive two chromosomes from the  $F_1$  generation, each of which is a combination of the two parental chromosomes.

Information on QTL analysis has accumulated quickly, and will eventually help the manipulation of the complex traits in cotton breeding (Tanksley, 1993). In cotton,

several QTL studies have been conducted using both intra- and inter-specific crosses mainly using RFLPs as markers to construct linkage maps. In a study investigating 19 agronomic and fiber traits, 100 QTL were mapped to 60 positions in 24 linkage groups. Several QTL influence more than one trait (Shappley, 1998a). The most frequent association of QTL with multiple traits was for fiber traits related to maturity and fineness (Shappley et al., 1998a). Jiang et al. (2000) used 180 F<sub>2</sub> plants from a cross of *G. hirsutum* and *G. barbadense* to map a total of 62 QTL for 14 different traits. He found that 38 (61.3%) QTL mapped to the D-genome. Several other studies suggest that the D-genome of tetraploid cotton has been subjected to a relatively greater rate of evolution than the A-genome, subsequent to polyploid formation. Twenty six QTL were detected on nine linkage groups constructed from 119 F<sub>2:3</sub> progeny from a cross between MD 567ne and Prema (Ulloa and Meredith, 2000). Two QTL were detected for lint yield and three for lint percentage, explaining from 5% to 20 % of the variation in each trait. Three QTL for fiber strength (explaining 10.6-24.6% of the phenotypic variation), four for Micronaire (explaining 6.2-21.7%), three for fiber strength (explaining 3.4-31.6%) and two for fiber 2.5%-span-length (explaining 11.5-44.6%) were detected. In a study of fiber quality traits, Kohel et al. (2001) used an F<sub>2</sub> population derived from an interspecific cross between TM-1 (*G. hirsutum*) and 3-79 (*G. barbadense*) to map 13 QTL. Four QTL influenced bundle fiber strength, three influenced fiber length and six influenced fiber fineness. These QTL collectively explained 30% to 60% of the total phenotypic variation.

## **2.6 Methods of QTL Mapping**

Modern analysis of the genetics of quantitative traits utilizes large sets of molecular markers for genome scanning that are capable of identifying genetic factor(s) associated with trait(s) in a mapping population. The basic principle underlining QTL detection is to partition the mapping population into different genotypic classes based on genotypes at the marker locus and then to use correlative statistics to determine whether the individuals of one genotype differ significantly compared with individuals of other genotype(s) with respect to the trait being measured. If the phenotypes differ significantly, the interpretation is that a gene(s) affecting the trait is (are) linked to the marker locus (loci) used to subdivide the population (Tanskley, 1993).

There are a large number of different methods for identifying the QTL segregating in a mapping population. These methods can be divided into methods that model a single QTL at a time (single QTL methods) and methods that model the effect of several QTL at once (multiple QTL methods). Single QTL methods include analysis of variance (t test or F test) (Soller et al., 1976), interval mapping (maximum likelihood using flanking markers (Lander and Botstein, 1989), and regression mapping (an approximation to interval mapping) (Knapp et al., 1990; Haley and Knott, 1992).

## **2.7 QTL X Environment Interaction**

Genotype X environment interactions are a challenge to plant breeders because they cause difficulties in selecting genotypes evaluated in different environments (Kang and Gorman, 1989). Problematically, it also leads to variable

levels of the significance of QTL effects across environments (Hayes et al., 1993 and Romagosa et al., 1996).

QTL detected in one environment but not in another might indicate interaction (Veldboom and Lee, 1996). Zhu (1998) proposed an indirect method to map QTL with QTL X environment effects using predicted total genotype X environment interaction effects. It was shown that some QTL had both genetic main effects and QTL X environment interaction effects, even though they could be detected in two environments (Yan et al., 1998). Recently, a new methodology was proposed to analyze QTL X environment interactions based on mixed linear model approaches (Wang et al., 1999).

## **2.8 Marker-Assisted Selection**

The detection of relationships between genetic markers and QTL could be valuable for several reasons: it may give us fundamental knowledge about the number of QTL and the magnitude of gene effects influencing the traits, and it may allow us to build more realistic models of phenotypic variation and of responses to selection (Haley, 1991). Moreover, marker information can be used for identification and possibly for introgression of genes of interest from foreign species. Once relationships between genetic markers and quantitative traits have been detected, and the marker allele substitution effects have been estimated, it will be possible to use MAS for such traits to increase the selection response, the accuracy of evaluation, or decrease the generation interval. In fact, many practical breeding situations are encountered in which trait-based selection index is very inefficient or impractical. In these instances, being able to use a marker-based selection index would be a very significant gain. With MAS, an increased emphasis can be put on

early selection and thus influence the selection response. Moreover, the relative efficiency of MAS is greatest for characters with low heritability if a large fraction of the additive genetic variance is associated with the marker loci (Lande and Thompson, 1990). Limitations that may affect the potential utility of MAS in applied breeding programs include: (i) the level of linkage disequilibrium in the populations, which affects the number of marker loci needed (ii) the sample size needed to detect QTL for traits with low heritability, and (iii) sampling errors in the estimation of relative weights in selection indices. In practice and through the development of RFLP markers in the early 1980s, indirect selection in plant breeding using markers became technically feasible. However, the laborious nature of the RFLP technique prevented a broad application of RFLP-marker-assisted breeding. In the 1980s and early 1990s, molecular diagnostic methods based on PCR technology, such as RAPD, AFLP and SSR emerged. The use of these markers to enhance plant breeding efforts has been described by many investigators (Paterson et al., 1991; Dudley, 1993; Bi et al., 1999; Stuber et al., 1999).

In cotton, Wright et al. (1999) identified, using a detailed RFLP linkage map from Reinisch et al. (1994), DNA markers diagnostic for five genes associated with the pubescence of cotton leaves and/or stems. Absence of trichomes reduces the attractiveness of the cotton plant to some major insect pests, reducing the need for pesticides. Using an  $F_{2:3}$  population derived from a cross between a *G. anomalum* introgression line (7235) and *G. hirsutum* (M-1), Zhang et al. 2003 identified nine molecular markers (three SSRs and six RAPDs) linked to two QTL for fiber strength. One was a major QTL detected both in Nanjing and Hainan, China and at College

Station, Texas. It was found to be associated with eight markers and explained more than 30% of the phenotypic variation (Zhang et al., 2003).

## 2.9 References

- Altaf, M. K., J. McD. Stewart, M. K. Wajahatullah, and J. Zhang. 1997. Molecular and morphological genetics of a trispecies F2 population of cotton. Proc. Beltwide Cotton Conf. 448-452.
- Avise, J. C. 1994. Molecular markers, natural history and evolution. Chapman & Hall, New York, NY.
- Bell, C.J., Jr. Ecker. 1994. Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis*. Genomics. 19: 137-144.
- Bendahmane, A. K. Kanyuka, D. C. Baulcombe. 1997. High-resolution genetical and physical mapping of the Rxgene for extreme resistance to potato virus X in tetraploid potato. Theor. Appl. Genet. 1997. 95: 153-162.
- Bernatzky, R., Tanksley, S.D. 1986. Towards a saturated linkage map in tomato based on isozyme and random cDNA sequences. Genetics. 112: 887-898.
- Bi, I. V., A. Maquet, J.-P. Baudoin, and P. Du. Jardin. 1999. Breeding for "low-gossypol seed and high-gossypol plants" in Upland cotton. Analysis of tri-species hybrids and backcross progenies using AFLPs and mapping RFLPs. Theor. Appl. Genet. 99: 1233-1244.
- Borner, A., M. S. Roder, O. Unger, and A. Meinel. 2000. The detection and molecular mapping of a major gene for non-specific adult-plant disease resistance against stripe rust (*Puccinia striiformis*) in wheat. Theor. Appl. Genet. 100: 1095-1099.
- Bradshaw, J. E., C. A. Hackett, R. C. Meyer, D. Milbourne, J. W. McNicol, M. S. Phillips, and R. Waugh. 1998. Identification of AFLP and SSR markers associated with quantitative resistance to *Globodera pallida* (Stone) in tetraploid potato (*Solanum tuberosum* subsp. *tuberosum*) with a view to marker-assisted selection. Theor. Appl. Genet. 97: 202-210.
- Breyne, P., D. Rombaut, A. Van Gysel, M. Van Montagu, T. Gerats. 1999. AFLP analysis of genetic diversity within and between *Arabidopsis thaliana* ecotypes. Mol. Gen. Genet. 261: 627-634.
- Broun, P., and S. D. Tanksley. 1996. Characterization and genetic mapping of simple repeat sequences in the tomato genome. Mol. Gen. Genet. 250: 39-49.

- Cho, Y. G., T. Ishii, S. Temnykh, X. Chen, L. Lipovich, S. R. McCouch, W. D. Park, N. Ayres, and S. Cartinhour. 2000. Original: Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 100: 713-722.
- Conkle, M. T., P. D. Hodgskiss, L. B. Nunnally, and S. C. Hunter. 1982. Starch gel electrophoresis of conifer seeds; a laboratory manual. USDA For. Ser. Gen. Tech. Report PSW-64. 18 pp.
- Cooke, R. J. 1995. Introduction: the reasons for variety identification. *In* Wrigley CW (eds.): *Identification of Food Grain Varieties*, p. 1-17, American Association of Cereal Chemists, St. Paul, MN, USA.
- Costa, J. M., A. Corey, P. M. Hayes, C. Jobet, A. Kleinhofs, A. Kopisch-Obusch, S. F. Kramer, D. Kudrna, M. Li, O. Riera-Lizarazu, K. Sato, P. Szucs, T. Toojinda, M. I. Vales, and R. I. Wolfe. 2001. Molecular mapping of the Oregon Wolfe Barleys: a phenotypically polymorphic doubled-haploid population. *Theor. Appl. Genet.* 103: 415-424.
- Crozier, R. H. 1993. Molecular methods for insect phylogenetics, pp. 164-221. *in* J. Oakeshott and M. Whitten (eds). *Molecular approaches to fundamental and applied entomology*. Springer-Verlag, New York.
- David, V. A., and M. Menotti-Raymond. 1998. Automated DNA detection with fluorescence-based technologies. *In*: Hoelzel, A. R. (editor). *Molecular Genetic analysis of populations: A practical approach*. Oxford University Press. USA. 445.
- Dib, C., S. Faure, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun, M. Lathrop, G. Gyapay, J. Morissette, and J. Wellssenbach. 1996. A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature.* 380: 152-154.
- Dietrich, W. F., J. Miller, R. Steen, M. A. Merchant, D. Damronboles, Z. Husain, R. Dredge, M. J. Daly, K. A. Ingalls, T. J. O'Connor, C. A. Evans, M. M. DEAngelis, D. M. Levinson, L. Kruglyak, N. Goodman, N. G. Copeland, N. A. Jenkins, T. L. Hawkins, L. Stein, D. C. Page, and E. S. Lander. 1996. A comprehensive genetic map of the mouse genome. *Nature.* 380: 149-152.
- Dudley, J. W. 1993. Molecular markers in plant improvement: manipulation of genes affecting quantitative traits. *Crop Sci.* 33: 660-668.
- Ellegran, H. 1993. Genome analysis with microsatellite markers. Ph. D. Dissertation. University of Agricultural Science. Swedish.
- El-Zik, K. M., and P. M. Thaxton. 1989. Genetic improvement for resistance to pests and stresses in cotton. p. 191-224. *In* R. E. Frisbie, K. M. El-Zik, and L. T.

- Wilson (eds) Integrated Pest Management Systems and Cotton Production. John Wiley & Sons, NY.
- Erschadi, S., G. Haberer, M. Schöniger, and R. A. Torres-Ruiz. 2000. Estimating genetic diversity of *Arabidopsis thaliana* ecotypes with amplified fragment length polymorphisms (AFLP). *Theor. Appl. Genet.* 100: 633-640.
- Flandez-Galvez, H., R. Ford, E. C. K. Pang, and P. W. J. Taylor. 2003. An intraspecific linkage map of the chickpea (*Cicer arietinum* L.) genome based on sequence tagged microsatellite site and resistance gene analog markers. *Theor. Appl. Genet.* 106: 1447-1456.
- Geldermann, H. 1975. Investigations on inheritance of quantitative characters in animals by gene markers. I. Methods. *Theor. Appl. Genet.* 46: 319-330.
- Gibson, J. R., E. Slater, J. Xerry, D. S. Tompkins, and R. J. Owen. 1998. Use of an amplified-fragment length polymorphism technique to fingerprint and differentiate isolates of *Helicobacter pylori*. *J. Clin. Microbiol.* 36: 2580-2585.
- Gyapay, G., J. Morissette, A. Vignal, C. Dip, C. Fizames, P. Millasseau, S. Marc, G. Bernardi, M. Lathrop, and J. Weissenbach. 1994. The 1993-39 Genethon human genetic linkage map. *Nature Genetics.* 7: 246-339.
- Haley, C. S. 1991. Use of DNA fingerprints for the detection of major genes for quantitative traits in domestic species. *Anim. Genet.* 22: 259-277.
- Haley, C. S., and S. A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using lanking markers. *Heredity.* 69: 315-324.
- Hartl, D. L. 1988. A primer of population genetics. Sinauer Associates. Sunderland, MA.
- Hashizume, T., I. Shimamoto, and M. Hirai. 2003. Construction of a linkage map and QTL analysis of horticultural traits for watermelon [*Citrullus lanatus* (THUNB.) MATSUM & NAKAI] using RAPD, RFLP and ISSR markers. *Theor. Appl. Genet.* 106:779-785.
- Hayes, P. M. , B. H. Liu, S. J. Knapp, F. Chen, B. Jones, T. Blake, J. Franckowiak, D. Rasmusson, M. Sorrells, S. E. Ullrich, D. Wesenberg, and A. Kleinhofs. 1993. Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. *Theor. Appl. Genet.* 87: 392-401.
- Jeuken, M., R. van Wijk, J. Peleman, and P. Lindhout. 2001. An integrated interspecific AFLP map of lettuce (*Lactuca*) based on two *L. sativa* *L. saligna* F2 populations. *Theor. Appl. Genet.* 103 :638-647.

- Jiang, C-X., R. J. Wright, S. S. Woo, T. A. Del Monte, and A. Paterson. 2000. QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). *Theor. Appl. Genet.* 100: 409-418.
- Jones, C. J., K. J. Edwaeds, S. Castaglione, M. O. Winfield, F. Sala, C. Van De Weil, G. Bredemeijer, B. Vosman, M. Matthes, A. Daly, R. Brettschneider, P. Bettini, M. Buitti, E. Maestri, A. Malcevschi, N. Marmiroli, R. Aert, G. Volckaert, J. Rueda, R. Linacero, A. Vazquez, and A. Karp. 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Mol Breed.* 3: 381-390.
- Kang, M. S., and D. P. Gorman. 1989. Genotype X environment in maize. *Agron. J.* 81: 662-664.
- Kardolus, J. P., H. J. van Eck, and R. G. van den Berg. 1998. The potential of AFLPs in biosystematics: a first application in *Solanum* taxonomy (*Solanaceae.* ) 210: 87-103.
- Karp, A., and K. J. Edwards. 1997. DNA markers: a global overview. In: Caetano-Anolles, G., and Gresshoff, P. M. (eds.). *DNA Markers: Protocols, Applications, and Overviews.* Wiley-Liss, Inc. USA. 364 p.
- Khan, M. A., J. McD. Stewart, J. Zhang, G. O. Myers, and R. G. Cantrell. 1999. Addition of new markers to the trispecific cotton map. In: *Beltwide Cotton Conference.* 439.
- Khan, M. A., J. Zhang, J. McD. Stewart, and R. G. Cantrell. 1998. Integrated molecular map based on a trispecific F<sub>2</sub> population of cotton. In: *Beltwide Cotton Conference.* 491-492.
- Knapp S. J., W. C. Jr. Bridges, and D. Birkes. 1990. Mapping quantitative trait loci using molecular marker linkage maps. *Theor. Appl. Genet.* 79: 583-592.
- Kohel, R. J., J. Yu. Y-H. Park, and G. R. Lazo. 2001. Molecular mapping and characterization of trait controlling fiber quality in cotton. *Euphatica.* 121: 163-172.
- Kojima, T., T. Nagaoka, K. Noda, and Y. Ogihara. 1998. Genetic linkage map of ISSR and RAPD markers in Einkorn wheat in relation to that of RFLP markers. *Theor. Appl. Genet.* 96: 37-45.
- Lahaye, T., S. Hartmann, S. Topsch, A. Freialdenhoven, M. Yano, and P. Schulze-Lefert. 1998. High-resolution genetic and physical mapping of the Rar1 locus in barley. *Theor. Appl. Genet.* 97: 526-534.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124: 743-756.

- Lander, E. S., and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185-199.
- Litt, M., and J. A. Luty. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* 44: 397.
- Liu, S., S. Saha, D. Stelly, B. Burr, and R. S. Cantrell. 2000. Chromosomal assignment of microsatellite loci in cotton. *The Journal of Heredity.* 91: 326-332.
- Lu, H. J., and G. O. Myers. 2002. Genetic Relationships and Discrimination of Ten Influential Upland Cotton Varieties using RAPD Markers. *Theor. Appl. Genet.* 105: 325-331.
- Maheswaran, M., P. K. Subudhi, S. Nandi, J. C. Xu, Parco, D. C. Yang, and N. Huang. 1997. Polymorphism, distribution, and segregation of AFLP markers in a double haploid rice population. *Theor. Appl. Genet.* 94: 39-45.
- Markert, C. L. and F. Moller. 1959. Multiple forms of enzymes. *Proc. Natl. Acad. Sci. U. S. A.* 45: 753-63.
- Martin, G. B., J. Williams, and S. D. Tanksley. 1991. Rapid identification of markers linked to a *Pseudomonas* resistance gene in tomato by using random primers and near-isogenic lines. *Proc. Natl. Acad. Sci. USA* 88: 2336-2340.
- McGregor, C. E., R. van Treuren, R. Hoekstra, Th. J. L. van Hintum. 2002. Analysis of the wild potato germplasm of the series *Acauliawith* AFLPs: implications for ex situ conservation. *Theor. Appl. Genet.* 104: 146-156.
- Mendel, G. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines.* 1866; 4:3-47.
- Meredith, W. R. 1992. RFLP association with varietal origin and heterosis. p. 607. In: D. Herber (ed.) *Proc. Beltwide Cotton Prod. Res. Conf.*, Nashville, TN. 6-10 Jan. 1992. *Natl. Cotton Council Am.*, Memphis, TN.
- Monna, L., A. Miyao, T. Inoue, S. Fukuka, M. Yamazaki, H. Sun Zhong, T. Sasaki, Y. Monobe. 1994. Determination of RAPD markers in rice and their conversion into sequence-tagged sites (STSs) and STS-specific primers. *DNA Res.* 1: 139-148.
- Murphy, R. W., J. W. S. Jr., D. G. Buth, and C. C. Hauffer. 1990. Proteins I: Isozyme electrophoresis, Pages 45-126 *in* D. M. Hillis, and C. Moritz, eds. *Molecular Systematics.* Sunderland, Mass., Sinaur Associates.

- Nagaoka, T., and Y. Ogiwara. 1997. Applicability of inter-simple sequence repeat polymorphisms in wheat for use as DNA markers in comparison to RFLP and RAPD markers. *Theor. Appl. Genet.* 94: 597–602.
- Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch, S. E. Lincoln, E. S. Lander, and S. D. Tanksley. 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics.* 127: 181–197.
- Powell, W., M. Morgante, C. Andre, M. Hanafey, J. Vogel, S. Tingey, and A. Rafalski. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol. Breed.* 2: 225-238.
- Prevost, A., and M. J. Wilkinson. 1999. A new system of comparing PCR primers applied to ISSR fingerprinting of potato cultivars. *Theor. Appl. Genet.* 98:107-112.
- Rafalski, J. A. 1997. Randomly amplified polymorphic DNA (RAPD) analysis. In: Caetano-Anolles, G., and Gresshoff, P. M. (eds.). *DNA markers: Protocols, Applications, and Overviews.* Wiley-Liss, Inc. USA. 364 p.
- Reinisch, M. J., J. Dong, C. L. Brubaker, D. M. Stelly, J. F. Wendel, and A. H. Paterson. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: Chromosome organization and evolution in a disomic polyploid genome. *Genetics.* 138: 829-847.
- Roder, M. S., V. Korzun, K. Wendehake, J. Plaschke, M. H. Tixier, P. Leroy, and M. A. Ganal. 1998. A microsatellite map of wheat. *Genetics.* 149: 2007-2023.
- Romagosa, I., S. E. Ullrich, F. Hann, and M. H. Hayes. 1996. Use of the additive main effects and multiplicative interaction model in QTL mapping for adaptation in barley. *Theor. Appl. Genet.* 93: 30-37.
- Saha, S. and D. M. Stelly. 1994. Chromosomal location of Phosphoglucosyltransferase7 locus in *Gossypium hirsutum*. *J. Hered.* 85:35-39.
- Sankar, A. A., and G. A. Moore. 2001. Evaluation of inter-simple sequence repeat analysis for mapping in Citrus and extension of the genetic linkage map. *Theor. Appl. Genet.* 102: 206-214.
- Saranga, Y., M. Menz, C-X. Jiang, R. L. Wright, D. Yakir, and A. H. Paterson. 2001. [www.genome.org](http://www.genome.org).
- Sax, K. 1923. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics.* 8: 552-560.

- Schlotterer, C. 1998. Microsatellite. In: Hoelzel, A. R. (editor). 1998. Molecular genetic analysis of populations: A practical approach. Oxford University Press. USA. 445 p.
- Schlotterer, C., and D. Tautz. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* 22: 285-288.
- Schmidt, T., and H. J. Heslop. 1998. Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends in Plt Sci.* 3: 195-198.
- Schork, N. J., M. Boehnke, J. D. Terwilliger, and J. Ott. 1993. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.* 53:1127-1136.
- Shappley, Z.W. 1994. RFLPs in cotton (*Gossypium hirsutum* L.): Feasibility of use, diversity among plants within a line, and establishment of molecular markers and linkage groups among two F2 populations. M.S. thesis. Mississippi State Univ., Mississippi State.
- Shappley, Z. W., J. N. Jenkins, C. E. Watson Jr., A. L. Kahler, and W. R. Meredith, Jr. 1996. Establishment of molecular markers and linkage groups in two F2 populations of Upland cotton. *Theor. Appl. Genet.* 92: 915-919.
- Shappley, Z. W., J. N. Jenkins, J. Zhu, and J. C. McCarty, Jr. 1998a. Quantitative trait loci associated with agronomic and fiber traits of Upland cotton. *The Journal of Cotton Sci.* 4: 153-163.
- Shappley, Z. W., J. N. Jenkins, W. R. Meredith, and J. C. McCarty, Jr. 1998b. An RFLP linkage map of Upland cotton, *Gossypium hirsutum* L. *Theor. Appl. Genet.* 97: 756-761.
- Soller, M., T. Brody, and A. Genizi. 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theor. Appl. Genet.* 47: 35-39.
- Stuber, C. W., M. Polacco, and L. Senior. 1999. Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. *Crop Sci.* 39: 1571-1583.
- Sverdlov, V. E., O. I. Dukhanina, B. Hoebee, and J. P. Rapp. 1998. Linkage mapping of fifty-eight new rat microsatellite markers. *Mammalian Genome.* 9: 816-821.
- Tanksley, S. D. 1983. Molecular markers in plant breeding. *Plant Mol. Biol. Rep.* 1: 3-8.
- Tanksley, S. D. 1993. Mapping polygenes. *Annu. Rev. Genet.* 27: 205-233.

- Thoday, J. M. 1961. Location of polygenes. *Nature*. 191: 368-370.
- Ulloa, M., and W. R. Meredith Jr. 2000. Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. *The Journal of Cotton Science*. 4: 161-170.
- Ulloa, M., R. G. Cantrell, R. G. Percy, E. Zeiger, and Z. Lu. 2000. QTL analysis of stomatal conductance and relationship to lint yield in an interspecific cotton. *The Journal of Cotton Sci.* 4: 10-18.
- Vaz Patto, M. C., D. Rubiales, A. Martin, P. Hernandez, P. Lindhout, R. E. Niks, and P. Stam. 2003. QTL mapping provides evidence for lack of association of the avoidance of leaf rust in *Hordeum chilense* with stomata density. *Theor. Appl. Genet.* 106:1283-1292.
- Veldboom, L. R., and M. Lee. 1996. Genetic mapping of quantitative trait loci in maize in stress and non stress environments. I. Grain yield and yield components. *Crop Sci.* 36: 1310-1319.
- Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Van de Lee, M. Hornes, A. Fritjters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23: 4407-4414.
- Wang, D. L., J. Zhu, Z. K. Li, and A. H. Paterson. 1999. Mapping QTLs with epistatic effects and QTL×environment interactions by mixed linear model approaches. *Theor. Appl. Genet.* 99: 1255-1264.
- Weller, J. I., M. Soller, and T. Brody. 1988. Linkage analysis of quantitative traits in an interspecific cross of tomato (*L. esculentum* x *L. pimpinellifolium*) by means of genetic markers. *Genetics*. 118: 329-339.
- Welsh, J., and M. McClelland. 1990. *Nucleic Acids Res.* 18: 7213-7218.
- Williams, J., A. Kubelik, J. L. Liviak, J. A. Rafalski, and S. V. Tingey. 1990. DNA polymorphisms amplified by random primers are useful as genetic markers. *Nucleic Acids Res.* 18: 6531-6535.
- Wright, R. J., P.M. Thaxton, K.M. El-Zik, and A. H. Paterson. 1999. Molecular mapping of genes affecting pubescence of cotton. *The American Genetic association.* 90: 215-219.
- Xu, K., X. Xu, P. C. Ronald, and D. J. Mackill. 2000. A high-resolution linkage map of the vicinity of the rice submergence tolerance locus Sub1. *Mol. Gen. Genet.* 263: 681-689.
- Yadav, R. S., C. T. Hash, F. R. Bidinger, G. P. Cavan, and C. J. Howarth. 2002. Quantitative trait loci associated with traits determining grain and stover yield

- in pearl millet under terminal drought stress conditions. *Theor. Appl. Genet.* 104: 67-83.
- Yan, J., J. Zhu, C. He, M. Benmoussa, and P. Wu. 1998. Molecular dissection of developmental behavior of plant height in rice (*Oryza sativa* L.). *Genetics*. 150: 1257-1265.
- Yu, Z. H., Y. H. Park, G. R. Lazo and R. J. Kohel. 1998. Molecular mapping of the cotton genome: QTL analysis of fiber quality characteristics. *Proc. of Plant Animal Genome VI*, Jan 18-22. 1998. San Diego California.
- Zabeau, M., and P. Vos, 1993. Selective restriction fragment amplification: a general method for DNA fingerprinting. European patent application number 92402629.7, Publication number 0-534-858 A1.
- Zhang, T., Y. Yuan, J. Yu, W. Guo, and R. J. Kohel. 2003. Molecular tagging of a major QTL for fiber strength in Upland cotton and its marker-assisted selection. *Theor. Appl. Genet.* 106: 262-268.
- Zhang, J., W. Guo, and T. Zhang. 2002. Molecular linkage map of allotetraploid cotton (*Gossypium hirsutum* L. X *Gossypium barbadense* L. with a haploid population. *Theor. Appl. Genet.* 105: 1166-1174.
- Zhu, J. 1998. Mixed model approaches for mapping quantitative trait loci. *Hereditas.* 20: 137-138.
- Zietkiewicz, E., A. Rafalski, and D. Labuda. 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics* 20: 176-183.

**CHAPTER 3**  
**THE DEVELOPMENT OF A GENETIC MAP FOR UPLAND COTTON**  
**(*GOSSYPIUM HIRSUTUM* L.) COMPRISED OF AMPLIFIED FRAGMENT LENGTH**  
**POLYMORPHISMS**

**3.1 Introduction**

*Gossypium hirsutum* (Upland cotton) is one of the four (two tetraploid and two diploid) cultivated cotton species. It accounts for 90% of world cotton production. The tetraploid species (*G. hirsutum* and *G. barbadense*) have 52 chromosomes ( $2n = 4x = 52$ ) with a relatively large genome size of 4700 cM. Cotton is grown in temperate zones as an annual crop, primarily for the production of fiber for the textile industry; although its other industrial and agricultural uses (oil and animal feed) are significant. Despite the importance of Upland cotton, information on its molecular genetics is sparse, especially in comparison with other major agricultural crops.

The molecular genetic information of a crop genome is usually presented in the framework of a genetic linkage map. Such maps are useful to locate or tag genes of interest, to facilitate MAS, and to enable map-based cloning. Thus, the addition of new markers to the Upland cotton linkage map will enhance our understanding of its genetics and also improve breeding efficiency, especially for quantitative traits.

Several types of DNA markers have been successfully used for genetic mapping in many species. Restriction fragment length polymorphisms (RFLP) and Southern blotting of restriction fragments have been a valuable source for the construction of linkage maps (Tanskley et al., 1993). However, the laborious steps involved limit its application. Randomly amplified polymorphic DNAs (RAPDs), based on the polymerase chain reaction, are a valuable tool in the toolkit of the molecular geneticist and have many advantages: non-radioactive detection, no prior DNA

sequence information for a genome is required, universal primers work in any genome, very small amount of genomic DNA is needed, experimental simplicity, and no need for expensive equipment beyond a thermocycler and a transilluminator (Rafalski, 1997). However, they are often criticized for their lack of reproducibility (Jones et al., 1997). Amplified fragment length polymorphisms (AFLPs) are a new, multilocus, high throughput method for detecting genetic polymorphisms. AFLP markers combine the accuracy of RFLP and the simplicity of PCR. The widespread utility of AFLPs as the marker of choice for molecular breeding and genomics research can be accredited to their high reproducibility, their informativeness, their universal presence, and the ease with which they can be automated (Vos et al., 1995).

As in other plant crops, RFLPs were used extensively for the construction of the first genetic linkage maps in Upland cotton. Shappley (1994) and Reinisch et al. (1994) separately reported on the first cotton linkage maps constructed with RFLP markers. Shappley's map (1994) consisted of five linkage groups in a cross of Upland cotton (*G. hirsutum* L.), while 41 linkage groups were constructed by Reinisch et al. (1994) in an interspecific cross between *G. hirsutum* and *G. barbadense*. Since then, several other genetic linkage maps for cotton have been developed (Table 3.1). Altaf et al. (1997) constructed a map with 11 linkage groups using both RAPD and AFLP markers in a trispecies cross of *G. hirsutum*, *G. barbadense*, and *G. trilobum*, but no details were given.

The majority of these genetic maps have been developed through interspecific hybridization, which overcomes the low genetic polymorphism in cotton but has little use in conventional breeding programs (Reinisch et al., 1994). In this

study, we employed the highly informative AFLP markers to construct an intraspecific cross-based genetic map of Upland cotton.

Table 3.1 Reported molecular marker based linkage maps in cotton for (a) intraspecific and (b) interspecific crosses.

Markers	Genome coverage (cM)	Number of linkage groups	References
<u>a) intraspecific</u>			
RFLP	43	05	Shappley (1994) Shappley et al. (1996)
RFLP	865	31	Shappley et al. (1998a, 1998b)
RFLP	700	17	Ulloa and Meredith (2000)
RFLP	1503	47	Ulloa et al. (2002)
<u>b) interspecific</u>			
RFLP	4675	41	Reinisch et al. (1994) Wright et al. (1999) Saranga et al. (2001)
RAPD, AFLP and Morphological	521.5	11	Altaf et al. (1997)
RFLP and isozyme	856	18	Brubaker et al. (1999)
RFLP and -	1486	17	Yu et al. (1998)
RAPD	-	50	Kohel et al. (2001)
RFLP	3664	26	Jiang et al. (2000)
RAPD and SSR	3315	43	Zhang et al. (2002)
RAPD and SSR	1058	28	Ulloa et al. (2000)
RFLP, RAPD, and SSR	1337	08	Zuo et al. (2000)

## 3.2 Materials and Methods

### 3.2.1 Mapping Population

The map was generated from AFLP analysis of an F<sub>2:3</sub> population of 138 individuals from an intraspecific cross between Paymaster 54 (Ramey, 1966) and

Pee Dee 2165 (Culp and Harrell, 1974) of *G. hirsutum*. These parents manifest different phenotypic characteristics for several agronomic and fiber quality traits as presented in Table 3.2. Paymaster 54 was bred by the private sector for high yield performance, while Pee Dee 2165 was specifically bred for high fiber quality and released as a parent for improvement of fiber quality by the USDA-ARS and South Carolina AES (Culp and Harrell, 1979).

Table 3.2 Mean performance for agronomic and fiber quality traits of Paymaster 54 and Pee Dee 2165 at Alexandria and Baton Rouge, LA, in 2002.

Trait	Paymaster 54	Pee Dee2165
<b>a) Agronomical traits</b>		
Lint weight per boll (LY) g	02.54	01.41
Boll number per plant (B/P)	08.15	04.02
Seedcotton weight per boll (BW) g	05.26	05.68
Lint Percentage (LP) %	39.75	36.35
<b>b) Fiber-quality traits</b>		
Elongation (E)	06.35	05.63
Length (L) mm	27.86	29.68
Uniformity (U)	83.85	85.17
Strength (S) g/tex	24.70	29.07
Micronaire (M)	04.91	04.69

F<sub>2</sub> seeds available from a previous study (Lu, 1999; Lu and Myers, 2002) of the two parents were planted in the field at the LSU AgCenter Dean Lee Research Station in Alexandria LA, on May 10, 2001 and were allowed to self-pollinate to generate F<sub>2:3</sub> progeny. Bulk samples of young leaves were collected from each F<sub>2</sub> plant for DNA extraction.

A total of 138 F<sub>2:3</sub> progeny and the two parents were planted the following year at the LSU AgCenter Dean Lee Research Station in Alexandria, LA and Central

Research Station in Baton Rouge, LA. These  $F_{2:3}$  seed were planted in single-row plots, 4.4 m long, spaced 1 m apart with seed sown 22 cm apart by hand. At each location, two replications of the entries, arranged in an incomplete block design, were used to determine agronomic and fiber quality traits. NPK fertilizer, weed control, irrigation and insect control followed standard practices and were applied as needed according to Louisiana Cooperative Extension Service recommendations.

### 3.2.2 DNA Preparation

Cotton DNA of each parent,  $F_1$ , and  $F_2$  was isolated from fresh young leaves harvested as a bulk sample from 4 to 5 plants (parents,  $F_1$ ) or from individual plants ( $F_2$ ). DNA was isolated according to the following protocol: Fresh leaf material (about 0.5 g) was homogenized in 1.0 mL extraction buffer (2% Hexadecyltrimethyl Ammonium Bromide, 100 mM Tris-HCl pH 8, 25 mM EDTA pH 8, 1 M NaCl, 1 mM 1,10-phenanthroline, 2% Polyvinyl Pyrrolidone, and 0.2%  $\beta$ -mercaptoethanol). The homogenization was carried out using an Omni General Lab Homogenizer (Omni International Inc. Warrenton, VA) that fits snugly into a 15 mL tube. Homogenization was followed by placing samples in a hot water bath (60 °C) for 45 min. Plant debris was then separated using centrifuging (15 min. at 960 g) and a chloroform gradient (900  $\mu$ L of 24:1 [Chloroform: iso-Amyl Alcohol]) was used to separate proteins and the supernatant was then transferred to new 1.5 mL snap-cap tube. Ice-cold iso-propanol (700  $\mu$ L) was added to each 800  $\mu$ L isolated upper part supernatant. Depending on the amount of the precipitated DNA, the solution was left at 4 °C for 30 to 120 min incubation. The remaining supernatant, after centrifugation at 5200 g for 5 minutes, was discarded and 500  $\mu$ L of 70% EtOH was added to pellet the DNA

before another centrifugation at 5200 g for 1 minute. The supernatant was then discarded. Fifty  $\mu\text{L}$  of TE buffer (10 mM Tris-HCl pH 8.0; 1.0 mM EDTA) and 2  $\mu\text{L}$  of Rnase-A (10 mg/mol) were added before storage at  $-20\text{ }^{\circ}\text{C}$ . The DNA samples were diluted to a concentration of 20 ng/ $\mu\text{L}$  with TE<sub>0.1</sub> (10 mM Tris-HCl pH 8.0; 0.1 mM EDTA) to be used as a working solution in AFLP marker analysis.

### 3.2.3 DNA Quantification

Three methods were used to quantify and identify the quality of DNA samples. An agarose gel method was used to provide information regarding both DNA quantity and quality. The concentration of genomic DNA was estimated by comparing the size and intensity of each sample band with those of sizing standard, DNA mass ladder (GIBCO<sup>®</sup>). Spectrophotometry was used for quantification and quality checking depending on A260/A280. The standard Hoechst-stain-fluorometer method was also used for DNA quantification.

### 3.2.4 Amplified Fragment Length Polymorphism Analysis

Twenty primer combinations were used to generate AFLP data (Table 3.3). The generation of the data was performed according to 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 enzyme restriction sites were ligated to the resulting fragments through incubation (150 min, 37  $^{\circ}\text{C}$ ) with DNA ligase. This step was carried out on GeneAmp PCR System 9600 (Perkin Elmer). The genomic DNA (20-40 ng) was digested with the restriction endonucleases in a 11  $\mu\text{L}$  reaction containing 3  $\mu\text{L}$  DNA, 3.5  $\mu\text{L}$  enzyme mix, and 4.5  $\mu\text{L}$  adapter mix

(Table 3.4). The reaction was incubated at 37 °C for 150 minutes, and then diluted with 89  $\mu$ L TE<sub>0.1</sub>.

Table 3.3 Adapters and primers used for pre-amplification and selective amplification of AFLP procedure.

Name of Primer/adaptor	Sequence (5'-3')
<i>EcoRI</i> adapter	CTCGTAGACTGCGTACC CATCTGACGCATGGTTAA
<i>MseI</i> adapter	GACGATGAGTCCTGAG TACTCAGGACTCAT
<i>EcoRI</i> primer	
E-A	GACTGCGTACCAATTCA
E- AAG	GACTGCGTACCAATTCAAG
E- AAC	GACTGCGTACCAATTCAAC
E-ACA	GACTGCGTACCAATTCACA
E-ACC	GACTGCGTACCAATTCACC
E-AGG	GACTGCGTACCAATTCAGG
E-ACG	GACTGCGTACCAATTCACG
E-ACT	GACTGCGTACCAATTCACT
E-AGC	GACTGCGTACCAATTCAGC
<i>MseI</i> primer	
M-C	GATGAGTCCTGAGTAAC
M-CAA	GATGAGTCCTGAGTAACAA
M-CTT	GATGAGTCCTGAGTAACTT
M-CAC	GATGAGTCCTGAGTAACAC
M-CAT	GATGAGTCCTGAGTAACAT
M-CTA	GATGAGTCCTGAGTAACTA
M-CTC	GATGAGTCCTGAGTAACTC
M-CTG	GATGAGTCCTGAGTAACTG
M-CAG	GATGAGTCCTGAGTAACAG

Double-stranded adapters were prepared by mixing individual synthetic oligonucleotides. *EcoRI*-adapter was prepared by mixing 7.0  $\mu$ L of the top strand oligonucleotide (2 $\mu$ g/ $\mu$ L) with 7.5  $\mu$ L of the bottom strand oligonucleotide (2 $\mu$ g/ $\mu$ L) in 486.1  $\mu$ L of TE. This gave 5 pmole/ $\mu$ L of *EcoRI*-adapter. *MseI* adapter was prepared by mixing 63.5  $\mu$ L of the top strand oligonucleotide (2 $\mu$ g/ $\mu$ L) with 54.5  $\mu$ L of the

bottom strand oligonucleotide (2µg/µL) in 382 µL of TE. This gave 50 pmole/µL of *MseI*-adapter.

Table 3.4 Protocol components for digestion and ligation of genomic DNA

a) Enzyme mix	µL	b) Adapter mix	µL
10X T4 Ligase buffer	0.350	10X T4 Ligase buffer	0.75
0.5 M NaCl	0.350	0.5 M NaCl	0.75
BSA (1mg/mL)	0.005	BSA (1mg/mL)	0.05
<i>MseI</i> enzyme (10U/µL)	0.050	<i>MseI</i> Adapter (50pmole/µL)	1.00
<i>EcoRI</i> enzyme (20U/µL)	0.250	<i>EcoRI</i> adapter (5pmole/µL)	1.00
T4 DNA Ligase (400 U/µL)	0.002		
H <sub>2</sub> O	2.492	H <sub>2</sub> O	0.95
	5		
<b>Total Volume</b>	<b>3.50</b>	<b>Total Volume</b>	<b>4.50</b>

DNA preparations needed to be of sufficient quality to allow complete digestion, since this step is crucial for the production of a good quality AFLP fingerprints (Figure 3.1). The pre-amplification step was performed using primers designed to amplify the DNA fragment between the adapter sequence and one additional nucleotide.

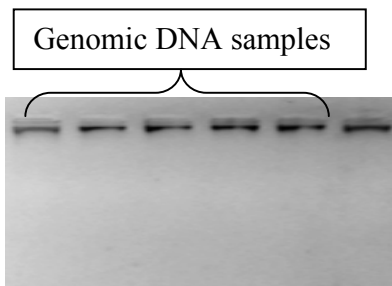


Figure 3.1 High quality undigested Upland cotton genomic DNA.

The pre-amplification reaction (20 µL total volume) consisted of 4 µL diluted (1:10) digestion ligation mixture, 1.0 µL of the *EcoRI* - primer+A (50uM) with 1.0 µL *MseI*-primer+C (50uM), 0.4 µL 10 mM dNTPs, 1.2 MgCl<sub>2</sub> (50uM), 0.2 µL *Taq*

polymerase (1 unit), 2.1  $\mu\text{L}$  10x PCR-buffer, and 10.1  $\mu\text{L}$  water (Table 3.5a). The mixture was pre-amplified for 20 cycles (30 seconds denaturation at 94  $^{\circ}\text{C}$ ; 60 seconds annealing at 56  $^{\circ}\text{C}$ ; 60 seconds extension at 72  $^{\circ}\text{C}$ ). After pre-amplification, 10  $\mu\text{L}$  of the reaction was used to run an agarose gel to check the quality of the digestion (Figure 3.2) and the rest (10  $\mu\text{L}$ ) was diluted with 190  $\mu\text{L}$  of low  $\text{TE}_{0.1}$  to 200  $\mu\text{L}$ , which was sufficient for 40 AFLP-reactions. The diluted reaction mix and the rest of the amplification reaction products were stored at  $-20^{\circ}\text{C}$ .

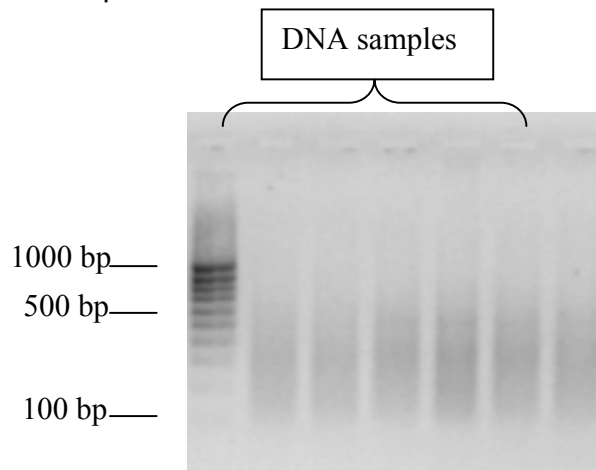


Figure 3.2 Preamplification products (10  $\mu\text{L}/\text{lane}$ ) create a visible smear in the 100 to 1000 bp range.

Duplex selective amplification was performed using the AFLP protocol developed by LiCor (AFLP Selective Amplification Kit, 2001), and the new *Mse1* and IRDye labeled *EcoR1* primers comprising three-nucleotide extensions. The reaction components (10.5  $\mu\text{L}$  total volume) included 1.2  $\mu\text{L}$  10X amplification buffer containing  $\text{MgCl}_2$ , 0.06  $\mu\text{L}$  Tag DNA polymerase [5 units/ $\mu\text{L}$ , Promega Inc.], 1.5  $\mu\text{L}$  diluted pre-amplification DNA, 2  $\mu\text{L}$  *Mse1* primer containing dNTPs, 0.25  $\mu\text{L}$  IRDye 700 labeled *EcoR1* primer-A, and 0.25  $\mu\text{L}$  IRDye 800 labeled *EcoR1* primer-B in 5.24  $\mu\text{L}$  deionized water (Table 3.5b).

The polymerase chain reaction was performed using a touchdown program: 13 cycles of subsequently lowering the annealing temperature 65 °C by 0.7 °C per cycle while keeping denaturation at 94 °C for 30 seconds and extension at 72 °C for 60 seconds. This was followed by 23 cycles of denaturation at 94 °C for 30 seconds, annealing at 56 °C for 30 seconds and extension at 72 °C for 60 seconds. After PCR, 4 µL of Blue Stop Solution was added immediately before storage at –20 °C.

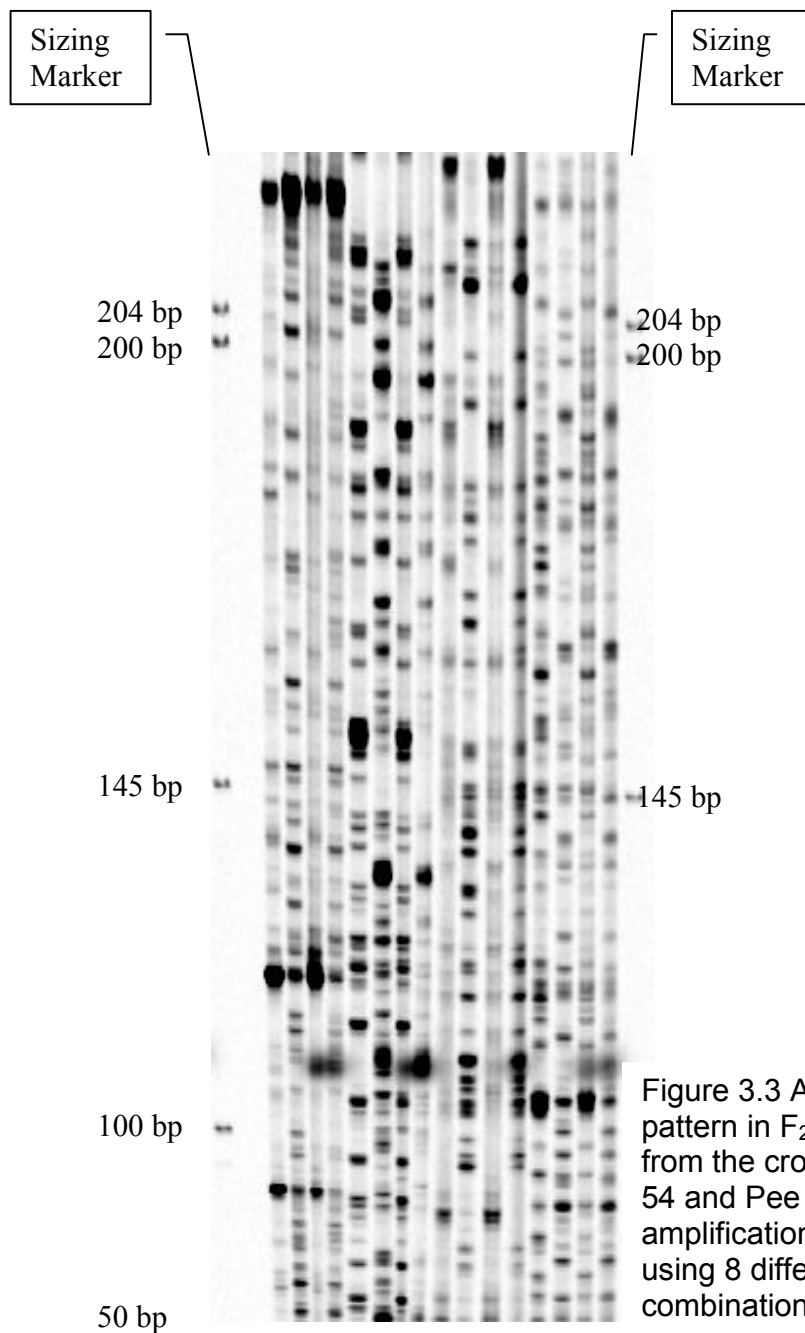
Table 3.5 Reagents used in the Pre-amplification step (a) and selective amplification step (b):

a) Pre-amplification step	µL	b) Selective amplification step	µL
10X PCR Buffer	2.1	10X PCR Buffer	1.20
MgCl <sub>2</sub> (50µM)	1.2	dNTPs (10µM)	
dNTPs (10µM)	0.4	Mse-Primer (containing dNTP)	2.00
Eco-Primer (50µM)	1.0	IRDye700 labeled <i>EcoRI</i> primer	0.25
Mse-Primer (50µM)	1.0	IRDye700 labeled <i>EcoRI</i> primer	0.25
Tag (5U/µl)	0.2	Tag (5U/ul)	0.06
H <sub>2</sub> O	10.1	H <sub>2</sub> O	5.24
Diluted DNA (after digestion and Ligation)	4.0	Diluted DNA (Pre-Amplified)	1.50
Total Volume	20.0	Total Volume	10.5

### 3.2.5 Gel Analysis

Amplified DNA fragments were separated on a 6% denaturing polyacrylamide gel (LiCor) that included 52.5 g urea, 7.12 g acrylamide, 0.375 g bis-Acrylamide, and 1.825 g 20x glycerol. The gels were cast at least 90 minutes before use and pre-run for 30 min just before loading the samples. Pre-running and running electrophoretic steps were performed using 16-bit data collection, 1500 V, 40 W, 40 mA, 45 °C, and 4 scan speed as recommended by LiCor. Tris-borate-EDTA (TBE) (1X) was used as running buffer. After the wells were completely flushed with a 20 cc syringe to remove urea precipitate or pieces of gel, 0.4 to 0.6 µL of each denatured sample (denaturation conducted at 94 °C for 3 minutes immediately before loading) was

added to a well using an 8-channel Hamilton syringe. Four molecular sizing standards (50-700 bp) were used in designated lanes. The real-time TIF images were automatically collected and recorded during electrophoresis (Figure 3.3). Loading the same gel three times, each run needed about 3 h to collect both channel images (700 and 800) resulting in a maximum of 6 images collected in a single day.



For automated data output, the images were transferred to Gene ImagIR (Bio-Rad, Hercules, CA) that scored, analyzed, and converted bands into numerical data files. The parameters specified for scoring were:

---

```

Median filter = YES
Band peak height threshold (% max intensity) = 2 (0.015-0.02)
Trace smoothing factor = 3
Tolerance = 0.2
Correct inflated bins automatically = YES
Extended PAUP absence/presence table = YES

```

---

Based on the number of polymorphic bands produced, present in one parent but not the other, and on their sharpness, 200 primers were selected for genetic linkage map construction. SAS (SAS Institute, Cary, NC) was used to check for segregation distortion in the 200 AFLP markers using Chi-square ( $\alpha = 0.01$ ) to test goodness-of-fit to the 3:1 ratio with one df (Table 3.6).

Table 3.6 Proc Freq. code to test for marker segregation distortion (SAS Version 9)

a) data step	b) analysis code step
data MolecularMarkers;	proc sort data=MolecularMarkers;
input marker \$ presence	by marker;
count;	
cards;	run;
C20_521      1      67	proc freq data=MolecularMarkers
.	order=data;
.	weight count;
.	tables presence/nocum testp=(0.75 0.25);
C20_521      0      70	by marker;
;	title 'Mendelian Segregation Test';
	run;
	quit;

### 3.2.6 Marker Naming

Each marker was assigned a two-part name consisting of the primer combination number followed by the band size (in base pairs) estimates from the mobility in the gel compared with the size standard.

### 3.2.7 Map Construction

Linkage analysis was performed using MAPMAKER/EXP version 3.0 (Lander et al., 1987; Lincoln et al., 1992), which estimates a maximum likelihood distance and a LOD score between two loci for a given number of F<sub>2</sub> individuals based on the presence or absence of AFLP fragments in relation to the parental lines (Morton, 1955; Mather, 1957). Linkage groups were established using the 'group' command with a LOD threshold of 4 and a maximum recombination frequency of 0.34. The 'three point' command determined the order of loci within each linkage group. However, for linkage groups with 3 to 8 markers, the 'compare' command was used as a two-point method of ordering. The orders were then verified by the 'ripple' command. The 'try' command was used to find possible linkages with unassigned loci and small linkage groups. Loci with segregation distortion were included in the grouping step but excluded from the framework order step. Distances between linked markers are presented in centimorgan (cM) map units, which were derived using the Haldane function (Haldane, 1919).

## **3.3 Result and Discussion**

Thirty two primer combinations were screened by selective amplification using parental template DNA. Of these, 20 gave reliable and reproducible polymorphism (Table 3.7). The other 12 combinations either gave less than three polymorphic bands or did not give scoreable bands at all. A total of 200 polymorphic and 1576

monomorphic bands was generated using the 20 primer pairings with a mean of about 10 polymorphisms and 79 monomorphisms per primer combination. The level of polymorphism ranged from 5.5% for the *EcoRI*- AAC/*MseI*-CAT primer combination to 16.5% for the *EcoRI*-ACG/*MseI*-CAC with an overall mean of 11.2% polymorphism. The polymorphism level detected in this study was similar to that observed in barley (11%) (Becker et al., 1995) but lower than that observed in rice (Mackill et al., 1996) and sugar beet (Schondelmaier et al., 1996), where 28% and 50% polymorphism, respectively, was reported.

Table 3.7 Number of monomorphic and polymorphic (total) and number of AFLP primer combinations between two lines (Pee Dee 2165 and Paymaster 54) of Upland cotton.

Name	Selective nucleotides		Number of bands	
	<i>EcoRI</i>	<i>MseI</i>	Total	Polymorphic
C01	AAG	CAA	115	10
C02		CTT	107	11
C03	AAC	CAT	103	06
C04		CTA	105	20
C05	ACA	CTC	092	11
C06		CTG	075	08
C07	ACC	CAC	071	06
C08		CAG	073	12
C09	AAG	CAC	102	06
C10	AAC	CAC	101	10
C11	AGG	CAA	111	08
C12		CTT	091	11
C13	ACG	CAT	065	04
C14		CTA	084	16
C15	ACT	CTC	089	08
C16		CTG	080	11
C17	AGC	CAC	078	09
C18		CAG	075	11
C19	ACT	CAC	083	07
C20	ACG	CAC	076	15
Grand total			1776	200

Upland cotton contains 26 chromosomes; however, out of a total of 200 markers analyzed, 143 markers were assigned to 13 major (Figure 3.4a) and 15 minor linkage groups (Figure 3.4b). This may be a consequence of the population size used in this study (138  $F_{2:3}$  lines). Kesseli et al. (1994) and Keim et al. (1997) suggested that increasing population size, and not the number of markers, would most likely reduce the number of linkage groups by helping identify key recombinants and fill remaining gaps. A linkage group is considered a major group if it has a total length of 50 cM or longer. The 13 major groups ranged from 50.3 to 205.1 cM in length and each group carried 3 to 19 markers (Table 3.8). The 15 minor groups ranged from 7.5 to 49.3 cM in length and each group carried 2 to 6 markers. Forty one markers were not linked to any group. The 28 linkage groups cover a genetic distance of 1773.2 cM. However, the total coverage for these 200 markers is 3066.2 assuming each unlinked locus and each pair of the 28 linkage group ends accounts for 20 cM on average (Weng, 2000). This gives a coverage of 65.2% of the cotton genome (4700 cM).

Segregation distortion, the deviation of segregation ratio from the expected Mendelian ratio, has been reported in a wide range of plant species (Jenczewski et al., 1997). A total of 88 (44%) of 200 markers showed segregation patterns skewed from the 3:1 ratio at  $P = 0.01$ . Sixty-seven of the 88 skewed markers were assigned to 17 linkage groups (Table 3.8). The skewed markers tended to be mapped towards the ends. This was true in linkage groups 1, 2, 4, 5, 13, 17, 21, and 24. In linkage group 15, four skewed markers mapped close to the group center. Segregation distortion may occur due to the presence of lethal genes and/or fragment complexes (overlapping fragments consisting of identically sized fragments) (Nikaido et al., 1999;

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). Distortion and the high proportion of RFLP markers in an intraspecific cotton population presumably resulted from polyploidy of cotton (Ulloa and Meredith, 2000).

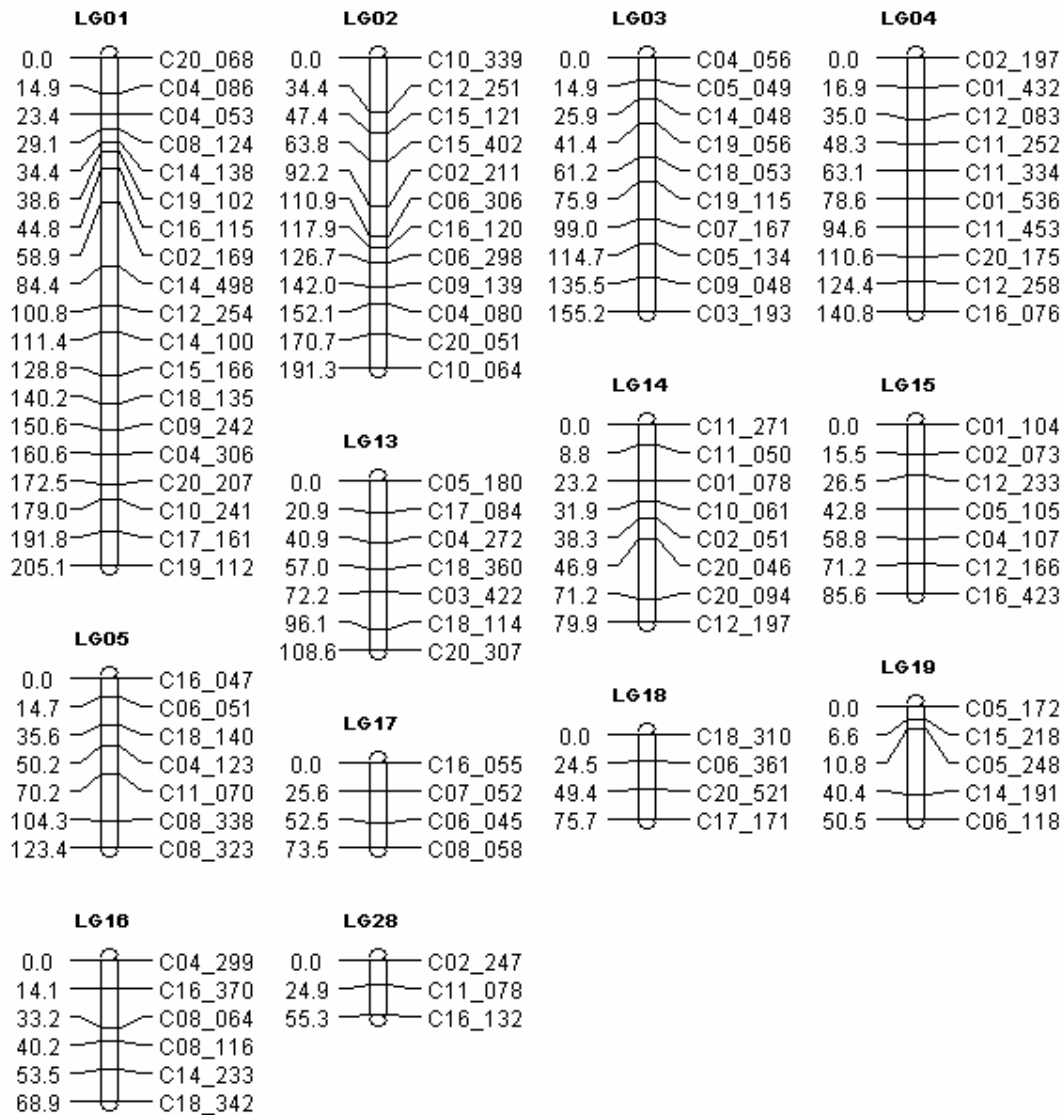


Figure 3.4a Major Genetic linkage groups of Upland cotton (*Gossypium hirsutum* L.) comprised of 102 amplified fragment length polymorphism markers.

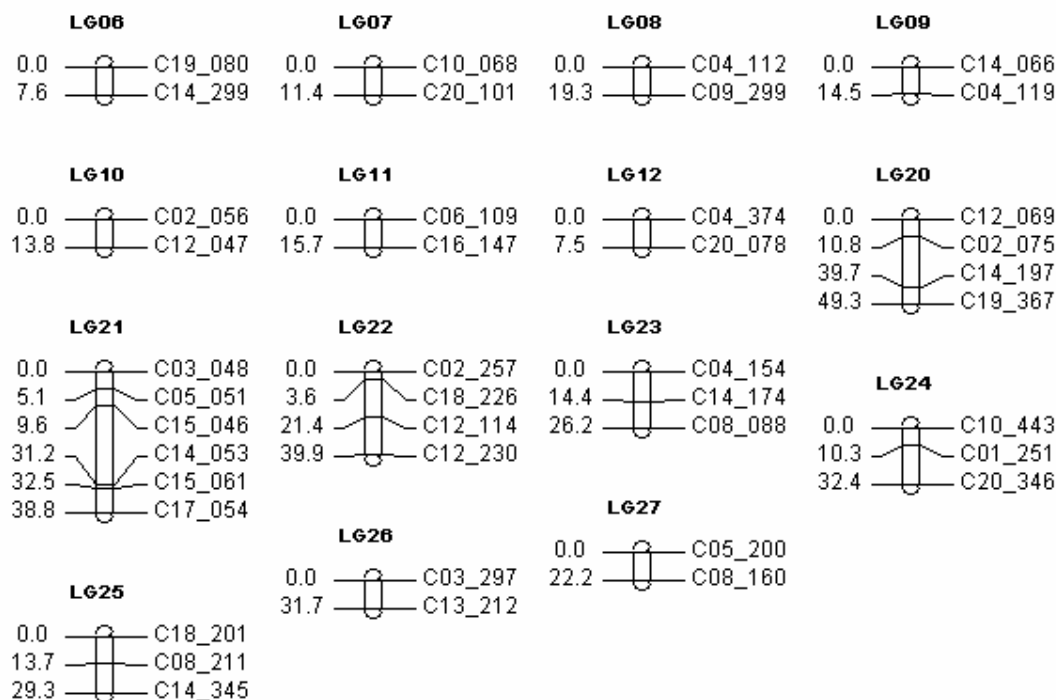


Figure 3.4b Minor Genetic linkage groups of Upland cotton (*Gossypium hirsutum* L.) comprised of 41 amplified fragment length polymorphism markers.

Except for linkage groups 1, 2 and 14, loci were well distributed within linkage groups (lack of clustering). This has also been reported, for example, in barley (Becker et al., 1995) and rice (Maheswaran et al., 1997). On the other hand, the fact that AFLP markers characteristically cluster in heterochromatin-rich centromeric regions has been reported in several plant species, for example *Arabidopsis* (Alonso Blanco et al., 1998), sugar beet (Schondelmaier et al., 1996), soybean (Keim et al., 1997), barley (Qi et al., 1998), and rice (Cho et al., 1998).

Adding more AFLP markers by screening different primer combinations of *EcoRI/MseI* and by assaying more enzyme combinations other than *EcoRI/MseI* will help saturate the map. With the addition of more markers, the smaller linkage groups may converge or join with other linkage groups. Such a saturated map could be directly used for marker assisted plant breeding and gene and QTL tagging.

Table 3.8 Marker distributions among the linkage groups of Upland cotton.

Linkage group	Marker number	Length (cm)	Average distance (cm)	Number of skewed markers	Number of markers not placed <sup>a</sup>
<b>a) Major groups</b>					
1	25	205.1	11.4	9	6
2	17	191.3	17.4	8	5
3	10	155.2	17.2	0	0
4	10	140.8	15.6	7	0
5	10	123.4	20.6	5	3
13	09	108.6	18.1	5	2
14	08	79.9	11.4	3	0
15	07	85.6	14.3	4	0
16	06	68.9	13.8	0	0
17	04	73.5	24.5	3	0
18	04	75.7	25.2	4	0
19	05	50.3	12.6	5	0
28	03	55.3	27.6	2	0
<b>b) Minor groups</b>					
6	02	07.6	07.6	0	0
7	02	11.4	11.4	0	0
8	02	19.3	19.3	1	0
9	02	14.5	14.5	0	0
10	02	13.8	13.8	0	0
11	02	15.7	15.7	0	0
12	02	07.5	07.5	2	0
20	04	49.3	16.5	0	0
21	06	38.8	07.8	4	0
22	04	40.0	13.3	0	0
23	03	26.2	13.0	0	0
24	03	32.4	16.2	2	0
25	03	29.3	14.7	3	0
26	02	31.7	31.7	1	0
27	02	22.2	22.2	0	0
Unlinked	41	NA	NA	21	NA
<b>Total</b>	<b>200</b>	<b>3066.2<sup>b</sup></b>	<b>15.5</b>	<b>88</b>	<b>16</b>

<sup>a</sup> Number of markers that cannot be confidently placed on the map were included in the total number of markers in each linkage group. Averages between adjacent markers were obtained from ordered markers.

<sup>b</sup> Assuming each unlinked locus and each pair of the 28 ends account for 20 cM on average.

### 3.4 References

- Alonso-Blanco, C., A. J. M. Peeters, M. Koornneef, C. Lister, C. Dean, N. Van den Bosch, J. Pot, and T. R. Kuiper. 1998. Development of an AFLP-based linkage map of L, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a L/Cvi recombinant inbred line population. *Plant Jour.* 14: 259-271.
- Altaf, M. K., J. McD. Stewart, M. K. Wajahatullah, and J. Zhang. 1997. Molecular and morphological genetics of a trispecies F<sub>2</sub> population of cotton. *Proc. Beltwide Cotton Conf.* p 448-452.
- Becker, J., P. Vos, M. Kuiper, F. Salamini, and M. Heun. 1995. Combined mapping of AFLP and RFLP markers in barley. *Mol. Gen. Genet.* 249: 65-73.
- Bert, P. F., G. Charmet, P. Sourdille, M. D. Hayward, and F. Balfourier. 1999. A high-density molecular map for ryegrass (*Lolium perenne*) using AFLP markers. *Theor. Appl. Genet.* 99: 445-452.
- Brubaker, C. L., A. H. Paterson, and J. F. Wendel. 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome.* 42: 184-203.
- Cho, Y. G., S. R. McCouch, M. Kuiper, M-R. Kang, J. Pot, J. T. M. Groenen, and M. Y. Eun 1998. Integrated map of AFLP, SSLP and RFLP markers using a recombinant inbred population of rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 97: 370-380.
- Culp, T. W., and D. C. Harrell. 1974. Breeding quality cotton at the Pee Dee Experiment Station, Florence, S. C. USDA Res. Rep. ARS-S-30.
- Culp, T. W., and D. C. Harrell. 1979. Registration of Pee Dee 0259 and Pee Dee 2165 germplasm lines of cotton. *Crop Sci.* 19: 418.
- Haldane, J. B. S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. of Genet.* 8: 299-309.
- Hansen, M., T. Kraft, M. Christiansson, and N-O. Nilsson. 1999. Evaluation of AFLP in *Beta*. *Theor. Appl. Genet.* 98: 845-852.
- Jenczewski, E., M. Gherardi, L. Bonnin, J. M. Prospero, I. Olivieri, and T. Huguet. 1997. Insight on segregation distortions in two intraspecific crosses between annual species of *Medicago* (*Leguminosae*). *Theor. Appl. Genet.* 94: 682-691.
- Jiang, C-X., R. J. Wright, S. S. Woo, T. A. Del Monte, and A. Paterson. 2000. QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). *Theor. Appl. Genet.* 100: 409-418.

- Jones, C. J., K. J. Edwaeds, S. Castaglione, M. O. Winfield, F. Sala, C. Van De Weil, G. Bredemeijer, B. Vosman, M. Matthes, A. Daly, R. Brettschneider, P. Bettini, M. Buitti, E. Maestri, A. Malcevschi, N. Marmiroli, R. Aert, G. Volckaert, J. Rueda, R. Linacero, A. Vazquez, and A. Karp. 1997. Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Mol. Breed.* 3: 381-390.
- Keim, P., J. M. Schupp, S. E. Travis, K. Clayton, T. Zhu, S. Liang, A. Ferreira, and D.M. Webb. 1997. A high-density soybean genetic map based on AFLP markers. *Crop Sci.* 37: 537-543.
- Kesseli, R.V., I. Paran, and R.W. Michelmore. 1994. Analysis of a detailed linkage map of *Lactuca sativa* (lettuce) constructed from RFLP and RAPD markers. *Genetics.* 136: 1435-1446.
- Kohel, R. J., J. Yu. Y-H. Park, and G. R. Lazo. 2001. Molecular mapping and characterization of trait controlling fiber quality in cotton. *Euphytica.* 121: 163-172.
- Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural population. *Genomics.* 1: 174-181.
- Li-Cor, 2001. IRDye fluorescent AFLP kit for large plant genome analysis. 15 p.
- Lincoln, S., M. Daly, and E. Lander. 1992. Mapping genes controlling quantitative traits with MAPMARKER/QTL. Whitehead Institute Technical Report. 2<sup>nd</sup> edition. Whitehead Institute, MA.
- Lu, H. 1999. Diallel Analysis and Molecular Genetics of Ten Influential Upland Cotton Cultivars. Ph. D. dissertation. Louisiana State University and Agricultural and Mechanical College.
- Lu, H. J., and G. O. Myers. 2002. Genetic Relationships and Discrimination of Ten Influential Upland Cotton Varieties using RAPD Markers. *Theor. Appl. Genet.* 105: 325–331.
- Mackill, D. J., Z. Zhang, E. D. Rodona, and P. M. Colowit. 1996. Level of polymorphism and genetic mapping of AFLP markers in rice. *Genome.* 39: 969-977.
- Maheswaran, M., P. K. Subudhi, S. Nandi, J.C. Xu, Parco, D.C. Yang, and N. Huang. 1997. Polymorphism, distribution, and segregation of AFLP markers in a double haploid rice population. *Theor. Appl. Genet.* 94: 39-45.

- Mather, K. 1957. The Measurement of Linkage in Heredity. John Wiley and Sons, New York.
- Morton, N. E. 1955. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* 7: 277-318.
- Nikaido, A., H. Yoshimaru, Y. Tsumura, Y. Suyama, M. Murai, and K. Nagasaka. 1999. Segregation distortion for AFLP markers in *Cryptomeria japonica*. *Genes Genet. Syst.* 74: 55-59.
- Qi, X., P. Stam, and P. Lindhout (1998) Use of locus-specific AFLP markers to construct a high-density molecular map in barley. *Theor. Appl. Genet.* 96: 376-384.
- Rafalski, J. A. 1997. Randomly amplified polymorphic DNA (RAPD) analysis. In Caetano-Anolles, G., and Gresshoff, P. M. (eds.). *DNA markers: Protocols, Applications, and Overviews*. Wiley-Liss, Inc. USA. 364 p.
- Ramey, H. H. 1966. Historical review of cotton variety development. p. 310-326. *Proc. 18<sup>th</sup> Cotton Improvement Conf.*, Memphis, TN.
- Reinisch, M. J., J. Dong, C. L. Brubaker, D. M. Stelly, J. F. Wendel, and A. H. Paterson. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: Chromosome organization and evolution in a disomic polyploid genome. *Genetics.* 138: 829–847.
- Saranga, Y., M. Menz, C-X. Jiang, R. L. Wright, D. Yakir, and A. H. Paterson. 2001. [www.genome.org](http://www.genome.org).
- SAS Institute. 2003. Version 9. Cary, N. C., USA.
- Schondelmaier, J., G. Steinrucken, and C. Jung. 1996. Integration of AFLP markers into a linkage map of sugar beet (*Beta vulgaris* L.). *Plant Breed.* 115: 231–237.
- Shappley, Z. W., J. N. Jenkins, C. E. Watson Jr., A. L. Kahler, and W. R. Meredith, Jr. 1996. Establishment of molecular markers and linkage groups in two F2 populations of Upland cotton. *Theor. Appl. Genet.* 92: 915-919.
- Shappley, Z. W., J. N. Jenkins, J. Zhu, and J. C. McCarty, Jr. 1998a. Quantitative trait loci associated with agronomic and fiber traits of Upland cotton. *The Journal of Cotton Sci.* 4: 153-163.
- Shappley, Z. W., J. N. Jenkins, W. R. Meredith, and J. C. McCarty, Jr. 1998b. An RFLP linkage map of Upland cotton, *Gossypium hirsutum* L. *Theor. Appl. Genet.* 97: 756-761.
- Shappley, Z.W. 1994. RFLPs in cotton (*Gossypium hirsutum* L.): Feasibility of use,

- diversity among plants within a line, and establishment of molecular markers and linkage groups among two F<sub>2</sub> populations. M.S. thesis. Mississippi State Univ., Mississippi State.
- Tanksley, S. D. 1993. Mapping polygenes. *Annu. Rev. Genet.* 27: 205-233.
- Ulloa, M., R. G. Cantrell, R. G. Percy, E. Zeiger, and Z. Lu. 2000. QTL analysis of stomatal conductance and relationship to lint yield in an interspecific cotton. *The Journal of Cotton Science.* 4: 10-18.
- Ulloa, M. and W. R. Meredith Jr. 2000. Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. *The Journal of Cotton Science.* 4: 161-170.
- Ulloa, M., W. R. Meredith Jr, and Z. W. Shappley. 2002. RFLP genetic linkage maps from four F<sub>2:3</sub> populations and a joinmap of *Gossypium hirsutum* L. *Theor. Appl. Genet.* 104: 200-208.
- Vos, P., R. Hogers, M. Bleeker, M.Reijans, T. Van de Lee, M. Hornes, A. Fritjters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau.1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23: 4407–4414
- Weng, C. 2000. Mapping quantitative trait loci controlling the early height growth of longleaf pine and slash pine. Ph. D. dissertation. Louisiana State Univ., Baton Rouge.
- Wright, R. J., P.M. Thaxton, K.M. El-Zik, and A. H. Paterson. 1999. Molecular mapping of genes affecting pubescence of cotton. *The American Genetic Association.* 90:215-219.
- Yu, Z. H., Y. H. Park, G. R. Lazo and R. J. Kohel. 1998. Molecular mapping of the cotton genome: QTL analysis of fiber quality characteristics. *Proc. of Plant Animal Genome VI*, Jan 18-22. 1998. San Diego California.
- Zhang, J., W. Guo, and T. Zhang. 2002. Molecular linkage map of allotetraploid cotton *Gossypium hirsutum* L. X *Gossypium barbadense* L. with a haploid population. *Theor. Appl. Genet.* 105:1166-1174.
- Zuo, K., J Sun, X Zhang, Y. Nie, J. Liu, and C. Feng. 2000. Constructing a linkage map of Upland cotton (*Gossypium hirsutum* L.) using RFLP, RAPD and SSR Markers. *Journal of Huazhong Agricultural University.* 19:190-193.

## CHAPTER 4 MOLECULAR QTL MAPPING FOR AGRONOMIC TRAITS IN UPLAND COTTON (*GOSSYPIUM HIRSUTUM* L.)

### 4.1 Introduction

Cotton (*Gossypium hirsutum* L.) is the world's major natural source of textile fiber and the second largest oilseed crop, with cotton seed and its meal also being used in food and feed products. Therefore, the main objectives of cotton breeding programs are the development of productive cultivars with high yields and fiber quality.

While yield itself is most typically used as a selection criteria this may be hindering future progress in developing highly productive cultivars since it is the sum of the contributions of several largely independently inherited yield components. It has been suggested (Lewis, 2001) that breeding efforts focused on improving individual yield components may be more efficient in raising yield. Lint weight per boll (LY), seedcotton weight per boll (BW), boll number per plant (B/P), and lint percentage (LP) are used to assess our understanding of cotton yield. The identification of chromosomal regions with effect on these agronomic traits would increase our understanding of the genetic control of these traits.

Compared to other crops, the importance of molecular markers and QTL identification in cotton genetic analysis was not demonstrated until Shappley et al. (1994), and Reinisch et al. (1994) provided the first linkage maps of QTL. Recently, several cotton QTL have been identified. For example, QTL for agronomic and fiber traits using RFLP markers have been identified (Shappley et al., 1998), for leaf morphology using RFLP markers (Jiang et al., 2000), for stomatal conductance using RAPD and SSR markers (Ulloa et al., 2000), for agronomic traits using RAPD and

AFLP markers (Khan et al., 1998), for density of leaf and stem trichomes using RFLP markers (Wright et al., 1999), and for cotton productivity, physiological and fiber quality traits using RFLP markers (Saranga et al., 2001).

The simplest approach for detecting a QTL is to use single-marker analysis (SMA). This method investigates the association between trait(s) and one marker at a time. The idea of single-marker analysis was placed into practice when Sax (1923) reported a positive association between seed size and seed coat pigmentation in beans. He concluded that the association was a linkage of a single gene controlling the seed color with gene(s) controlling the seed coat.

In SMA, the mapping population is partitioned into different genotypic classes that reflect genotypes at the marker locus. ANOVA is then used to determine whether the individuals of one genotype differ significantly from the individuals of other genotypes with respect to the trait being measured. If the phenotypes differ significantly, a gene(s) affecting the trait is said to be linked to the marker locus used to subdivide the population (Tanskley, 1993).

Because SMA does not require a linkage map, it is the analysis of choice whenever information about linkage maps is not available. This fact also explains why SMA was widely employed in earlier studies (Soller et al., 1976; Weller et al., 1988). It is the only method for researchers can use for unlinked markers that cannot be included in their linkage maps.

Although SMA captures the basic idea of QTL mapping, Lander and Botstein (1989) stated several drawbacks of SMA: (1) If the trait does not lie near the marker, its phenotypic effect may be seriously underestimated, (2) If the trait does not lie at the marker locus, substantially more progeny may be required, and (3) The approach

does not define the likely position of the trait. In particular, it cannot distinguish between tight linkage to a QTL with small effect or loose linkage to a marker with large effect. (4) The suggested false positive rate of  $\alpha = 0.05$  neglects the fact that many markers are being tested. While the chance of a false positive at any given marker is only 5%, the chance that at least one false positive will occur somewhere in the genome is much higher.

Where information is available for several genetic markers, interval-marker analysis (IMA) interval mapping (IM) and composite interval mapping (CIM) procedures are the most accepted and used methods. IM is based on an Expectation Maximization (EM) algorithm (Dempster et al., 1977) that maximizes the likelihood ratio tests of a single QTL by averaging it across the possible states of the unknown genotype at flanking markers (Lander and Botstein, 1989). The LOD score, which is the log likelihood ratio comparing the hypothesis of the presence of a single QTL at any locus to the null hypothesis of no segregating QTL at that locus, is scanned against linkage groups and is compared to a threshold, usually set to a value of two, to ensure a 0.05 overall false positive error rate. A one or two LOD support interval is used as an interval estimate for QTL location.

There are, however, some problems with IM. The more serious of these include: (1) If there are more than one QTL on a linkage group, interval mapping can be seriously biased and the position being tested will be affected by all other QTL on the same linkage group. (2) It is not efficient to use only two markers at a time to do the test, as the information from other markers is not utilized (Zeng and Weir, 1996).

Similar to IM, CIM (Zeng, 1993; Jiang and Zeng, 1995) evaluates the presence of a putative QTL at flanking markers. However, CIM uses the multiple

regression method. In multiple regression, the partial regression coefficient of a trait on a marker is expected to depend only on those QTL that are located on the interval bracketed by the two neighboring markers and to be independent of any other marker (Zeng and Weir, 1996). The main problem in this method is the number of regressor markers (background markers). Using too many background markers will increase the variance of the LOD score, and thus will decrease the power for detecting QTL. Basten et al. (1997) recommended using forward selection up to a fixed number of markers and then dropping any markers that are within 10 cM of the putative QTL. Many previous studies have used five markers, the default in QTL-CARTOGRAPHER, (Wang et al., 2000; Wang et al., 2001; Marques et al., 1999); others have used 10 (Johnson et al., 2000) and 15 markers (Flores-Berrios et al., 2000).

Multiple QTL methods are an improvement over single QTL methods because of their ability to separate linked QTL on the same linkage group and to detect interacting QTL that may otherwise be undetected. These methods provide an increased power to detect QTL and can eliminate bias in the estimates of effect size and location that can be introduced by using single QTL methods (Schork et al., 1993).

QTL X environment interaction has been discussed in many studies. The result of these studies indicated either significant QTL effects being detected only in a subset of all environments, or changes in the magnitude of the QTL effect (Paterson et al., 1991; Wang et al., 1999; Cao et al., 2001; Yadav et al., 2002). In cotton, 61 of 161 QTL for 16 measured traits showed significant differences in their

effect estimates between well-watered and water-limited conditions. These results indicated a significant QTL X environment interaction.

In the present chapter, we used an AFLP linkage map (28 linkage groups comprised of 143 markers) resulting from the cross of two *Gossypium hirsutum* parents (Paymaster 54 X Pee Dee 2165) to identify QTL linked with four agronomic traits using both interval and composite interval analysis. QTL X environment interaction was also investigated.

## **4.2 Materials and Methods**

### **4.2.1 Plant Material**

The QTL mapping population was initiated by an intraspecific cross between two parents of the species *G. hirsutum* (Paymaster 54 and Pee Dee 2165). Crosses were made between them, and 138 F<sub>2:3</sub> progeny lines were used in this study. Parents available from previous study (Lu and Myers, 2002) and F<sub>2:3</sub> seeds were planted in the field on May 10, 2002 at the LSU AgCenter Dean Lee Research Station in Alexandria, LA and Central Research Station in Baton Rouge, LA. These F<sub>2:3</sub> seed were planted in single-row plots, 5 m long, spaced 1 m apart with seed sown by hand, 15 cm apart. At each station, two replications of the entries, arranged in an incomplete block design, were used to evaluate agronomic traits.

### **4.2.2 Phenotypic Measurement**

A sample of 20 to 25 bolls was handpicked from each F<sub>2:3</sub> row and phenotypic data were collected on the following: Lint weight per boll (LY), seedcotton weight per boll (BW), boll number per plant (B/P), and lint percentage (LP).

#### 4.2.3 Linkage Analysis

AFLP polymorphic bands were scored as present (1) or absent (0) on GenElmagelR. Data coded (0) and (1) were transformed to A, B, C, D genotype codes, according to the presence of the band for the two parents, following the MAPMAKER convention. A molecular linkage map consisting of 143 AFLPs was constructed (see chapter 3) using MAPMAKER 3.0 (Lander et al., 1987). Linkage groups were obtained with a LOD score of 4 and a maximum recombination frequency of 0.34. The Haldane function was used to transform the recombination frequencies to genetic distances (Haldane, 1919).

#### 4.2.4 QTL Analysis

Two different SMA methods (simple and logistic regression) were used to study the degree of association between the four agronomic traits and each marker. Simple regression/ANOVA (SAS Version 9) was performed using marker genotype as a class variable. To reduce the false positive error rate, an association was considered to be significant only when the p value was less than or equal to 0.01. Logistic regression was used as a second single marker method. Using logistic regression, marker genotype was used as a dependent variable.

Two different IMA methods (IM and CIM) were used to study the degree of association between the four agronomic traits and each marker interval. IM was performed using MAPMAKER/QTL V2.0 (Paterson et al., 1988; Lander and Botstein, 1989). A LOD threshold of 2 was set to declare the presence of putative QTL. Estimates of the percent explained variation (PEV), the additive effect, and the dominance effect were obtained from the output of MAPMAKER/QTL.

CIM was carried out with the software package QTL-CARTOGRAPHER/Zmapqtl V2.0 (Zeng, 1994; Zeng and Weir, 1996). Program options included a maximum of five background markers based on forward-backward regression method of selection and a default window size of 10 cM. The Zmapqtl program provides estimates for the square of the partial correlation coefficient ( $R^2$ ), the additive effect, and the dominance effect.  $R^2$  is used to estimate the phenotypic variance explained by QTL. Different algorithms such as multiple linear regression, the maximum likelihood function (Jansen, 1993; Zeng, 1993; Zeng, 1994) and the Markov Chain Monte Carlo (MCMC) approach were applied. For non-normally distributed traits, results were obtained by performing 1000 permutations of each trait using QTL-CARTOGRAPHER (Churchill and Doerge, 1994) which can handle non-normality in both the marker and the trait data.

The multiple interval mapping method (MIM) of QTL-CARTOGRAPHER was employed whenever IM and/or CIM detected more than one QTL on the same linkage group to verify their significance.

#### 4.2.5 QTL X Environment Interaction

Module Jzmapqtl of QTL-CARTOGRAPHER was used to investigate QTL X environment interaction. This module uses the multitrait mapping method of Jiang and Zeng (1995). Herein, each trait from both locations is analyzed simultaneously. A joint LOD score of 2 or higher was considered significant.

### **4.3 Results and Discussion**

To identify QTL controlling each of the four agronomic traits, a molecular map consisting of 28 linkage groups containing 143 AFLP markers was employed; the distance between two markers differed from 1.2 cM to 34.4 cM (see chapter 3). Both

SMA analyses (simple and logistic regression) and IMA analyses (IM and CIM) were used to study the degree of association between the traits measured and marker loci.

#### 4.3.1 Quantitative Traits

Frequency plots (Figure 4.1) showed a continuous normal distribution for 3 of the 4 traits at both locations (Baton Rouge and Alexandria) or at least at one location. Lint percentage was not normally distributed at either location. Boll number per plant and BW were not normally distributed only at Alexandria (Table 4.1). These traits were analyzed based on a non-normal distribution.

The log transformation was used to normalize lint weight per boll at Alexandria and the square root transformation was used to normalize boll number per plant at Baton Rouge (Table 4.1).

Table 4.1 Normality tests for Upland cotton agronomic traits: lint weight per boll (LY), Seedcotton weight per plant (BW), Boll number per plant (B/P), and lint percentage (LP) at Baton Rouge (B) and Alexandria (A).

Trait	Location	N	Mean	SD	Kurtosis	Skewness	Pr<W
LY	B	122	02.197	0.496	01.480	-0.315	0.1310
	A*	131	00.687	0.163	01.473	0.478	0.0250
BW	B	121	03.410	0.746	00.619	0.023	0.5400
	A	131	03.177	0.517	01.997	0.359	0.0004
B/P	B*	121	03.007	0.386	00.217	0.118	0.5890
	A	132	11.072	2.205	04.593	1.169	0.0001
LP	B	121	00.620	0.033	23.464	-4.076	0.0001
	A	132	00.385	0.024	10.843	0.809	0.0001

† Number of lines

‡ Standard deviation

\* After square root, square, or log transformation.

The correlation coefficients for the 4 agronomic traits were analyzed, using the SAS procedure CORR, and are given in Table 4.2. The correlation coefficients were 73.5%, 29.4%, and 20.4% for LY and BW, LY and LP, and B/P and LP, respectively, with P value less than 0.05. These results are similar to those of Lu and Myers (2002).

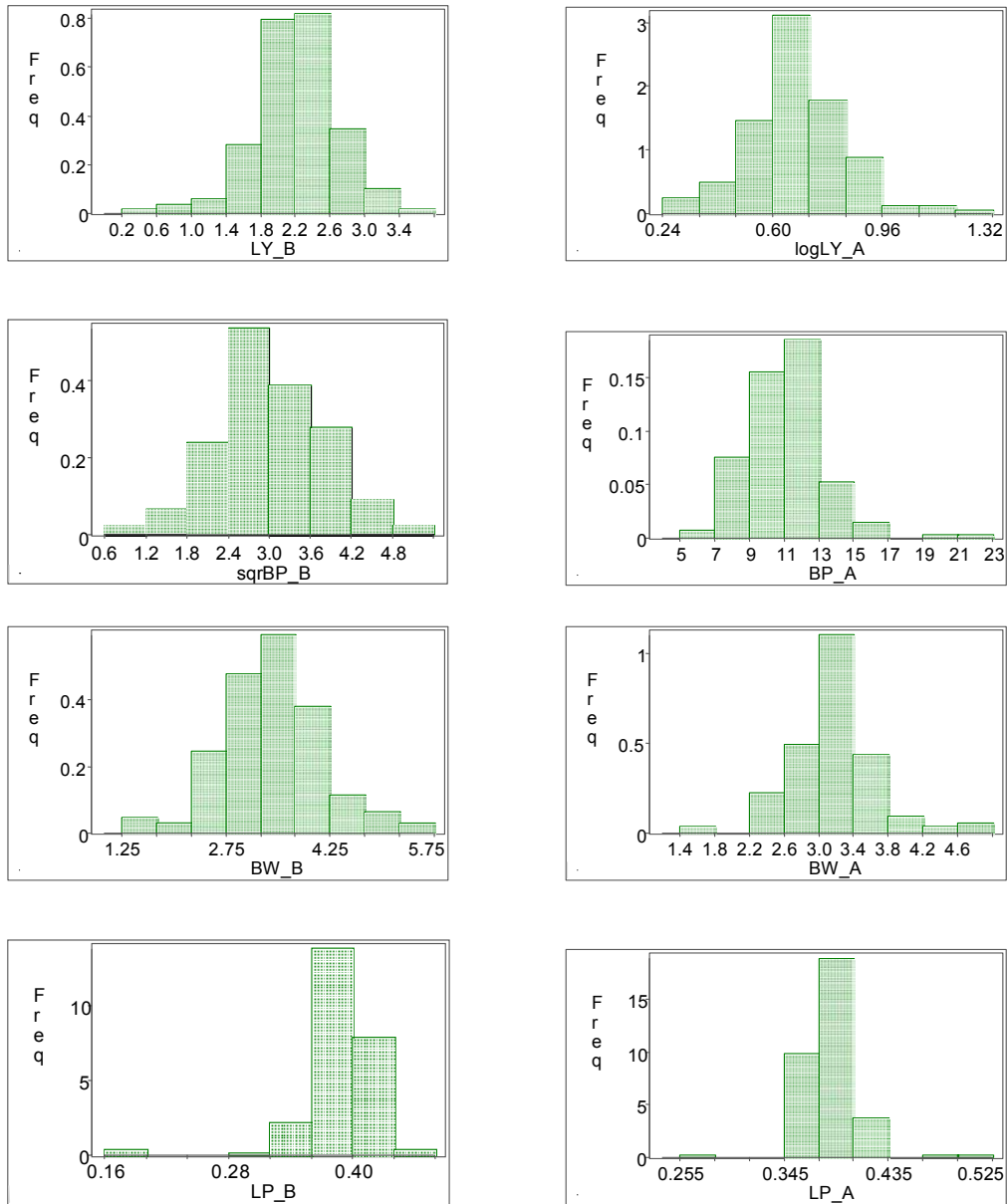


Figure 4.1 Frequency distribution for each Upland cotton agronomic trait in the F<sub>2:3</sub> population at Baton Rouge (B) and Alexandria (A). The data shown for boll number per plant at Baton Rouge (BP\_B) and lint weight per boll at Alexandria (LY\_A) were transformed (square root and log transformation, respectively).

Table 4.2 The correlation among Upland cotton agronomic trait. Data combined for two locations (Alexandria and Baton Rouge). The number on the top is the correlation coefficient and the number below is its correspondent P value.

	BW	B/P	LP <sup>†</sup>
LY <sup>†</sup>	0.74 0.000	0.01 0.890	0.29 0.001
BW <sup>‡</sup>		-0.02 0.800	0.13 0.155
B/P <sup>§</sup>			0.20 0.026

<sup>†</sup> Lint weight per boll                      <sup>‡</sup> Seedcotton weight per plant  
<sup>§</sup> Boll number per plant                      <sup>††</sup> Lint percentage

#### 4.3.2 QTL for Lint Weight Per Boll (LY)

For lint weight per boll, two marker intervals (C08\_211-C14\_345 and C09\_242-C04\_306) were identified in IM, and also in CIM. C08\_211-C14\_345, detected in IM and CIM, accounted for 17.6% and 19% of the phenotypic variation, respectively (Table 4.3). C09\_242-C04\_306, detected in IM and CIM, accounted for 14.9% and 18.5% of the phenotypic variation, respectively. Of the six marker intervals detected using CIM, two intervals (C09\_242-C04\_306 and C20\_207-C10\_241) were located on linkage group 1. MIM was used to test whether or not this linkage group was due to two separate putative QTL (ghost QTL). The results showed the presence of only one QTL located in the C09\_242-C04\_306 interval. The six different QTL explained variation ranging from 11% for LY\_21\_B3 to 19% for LY\_25\_B4. The additive effects ranged from -0.001 to -0.37. QTL detected in both IM and CIM collectively explained about 32.5% and 37.5% of the phenotypic variation in the F<sub>2:3</sub> population, respectively. Ulloa and Meredith (2000) identified two

QTL for LY that collectively explained about 25% of the phenotypic variance in an intraspecific F<sub>2:3</sub> population.

Table 4.3 Putative QTL and their interval position influencing Upland cotton lint weight per boll trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
<u>IM</u>								
LY_25_B1	B	25	C08_211- C14_345	23.7	0.300	0.55	17.6	2.31
LY_01_A1	A	01	C09_242- C04_306	150	-0.001	-0.05	14.9	1.96
<u>CIM</u>								
LY_11_B1	B	11	C06_109- C16_147	14.0	0.290	-0.09	16.4	2.28
LY_19_B2	B	19	C15_218- C05_248	12.5	0.005	0.45	17.8	2.28
LY_21_B3	B	21	C14_053- C15_061- C17_054	63.0	-0.180	-0.33	11.0	2.69
LY_25_B4	B	25	C18_201- C08_211- C14_345	21.8	-0.370	0.61	19.0	2.95
LY_1_A1	A	01	C09_242- C04_306	150.6	-0.020	-0.27	18.5	2.97
LY_1_A2	A	01	C20_207- C10_241	172.5	0.250	-0.14	27.0	2.47

<sup>†</sup> Additive effect

<sup>‡</sup> Dominance effect

<sup>§</sup> Percent explain variation

DNA markers significantly associated with LY using simple regression and logistic regression are listed in Table 4.4.

Table 4.4 AFLP markers that were associated with putative QTL influencing Upland cotton lint weight per boll trait using simple and logistic regression at Baton Rouge (B).

Marker	Location	LG	F value/ $\chi^2$	Pr>F/ Pr> $\chi^2$	R <sup>2</sup> %
<u>Simple regression</u>					
C17_161	B	01	7.88	0.0062	8.67
C15_218	B	19	7.29	0.0084	7.98
C14_053	B	21	7.82	0.0064	8.61
C15_061	B	21	10.2	0.0020	11.0
<u>Logistic regression</u>					
C14_053	B	21	8.32	0.0039	7.90
C15_061	B	21	10.2	0.0014	10.7

#### 4.3.3 QTL for Seedcotton Weight Per Boll (BW)

After MIM dropped the interval C18\_201-C08\_211 (that was detected to be associated with BW in CIM), C15\_061-C17\_054 and C08\_211-C14\_345 (located on linkage group 21 and 25, respectively) were the only marker intervals that were detected (using a LOD threshold of 2) both in IM and CIM analyses. These two intervals were expected to carry putative QTL that explained 8.9% and 19.7% of the phenotypic variation in IM analysis and 18.8% and 25.7% in CIM analysis, respectively (Table 4.5). QTL detected in both IM and CIM collectively explained about 28.6% and 44.9% of the phenotypic variation in the F<sub>2:3</sub> population, respectively. Previous research has indicated that 15 QTL identified for BW explained variation ranging from 4.4 to 23.1% (Saranga et al., 1998).

DNA markers significantly associated with BW using simple and logistic regression are listed in Table 4.6.

Table 4.5 Putative QTL and their interval position influencing Upland cotton seedcotton weight per plant trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
<u>IM</u>								
BW_21_B 1	B	21	C15_061- C17_054	32.4	0.33	-0.33	8.90	2.45
BW_25_B 2	B	25	C08_211- C14_345	25.7	0.29	0.91	19.7	2.27
BW_03_A 1	A	03	C19_115- C07_167	75.9	0.05	-0.32	10.1	1.92
<u>CIM</u>								
BW_21_B4	B	21	C14_053- C15_061- C17_052	63.5	-0.98	-0.71	18.8	2.17
BW_25_B4	B	25	C18_201- C08_211	12.0	-0.67	0.83	18.6	2.52
BW_25_B4	B	25	C08_211- C14_345	23.8	-0.5	0.93	25.7	2.54
† Additive effect				‡ Dominance effect				
§Percent explain variation								

Table 4.6 AFLP markers that were associated with putative QTL influencing Upland cotton seedcotton weight per plant trait using simple and logistic regression at Baton Rouge (B).

Marker	Location	LG	F value/ $\chi^2$	Pr>F/ Pr> $\chi^2$	R <sup>2</sup> %
<u>Simple regression</u>					
C17_161	B	01	8.78	0.0040	8.78
C15_061	B	21	10.5	0.0017	11.2
<u>Logistic regression</u>					
C14_053	B	21	7.13	0.0076	6.31
C15_061	B	21	9.23	0.0018	9.14

#### 4.3.4 QTL for Boll Number Per Plant (B/P)

Two intervals for B/P were detected using IM. The interval C12\_233-C05\_105, on linkage group 15, explained 19.8% of the phenotypic variation and showed a positive additive effect of 0.19 while the interval C20\_094-C12\_197, on linkage group 14, explained 10.1% of the phenotypic variation and showed a negative additive effect of -0.5 (Table 4.7). Three different intervals were detected on three different linkage groups using CIM, which explained phenotypic variation ranging from 7.58% to 19.2%. Four QTL were identified for B/P and the variation explained ranged from 4.0 to 17.7% (Saranga et al., 1998). While the additive effect was negative for the QTL B/P\_05\_A1 (-1.39), the additive effects for the QTL B/P\_27\_B4 and B/P\_23\_A4 were positive (4.36 and 0.61, respectively).

DNA markers significantly associated with B/P using simple and logistic regression are listed in Table 4.8.

Table 4.7 Putative QTL and their interval position influencing Upland cotton boll per plant trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
<u>IM</u>								
B/P_08_B 1	B	08	C04_112- C09_299	14.0	0.4	0.17	14.1	1.89
B/P_15_B 2	B	15	C12_233- C05_105	26.5	0.19	0.74	19.8	2.26
B/P_14_A 1	A	14	C20_094- C12_197	79.2	-0.51	1.53	10.1	2.03

(Table cont'd)



37% of the phenotypic variance (Ulloa and Meredith, 2000). In a study conducted by Shappley et al. (1998), a total of five QTL was detected on five different linkage groups. The additive effects for these QTL were of minor importance ranging from 0.012 to 0.1.

Table 4.9 Putative QTL and their interval position influencing Upland cotton lint percentage trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
<u>IM</u>								
LP_25_B1	B	25	C08_211- C14_345	13.7	0.100	0.100	4.40	13.1
<u>CIM</u>								
LP_25_B4*	B	25	C08_211- C14_345	0.00	0.100	0.100	5.40	2.84
LP_21_A4	A	21	C14_053- C15_061- C17_054	53.6	0.012	-0.002	16.3	2.71

<sup>†</sup> Additive effect

<sup>‡</sup> Dominance effect

<sup>§</sup> Percent explain variation

\*The regression was not performed by Module Jzmapqtl

DNA markers significantly associated with LP using simple and logistic regression are listed in Table 4.10.

Table 4.10 AFLP markers associated with putative QTL influencing Upland cotton lint percentage trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A).

Marker	Location	LG	F value/ χ <sup>2</sup>	Pr>F/ Pr>χ <sup>2</sup>	R <sup>2</sup> %
<u>Simple regression</u>					
C15_218	B	19	7.43	0.0078	8.13
C05_049	A	03	11.5	0.0010	11.2
C19_056	A	03	9.10	0.0033	9.09
C19_115	A	03	7.06	0.0093	7.20

(Table cont'd)

Marker	Location	LG	F value/ $\chi^2$	Pr>F/ Pr> $\chi^2$	R <sup>2</sup> %
<u>Logistic regression</u>					
C01_251	A	24	6.65	0.0099	5.32
C03_193	A	03	6.83	0.0090	5.44
C05_049	A	03	10.8	0.0010	9.13
C14_048	A	03	7.33	0.0068	6.02
C15_061	A	21	7.60	0.0058	6.45
C19_056	A	03	10.9	0.0010	9.25
C19_115	A	03	9.63	0.0019	8.04

#### 4.3.6 QTL X Environment Interaction.

Molecular markers offer the opportunity to study QTL X environment interaction (Paterson et al., 1991; Dudley, 1993; Beavis and Keim, 1996). Although there are no statistical methods available to test the QTL X environment interaction in MapMaker/QTL, interval mapping in QTL-CARTOGRAPHER was used to analyze the QTL X environment interaction using the joint analysis method (module JZmapqtl). Only two QTL for LY and BW were shown to have a significant interaction effect at a LOD threshold of 2 between the two locations (LY\_25\_B1 at 2.49 LOD and BW\_21\_B1 at 2.67 LOD, respectively (Table 4.11). The present study supported the general conclusion made by Tanksley (1993) that different QTL affecting a trait may be found under varying environmental conditions.

Table 4.11 QTL X environment interaction LOD using Module Jzmapqtl of QTL-CARTOGRAPHER in Upland cotton.

QTL	LOD		Joint LOD
	Baton Rouge	Alexandria	
LY_25_B1	02.31	0.22	2.49
BW_21_B1	02.45	-	2.67
BW_25_B2	02.27	-	1.87
B/P_15_B2	02.26	0.80	1.95
B/P_14_A1	00.45	2.03	0.11
LP_25_B1	13.14	0.08	-

The segregation in the  $F_{2:3}$  population for the four agronomic traits could be largely explained by several QTL and their complex interactions with the environment. Based on the IM method of QTL detection, the total number of QTL detected (number of marker intervals that were found to be significantly associated with the four agronomic traits) at both locations was five at Baton Rouge and one at Alexandria. Results obtained from composite interval mapping were similar when compared with the results obtained from interval mapping (Table 4.12 and Figure 4.2). The total number of QTL detected was nine at Baton Rouge and five at Alexandria. In total, five and 10 different QTL were detected using IM and CIM, respectively. A range of small to medium proportions of the trait phenotypic variance (4.4 to 32.5%) explained by QTL was common in our study and supports a model for quantitative inheritance for all the agronomic traits studied (Lande and Thompson, 1990; Ulloa and Meredith, 2000).

In IM, the same marker interval (C08\_211-C14-345) for LY, BW, and LP was detected and is likely due to either linkage or pleiotropic effects on multiple traits. However, the exact putative QTL positions for the three traits, from the upper start on linkage group 25, were 23.7, 25.7, and 13.7 cM, respectively. These results explain the strong correlation between lint weight per boll and seedcotton weight per plant ( $r = 0.73$  with  $p < 0.0001$ ) and between lint weight per boll and lint percentage ( $r = 0.29$  with  $p < 0.001$ ). Similar to IM results, CIM analysis indicated that the interval C08\_211 contained a QTL associated with LY, BW, and LP with an explained variation of 19%, 25.7%, and 4.4% respectively. With the sole exception for the trait B/P, CIM detected the same QTL that IM detected plus a few more. Five marker intervals were detected using both CIM and IM.

Generally, all QTL that had a low magnitude of effect was consistent with the low heritability of the studied traits. Another reason may be a relatively small difference between the two parents. However, we have been able to identify QTL that could not be found using traditional methods, by using interval mapping, and confirmed their occurrence with composite interval mapping. Knowledge of the number and the likely position of QTL can provide information required to select optimal combinations of alleles by the use of MAS.

Table 4.12 The QTL summary for Upland cotton agronomic traits.

Trait	Number of QTL detected in						Accumulative percent explained variation *
	IM		CIM		Both method		
	B	A	B	A	B	A	
LY	1	1	4	2	1	1	32.5
BW	2	1	3	0	2	0	28.6
B/P	2	1	1	2	0	0	-
LP	1	0	1	1	1	0	04.4
Total	6	3	9	5	4	1	NA

\* Using QTL identified by both IM and CIM methods.

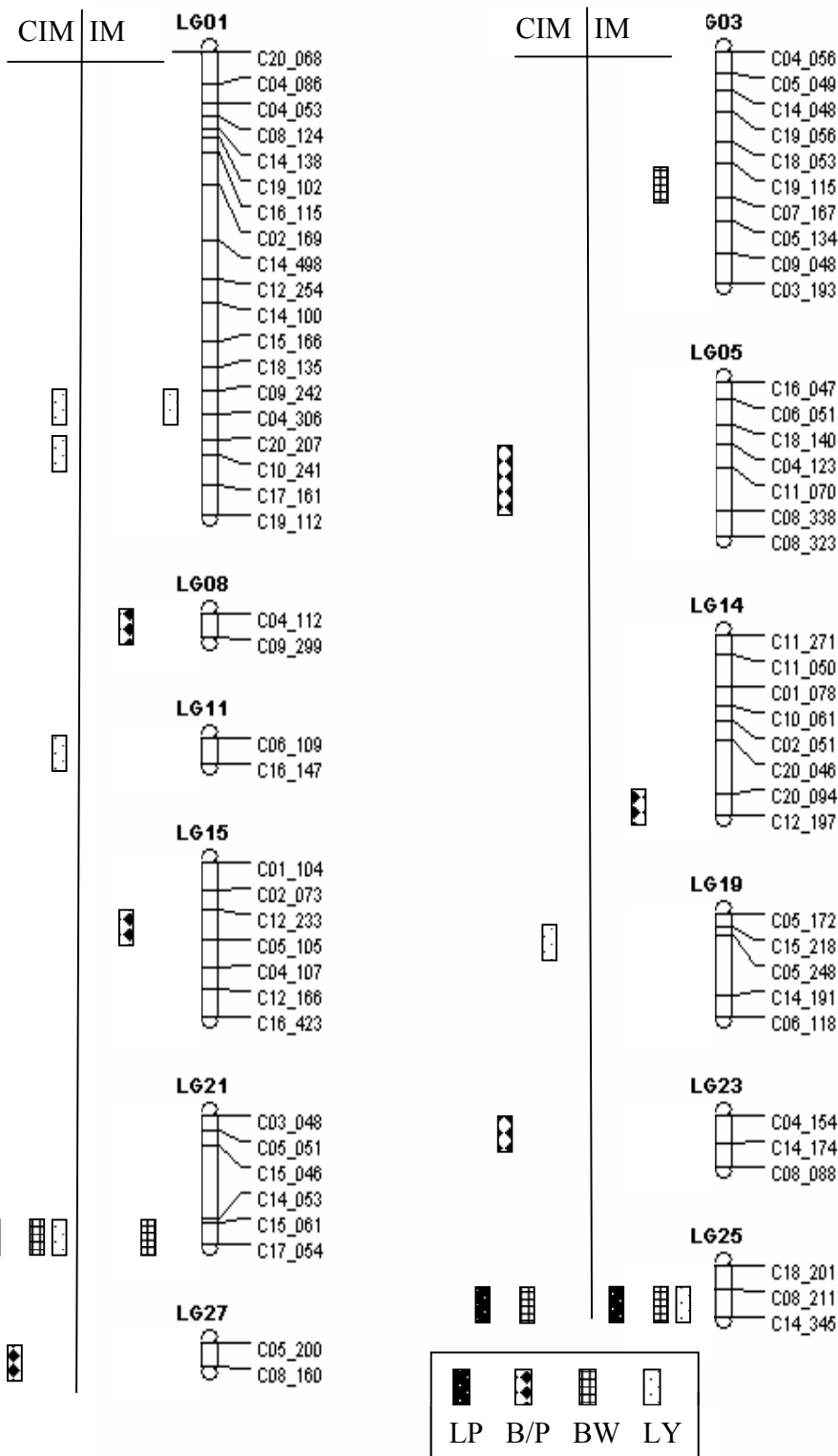


Figure 4.2 A comparison of QTL positions for Upland cotton lint weight per boll (LY), seed cotton weight per boll (BW), bolls number per plant (B/P), and lint percentage (LP) using composite interval mapping (CIM) and interval mapping (IM).

#### 4.4 References

- Basten, C. J., B. S. Weir, and Z-B. Zeng. 1997. QTL Cartographer: A reference manual and tutorial for QTL mapping. Department of Statistics, North Carolina State University, Raleigh, NC.
- Beavis, W. D., and P. Keim. 1996. Identification of QTL that are affected by environment. In: Kang M (ed.) New Perspectives on Genotype by Environment Interactions. CRC Press.
- Cao, G., J. Zhu, C. He, Y. Gao, J. Yan, and P. Wu. 2001. Impact of epistasis and QTL×environment interaction on the developmental behavior of plant height in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 103:153-160.
- Churchill, R. W., and G. A. Doerge. 1994. Empirical threshold values for quantitative trait mapping. *Genetics.* 138: 963-971.
- Dempster, A., A. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B.* 39: 1-38.
- Dudley, J. W. 1993. Molecular markers in plant improvement: manipulation of genes affecting quantitative traits. *Crop Sci.* 33: 660-668.
- Ensminger, M. E., J. E., Oldfield and W. W. Heinemann. 1990. Excerpts with reference to cottonseed components. In Ensminger Publishing Company, USA.
- Flores-Berrios, E. L. Gentsbittel, L. Mokrani, G. Alibert, and A. Sarrafi. 2000. Genetic control of early events in protoplast division and regeneration pathways in sunflower. *Theor. Appl. Genet.* 101: 606–612.
- Haldane, J. B. S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. of Genet.* 8: 299-309.
- Jansen, R. C. 1993. Interval mapping of multiple quantitative trait loci. *Genetics.* 135: 205-211.
- Jiang, C., and Z-B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics.* 140: 111-1127.
- Jiang, C-X., R. J. Wright, S. S. Woo, T. A. Del Monte, and A. Paterson. 2000. QTL analysis of leaf morphology in tetraploid *Gossypium* (cotton). *Theor. Appl. Genet.* 100: 409-418.
- Johnson, W. C., L. E. Jackson, O. Ochoa, R. Van Wijk, J. Peleman, D. A. St.Clair, and R. W. Michelmore. 2000. Lettuce, a shallow-rooted crop, and *Lactuca*

- serriola, its wild progenitor, differ at QTL determining root architecture and deep soil water exploitation. *Theor. Appl. Genet.* 101: 1066-1073.
- Khan, M. A., J. Zhang, J. McD. Stewart, and R. G. Cantrell. 1998. Integrated molecular map based on a trispecific F<sub>2</sub> population of cotton. In: Beltwide Cotton Conference. 491-492.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics.* 124: 743-756.
- Lander, E. S., and D. Botstein, 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics.* 121: 185-199.
- Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural population. *Genomics.* 1: 174-181.
- Lewis, H. 2001. A review of yield and quality trends and components in American Upland cotton. p. 1447-1453. In: 2001 Proc. Beltwide Cotton Conferences. National cotton Council, Memphis, TN.
- Lu, H. J., and G. O. Myers. 2002. Genetic Relationships and Discrimination of Ten Influential Upland Cotton Varieties using RAPD Markers. *Theor. Appl. Genet.* 105: 325-331.
- Marques, C. M., J. Vasquez-Kool, V. J. Carocha, J.G. Ferreira, D. M. O'Malley, B-H. Liu, and R. Sederoff. 1999. Genetic dissection of vegetative propagation traits in *Eucalyptus tereticornis* and *E. globules*. *Theor. Appl. Genet.* 99: 936-946.
- Paterson, A., E. Lander, S. Lincoln, J. Hewitt, S. Peterson, and S. Tanksley. 1988. Resolution of quantitative traits into mendelian factors using a complete RFLP linkage map. *Nature.* 335: 721-726.
- Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch, S. E. Lincoln, E. S. Lander, and S. D. Tanksley. 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics.* 127: 181-197.
- Reinisch, M. J., J. Dong, C. L. Brubaker, D. M. Stelly, J. F. Wendel, and A. H. Paterson. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: Chromosome organization and evolution in a disomic polyploid genome. *Genetics.* 138: 829-847.
- Saranga, Y., M. Menz, C-X. Jiang, R. L. Wright, D. Yakir, and A. H. Paterson. 2001. [www.genome.org](http://www.genome.org).

- Sax, K. 1923. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics* 8: 552-560.
- Schork, N., M. Boehnke and J. Terwilliger. 1993. Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.* 53: 1127-1136.
- Shappley, Z.W. 1994. RFLPs in cotton (*Gossypium hirsutum* L.): Feasibility of use, diversity among plants within a line, and establishment of molecular markers and linkage groups among two F2 populations. M.S. thesis. Mississippi State Univ., Mississippi State.
- Shappley, Z. W., J. N. Jenkins, J. Zhu, and J. C. McCarty, Jr. 1998. Quantitative trait loci associated with agronomic and fiber traits of Upland cotton. *The Journal of Cotton Sci.* 4: 153-163.
- Soller, M., T. Brody, and A. Genizi, 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theor. Appl. Genet.* 47: 35-39.
- Tanksley, S. D. 1993. Mapping polygenes. *Annu. Rev. Genet.* 27: 205-233.
- Ulloa, M., and W. R. Meredith Jr. 2000. Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. *The Journal of Cotton Science.* 4: 161-170.
- Wang, D., R. Karle, and A. F. Iezzoni. 2000. QTL analysis of flower and fruit traits in sour cherry. *Theor. Appl. Genet.* 100: 535-544.
- Wang, D. L., J. Zhu, Z. K. Li, and A. H. Paterson. 1999. Mapping QTLs with epistatic effects and QTL X environment interactions by mixed linear model approaches. *Theor. Appl. Genet.* 99: 1255-1264.
- Wang, D., P. R. Arelli, R. C. Shoemaker, and B.W. Diers. 2001. Loci underlying resistance to Race 3 of soybean cyst nematode in *Glycine soja* plant. *Theor. Appl. Genet.* 103: 561-566.
- Weller, J. I., M. Soller, and T. Brody. 1988. Linkage analysis of quantitative traits in an interspecific cross of tomato (*L. esculentum* x *L. pimpinellifolium*) by means of genetic markers. *Genetics.* 118: 329-339.
- Wright, R. J., P. M. Thaxton, K. M. El-Zik, and A. H. Paterson. 1999. Molecular mapping of genes affecting pubescence of cotton. *The American Genetic Association.* 90: 215-219.
- Yadav, R. S., C. T. Hash, F. R. Bidinger, G. P. Cavan, and C. J. Howarth. 2002. Quantitative trait loci associated with traits determining grain and stover yield

- in pearl millet under terminal drought stress conditions. *Theor. Appl. Genet.* 104: 67-83.
- Zeng, Z. B. 1993. Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA.* 90: 10972-10976.
- Zeng, Z-B. 1994. Precision mapping of quantitative trait loci. *Genetics.* 136: 1457-1468.
- Zeng, Z-B., and B. S. Weir. 1996. Statistical methods for mapping Quantitative Trait Loci. *Acta Agronomica Sinica.* 22: 535-549.

## **CHAPTER 5**

### **MOLECULAR QTL MAPPING FOR FIBER QUALITY TRAITS IN UPLAND COTTON (*GOSSYPIUM HIRSUTUM* L.)**

#### **5.1 Introduction**

Cotton (*Gossypium hirsutum* L.) is the world's most important fiber crop and also a major oilseed crop. It is grown commercially in the temperate and tropical regions of more than 50 countries including the United States, India, China, Central and South America, the Middle East, and Australia (Smith, 1999; Fryxell, 1979). Traditional plant improvement efforts have been largely successful in modifying the crop to meet the needs of both producers and consumers. Genetic engineering has been used in recent years as well in cotton to address several important pest problems such as weeds (Murdock et al., 2001) and lepidopterous pests (Perlak et al., 2001). Future improvements in cotton will depend upon the concerted application of traditional plant breeding, genetic engineering, and molecular genetic tools to increase yield and fiber quality. Modern spinning technologies require strong fibers that hold up to the rigors of ginning, opening, cleaning, combing and drafting (Zhang et al., 2003).

With the availability of molecular markers as well as genetic maps, it is now possible to identify quantitative trait loci (QTL) for cotton phenotypic traits. Mapping is an efficient method to scan the genome for putative QTL. The determination of the locations of QTL should increase selection efficiency through the use of marker-assisted selection (MAS), and open the door for their future genetic manipulation and possible transfer among different plant species.

Several molecular marker technologies have been applied to cotton in an attempt to identify QTL. Shappley et al. (1998) identified 100 QTL for 19 agronomic and fiber traits in cotton by using RFLP markers. Saranga et al. (2001) identified 161 QTL for 16 cotton productivity, physiological, and fiber quality traits by using RFLP markers. Kohel et al. (2001) also identified 13 QTL for three fiber quality traits by using RFLP and RAPD markers. However, no QTL involved in the expression of the five standard fiber quality traits (fiber length, elongation, strength, uniformity and micronaire) have been identified by means of amplified fragment length polymorphism (AFLP) markers. In cotton, AFLPs may be the marker system of choice due to the low amount of polymorphism detectable by other DNA marker technologies.

There are several different methods available for identifying QTL that segregate in a mapping population. The regression of a trait on a single marker (single-marker analysis (SMA)) is the simplest single-QTL method. The loci are tested one at a time for the presence of a single QTL. Generally, the significance level is adjusted to account for the multiple tests performed. Locations on the genome that show significant results are indicated to contain a QTL. Tanksley (1993) discussed several drawbacks, the most serious one was the confounding of the QTL effects with recombination frequencies, which will lead to underestimation of the QTL effect, especially if the QTL is far from the locus under investigation.

Compared with SMA, interval-marker analysis (IMA) (including interval mapping (IM) and composite interval mapping (CIM)) have several advantages. These advantages include, (1) the probable position of the QTL is inferred by the support interval, (2) the estimated locations and QTL effects can be asymptotically

unbiased if there is only one segregating QTL on a linkage group, and (3) the method requires fewer individuals than single-marker analysis for the detection of QTL (Zeng and Weir, 1996).

There are, however, some problems with IM. The more serious of these include: (1) If there are more than one QTL on a linkage group, interval mapping can be seriously biased and the position being tested will be affected by all other QTL on the same linkage group, and (2) It is not efficient to use only two markers at a time to do the test as the information from other markers is not used (Zeng and Weir, 1996).

Similar to IM, CIM (Zeng, 1993; Jiang and Zeng, 1995) evaluates the presence of a putative QTL using flanking markers. However, CIM relies on the use of multiple regression methods. This has the advantage in that the partial regression coefficient of a trait on a marker is expected to depend only on those QTL that are located in the interval bracketed by the two neighboring markers and to be independent of any other marker (Zeng and Weir, 1996). The main problem in this method is the number of regressor markers (background markers). Using too many background markers will increase the variance of the LOD score, and thus will decrease the power for detecting QTL. Basten et al. (1997) recommended using forward selection up to a fixed number of markers and then dropping any markers that are not within 10 cM of the putative QTL. Many previous studies have used five markers which is the default for one of the most common QTL mapping programs, QTL-CARTOGRAPHER, (Wang et al., 2000, Wang et al., 2001, Marques et al., 1999), Johanson et al. (2000) and Flores-Berrios et al. (2000), respectively, have used 10 and 15 markers.

Because of their ability to separate linked QTL on the same linkage group and to detect interacting QTL that may otherwise be undetected, multiple QTL methods are an improvement over single QTL methods. Multiple-QTL methods provide increased power to detect QTL and eliminate any biased estimates of effects of size and location that can be introduced by using a single-QTL methods (Schork et al., 1993).

The interaction of QTL with the environment has been discussed in many studies. These studies have found either significant QTL effects being detected only in a subset of all the environments, or changes in the magnitude of the QTL effect (Paterson et al., 1991; Wang et al., 1999; Cao et al., 2001; Yadav et al., 2002). In a cotton study, 61 of 161 QTL, detected for 16 measured traits, showed significant differences in their effect estimate between well-watered and water-limited conditions, indicating a significant QTL X environment interaction (Saranga et al., 1998).

In the present chapter, an AFLP linkage map (28 linkage groups comprised of 143 markers) for two *Gossypium hirsutum* parents (Paymaster 54 X Pee Dee 2165) was used to identify QTL for five fiber quality traits using both IM and CIM. QTL X environment interaction was also investigated.

## **5.2 Materials and Methods**

### **5.2.1 Plant Material**

The QTL mapping population was developed for an intraspecific cross between two parents of *G. hirsutum* (Paymaster 54 and Pee Dee 2165). Crosses were made between them and 138 F<sub>2:3</sub> progeny lines were used in this study. Parents available from a previous study (Lu and Myers, 2002) and F<sub>2:3</sub> seeds were planted in the field on May 10, 2002 at the LSU AgCenter Dean Lee Research

Station in Alexandria, LA and Central Research Station in Baton Rouge, LA. These F<sub>2:3</sub> seed were planted in single-row plots, 5 m long, spaced 1 m apart with seed sown by hand 15 cm apart. At each station, two replications of the entries, arranged in an incomplete block design, were used to determine fiber quality traits.

### 5.2.2 Phenotypic Measurement

Traits data were collected from 138 F<sub>2:3</sub> samples (20 to 25 bolls per row). Samples were then ginned on a 7-saw laboratory gin to separate lint from fussy seeds. Fiber strength (in grams per tex), length (upper half means in inches), elongation, length uniformity index, and micronaire (expressed in standard micronaire units) were measured as fiber quality measurement standards. These traits were determined using High Volume Instrumentation (HVI) equipment (Uster Technologies, Inc. Knoxville, TN) at the LSU Cotton Fiber Lab (Louisiana State University, Baton Rouge, LA).

### 5.2.3 Linkage Analysis

AFLP polymorphic bands were scored as present (1) or absent (0) on GenElmagIR. Data coded (0) and (1) were transformed to A,B,C,D genotype codes, according to the presence of the band for the two parents, following the Mapmaker convention. A molecular linkage map consisting of 143 AFLPs was constructed using MAPMAKER 3.0 (Lander et al., 1987). Linkage groups were obtained with a LOD score of 4 and a maximum recombination frequency of 0.34. The Haldane function was used to transform the recombination frequency to genetic distances (Haldane, 1919).

#### 5.2.4 QTL Analysis

To identify QTL that control each of the five fiber quality traits, a molecular map that had 28 linkage groups based upon 143 AFLP markers was used. SMA (simple and logistic regression) and IMA (IM and CIM) methods were used to study the degree of association. Two SMA methods were used to study the degree of association between the five traits and each marker. The first single-marker analysis method was based on simple regression/ANOVA (SAS Institute, Cary, NC) in which marker genotype was used as a class variable. To reduce the chance of a false positive error rate, an association was considered to be significant whenever the probability value was less than or equal to 0.01. Logistic regression was used as a second single marker method, again using PC-SAS Version 9.0 (SAS Institute, Cary, NC). Using logistic regression, marker genotype was used as the dependent variable.

Interval mapping was performed using MAPMAKER/QTL V2.0 (Paterson et al., 1988; Lander and Botstein, 1989). An LOD threshold of 2 was set to declare the presence of putative QTL. An estimate of the percent explained variation (PEV), the additive effect, and the dominance effect were obtained from the output of MAPMAKER/QTL

Composite interval mapping was carried out using QTL-CARTOGRAPHER/Zmapqtl V2.0 (Zeng, 1994; Zeng and Weir, 1996). Program parameter used set the maximum number of background markers to 5 with a forward-backward regression method of selection and a default window size of 10 cM. The Zmapqtl program provides estimates for the square of the partial correlation coefficient ( $R^2$ ), the additive effect, and the dominance effect.  $R^2$  is used to estimate the phenotypic variance explained by QTL. Different algorithms, such as multiple

linear regression, maximum likelihood function (Jansen, 1993; Zeng, 1993; Zeng, 1994) and the Markov Chain Monte Carlo (MCMC) approach, were applied to CIM.

For non-normally distributed traits, results were determined by performing 1000 permutations of each trait using QTL-CARTOGRAPHER (Churchill and Doerge, 1994), which can handle non-normality in both the marker and the trait data. The likelihood value for the presence of a QTL was expressed as a LOD score  $\log_{10}(L1/L0)$ , where L1 is the likelihood of the model with the putative QTL and L0 is the likelihood of the model without the QTL.

Multiple interval mapping method (MIM) using the program QTL-CARTOGRAPHER was used whenever IM and/or CIM detected more than one QTL on the same linkage group to verify their significance.

#### 5.2.5 QTL X Environment Interaction

Module Jzmapqtl of QTL-CARTOGRAPHER was used to investigate the QTL X environment interaction. This module uses the multitrait mapping method of Jiang and Zeng (1995). Each trait from both locations was analyzed simultaneously.

### **5.3 Results and Discussion**

#### 5.3.1 Quantitative Traits

The frequency plots (Figure 5.1) showed a continuous distribution following normality for four traits in both locations (Baton Rouge and Alexandria) or at least in one location. Fiber elongation was not normally distributed in either location. Micronaire was not normally distributed only at Alexandria (Table 5.1) and was analyzed based on a non-normal distribution.

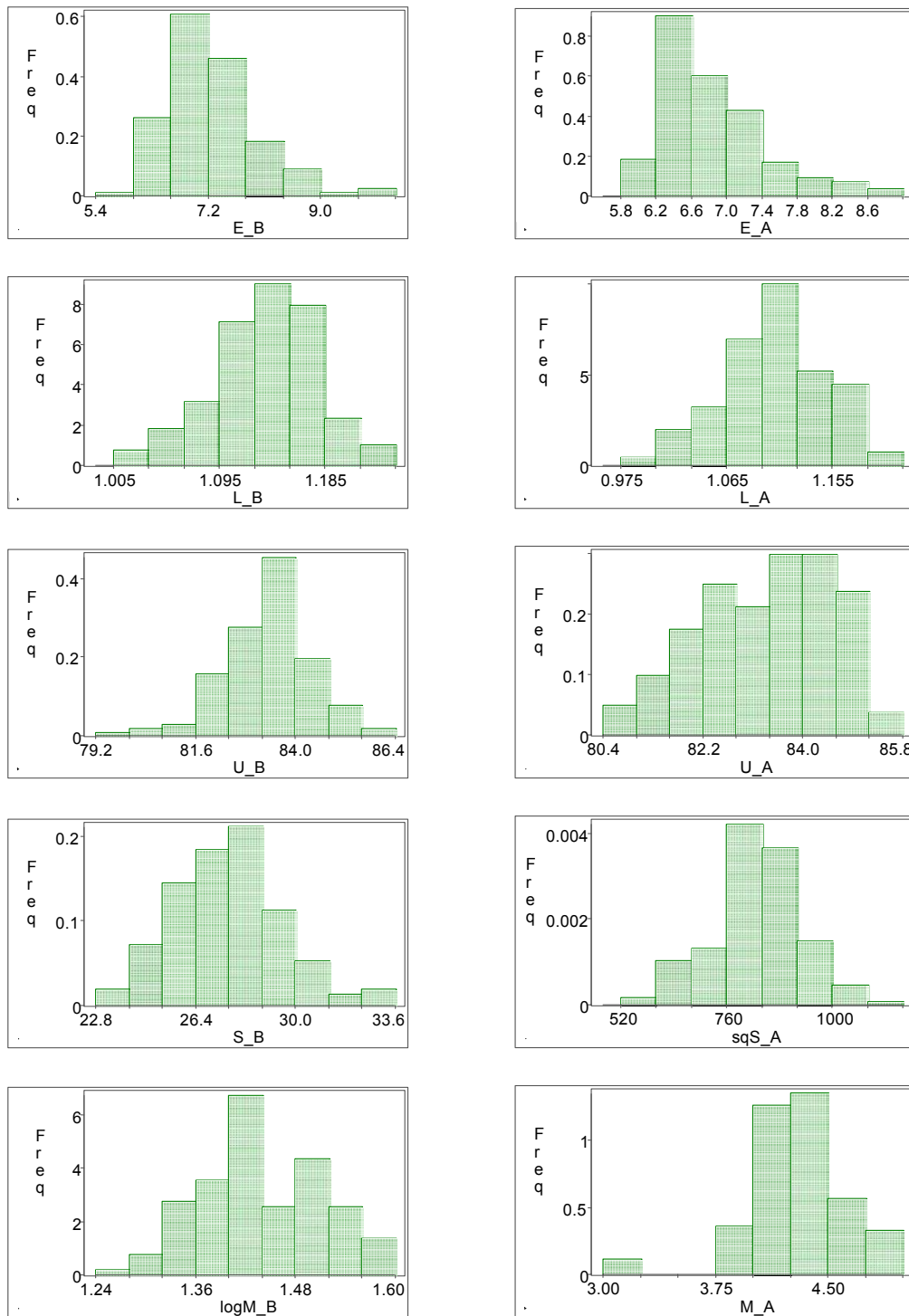


Figure 5.1 Frequency distribution for each Upland cotton fiber quality trait in the F<sub>2:3</sub> population at Baton Rouge (B) and Alexandria (A). The data shown for micronaire at Baton Rouge (M<sub>B</sub>) and strength at Alexandria (S<sub>A</sub>) were transformed (square root and log transformation, respectively).

To normalize micronaire at Baton Rouge a log transformation was used and a square transformation was used to normalize fiber strength at Alexandria (Table 5.1).

Table 5.1 Normality tests for Upland cotton fiber quality traits: elongation (E), length (L), uniformity (U), strength (S), and micronaire (M) at Baton Rouge (B) and Alexandria (A).

Trait	Location	N <sup>†</sup>	Mean	SD <sup>‡</sup>	Kurtosis	Skewness	Pr<W
E	B	126	7.230	0.725	1.893	1.031	0.0001
	A	133	6.793	0.580	1.345	1.139	0.0001
L	B	126	1.135	0.044	0.258	-0.242	0.1490
	A	133	1.105	0.045	-0.163	-0.266	0.1600
U	B	126	83.30	1.064	0.710	-0.321	0.2940
	A	133	83.11	1.030	-0.201	-0.396	0.0513
S	B	126	27.61	1.983	0.292	0.342	0.2350
	A*	133	828.6	106.2	0.446	-0.203	0.3627
M	B*	126	1.434	0.071	-0.347	0.034	0.0140
	A	133	4.260	0.338	2.810	-0.922	0.0001

† Number of lines

‡ Standard deviation

\* After square root, square, or log transformation.

The correlation analysis for the 5 fiber quality traits are summarized in Table 5.2. Four different correlation coefficients were highly significant ( $P < 0.0001$ ). The correlation coefficient was 47% between length and uniformity, 39% between length and strength, 34.8% between length and micronaire, 48.7% between uniformity and strength. These results are similar to those reported by Lu and Myers (2002).

Table 5.2 The correlation among Upland cotton fiber quality traits. Data combined the two locations. The number on the top is the correlation coefficient and the number below is its correspondent P value.

	Length	Uniformity	Strength	Micronaire
Elongation	-0.16	-0.07	-0.04	0.07
	0.08	0.43	0.64	0.49
Length		0.47	0.39	-0.35
		0.00	0.00	0.00
Uniformity			0.49	0.12
			0.00	0.19
Strength				0.07
				0.47

### 5.3.2 QTL for Fiber Elongation (E)

The marker interval (C15\_061-C17\_054) was found to be significantly associated with E in both locations (Table 5.3). The QTL, E\_21, detected within this interval explained 54.7% and 47% of the phenotypic variation in Baton Rouge and Alexandria, respectively. Ulloa and Meredith (2000) detected three QTL that collectively explained 47% of the phenotypic variation for E using an RFLP linkage map based on 119 F<sub>2:3</sub> progeny from an intraspecific cross between MD5678ne X Prema. However, other research has indicated that fiber elongation is controlled by as many as 18 QTL (Shappley et al., 1998). One copy of the Paymaster 54 alleles at E\_21 in the Pee Dee 2165 background decreased E by 0.4 to 0.5 units. No results were obtained by CIM since the required multiple regression step failed to find the required markers.

Table 5.3 Putative QTL and their interval position influencing Upland cotton fiber elongation trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
IM								
E_21_B1	B	21	C15_061- C17_054	38.4	-0.51	-0.81	54.7	2.34
E_21_A1	A	21	C15_061- C17_054	38.4	-0.40	0.63	47.0	2.27

<sup>†</sup> Additive effect

<sup>‡</sup> Dominance effect

<sup>§</sup> Percent explain variation

DNA markers significantly associated with E using simple regression and logistic regression are listed in Table 5.4.

Table 5.4 AFLP markers that were associated with putative QTL influencing fiber elongation trait using logistic regression at Baton Rouge (B).

Marker	Location	LG	F value $\chi^2$	Pr>F Pr> $\chi^2$	R <sup>2</sup> %
<u>Logistic regression</u>					
C20_293	B	*	7.07	0.0078	5.99

\* Unlinked markers

### 5.3.3 QTL for Fiber Length (L)

Using IM method, five marker intervals were associated with L (four in Alexandria and one in Baton Rouge) and located on five different linkage groups (Table 5.5). The explained variation accounted for by the QTL with these intervals ranged from 12.3% for L\_13\_A4 to 18.7% for L\_02\_A3. However, the same marker interval (C05\_180-C17\_084) was significantly detected in both locations. The marker interval C20\_051-C10\_064 was the only one that also was detected by CIM. In CIM analysis, seven QTL were associated with fiber length, three in Baton Rouge and four in Alexandria, with an explained variation ranging from 9.05% for L\_02\_A1 to 23% for L\_02\_A2. Each Linkage group (4 and 2) has two QTL. MIM analysis dropped L\_04\_B1 from linkage group 4 but did not drop any of the two QTL on linkage group 2. QTL detected with both IM and CIM collectively explained about 18.7% and 9.05% of the phenotypic variation in the F<sub>2:3</sub> population, respectively. Kohel et al. (2002) identified three QTL for fiber length that collectively explained about 30% of the phenotypic variation in an F<sub>2</sub> population. The additive effects of these QTL were of minor importance, ranging from -0.009 to -0.2 in IM and 0.008 to 0.032 in CIM.

Table 5.5 Putative QTL and their interval position influencing Upland cotton fiber length trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	$a^{\dagger}$	$d^{\ddagger}$	PEV $^{\S}$	LOD
<u>IM</u>								
L_13_B1	B	13	C05_180- C17_084	04.0	-0.2	0.01	12.6	2.98
L_18_A1	A	18	C20_521- C17_171	75.4	0.004	0.04	16.9	2.47
L_28_A2	A	28	C11_078- C16_132	42.9	-0.02	0.13	13.7	2.47
L_02_A3	A	02	C20_051- C10_064	188.6	-0.009	0.04	18.7	2.26
L_13_A4	A	13	C05_180- C17_084	08.0	-0.02	0.01	12.3	2.65
<u>CIM</u>								
L_04_B1	B	04	C12_083- C11_252- C11_334	52.3	0.026	-0.01	13.4	2.17
L_04_B2	B	04	C11_334- C11_453	75.1	0.031	-0.005	14.9	2.60
L_16_B3	B	16	C04_299- C16_370	00.0	-0.025	0.002	11.0	2.10
L_02_A1	A	02	C10_339- C12_251- C15_121	36.4	-0.002	0.043	9.05	0.25
L_02_A2	A	02	C20_051- C10_064	188.7	0.008	0.047	23.0	2.30
L_05_A3	A	05	C16_047- C06_051- C18_140	14.8	-0.013	0.045	22.6	2.08
L_13_A4	A	13	C18_114- C20_307	68.1	0.032	0.027	13.7	2.32

$^{\dagger}$  Additive effect

$^{\ddagger}$  Dominance effect

$^{\S}$  Percent explain variation

DNA markers significantly associated with L using simple regression and logistic regression are listed in Table 5.6.

Table 5.6 AFLP markers that were associated with putative QTL influencing Upland cotton fiber length trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A).

Marker	Location	LG	F value $\chi^2$	Pr>F Pr> $\chi^2$	R <sup>2</sup> %
<u>Simple regression</u>					
C12_083	B	04	10.1	0.0021	10.9
C05_180	B	13	10.5	0.0016	10.4
C18_114	B	13	12.8	0.0006	12.3
C08_116	B	16	7.30	0.0082	7.43
C11_078	A	28	7.32	0.0083	8.11
C16_132	A	28	10.6	0.0016	11.3
C01_432	A	04	8.60	0.0043	9.39
C12_083	A	04	13.1	0.0005	13.6
C08_230	A	*	7.99	0.0059	8.69
C05_180	A	13	9.20	0.0032	9.18
C17_084	A	13	8.75	0.0039	8.77
C18_114	A	13	11.3	0.0012	11.0
C17_171	A	18	8.67	0.0041	8.70
C08_116	A	16	10.4	0.0017	10.3
C10_064	A	02	7.83	0.0064	8.63
<u>Logistic regression</u>					
C05_180	B	13	11.2	0.0008	9.98
C18_114	B	13	9.23	0.0024	8.12
C05_180	A	13	8.44	0.0037	6.77
C06_272	A	*	8.83	0.0030	7.24
C08_116	A	16	7.00	0.0081	5.51
C08_230	A	*	7.61	0.0058	6.06
C10_064	A	02	8.85	0.0029	7.10
C16_132	A	28	7.85	0.0051	6.43
C17_084	A	13	7.24	0.0071	5.72
C17_171	A	18	8.37	0.0036	6.81
C18_114	A	13	8.47	0.0036	6.87
C20_307	A	13	7.22	0.0072	5.72

\* Unlinked markers

#### 5.3.4 QTL for Fiber Uniformity (U)

A total of three and eight marker intervals were detected to be associated with U in IM and CIM respectively (Table 5.7). Of the three intervals detected using IM, two (02\_247-C11\_078 on linkage group 28 and c15\_061-C17\_054 on linkage group

21) were also detected by CIM. These two intervals explained 15.3% and 53.7%, using IM, and 21.4% and 52.4%, using CIM, of the total phenotypic variation, respectively. The explained variation ranged from 12.8% for the QTL U\_13\_B1 to 53.7% for the QTL U\_21\_A1 in IM and 9.6 percent for the QTL U\_13\_B3 to 56.9% for the QTL U\_21\_A3 in CIM. Out of the three QTL (U\_21\_A2, U\_21\_A3, and U\_21\_A4) that were located on linkage group 21, MIM dropped both U\_21\_A3, and U\_21\_A4 after testing their effects. QTL detected with both IM and CIM collectively explained about 69% and 62% of the phenotypic variation in the F<sub>2.3</sub> population, respectively. This study was the first attempt to discuss QTL for fiber uniformity. The additive effect estimates ranged from a positive 0.14 for the QTL U\_06\_B2 to a negative 0.84 for the QTL U\_15\_A1.

Table 5.7 Putative QTL and their interval position influencing Upland cotton fiber uniformity trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
<u>IM</u>								
U_13_B1	B	13	C17_084- C04_272	20.9	-0.54	0.13	12.8	2.50
U_28_B2	B	28	C02_247- C11_078	22.0	0.18	-0.74	15.3	2.83
U_21_A1	A	21	C15_061- C17_054	36.4	0.57	1.35	53.7	2.90
<u>CIM</u>								
U_03_B1	B	03	C19_056- C18_053- C19_115	63.2	-0.17	-1.10	25.8	3.80
U_06_B2	B	06	C19_080- C14_299	0.0	0.14	0.71	10.7	2.26
U_13_B3	B	13	C18_360- C03_422- C18_114	37.0	0.48	-0.12	09.6	2.60

(Table cont'd)

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
U_28_B4	B	28	C02_247- C11_078- C16_132	20.0	-0.24	-0.86	21.4	4.36
U_15_A1	A	15	C02_073- C12_233- C05_105- C04_107	31.8	-0.84	0.54	21.2	2.28
U_21_A2	A	21	C03_048- C05_051- C15_046	0.0	-0.78	1.10	52.9	3.58
U_21_A3	A	21	C15_046- C14_053	40.7	-0.23	1.66	56.9	2.39
U_21_A4	A	21	C15_061- C17_054	65.5	-0.48	1.37	52.4	2.47

<sup>†</sup> Additive effect

<sup>‡</sup> Dominance effect

<sup>§</sup> Percent explain variation

DNA markers significantly associated with U using simple regression and logistic regression are listed in Table 5.8.

Table 5.8 AFLP markers that were associated with putative QTL influencing Upland cotton fiber uniformity trait using simple and logistic regression at Baton Rouge (B).

Marker	Location	LG	F value $\chi^2$	Pr>F Pr> $\chi^2$	R <sup>2</sup> %
<u>Simple regression</u>					
C11_078	B	28	8.12	0.0055	8.91
C17_084	B	13	9.03	0.0034	9.03
C18_114	B	13	7.78	0.0064	7.88
C04_072	B	*	8.33	0.0050	9.12
<u>Logistic regression</u>					
C11_078	B	28	10.1	0.0015	9.15
C17_084	B	13	8.05	0.0046	6.99

\* Unlinked markers

### 5.3.5 QTL for Fiber Strength (S)

Three marker intervals (C04\_154-C14\_174, C02\_247-C11\_078, and C19\_056-C18\_053 with explained variations ranging from 14% to 31.4%, were detected using both IM and CIM (Table 5.9). In total, IM detected five marker intervals (four at Baton Rouge and one at Alexandria) while CIM was able to detect six marker intervals (four at Baton Rouge and two at Alexandria). The number of QTL identified in CIM was similar to the results obtained by Shappley et al. (1998) who identified six QTL for fiber strength. Fiber strength had QTL with an explained variation ranging from 8.4 % for S\_04\_B3 to 18.2% for S\_28\_B2 and from 14% for S\_23\_B4 to 35.5% for S\_09\_B1 in both IM and CIM, respectively. Zhang et al. (2003) identified three QTL for S with an explained variation ranging from 18.5% to 53.8%. QTL detected with both IM and CIM collectively explained about 49.6% and 72.2% of the phenotypic variation in the F<sub>2:3</sub> population, respectively. Four QTL (Yu et al., 1998), three QTL (Jiang et al., 1998), four QTL (Kohel et al., 2001), and three QTL (Ulloa and Meredith, 2000) for fiber strength were identified collectively, which explained 68.8, 30.9, 35, 46.5% of the phenotypic variance, respectively. The eight different QTL had both positive and negative additive effects that ranged from +1.1 to -0.35.

DNA markers significantly associated with S using simple regression and logistic regression are listed in Table 5.10.

Table 5.9 Putative QTL and their interval position influencing Upland cotton fiber strength trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
<b>IM</b>								
S_23_B1	B	23	C04_154- C14_174	8.00	0.59	-1.51	16.7	2.69
S_28_B2	B	28	C02_247- C11_078	180	0.64	-1.18	18.2	3.54
S_04_B3	B	04	C20_175- C12_258	115	0.59	-0.91	12.5	2.80
S_01_B4	B	01	C15_166- C18_135	129	0.53	-0.97	8.40	2.22
S_03_A1	A	03	C19_056- C18_053	41.3	0.98	0.32	14.7	2.21
<b>CIM</b>								
S_09_B1	B	09	C14_066- C04_119	14.0	0.71	-2.14	35.5	2.37
S_13_B2	B	13	C18_114- C20_307	78.1	1.10	-0.74	15.8	2.32
S_23_B3	B	23	C04_154- C14_174	8.00	-0.52	-1.51	14.0	2.30
S_28_B4	B	28	C02_247- C11_078- C16_132	18.0	-0.35	-2.09	31.4	5.01
S_03_A1	A	03	C14_048- C19_056- C18_053	38.0	-1.3	0.59	26.8	3.58
S_19_A2	A	19	C05_248- C14_191- C06_118	18.4	-1.24	1.43	24.7	2.32

<sup>†</sup> Additive effect

<sup>‡</sup> Dominance effect

<sup>§</sup> Percent explain variation

Table 5.10 AFLP markers that were associated with putative QTL influencing Upland cotton fiber strength trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A).

Marker	Location	LG	F value $\chi^2$	Pr>F Pr> $\chi^2$	R <sup>2</sup> %
<u>Simple regression</u>					
C11_078	B	28	8.06	0.0057	8.85
C02_115	B	*	7.38	0.0080	8.07
C19_056	A	03	7.15	0.0089	7.28
<u>Logistic regression</u>					
C04_154	B	23	6.77	0.0093	5.61
C08_230	B	*	8.26	0.0041	7.14
C11_078	B	28	12.6	0.0004	11.5
C11_252	B	04	7.15	0.0075	6.15
C11_334	B	04	6.72	0.0095	5.76
C12_258	B	04	6.92	0.0085	5.90
C15_166	B	01	8.19	0.0042	7.01
C20_175	B	04	9.34	0.0022	8.23
C14_048	A	03	7.06	0.0079	6.00
C19_056	A	03	9.62	0.0019	8.45
C20_046	A	14	6.99	0.0082	6.06

\* Unlinked markers

### 5.3.6 QTL for Fiber Micronaire (M)

Three (two in Baton Rouge and one in Alexandria) and six (one in Baton Rouge and five in Alexandria) marker intervals were found to be associated with M using both IM and CIM, respectively (Table 5.11). C02\_247-C11\_078 and C01\_104-C02\_073 were detected using both methods. These two intervals accounted for an explained variation ranging from 3.6% to 21.7% and from 13.6% to 37.3% in both IM and CIM, respectively. This was the same range accounted for by all marker intervals. CIM detected two intervals on linkage group 21. However, MIM dropped both. QTL detected in both IM and CIM collectively explained about 25.3% and 50.9% of the phenotypic variation in the F<sub>2:3</sub> population, respectively. Ulloa and

Meredith (2000) identified four QTL for M that collectively explained 56.3% of the total phenotypic variation. Shappley et al. (1998) identified as many as 15 QTL for M. The additive effect for all micronaire QTL was of minor to small importance ranging from 0.01 to 0.22 using IM and 0.032 to -0.34 in CIM.

Table 5.11 Putative QTL and their interval position influencing Upland cotton fiber micronaire trait. IM and CIM were used under MapMaker/QTL and QTL-CARTOGRAPHER, respectively, with a LOD threshold of 2.0 at Baton Rouge (B) and Alexandria (A).

QTL	Loc	LG	Interval	Position	a <sup>†</sup>	d <sup>‡</sup>	PEV <sup>§</sup>	LOD
<u>IM</u>								
M_04_B1	B	04	C12_083- C11_252	39.0	0.01	-0.01	3.60	2.58
M_28_B2	B	28	C02_247- C11_078	22.0	0.02	-0.01	3.60	2.92
M_15_A1	A	15	C01_104- C02_073	10.0	0.22	0.13	21.7	3.35
<u>CIM</u>								
M_28_B1	B	28	C02_247- C11_078- C16_132	16.0	-0.06	-0.34	37.3	6.58
M_05_A1	A	05	C04_123- C11_070- C08_338	60.2	-0.06	-0.19	12.8	2.41
M_14_A2	A	14	C20_046- C20_094- C12_197	50.6	0.13	0.12	9.47	2.32
M_15_A3	A	15	C01_104- C02_073- C12_233- C05_105	6.00	-0.18	0.03	13.6	2.47
M_21_A4	A	21	C05_051- C15_046- C14_053	27.8	-0.00	-0.33	28.6	2.34
M_21_A5	A	21	C14_053- C15_061- C17_054	63.5	0.23	-0.09	27.8	2.21

<sup>†</sup> Additive effect

<sup>‡</sup> Dominance effect

<sup>§</sup> Percent explain variation

DNA markers significantly associated with M using simple regression and logistic regression are listed in Table 5.12.

Table 5.12 AFLP markers that were associated with putative QTL influencing Upland cotton fiber micronaire trait using simple and logistic regression at Baton Rouge (B) and Alexandria (A).

Marker	Location	LG	F value $\chi^2$	Pr>F Pr> $\chi^2$	R <sup>2</sup> %
<u>Simple regression</u>					
C11_078	B	28	24.2	0.0001	22.6
C12_083	B	04	9.45	0.0029	10.2
C02_075	A	20	10.7	0.0015	11.3
C05_105	A	20	8.60	0.0043	9.28
<u>Logistic regression</u>					
C02_197	B	04	6.84	0.0089	5.74
C02_247	B	28	9.79	0.0018	8.40
C11_252	B	04	8.80	0.0030	7.56
C12_083	B	04	10.1	0.0015	8.80
C15_046	A	21	6.77	0.0093	5.42

### 5.3.7 QTL X Environment Interaction.

Although there are no statistical methods available to test the QTL X environment interaction in MapMaker/QTL, IM in QTL-CARTOGRAPHER was used to analyze the QTL X environment interaction using the impeded joint analysis method (module JZmapqtl). Nine different QTL were found to had significant interaction effects at a LOD threshold of 2 (Table 5.13). Strength and micronaire have seven QTL that interacted significantly between the two locations. Uniformity had one QTL with significant interaction (U\_13\_ B1 at 3.15 LOD). Among a total of 161 QTL detected for 16 measured traits (plant productivity, physiological and fiber quality traits) in a study conducted by Saranga et al. (1998), 59 (37%) had significant differences in their

effects between two different environments (well-watered and water-limited environments).

Table 5.13 QTL X environment interaction LOD using Module Jzmapqtl of QTL-CARTOGRAPHER in Upland cotton.

QTL	LOD		
	Baton Rouge	Alexandria	Joint LOD
E_18	2.34	2.27	2.01
L_13	2.98	2.65	0.00
L_18_A1	0.38	2.47	0.88
L_28_A2	0.22	2.47	1.58
L_02_A3	0.59	2.26	0.72
U_13_B1	2.50	0.36	3.15
U_28_B2	2.83	0.36	1.50
U_21_A1	0.31	2.90	1.84
S_23_B1	2.69	-	1.97
S_28_B2	3.54	0.53	2.46
S_04_B3	2.80	0.38	3.17
S_01_B4	2.22	-	2.67
S_03_A1	0.52	2.21	3.09
M_04_B1	2.58	0.14	2.83
M_28_B2	2.92	1.59	2.67
M_15_A1	3.35	0.71	3.27

In general, CIM performs the analysis in the same way as IM does. In CIM, the variance from other QTL is accounted for by including partial regression coefficients from markers in other regions of the genome that reduce noise and increase detection power. Using simulation, Zeng (1994) showed that CIM had higher resolution and detection power than IM. The power of a QTL-detection experiment, defined as the probability of detecting a QTL at a given level of statistical significance, depends on the strength of the QTL and the number of progeny in the population (Manly and Olson, 1999). The marker interval from Baton Rouge, C02\_247-C11\_078 interval, was detected to be significant for three fiber traits (uniformity, micronaire, and strength). This explains the moderate observed correlation between uniformity and strength ( $r = 0.49\%$ ;  $p < 0.0001$ ) and between

uniformity and micronaire ( $r = 0.12\%$ ;  $p < 0.11887$ ). In previous studies, it has been shown that QTL for related traits were frequently detected in the same interval. One QTL was detected in the marker interval C 15\_061-C17\_054 for E at both Baton Rouge and Alexandria, E\_21\_B1 with 54.7% explained variation and E\_21\_A1 with 47% explained variation, indicating a major and stable QTL. A major QTL can overshadow the effects of minor independently segregating QTL by increasing the total phenotypic variation, and thus genes with lesser effects might fall below the threshold for detection (Zhang et al., 2003).

In this study, and based on the IM method of QTL detection and LOD threshold of 2, the total number of QTL detected (number of marker intervals that were found to be significantly associated with the five fiber quality traits) was 10 at Baton Rouge and eight at Alexandria (Table 5.14 and Figure 5.2). Results using CIM were similar when compared with the results from IM. The total number of QTL detected were 13 at Baton Rouge and 16 at Alexandria. A range of small to moderately high accumulative proportions of the trait phenotypic variance (18.7 to 69%) was common in our study and supported a model for quantitative inheritance for the five fiber quality traits studied (Lande and Thompson, 1990; Ulloa and Meredith, 2000)

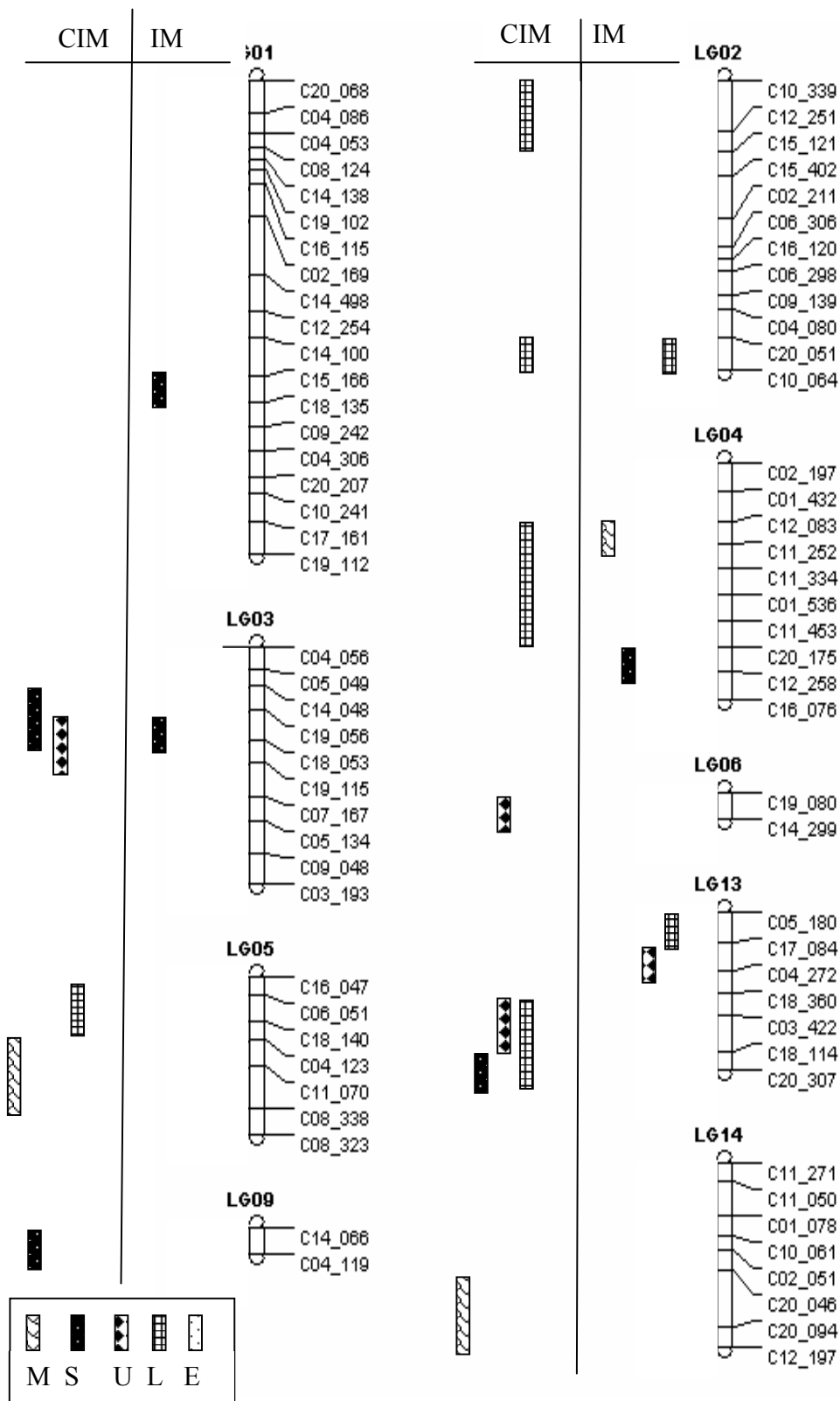


Figure 5.2 A comparison of QTL positions for Upland cotton micronaire (M), strength (S), uniformity (U), length (L), and elongation (L) using composite interval mapping (CIM) and interval mapping (IM).

Fig. cont'd

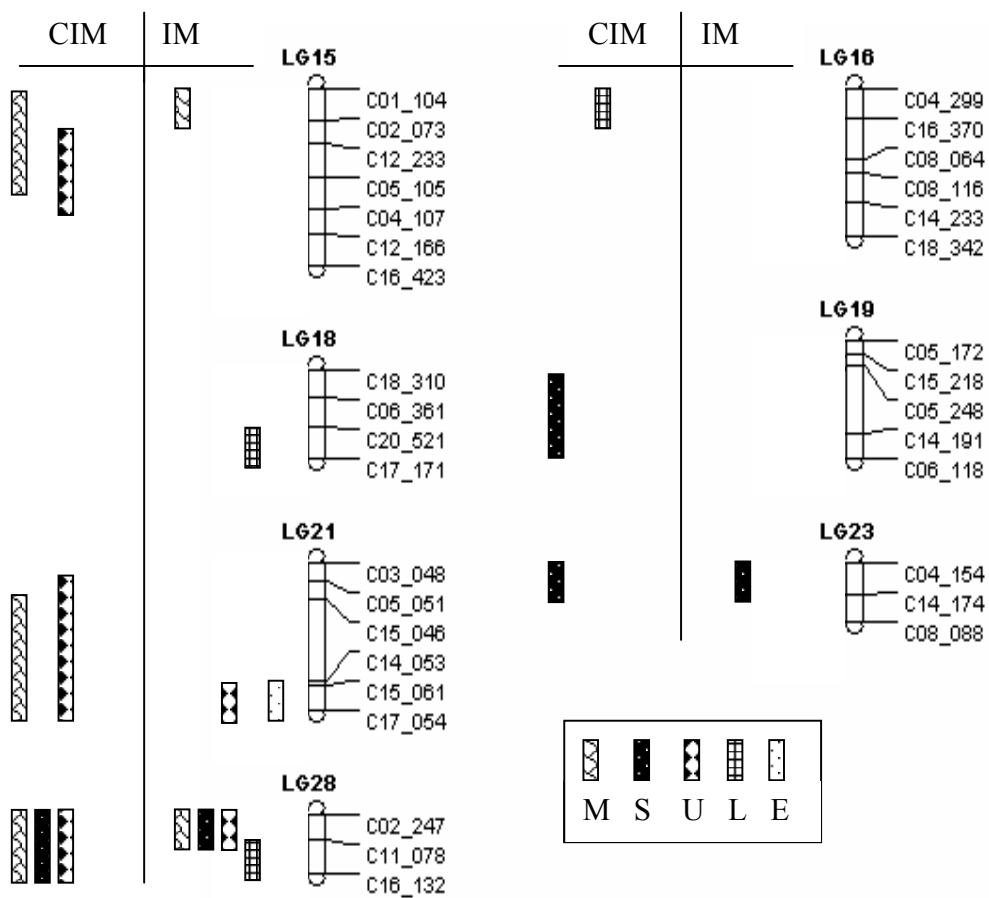


Table 5.14 The QTL summary for Upland cotton fiber quality traits.

Trait	Number of QTL detected in						Accumulative percent explained variation *
	IM		CIM		Both method		
	B	A	B	A	B	A	
E	1	1	1	1	1	1	47.0-54.7
L	1	4	3	4	0	1	18.7
U	2	1	4	4	1	1	69.0
S	4	1	4	2	2	1	49.6
M	2	1	1	5	1	1	25.3
Total	10	8	13	16	5	5	NA

\* Using QTL identified by both IM and CIM methods.

The identification of the QTL and marker-assisted selection should become more feasible as more molecular markers are developed and the map is supplemented with finely scaled increments. However, the putative locations of the QTL do not necessarily represent physical distances. Thus, a physical map of the linkage groups is very much needed and would be of great value in cloning selected QTL in cotton (Shappley et al., 1998).

#### 5.4 References

- Basten, C. J., B. S. Weir, and Z-B. Zeng. 1997. QTL Cartographer: A Reference manual and Tutorial for QTL Mapping. Department of Statistics, North Carolina State University, Raleigh, NC.
- Cao, G., J. Zhu, C. He, Y. Gao, J. Yan, and P. Wu. 2001. Impact of epistasis and QTL×environment interaction on the developmental behavior of plant height in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 103:153-160.
- Churchill, R. W., and G. A. Doerge. 1994. Empirical threshold values for quantitative trait mapping. *Genetics.* 138: 963-971.
- Flores-Berrios, E. L. Gentzbittel, L. Mokrani, G. Alibert, and A. Sarrafi. 2000. Genetic control of early events in protoplast division and regeneration pathways in sunflower. *Theor. Appl. Genet.* 101: 606-612.
- Fryxell, P. A. 1979. *The Natural History of the Cotton Tribe.* Texas A&M University Press, Collage Station, Texas.
- Haldane, J. B. S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. of Genet.* 8: 299-309.
- Jansen, R. C. 1993. Interval mapping of multiple quantitative trait loci. *Genetics.* 135: 205-211.
- Jiang, C., and Z-B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics.* 140: 111-1127.
- Jiang, C-X., R. J. wright, K. M. El-Zik, and A. Paterson. 1998. Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). *Proc. Natl. Acad. Sci.* 95: 4419-4424.
- Johnson, W. C., L. E. Jackson, O. Ochoa, R. Van Wijk, J. Peleman, D. A. St.Clair, and R. W. Michelmore. 2000. Lettuce, a shallow-rooted crop, and *Lactuca*

- serriola, its wild progenitor, differ at QTL determining root architecture and deep soil water exploitation. *Theor. Appl. Genet.* 101: 1066-1073.
- Kohel, R. J., J. Yu. Y-H. Park, and G. R. Lazo. 2001. Molecular mapping and characterization of trait controlling fiber quality in cotton. *Euphytica*. 121: 163-172.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*. 124: 743-756.
- Lander, E. S., and D. Botstein, 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. 121: 185-199.
- Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural population. *Genomics*. 1: 174-181.
- Lu, H. J., and G. O. Myers. 2002. Genetic Relationships and Discrimination of Ten Influential Upland Cotton Varieties using RAPD Markers. *Theor. Appl. Genet.* 105: 325-331.
- Manly, K. F. and J. M. Olson. 1999. Overview of QTL mapping software and introduction to MAP Manager QT. *Mammalian Genome*. 10: 327-334.
- Marques, C. M., J. Vasquez-Kool, V. J. Carocha, J.G. Ferreira, D. M. O'Malley, B-H. Liu, and R. Sederoff. 1999. Genetic dissection of vegetative propagation traits in *Eucalyptus tereticornis* and *E. globules*. *Theor. Appl. Genet.* 99: 936-946.
- Murdock, S. W., D. S. Murray, and J. W. Moore. 2001. Weed control and net returns with transgenic cotton using DSS and human recommendations. *Proceedings of the Annual Meeting- Southern Weed Science Society*. 54: 31-31.
- Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch, S. E. Lincoln, E. S. Lander, and S. D. Tanksley. 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics*. 127: 181-197.
- Paterson, A., E. Lander, S. Lincoln, J. Hewitt, S. Peterson, and S. Tanksley. 1988. Resolution of quantitative traits into Mendelian factors using a complete RFLP linkage map. *Nature*. 335: 721-726.
- Perlak, F. J., M. Oppenhuizen, K. Gustafson, R. Voth, S. Sivasupramaniam, D. Heering, B. Carey, R. A. Ihrig, and J. K. Roberts. 2001. Development and commercial use of Bollgard cotton in the USA early promises versus today's reality. *Plant J.* 27: 489-501.

- Saranga, Y., M. Menz, C-X. Jiang, R. L. Wright, D. Yakir, and A. H. Paterson. 2001. Genomic dissection of genotype x environment interactions conferring adaptation of cotton to arid conditions. [www.genome.org](http://www.genome.org).
- SAS Institute. 2003. Version 9. Cary, N. C., USA.
- Schork, N., M. Boehnke and J. Terwilliger, 1993 Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. *Am. J. Hum. Genet.* 53: 1127-1136.
- Shappley, Z. W., J. N. Jenkins, J. Zhu, and J. C. McCarty, Jr. 1998. Quantitative trait loci associated with agronomic and fiber traits of Upland cotton. *The Journal of Cotton Sci.* 4: 153-163.
- Smith, W. C. 1999. Production Statistics. In Smith, W. C. and J. T. Cothorn (eds) . *Cotton: Origin, History, Technology, and Production*. John Wiley and Sons, Inc.
- Tanksley, S. D. 1993. Mapping polygenes. *Annu. Rev. Genet.* 27: 205-233.
- Ulloa, M., and W. R. Meredith Jr. 2000. Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. *The Journal of Cotton Science.* 4: 161-170.
- Wang, D., R. Karle, and A. F. Iezzoni. 2000. QTL analysis of flower and fruit traits in sour cherry. *Theor. Appl. Genet.* 100: 535-544.
- Wang, D. L., J. Zhu, Z. K. Li, and A. H. Paterson. 1999. Mapping QTLs with epistatic effects and QTL X environment interactions by mixed linear model approaches. *Theor. Appl. Genet.* 99: 1255-1264.
- Wang, D., P. R. Arelli, R. C. Shoemaker, and B.W. Diers. 2001. Loci underlying resistance to Race 3 of soybean cyst nematode in *Glycine soja* plant. *Theor. Appl. Genet.* 103: 561-566.
- Yadav, R. S., C. T. Hash, F. R. Bidinger, G. P. Cavan, and C. J. Howarth. 2002. Quantitative trait loci associated with traits determining grain and stover yield in pearl millet under terminal drought stress conditions. *Theor. Appl. Genet.* 104: 67-83.
- Yu, Z. H., Y. H. Park, G. R. Lazo and R. J. Kohel. 1998. Molecular mapping of the cotton genome: QTL analysis of fiber quality characteristics. *Proc. of Plant Animal Genome VI*, Jan 18-22. 1998. San Diego California.
- Zeng, Z. B. 1993. Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA.* 90: 10972-10976.

Zeng, Z-B. 1994. Precision mapping of quantitative trait loci. *Genetics*. 136: 1457-1468.

Zeng, Z-B., and B. S. Weir. 1996. Statistical methods for mapping Quantitative Trait Loci. *Acta Agronomica Sinica*. 22: 535-549.

Zhang, T., Y. Yuan, J. Yu, W. Guo, R. J. Kohel. 2003. Molecular tagging of a major QTL for fiber strength in Upland cotton and its marker-assisted selection. *Theor. Appl. Genet.* 106: 262-268.

## **CHAPTER 6 MULTIPLE IMPUTATION FOR MISSING DATA IN MOLECULAR PLANT BREEDING STUDIES**

### **6.1 Introduction**

Molecular plant breeding, especially the quantitative trait loci (QTL) mapping procedure allows for the discovery of important putative genes for plant improvement, which may be needed to meet the demands of today and the future. Commonly used analysis methods include both univariate and multivariate methods that require complete matrices. QTL mapping data are usually in the form of large matrices where the plants under study (rows) on which markers and traits (columns) have been scored.

Until recently, incomplete data were handled primarily either by ignoring subjects with missing information or by substituting plausible values, such as means or regression predictions. These approaches may do more harm than good, producing answers that are biased, inefficient, or unreliable (Shafer and Graham, 2002). Unfortunately, most commonly used software relies on such simple procedures. For example, most SAS statistical procedures exclude subjects with any missing values (SAS\_V9 On line Doc.). This means that in the end, we may not have enough data to perform the analysis. Another strategy for handling missing data is multiple imputation (MI) (Rubin, 1987), which relies on different methods of imputation, such as propensity score, regression, logistic regression, discriminant function, markov chain monte carlo (MCMC), full-data imputation, and MCMC monotone-data imputation. The method of specification depends on the missingness pattern and on the type of imputed variable. MI imputes the values multiple times. The result is multiple data sets with identical values for all of the non-missing values

and slightly different values for the imputed values in each data set. The statistical analysis of interest, such as ANOVA, discriminant analysis, or logistic regression is performed separately on each data set, and the results are then combined. MI reflects the uncertainty associated with the missing observations, providing unbiased estimates for the parameters of interest and their variances (Rubin, 1996)

In this study, our objective is to give a brief overview of missing data handling concepts and several popular methods for handling incomplete data. We then explain how these methods apply to the problem of imputing reasonable values for incomplete QTL mapping data.

### 6.1.1 Patterns of Missing Data

Consider Table 6.1b, in which missing values occur for markers  $M_1, M_2, \dots, M_n$  ordered in such a way that if  $M_j$  is missing for a plant, then  $M_{j+1}, M_{j+2}, \dots, M_n$  are missing as well; this is called a monotone pattern in which ordering of markers is important. Table 6.1a shows an arbitrary pattern in which any set of markers may be missing for any plant. In this pattern, ordering of markers is not important.

Table 6.1 Pattern of missing data. A: arbitrary pattern and B: monotone pattern. Here, an "X" means that the variable is observed and a "." means that the variable is missing.

<u>A) Arbitrary</u>					<u>B) Monotone</u>				
Plant	M1	M2	--	Mn	Plant	M1	M2	--	Mn
P1	X	X	--	X	P1	X	X	--	X
P2	.	X	--	.	P2	X	.	--	.
P3	X	.	--	X	P3	.	.	--	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
P4	.	X	--	X	P4	X	X	--	X

### 6.1.2 Types of Missing Data

Define indicator variable  $R$  that identify what is observed and what is missing. In modern missing data procedures,  $R$  is regarded as a probabilistic phenomenon (Rubin, 1976) and its probability distribution is the distribution of missingness or the probability of missingness.

#### 6.1.2.1 Missing at Random (MAR)

Let  $Y_{\text{com}}$ ,  $Y_{\text{obs}}$ , and  $Y_{\text{mis}}$  denote the complete, observed, and missing data, respectively. When the distribution of missingness does not depend on  $Y_{\text{mis}}$

$$P(R/Y_{\text{com}}) = P(R/Y_{\text{obs}})$$

The missing data are said to be missing at random or ignorable nonresponse.

#### 6.1.2.2 Missing not at Random (MNAR)

When the above equation is violated and the distribution depends on  $Y_{\text{mis}}$ , the missing data are said to be MNAR or nonignorable missing data.

$$P(R/Y_{\text{com}}) = P(R/Y_{\text{mis}})$$

#### 6.1.2.3 Missing Completely at Random (MCAR)

If the distribution of missingness does not depend on either  $Y_{\text{obs}}$  or  $Y_{\text{mis}}$ , then the missing data are said to be MCAR.

$$P(R/Y_{\text{com}}) = P(R)$$

For illustration, let  $Y$  have  $T$  and  $K$  variables be drawn from a standard normal distribution, each with means of about 0 and standard deviations of about 1.0. Then we force half of  $T$  values to be missing according to the three different missing types:

1. MAR:  $T$  missing if  $K < 0$ .
2. MNAR:  $T$  missing if  $T < 0$ .
3. MCAR:  $T$  missing with probability 0.5, independent of  $K$ .

Note that in situation 1, we forced values in the T variable to be missing only if values of the observed K variable are negative. However, in situation 2, we randomly discarded values from the T variable without any dependence upon T or K. In the third situation we depend on the T values themselves to decide whether or not to discard.

### 6.1.3 Methods for Handling Missing Data

#### 6.1.3.1 Case Deletion

Case deletion, also known as listwise deletion (LD) and complete case analysis, is performed by discarding subjects whose information is incomplete. It is a default method in many statistical programs (Schafer and Graham, 2002). Most SAS statistical procedures, for example, use complete case analysis to handle missing data. Some SAS procedures use different sets of sample units for different parameters; this is called Available Case (AC) analysis. For example, the PROC CORR procedure estimates a correlation by using all subjects with no missing value for this pair of variables. This makes better use of the available data than using only the complete subjects. However, it is difficult to compute standard errors or other measures of uncertainty since parameters are estimated from different sets of subjects (Shafer and Graham, 2002).

While analyzing only complete subjects has its simplicity, it is only valid under MCAR and the information contained in the incomplete subjects is lost. This approach ignores possible systematic differences between the complete subjects and the incomplete subjects. Therefore, standard errors will generally be larger in the reduced sample because less information is used and consequently produce biased estimates if the reduced sample is not a random sub-sample of the original data.

Also, the resulting inference may not be applicable to the population, especially with a small number of complete subjects (V9 online SAS doc.)

#### 6.1.3.2 Single Imputation

Another strategy for handling missing data is single imputation, which replaces the missing data with plausible values and proceeds with the desired analysis rather than discarding the subject entirely. Here we briefly list and review some popular single imputation methods that have been extensively discussed by different authors: Schafer and Graham (2002), Little and Rubin (1987), and Rubin (1987).

Unconditional Mean Estimation: In this popular type of estimation, each missing value can be imputed with the mean of non-missing values. Although popular, this procedure underestimates the standard deviation and standard error, and it also distorts covariances and correlations between variables.

Unconditional Distribution Estimation: It is generally more desirable to preserve a variable's distribution than preserve its mean. One popular class of example is hot deck imputation which fills in missing data with values from the observed data. This method still distorts correlations and other measures of association.

Conditional Mean Estimation: In this method, we first estimate a regression model in which the dependent variable has missing values for some observations, then the estimated regression coefficients are used to predict missing values of that variable. This method is not recommended for analyses of covariances or correlations, because it overstates the strength of the relationship between the dependent and the independent variables. Also, if there is no association between

variables, the method reduces to the unconditional mean estimation (Schafer and Graham, 2002).

Conditional Distribution Estimation: Distortion of covariance, the main disadvantage of conditional mean estimation strategy, can be eliminated if each missing value is replaced not by a regression prediction but by a random draw from the conditional or prediction distribution of the dependent variable, given the independent variable. In other words, the predicted value comes from a regression plus a random residual value.

In general, single imputation treats missing values as if they were known in the complete data analysis, which does not reflect the uncertainty about the prediction of the unknown missing value (Rubin, 1987; Rubin and Schenker, 1986).

Maximum Likelihood Estimation: The maximum likelihood estimate of a parameter is the value of the parameter that is most likely to have resulted in the observed data (Dempster et al., 1977). This technique, called estimation maximization (EM) algorithm, consists of an iterative calculation involving two steps: First, a prediction step that predicts the contribution of any missing observation to the complete data sufficient statistics. Second, an estimation step that uses the predicted sufficient statistics to compute a revised estimate of the parameters. The iteration between the two steps continues until the parameter estimates remain essentially unchanged.

Although this method gives unbiased parameter estimates and standard errors, it is limited to linear models. The ML estimation algorithm is available in SPSS (Version 10.0), EMCOV (Graham and Hofer, 1991), NORM, SAS (Yuan, 2000), Amelia (King et. al., 2001), and S-Plus (Schimert et. al., 2001).

### 6.1.3.3 Multiple Imputation (MI)

MI is the most attractive method for general purpose handling of missing data in multivariate analysis. The basic idea, first proposed by Rubin (1977), is to fill in estimates for the missing data. However, to capture the uncertainty in those estimates, MI imputes the values multiple times. The result is multiple data sets with identical values for all of the non missing values and slightly different values for the imputed values in each data set. The statistical analysis of interest, such as ANOVA, discriminant analysis, or logistic regression, is performed separately on each data set, and the results are then combined (Little and Rubin, 1989). MI reflects the uncertainty associated with the missing observations, providing unbiased estimates for the parameters of interest and their variances. Also, MI can be used with any kind of data or analysis without the need for specialized software.

#### 6.1.3.3.1 The MI Procedure

MI assumes that the missing data are missing at random (MAR) and therefore has a limitation in not being able to handle data that is MNAR. To begin with, the method used to generate the imputed values must be correctly specified. The method specified depends on the pattern of missingness in the data and the type of the imputed variable, as summarized in the following Table (SAS\_V9 On line Doc.).

Propensity score method generates a propensity score for each variable missing value to indicate the probability of being missing. The observations are then grouped based on these scores and an approximate Bayesian bootstrap imputation is applied to each group (Lavori et al., 1995; Rosenbaum and Rubin, 1983; Rubin, 1987).

Table 6.2 Imputation Methods in SAS PROC MI (SAS V9 On line Doc.)

Pattern of missingness	Type of imputed variable	Recommended methods	Pattern of missingness	Type of imputed variable
Monotone	Continuous	Regression Predicted mean matching Propensity score	Monotone	Continuous
Monotone	Classification (ordinal)	Logistic regression	Monotone	Classification (ordinal)
Monotone	Classification (nominal)	Discriminant function method	Monotone	Classification (nominal)
Monotone	Binary	Logistic regression Discriminant function method	Monotone	Binary
Arbitrary	Continuous	MCMC full-data imputation MCMC monotone-data imputation	Arbitrary	Continuous

The MCMC technique is applied to substitute missing observations with plausible pseudorandom samples from the conditional probability distribution of the missing data given the observed values. By repeated iteration steps, it simulates draws from the stationary distribution. Stationary distribution means the mean vector and the covariance matrix remain unchanged through the iterations. The goal is to have the iterations converge to their stationary distribution and then to simulate an approximately independent draw of the missing values (Schafer, 1997).

#### 6.1.3.3.2 MI Efficiency

MI estimation does not need a large number of repetitions for precise estimates. Rubin (1987) showed that the relative efficiency (RE) of using the finite  $m$  imputation estimator, rather than using an infinite number, is

$$RE = \left(1 + \frac{\lambda}{m}\right)^{-1}$$

where  $\lambda$  is the rate of missing information. (Table 6.3) (SAS\_V9 On line Doc.).

Table 6.3 Percent efficiency of MI estimation by number of imputation  $m$  and percentage of missing data  $\lambda$

$m$	$\lambda$				
	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

### 6.1.3.3.3 The MIANALYZE Procedure

The MIANALYZE procedure combines the results of the analyses of the MI multiple imputation and generates valid statistical inferences; it reads parameter estimates and associated standard errors that are computed by the standard statistical procedure for each imputed data set.

Rubin's (1987) method for a scalar (one-dimensional) parameter proceeds as follows: Letting  $Q$  represent a population quantity of interest and  $U$  its variance, then  $\hat{Q}$  and  $\sqrt{U}$ -hat denote the estimate of  $Q$  and the standard error that one would use if no data were missing. With  $m$  imputations, we have  $m$  equally plausible estimates:  $Q_1$ -hat,  $Q_2$ -hat, ...,  $Q_m$ -hat and their corresponding standard errors  $\sqrt{U_1}$ -hat,  $\sqrt{U_2}$ -hat, ...,  $\sqrt{U_m}$ -hat. The combined point estimate for  $Q$  from multiple imputation is the mean of the complete data estimates:

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i$$

The uncertainty in  $\bar{Q}$  has two components: The within-imputation variance,

$$\bar{U} = \frac{1}{m} \sum_{i=1}^m \hat{U}_i$$

and the between-imputation variance (B),

$$B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2$$

The total variance estimate associated with  $\bar{Q}$  is a modified sum of the two components.

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B$$

For confidence limits and tests, the statistic  $(Q - \bar{Q}) * T^{-1/2}$  is approximately distributed as t with  $v_m$  degrees of freedom (Rubin, 1977), where the degrees of freedom are given by

$$v_m = (m-1) \left[ 1 + \frac{\bar{U}}{(1+m^{-1})B} \right]^2$$

The degrees of freedom may vary from  $m-1$  to infinity depending on  $m$  and the relative increase in variance due to the nonresponse ratio,

$$r = \frac{(1+m^{-1})B}{\bar{U}}$$

Using these formulas, one can combine almost all known standard analyses of interest; however, SAS On-Line help provides ready to use codes for several types of analysis, such as regression, mixed model, generalized linear model, logistic regression, and correlation.

## **6.2 Materials and Methods**

### **6.2.1 Data Preparation**

Two hundred Amplified Fragment Length Polymorphism (AFLP) molecular markers were used to map and characterize quantitative trait loci (QTL) to determine Upland cotton agronomic and fiber quality traits. In addition, 138 F<sub>2,3</sub> bulked sampled rows from an intraspecific cross between Paymaster 54 and Pee Dee2165 were used.

In this study, the data were reduced into three markers (C12\_166, C06\_361, and C18\_114) and four traits (lint weight per boll (LY), lint percentage (LP), seedcotton weight per boll (BW), boll number per plant (B/P)).

A total of 19 subjects (plants) were deleted since they had missing values either in the markers or in the traits. The objective was to start with a complete matrix with no missing data. This matrix was used as a starting point and we then proceeded to create eight different data sets where we randomly removed 5%, 10%, 20%, and 40% of the data points creating both monotone and arbitrary missing patterns that represent MAR and MCAR, respectively. The monotone pattern of missingness was created by randomly selecting values from the data matrix, then the pattern was created by discarding values that lie next to the selected ones. To study the performance of MI methods for MNAR type of missingness, four data sets were created to represent MNAR with 5%, 10%, 20%, and 40% missing data. MNAR was created by discarding the extreme values from each variable.

### **6.2.2 MI Methods**

Using the MI Procedure (SAS V9), six different methods were applied. These included the propensity score and regression methods for monotone missing

patterns of a continuous variable, logistic regression and discriminant function for monotone missing patterns of a binary variable, and the MCMC full-data imputation and MCMC monotone-data imputation for arbitrary missing patterns of a continuous variable.

### 6.2.3 Data Analysis

For the complete data, correlation and logistic regression were used as the standard statistical analyses without the use of the MI procedure. However, six different PROC MI methods, representing the six different methods being used in this study, were written for each of the 5%, 10%, 20%, and 40% missing data sets. This was followed by performing the standard analysis (correlation or logistic regression) for each imputed data set (number of imputations ( $m$ ) was set to 5 in all analyses). Then the PROC MIANALYZE procedure was used to combine the results of the analyses of imputations and to generate valid statistical inferences.

## **6.3 Results and Discussion**

### 6.3.1 Complete Data Analysis

Table 6.4 shows a significant positive correlation between LY and BW ( $r = 0.66$ ;  $p < 0.0001$ ) and between LY and LP ( $r = 0.24$ ;  $p = 0.0097$ ). However, a significant negative correlation between BW and LP ( $r = -0.21$ ;  $p = 0.0203$ ) and between BW and B/P ( $r = 0.22$ ;  $p = 0.0182$ ). These results are consistent with the previous study of Lu and Myers (2002).

The boll number per plant trait maps close to C06\_361 and at least two putative QTL for seedcotton weight per plant map close to C06\_361 and C18\_114 (Table 6.5). In contrast to interval analysis, this approach, which analyzes each marker separately, does not allow us to determine the exact location of the putative

QTL. However, knowing the significant markers allows us to use them in MAS, which facilitates more efficient plant improvement programs.

Table 6.4. Complete data analysis showing Pearson correlation coefficients and corresponding P-values for Upland cotton lint weight per boll (LY), Lint Percentage (LP), seedcotton weight per boll (BW), boll number per plant (B/P).

	BW	LP	B/P <sup>§</sup>
LY <sup>†</sup>	0.66 <0.0001	0.24 0.0097	0.02 0.8718
BW <sup>‡</sup>		-21.3 0.0203	-0.22 0.0182
LP <sup>¶</sup>			0.04 0.6985

† Lint weight per boll    ‡ Seedcotton weight per plant  
<sup>§</sup>Boll number per plant    ¶ Lint percentage

Table 6.5. Complete Upland cotton data analysis showing logistic regression parameter estimates and their associated confidence limits and P-values.

Parameter	Estimates	Missingness percentage		
		C12_166	C06_361	C18_114
Intercept	Value:	-5.39	1.3637	7.5992
	Conf. Limits:	(-18,7.3)	(-9.93,12.7)	(-3.8,19)
	P value	0.405	0.813	0.192
LY	Value:	-1.69	-1.581	0.6288
	Conf. Limits:	(-4.53,1.14)	(-4.12,0.96)	(-1.8,3.1)
	P value:	0.2431	0.222	0.6181
BW	Value:	0.8934	1.6987	-1.942
	Conf. Limits:	(-0.88,2.67)	(0.035,3.36)	(-3.59,-0.29)
	P value:	0.32	0.0454	0.0211
LP	Value:	20.216	-11.39	-3.5612
	Conf. Limits:	(-8.9,49.4)	(-38.1,15.39)	(-29.8,22.74)
	P value	0.1742	0.4045	0.7907
B/P	Value:	-0.3619	0.1429	-0.1941
	Conf. Limits:	(-0.69,-0.04)	(-0.12,0.41)	(-0.46,0.078)
	P value:	0.0294	0.2931	0.1625

Because it is to the benefit of molecular plant breeders to know if missing data are affecting their ability to detect putative QTL, we decided to focus only on those results with significant P-values.

### 6.3.2 Propensity Score and Regression Methods

Both the propensity score and the regression methods were able to correctly estimate correlation coefficients, the estimates falling within the confidence limits of the reference analysis, even in the 40% missing data set (Table 6.6 and 6.7).

However, the regression method failed to give confidence intervals and P-values for two correlation coefficients for LY with LP and BW with LP because of setting zero variance between imputations. In contrast to the regression method, it is noticed (by tracing significant P-values) that the propensity score method resulted in an insignificant P-value most noticeably in the 40% missing data set (Table 6.7).

Table 6.6 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Regression Method.

Parameter	Estimates	Missingness percentage				
		0.0%	5%	10%	20%	40%
LY & BW	Correlation:	0.664	0.801	0.782	0.787	0.792
	Conf. Limits:	(52.4,80.4)	(62,98)	(60,96.8)	(56,99.6)	(57.8,100)
	P value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
LY & LP	Correlation:	0.236	0.252	0.243	0.234	0.235
	Conf. Limits:	(5.9,39)	(--,--)	(--,--)	(--,--)	(--,--)
	P value:	0.0097	.	.	.	.
LY & B/P	Correlation:	0.015	-0.021	-0.001	-0.062	-0.004
	Conf. Limits:	(-16.5,19)	(-21,16)	(-20,18)	(-28,15.5)	(-27,26)
	P value:	0.8718	0.8267	0.9366	0.5756	0.9749
BW & LP	Correlation:	-0.213	-0.220	-0.223	-0.180	-0.106
	Conf. Limits:	(-38,-3.4)	(--,--)	(--,--)	(--,--)	(--,--)
	P value:	0.0203	.	.	.	.
BW & B/P	Correlation:	-0.216	-0.214	-0.230	-0.283	-0.275
	Conf. Limits:	(-38,-3.8)	(-40,-2)	(-43,-3.6)	(-50,-6.5)	(-56,1.6)
	P value	0.0182	0.0267	0.0224	0.0108	0.064
LP & B/P	Correlation:	0.036	-0.003	-0.013	-0.026	-0.014
	Conf. Limits:	(-14.5,21)	(-22,16)	(-21,18.4)	(-24,19)	(-29,26)
	P value:	0.6985	0.7442	0.8961	0.8142	0.9174

Table 6.7 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Propensity Score.

Parameter	Estimate	Missingness percentage				
		0.0%	5%	10%	20%	40%
LY & BW	Correlation:	0.664	0.807	0.787	0.762	0.711
	Conf. Limits:	(52.4,80.4)	(62.4,99)	(60.1,97)	(55.9,96)	(47.1,95)
	P value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
LY & LP	Correlation:	0.236	0.243	0.238	0.298	0.126
	Conf. Limits:	(5.9,39)	(5.1,43.4)	(4.5,43.1)	(9.5,50)	(-23,48.5)
	P value:	0.0097	0.030	0.0154	0.0042	0.456
LY & B/P	Correlation:	0.015	-0.062	0.024	-0.029	-0.013
	Conf. Limits:	(-16.5,19)	(-26,14)	(-17,21.9)	(-30,24.7)	(-32,30)
	P value:	0.8718	0.54	0.81	0.826	0.932
BW & LP	Correlation:	-0.213	-0.194	-0.199	-0.127	-0.216
	Conf. Limits:	(-38,-3.4)	(38,-1.1)	(-39,-0.6)	(-38,13)	(-53,10.4)
	P value:	0.0203	0.03	0.043	0.3207	0.172
BW & B/P	Correlation:	-0.216	-0.205	-0.177	-0.194	-0.189
	Conf. Limits:	(-38,-3.8)	(-41,-0.5)	(-36,1.3)	(-45,6.16)	(-50,12.6)
	P value	0.0182	0.045	0.684	0.131	0.22
LP & B/P	Correlation:	0.036	-0.032	-0.016	0.012	0.044
	Conf. Limits:	(-14.5,21)	(-23,16)	(-20,17.4)	(-18,20.8)	(-16.3,25)
	P value:	0.6985	0.7457	0.8722	0.9	0.6744

### 6.3.3 MCMC Monotone-Data and MCMC Full-Data Imputation Methods

Tables 6.8 and 6.9 showed that the full imputation method was superior to the monotone imputation method since the monotone method failed to correctly estimate correlation coefficients in the 40% missing data, and it also failed to estimate the confidence interval and P-value for LP with B/P parameter. However, both incorrectly estimated the correlation coefficients and their P-values for the correlation estimate between BW and LP.

### 6.3.4 Logistic Regression and Discriminant Function Methods

Similar results were obtained from both methods. Correct estimates were calculated in most cases including 40% missing data (Table 6.10 and 6.11). However, the 40% data analysis clearly illustrates the problem of false, insignificant P-values especially for an original P-value higher than 2.5% as noticed in C06\_361 and BW.

Table 6.8 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Monotone-Data Imputation.

Parameter	Estimate	Missingness percentage				
		0.0%	5%	10%	20%	40%
LY & BW	Correlation:	0.664	0.791	0.760	0.503	0.161
	Conf. Limits:	(52.4,80.4)	(59.7,98.3)	(57.4,94.6)	(20.7,79.9)	(-20,52)
	P value	<0.0001	<0.0001	<0.0001	0.0027	0.3503
LY & LP	Correlation:	0.236	0.121	0.166	0.141	-0.067
	Conf. Limits:	(5.9,39)	(-12,36)	(--,--)	(-5.2,33.4)	(-30,16)
	P value:	0.0097	0.316	.	0.1522	0.5628
LY & B/P	Correlation:	0.015	0.009	0.090	0.156	0.203
	Conf. Limits:	(-16.5,19)	(-20,21.5)	(--,--)	(-7.3,38.5)	(-6.2,47)
	P value:	0.8718	0.935	.	0.179	0.1302
BW & LP	Correlation:	-0.213	0.035	0.096	0.002	-0.091
	Conf. Limits:	(-38,-3.4)	(-15,22)	(-9.7,29)	(-23,23.1)	(-34,16)
	P value:	0.0203	0.709	0.329	0.989	0.4668
BW & B/P	Correlation:	-0.216	-0.160	-0.196	-0.099	-0.013
	Conf. Limits:	(-38,-3.8)	(-34,2.89)	(-38.5,0.6)	(-32,12.5)	(27.5,25)
	P value	0.0182	0.0968	0.431	0.382	0.9242
LP & B/P	Correlation:	0.036	0.005	0.060	0.057	0.045
	Conf. Limits:	(-14.5,21)	(--,--)	(--,--)	(--,--)	(--,--)
	P value:	0.6985	.	.	.	.

Table 6.9 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Full-Data Imputation.

Parameter	Estimate	Missingness percentage				
		0.0%	5%	10%	20%	40%
LY & BW	Correlation:	0.664	0.775	0.776	0.710	0.566
	Conf. Limits:	(52.4,80.4)	(59,96)	(58.6,96)	(44.5,98)	(4.67,1.0)
	P value	<0.0001	<0.0001	<0.0001	0.0002	0.037
LY & LP	Correlation:	0.236	0.072	0.191	0.167	-0.071
	Conf. Limits:	(5.9,39)	(--,--)	(--,--)	(-2.7,36.1)	(-64,50)
	P value:	0.0097	.	.	0.0911	0.6937
LY & B/P	Correlation:	0.015	0.043	0.088	0.105	0.293
	Conf. Limits:	(-16.5,19)	(-14.4,23)	(-23.7,41)	(-17.1,38)	(-99,99)
	P value:	0.8718	0.652	0.5168	0.4011	0.531
BW & LP	Correlation:	-0.213	0.031	0.106	0.023	-0.163
	Conf. Limits:	(-38,-3.4)	(-15,23.3)	(-7.6,29)	(-27,31)	(-65,33)
	P value:	0.0203	0.736	0.255	0.852	0.36
BW & B/P	Correlation:	-0.216	-0.190	-0.168	-0.167	-0.056
	Conf. Limits:	(-38,-3.8)	(-38,-0.5)	(-63.7,30)	(-53,19.4)	(-32,21)
	P value	0.0182	0.435	0.3321	0.27	0.6448
LP & B/P	Correlation:	0.036	-0.009	0.033	0.058	-0.023
	Conf. Limits:	(-14.5,21)	(-19,17.4)	(-20.3,26)	(-16.8,28)	(-21,17)
	P value:	0.6985	0.9254	0.769	0.597	0.8138

Table 6.10 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Discriminant Function.

Parameter	Estimate	Missingness percentage				
		0.0%	5%	10%	20%	40%
C12_166	Value:	-0.3619	-0.33	-0.401	-0.43	-0.328
&	Conf. Limits:	(-0.69,-0.04)	(-0.7,0)	(-0.7,0)	(-0.8,-0.1)	(-0.8,0.9)
B/P	P value	0.0294	0.05	0.0267	0.01	0.1234
C06_361	Value:	1.6987	1.55	1.61	1.26	2.476
&	Conf. Limits:	(0.04,3.36)	(-0.2,3)	(-0.2,3)	(-0.3,2.9)	(-0.5,5.4)
BW	P value:	0.0454	0.076	0.178	0.12	0.0956
C18_114	Value:	-1.942	-1.633	-1.69	-2.2	-0.867
&	Conf. Limits:	(-3.6,-0.3)	(-3,-0)	(-28,0)	(-4.8,0.4)	(-3.3,1.6)
BW	P value:	0.0211	0.047	0.045	0.088	0.46

Table 6.11 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using Logistic Regression.

Parameter	Estimate	Missingness percentage				
		0.0%	5%	10%	20%	40%
C12_166	Value:	-0.3619	-0.342	-0.397	-0.455	-0.394
&	Conf. Limits:	(-0.7,-0.04)	(-0.7,0)	(-0.7,-1)	(-1,-0.1)	(-0.8,0.1)
B/P	P value	0.0294	0.0431	0.0205	0.0123	0.0854
C06_361	Value:	1.6987	1.55	1.668	1.604	3.477
&	Conf. Limits:	(0.04,3.4)	(-0.19,3.3)	(0,3.37)	(-0.3,4)	(-0.2,7.2)
BW	P value:	0.0454	0.0802	0.054	0.0923	0.06
C18_114	Value:	-1.942	-1.5233	-2.219	-2.72	-0.997
&	Conf. Limits:	(-3.6,-0.3)	(-3.7,0.63)	(-4.1,-0.3)	(-5,-0.3)	(-2.7,0.7)
BW	P value:	0.0211	0.159	0.02	0.03	0.25

### 6.3.5 Missing not at Random (MNAR)

Tables 6.12 and 6.13 showed that the full imputation method performed better than the monotone imputation method since the monotone method failed to correctly estimate correlation coefficients in the 20% and 40% missing data. The full imputation method failed to correctly estimate correlation coefficients only in the 40% missing data. However, both incorrectly estimated the correlation coefficients and their P-values for LY with LP.

Table 6.12 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Full-Data Imputation for MNAR type of missingness.

Parameter	Estimate	Missingness percentage				
		0.0%	5%	10%	20%	40%
LY & BW	Correlation:	0.664	0.664	0.581	0.338	0.309
	Conf. Limits:	(52.4,80.4)	(45,87)	(39,77)	(14,54)	(11.8,50)
	P value	<0.0001	<0.0001	<0.0001	0.0008	0.0015
LY & LP	Correlation:	0.236	0.024	-0.081	-0.015	0.011
	Conf. Limits:	(5.9,39)	(-16,20.7)	(-26,10)	(--,--)	(--,--)
	P value:	0.0097	0.7959	0.3861	.	.
BW & LP	Correlation:	-0.213	-0.149	-0.106	-0.054	0.043
	Conf. Limits:	(-38,-3.4)	(-33,3.3)	(-29,7.7)	(-32.5,22)	(-15,24)
	P value:	0.0203	0.1081	0.2546	0.6786	0.6647
BW & B/P	Correlation:	-0.216	-0.157	-0.150	-0.052	0.010
	Conf. Limits:	(-38,-3.8)	(-34,28)	(-39,8.9)	(-28,17)	(-29,31)
	P value	0.0182	0.0967	0.2106	0.6447	0.9439

Table 6.13 The MIANALYZE combined estimates, associated confidence limits and P-values for the MI procedure using MCMC Monotone-Data Imputation for MNAR type of missingness.

Parameter	Estimate	Missingness percentage				
		0.0%	5%	10%	20%	40%
LY & BW	Correlation:	0.664	0.678	0.520	0.321	0.179
	Conf. Limits:	(52.4,80.4)	(50,86)	(34,71)	(9.36,55)	(-3.2,39)
	P value	<0.0001	<0.0001	<0.0001	0.0068	0.0962
LY & LP	Correlation:	0.236	0.035	-0.004	0.031	0.122
	Conf. Limits:	(5.9,39)	(--,--)	(--,--)	(-15,22)	(--,--)
	P value:	0.0097	.	.	0.7408	.
BW & LP	Correlation:	-0.213	-0.109	-0.106	-0.020	0.000
	Conf. Limits:	(-38,-3.4)	(-29,7.4)	(-26,7.9)	(-24,20)	(-23,27)
	P value:	0.0203	0.2424	0.2604	0.8522	0.9994
BW & B/P	Correlation:	-0.216	-0.099	-0.159	-0.001	0.199
	Conf. Limits:	(-38,-3.8)	(-29,9.2)	(-36,4.3)	(-22.8,2)	(-4.9,45)
	P value	0.0182	0.3105	0.1219	0.9901	0.116

In this study, Regression and MCMC monotone-data imputation methods failed to give confidence intervals and P-values for some of the missing data sets. The propensity score method for continuous variables with a monotone pattern of missingness and MCMC full-data imputation method for continuous variables with an arbitrary pattern of missingness performed well with data less than 40% missingness. Both logistic regression and discriminant function methods for binary variables with a

monotone pattern of missingness gave correct estimates in most cases. We highly recommend to researchers that they pay attention to the fact that the estimated P-value tends to get higher with an increasing proportion of missingness. For MNAR data, MCMC full-data imputation started to give incorrect estimations at 20% and 40% missingness. However, this method performed better than monotone-data imputation, which gave a correct estimation only at 5% missingness.

#### **6.4 References**

- Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from Incomplete Data via the EM Algorithm.(with discussion). *Journal of the Royal Statistical Society.* 39: 1-38.
- Graham, J. W. and S. M. Hofer. 1991. EMCOV.EXE Users Guide [Computer software manual]. Unpublished manuscript, University of Southern California, Los Angeles.
- King, G., J. Honaker, A. Joseph, and K. Scheve. 2001. Analyzing Incomplete Political Science Data: An alternative algorithm for multiple imputation. *American Political Science Review.* 95: 49-69.
- Lavori, P., R. Dawson and D. Shera. 1995. A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. *Statistics in Medicine.* 14: 1913-1925.
- Little, R. J. A. and D. B. Rubin. 1987. *Statistical Analysis with Missing Data.* New York, Academic Press.
- Little, R. J. A. and D. B. Rubin. 1989. The Analysis of Social Science Data with Missing Values. *Sociological Methods and Research.* 18: 292-326.
- Lu, H. J., and G. O. Myers. 2002. Genetic Relationships and Discrimination of Ten Influential Upland Cotton Varieties using RAPD Markers. *Theor. Appl. Genet.* 105: 325-331.
- Rosenbaum, P. R. and D. B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika.* 70: 41-55.
- Rubin, D. B. 1996. Multiple Imputation after 18+ Years (with discussion). *Journal of the American Statistical Association.* 91: 473-489.
- Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley and Sons.

- Rubin, D. B. 1977. Formalizing Subjective Notion about the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association* 72: 538-543.
- Rubin, D. B. 1976. Inference and Missing Data. *Biometrika* 63: 581-592.
- Rubin, D. B. and N. Schenker. 1986. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association* 81: 366-374.
- SAS Institute. 2003. Version 9. Cary, N. C., USA.
- Schafer, J. L. and J. W. Graham. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods*. 7:147-177.
- Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.
- Schimert, J., J. L. Schafer, T. Hesterberg, C. Fraley, and D. Clarkson. 2001. *Analyzing Missing Values in S-PLUS*. Seattle, WA.
- Yuan, Y. C. 2000. Multiple Imputation for Missing Data: Concepts and new development. In: *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute.

## CHAPTER 7 SUMMARY AND CONCLUSIONS

While cottonseed oil is one of the most important crop oils, cotton fiber is the most important textile fiber crop. Cotton is grown commercially in the temperate and tropical regions of more than 50 countries, including the United States, India, China, Central and South America, The Middle East, and Australia (Smith, 1999; Fryxell, 1979). Producer, manufacturer, and consumer demands are driving the development of cotton varieties that yield greater cotton fiber quantity and improved quality. Therefore, a better understanding of the genetic basis of agronomic and fiber quality traits by mapping QTL with molecular markers is an important objective in cotton breeding. An  $F_{2:3}$  population composed of 138 lines, derived from the intraspecific (*G. hirsutum*) cross between Paymaster 54 and Pee Dee 2165, was developed and a linkage map including 143 AFLP markers was constructed. The  $F_{2:3}$  population was grown in two locations, Alexandria and Baton Rouge in LA. Single-marker analysis (SMA), including simple and logistic regression, and interval-marker analysis (IMA), including interval mapping (IM) and composite interval mapping (CIM), was used for mapping agronomic and fiber quality QTL. Interval mapping was used to study QTL interaction effects with the environment.

Upland cotton contains 26 chromosomes; however, the 143 linked markers were assigned to 13 major and 15 minor linkage groups. A linkage group is considered a major group if it has a total length of 50 cM or longer. The 13 major groups ranged from 50.3 to 205.1 cM in length and each group carried 3 to 19 markers. The 15 minor groups ranged from 7.5 to 49.3 cM in length and each group carried 2 to 6 markers, the 28 linkage groups cover a genetic distance of 1773.2 cM.

An additional 57 unlinked markers were also detected. The total coverage for these 200 markers is 3066.2 cM assuming each unlinked locus and each pair of the 28 linkage group ends accounts for 20 cM on average (Weng, 2002). This gives a coverage of 65.2% of the cotton genome (4700 cM).

For the agronomic traits, the same five QTL were detected, using a significant threshold of 2 LOD, in both IM and CIM. These include two for lint weight per boll (LY), two for Seedcotton weight per plant (BW), and one for lint percentage (LP), which collectively, based on IM analysis, explained 32.5%, 28.6%, and 4.4% of the phenotypic variation, respectively. In total, seven and nine different QTL were detected by IM and CIM, respectively. This range of explained variation was common in our study and supports a model for quantitative inheritance for the agronomic traits studied (Lande and Thompson, 1990; Ulloa and Meredith, 2000). Two QTL for LY and BW were shown to have significant interaction effect with the two locations (Alexandria and Baton Rouge, LA) at a LOD threshold of two.

For the fiber quality traits, the same nine QTL were detected, using a significance threshold of 2 LOD, in both IM and CIM. These include one for fiber elongation (E), one for length (L), two for uniformity (U), three for strength (S), and two for micronaire (M), which collectively, based on IM analysis, explained 50.9%, 18.7%, 69%, 49.6%, and 25.3% of the phenotypic variation, respectively. In total, nine and 19 different QTL were detected in IM and CIM, respectively. Nine QTL were found to have significant interaction effects with the two locations (Alexandria and Baton Rouge) at a LOD threshold of two. The present study supports the general conclusion made by Tanksley (1993), i.e., a substantial proportion of QTL affecting a trait can be identified under different environments.

Adding more AFLP markers by screening different primer combinations of *EcoRI/MseI* and by assaying more enzyme combinations other than *EcoRI/MseI* will help saturate the map. An ongoing project at Louisiana State University is screening of simple sequence repeats (SSRs) to add anchored markers onto the map for further comparative mapping. With the addition of more markers, the smaller linkage groups may converge or join with other linkage groups. Such a saturated map could be directly used for marker-assisted plant breeding, and gene and QTL tagging. Future efforts in QTL mapping should focus on developing more saturated maps, using larger population sizes, and more powerful statistical algorithms and theories for identifying QTL and elucidating QTL X environment interactions.

#### **7.1 References**

- Fryxell, P. A. 1979. The Natural History of the Cotton Tribe. Texas A&M University Press, Collage Station, Texas.
- Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124: 743-756.
- Smith, W. C. 1999. Production Statistics. In: Smith, W. C. and J. T. Cothorn (eds). Cotton: Origin, history, technology, and production. John Wiley and Sons, Inc.
- Tanksley, S. D. 1993. Mapping polygenes. *Annu. Rev. Genet.* 27: 205-233.
- Ulloa, M. and W. R. Meredith Jr. 2000. Genetic linkage map and QTL analysis of agronomic and fiber quality traits in an intraspecific population. *The Journal of Cotton Science.* 4: 161-170.
- Weng, C., T. L. Kubisiak, C. D. Nelson, and M. Stine. 2002. Mapping quantitative trait loci controlling growth in a (longleaf pine x slash pine) x slash pine BC<sub>1</sub> family. *Theor. Appl. Genet.* 104: 852-859.

## **VITA**

Muhanad Walid Akash was born August 05, 1973, in Amman, Jordan. He graduated from Al-Hussein Collage High School in Amman in June, 1991. Muhanad received his Bachelor of Science degree in plant production from The University of Jordan in May, 1995 and his Master of Science degree in Crop Production from The University of Jordan in July, 1997.

He was awarded the Mott Meritorious graduate award in recognition of outstanding achievements and contribution, in 2002. Muhanad earned his Master of Science degree in applied statistics from The Department of Experimental Statistics, Louisiana State University in May, 2003. He will earn the degree of Doctor of Philosophy in December, 2003.