HIGH-DENSITY SNP GENOTYPING APPLIED TO INTERSPECIFIC GERMPLASM IN UPLAND COTTON (*GOSSYPIUM HIRSUTUM* L.): (I.) CS-B17 CHROMOSOME-SPECIFIC RIL ANALYSIS AND (II.) *G. MUSTELINUM* (MIERS EX WATT) LINKAGE MAPPING

A Thesis

by

YU-MING LIN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	David M. Stelly
Committee Members,	Steve Hague
	Michael Thomson
	Mauricio Ulloa
Head of Department,	David Baltensperger

December 2017

Major Subject: Plant Breeding

ABSTRACT

The narrow genetic base of Upland cotton has slowed growth of its productivity as a crop and perhaps also its use in the world. The need to broaden genetic diversity of Upland cotton is urgent, especially given the contemporary need to improve competitiveness of the fiber attributes, productivity and sustainability. The advent of high-density high-throughput molecular marker genotyping in cotton using the CottonSNP63K array has revolutionized the resolution and accuracy of genetic analysis in cotton. In this study, the CottonSNP63K array is used to analyze two populations, one a set of interspecific chromosome-specific RILs and the other an early-generations interspecific mapping population, both at the 52-chromosome level.

A chromosome linkage map was derived from 50 isogenic chromosome-specific recombinant inbred lines, which were derived from a cross between a disomic substitution line CS-B17 homozygous for *G. barbadense* '3-79' chromosome 17 and its recurrent parent TM-1. Fiber quality and Fusarium wilt race 4 resistance data on the lines were subjected to quantitative trait locus (QTL) analysis. Results indicated that the CS-RIL approach affords high sensitivity, in that it detected seven fiber quality QTLs in chromosome 17, whereas none had been found previously by analysis of conventional TM-1/3-79 populations. However, one lint% QTL was detected previously using a similarly interspecific population. A single locus accounted for multiple FOV4 resistance trait QTLs and corresponded to previous research. In this CS-RIL study,

QTLs exhibited exceptionally high R^2 values and consistency across experiments, reflecting avoidance of genetic background noise and GxE interactions.

The first high-resolution SNP-based genetic map between *G. hirsutum* and *G. mustelinum* was constructed from a 59 individuals of BC_1F_1 population. The map was highly collinear with the *G. hirsutum – G. barbadense* map and the *G. hirsutum* reference genome. In certain chromosomes, some markers exhibited segregation distortion. Co-segregation difference between genetic maps revealed possible chromosomal structure changes among species. Possible errors in the genome assembly were found by alignments of 1,996 low-specificity SNP markers to their homeologs in the reference genome. The genetic map can help guide genome assembly corrections, and facilitate many sorts of future studies, e.g., genetic dissection of complex traits and marker-assisted breeding.

DEDICATION

I dedicate my thesis research work to all people who have encouraged and supported me for my scientific American dream. Particularly thanks to my wonderful family, who always give me great support mentally and financially for my study. Thanks to all the great friends I met here for sharing their knowledge and experience. I could not have accomplished this work without you.

ACKNOWLEDGEMENTS

First, I would like to acknowledge my greatest appreciation to my advisor, Dr. Stelly, for providing me with this wonderful opportunity to pursue my dream in this excellent program of a well-known university. Thank you for all the guidance for my research and my skill for critical thinking, writing, and presentation. I sincerely appreciate your support for attending numerous important meetings and conference to present my research, and also the precious chances to interact with many distinguished scientists in the world. Truly thank you for all you have done for me, I am really grateful to study in your lab.

I would like to thank my committee members, Dr. Steve Hague, Dr. Michael Thomson, and Dr. Mauricio Ulloa for your advice and assistance to my research. A special thanks to Dr. Ulloa for arranging the field evaluation for FOV4 in California, which helped me to understand more thoroughly my projects.

Thanks to Dr. Alois A Bell and Dr. Xiuting Zheng for the knowledge of FOV, the FOV4 bioassays in growth chambers, and DNA extractions. Thanks to Dr. Amanda M. Hulse-Kemp for teaching me how to handle SNP genotypic data and some bioinformatic tools for analysis. I am very thankful to Dr. Robert Vaughn for taking care for my plants in the field and greenhouse and all kinds of experiments in the lab, and Luis de Santiago for teaching me command line and scripts for recombination analysis as well as linkage mapping analysis. Thanks to all the members of Stelly lab for the encouragement and scientific questions: Dr. Bo Liu, Ernesto Elizalde, Bree Vculek,

Junaid Jamshaid, Christian Hitzelberger, Velioğlu Kübra, Fisher Cherry, David O'Krafka, Andrea Maeda, Mariana Machado, and Ammani Kyanam.

I would like to acknowledge the support from Agri-Genomics Laboratory and Texas A&M Genome Science and Society (TIGSS) for generating genotypic data here: Dr. Fei Wang and Dr. Nithya Subramanian for teaching me DNA extractions and other genotyping methods, Kelli Kochan and Dr. Andrew Hillhouse for running cotton SNP chips and high-throughput genotyping.

The development of CS-B17-RIL population and the further phenotypic evaluations were the most essential works for my QTL analysis project; therefore, I would like to acknowledge this critical contribution to Dr. Saha Sukumar, Dr. Johnie N. Jenkins, Dr. Jack C. McCarty, Dr. Russell W. Hayes, and Dr. Todd Campbell for your time and labor in CS-B17-RILs development and the fiber traits phenotyping

A special thanks to the research team in California for scanning valuable resistance resource to FOV4 and increase the value of the CS-B17-RILs. This study was funded by USDA-ARS (Project 3096-21000-019-00) (MU), Cotton Incorporated (Core and CA State Support Committee) (RH, PR, DS, MU), Cary, NC., and California Cotton Ginners & Growers Association (RH and MU). I also would like to thank RB, Hutmacher, PA, Roberts, M. Keeley, TL. Frigulti, and T. Doung, for their help in evaluations. Use of greenhouse facilities of the University of California Kearney Research and Extension Center (Parlier CA) is gratefully acknowledged. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U. S.

Department of Agriculture or University of California. The U. S. Department of Agriculture is equal opportunity provider and employer. (FOV4)

Finally, thanks to my parents and sisters, who always keep faith with me and encourage me for chasing my dream. Thank you for listening the difficulties that I encountered during analysis, and for writing when you were either sleepy in the evening or in a trance in the early morning. A special thanks to my friend, Amy, for always standing by my side, and for providing me with some constructive thoughts when I had question in writing. Thank you for discussing my research problem and always trying to return me some helpful idea even though you are not working in a biology area.

Thank you all the people who I have met during this wonderful scientific trip for spending your time helping me to overcome every problem I had. Without you, I couldn't finish this work. Thank you all.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor David M. Stelly and Professors Steve Hague and Michael Thomson of the Department of Soil and Crop Sciences and Professor Mauricio Ulloa of the Department of Plant Stress and Germplasm Development Research, USDA.

The fiber quality and quantity phenotypic data of CS-B17-RILs for Chapter II were provided by Dr. Saha Sukumar, Dr. Johnie N. Jenkins, Dr. Jack C. McCarty, Dr. Russell W. Hayes, and Dr. Todd Campbell. The FOV4 resistance phenotypic data of CS-B17-RILs for Chapter II were provided by Dr. Mauricio Ulloa, Dr. RB, Hutmacher, Dr. PA, Roberts, M. Keeley, TL. Frigulti, and T. Doung.

The linkage disequilibrium analyses via CheckMatrix and the SNP markers BLAST analysis on JGI *G. hirsutum* reference genome database in Chapter II and III were assisted in part by Luis De. Santiago of the Department of Soil and Crop Sciences.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from Cotton Incorporated, and the FOV4 evaluation in California was funded by USDA-ARS, Cotton Incorporated, and California Cotton Ginners & Growers Association.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

Introduction of Cotton

Cotton produces seed-borne fibers with excellent absorbency, softness and versatility, for which it has long been one of the most important natural fiber crops and probably the most widely employed plant in human daily life. As a factor in human history, agriculture for food production has been considered as one of the cornerstones for civilization in ancient societies and also a catalyst of miscellaneous advanced technologies, social institutions, and language development (Diamond, 1998). Of course, agriculture also plays important roles in producing other essential material resources, such as biofuels, biodegradable plastics and textiles. Cotton, one of the major economic crops worldwide, is mainly cultivated for its fiber, which is manufactured into clothing, bed accessories, furniture upholstery and so on. Furthermore, the market demand for cotton fiber usage in high quality hygienic and medical products has significantly increased because of its favorable attributes, such as absorbency and softness (Luitel, Hudson and Ethridge, 2015). In addition, the cottonseed can be processed into livestock feed industry as a protein resource, as well as a source of vegetable oil, e.g., for cooking, salad dressing, and other food products (Cherry and Leffler, 1984). In 2016, 105 millions bales of cotton were produced in the world, and the United States, the third largest cotton production country right after China and India, contributed 17 million bales, which was worth about $5.6 billion, and led to an overall

economic revenue of approximately $30 billion (National Agricultural Statistics Service, https://www.nass.usda.gov). Even though synthetic fibers are now used widely and in many fabric products, the demand for cotton fiber remains huge, and world cotton production has gradually risen from 70 million bales in 1980's to 113 million bales today (USDA, Production, Supply and Distribution Database, http://apps.fas.usda.gov/psdonline/psdHome.aspx). Although four species of the cotton genus *Gossypium* are cultivated to produce cotton, most production (95%) relies on the tetraploid (2n = 52) species *Gossypium hirsutum* L., while another tetraploid species, *Gossypium barbadense* L. accounts for the most of the remaining proportion. Very small amounts of production, mainly in environments with very high biotic stress, are based on two diploid species, *G. herbaceum* and *G. arboretum*.

Overall, the genus *Gossypium* comprises 52 recognized species, including 45 diploid species (2n = 2x =26) and 7 allotetraploid species (2n = 4x = 52) (Hutchinson, Silow and Stephens, 1947; Saunders, 1961; Fryxell, Craven and Stewart, 1992; Cronn *et al.*, 2002; Wendel and Grover, 2015). Although diploid cotton species share the same chromosome number (2n = 26), their relative chromosome and genome sizes vary considerably (Hendrix and Stewart, 2005). By careful comparisons of chromosome sizes, meiotic behavior and fertility of diploid cottons, and their interspecific F1 hybrids, closely related species were clustered together into eight different genomic groups (A - G & K) (Beasley, 1941; Endrizzi, Turcotte and Kohel, 1985; Stewart, 1995; Wendel *et al.*, 2009). Proven by many phylogenetic analyses (Senchina *et al.*, 2003; Flagel, Wendel and Udall, 2012), the diploid genomic groups are now categorized into three

geographically related lineages: the African-Arabian clade, also known as Old World cottons, including A, B, E, and F genomes, the New World clade (D genome) originated from the western coast of Mexico, and the Australian clade, consist of C, G, and K genomes. Among the diploid species, most bear very short lint or even none; only A-genome diploid species possess long fiber. The species-specific genomes within each genome group are identified according to subscripts. For example, we find that the two cultivated diploid species, *G. herbaceum* and *G. arboreum* have genomes designated A_1 and A_2.

These three diploid lineages of *Gossypium* are estimated to have begun diverging around 5 to 10 million years ago (mya) and underwent the speciation in their presumptive diversity centers, African-Arabian, Australian, and Central America (Senchina *et al.*, 2003). "Molecular clock" analyses of sequence data indicate that a remarkable transoceanic hybridization and polyploidization event occurred about 1 to 2 mya between two diploid species, one having an A-like genome and one with a D-like genome, thus giving rise to a new "AD" genome (Wendel, 1989; Wendel and Cronn, 2003). This hybridization far pre-dates evolutionary origins of modern humans (*Homo sapiens*), and it is not known how polyploidization occurred, nor whether it occurred concomitantly with initial hybridization or afterwards. But molecular evidence indicates all of the known extant natural polyploids arose monophyletically, indicating that the single/rare allotetraploid cotton subsequently evolved into new species and dispersed throughout large areas of the New World (Grover *et al.*, 2012). Currently, seven allotetraploid species are recognized: *G. hirsutum* $[AD]_1$, *G. barbadense* $[AD]_2$, *G.*

tomentosum Nuttall $[AD]_3$, *G. mustelinum* $[AD]_4$, *G. darwinii* $[AD]_5$, *G. ekmamianum* $[AD]_6$ (Krapovickas and Seijo, 2008; Grover *et al.*, 2014) and *G. stephensii* (Gallagher et al. 2017). Surprisingly, domestication of *G. hirsutum* $[AD]_1$ and *G. barbadense* $[AD]_2$ accounts for their extensive intraspecific morphological variation and wide habitat distributions, compared with other allotetraploid species (Brubaker and Wendel, 1993; Brubaker, Bourland and Wendel, 1999). Cultivated tetraploid cottons are distributed across the tropics and some part of subtropics in the Central America; in contrast, the wild tetraploid cotton species mainly colonized relative small regions, like Pacific islands, coastal areas, or arid regions.

Genetic Diversity in Cotton Breeding

When breeding for genetic improvement, genetic diversity is always regarded as the essential ingredient. The idea of using genetically diverse parents to create superior progeny has become the central dogma in breeding (Duvick, 1984; Cox, Murphy and Rodgers, 1986), and great successes have been reported in many crops, like soybean (Manjarrez-Sandoval *et al.*, 1997), and oat (Cowen and Frey, 1987). However, some interspecific hybrids of other crops have worse agronomic performance than parents instead, for example, barley (Rasmusson and Phillips, 1997), wheat (Souza and Sorrells, 1991), and cotton (Esbroeck and Bowman, 1998). Because of the differences in the ploidy levels, meiotic affinity and chromosomal structure, relatively few of the diploid cotton species can be utilized as introgression resources for major cultivar species, *G. hirsutum*. Given that all seven of the extant 52-chromosome AD *Gossypium* species

descend from a common evolutionary AD polyploid ancestor, it would seem likely all of the AD tetraploid species would be fine resources for interspecific introgression, but in terms of their end products, i.e., cultivars, conventional Upland cotton breeding programs have relied heavily upon the cross between genetic closely related elite lines for new cultivar development (Esbroeck and Bowman, 1998). Another interpretation might be that some breeding programs, if not many, attempted to derive useful germplasm following interspecific introgression, e.g., from crosses between *G. barbadense* and *G. hirsutum*, but that extraordinarily few of those genetic lineages led to the development of released cultivars.

Although genetic diversification of Upland cottons by introgression of germplasm from other AD tetraploid species seems like it would be a useful approach because it avoids major physical impediments caused by major differences in ploidy, genome size, and chromosome structur, AD species introgression has nevertheless been hampered by some significant challenges. Sterility, distorted segregation, and poor agronomic quality have often occurred during tetraploid interspecific introgression, and been attributed to genetic incompatibilities (Stephens, 1949; McKenzie, 1970; Reinisch *et al.*, 1994). Furthermore, linkage drag between favorable and unfavorable syntenic loci seems to frequently plague interspecific germplasm. The close association of traits and the limited understanding of the genetic architecture of beneficial and undesired loci constrained efforts to address these challenges facing exotic elite trait introgression and efforts to augment the genetic diversity available among elite types of cultivated cotton.

Interspecific breeding inferably requires more recombination events to remove deleterious alleles than conventional intraspecific breeding. Given that agriculturally successful cultivars and other elite lines have genomes that necessarily feature large numbers of loci with beneficial alleles and inter-locus allele combinations, and very few loci with deleterious alleles and inter-locus allele combinations, intraspecific crosses between varieties and other elite lines of *G. hirsutum* L. allow for the most time-effective means to create genetically elite and potentially agriculturally competent products. The difficulty of using wild germplasm and interspecific germplasm may explain why almost one fourth of cotton cultivars arose by reselection within released germplasm or cultivars (Esbroeck and Bowman, 1998). When viewed across time in cotton breeding, extensive reliance on this "closed" genetic resource for parents bred to create successful cultivars has led to the fixation of many alleles in modern elite Upland cotton germplasm and will continue to progressively decrease allelic diversity, which will reduce the rates at which traits can be improved. The only way to solve this dilemma is to expand the gene pool, e.g., by utilizing exotic germplasm (Tanksley and McCouch, 1997). Complementary approaches include natural and induced mutagenesis epigenetic modification and genetic engineering, including transgenics, cisgenics and gene-edited products (John and Stewart, 1992; Perlak *et al.*, 2001; Patel *et al.*, 2014).

High-throughput Genotyping in Cotton

The development of molecular marker technology continues to advance genomic research as well as the efficient genetic improvement of cotton and other economically

important plants. Conventional breeding programs that develop successful varieties often feature use of large populations to create new breeding materials, pedigree breeding to develop the most elite lines, the strategic application of time- and cost-efficient experimental designs, and graduated use of replicated and multi-environment testing. While time-efficient methods are adopted as much as possible, breeding programs also require time-consuming labor-intensive activities and experienced breeders. Modern phenomic methods are being developed and implemented to enhance throughput, increase useful data acquisition and reduce costs. Molecular markers can be used to deduce major features about the genomic architecture of critical traits, and then used to facilitate introgression of single loci or concerted introgression of multiple loci into target cultivars. In this manner, some breeding time can be saved, and the important traits can also be identified for other applications (Barone and Frusciante, 2007; Eathington *et al.*, 2007; Ibitoye and Akin-Idowu, 2010). A strategy widely used to enhance crop marker-based breeding methods has been to increase the representation of molecular markers that can be closely associated with crop traits of interest; a part of the strategy is to develop large numbers of markers. For most crops, these include multiple types of markers and genotyping systems.

In cotton, the first linkage map was constructed based on 705 restriction fragment length polymorphism (RFLP) loci assembled into 41 linkage groups with total length of 4675 cM (Reinisch *et al.*, 1994). After the invention of the polymerase chain reaction (PCR), PCR-based DNA markers, like RAPD, AFLP, RGA, and SSR, gradually replaced RFLPs; they generally allowed for increasing marker numbers and

reproducibility, the avoidance of radioactive probes, and reducing need for large amounts of sample DNA. The low level of genetic diversity in *G. hirsutum* L. impeded cotton marker development. One strategy for increasing density of linkage maps was to use multiple types of markers. For example, a combination of RFLP, SSR, and AFLP markers was used on an interspecific backcross population between *G. hirsutum* and *G. barbadense* to construct a linkage map with 888 markers that assembled into 37 linkage groups with a combined length of 4400 cM (Lacape *et al.*, 2003). Compared with other types of markers, SSRs have been extensively developed and employed in genetic and germplasm analysis because of their relatively high level of polymorphism, accessibility, and ability to be easily shared among labs. Many SSR linkage maps have been made (CottonGen Map Data Summary: https://www.cottongen.org/find/featuremap/summary), some with high-resolution (Han *et al.*, 2006; Guo *et al.*, 2007). Nonetheless, the utility of SSR markers seemed to reach a bottleneck, as SSR genotyping is not very amenable to automation and each datapoint is relatively expensive. The maximum number of SSRs in a single-population map was barely over 3,500. To attain a higher resolution map, a consensus mapping strategy was employed by combining 6 mapping populations and a overall (collective) set of 6,669 different markers or primer sets; this effort yielded a consensus map spanning 4,070 cM, with 26 linkage groups and 8,254 loci (Blenda *et al.*, 2012).

As DNA sequencing became popular and affordable, single nucleotide polymorphism (SNP) markers attracted people's attention because of their abundance genomic variation, sequence-based nature, amenability to automation, and inexpensive

high-throughput genotyping. By increasing the number of molecular markers, combined with significant improvements in speed and affordability, SNP genotyping has in many organisms improved the precision of genomic analysis and expanded the range of research applications, such as genomic diversity analysis, genome assembly validation or refinement, and various types of marker-assisted selection. As for other markers, the development of SNP markers for cotton germplasm was significantly impeded by low diversity of the elite germplasm, as well as by complexities of the cotton genome, which is both polyploid and paleopolyploid (Paterson *et al.*, 2012). Prior to the evolutionary divergence and reunion between lineages of the tetraploid species' A- and D-subgenome subgenome ancestors, the cotton lineage actually underwent a five- to six-fold ploidy increase events after separating from grape and cacao lineage (Carvalho *et al.*, 2011; Paterson *et al.*, 2012). To be successful, a SNP marker in cotton should be capable of differentiating homeologous sequence variants (HSVs) as well as paralogous sequence variants (PSVs) (Flagel, Wendel and Udall, 2012; Kaur, Francki and Forster, 2012)

Until recently, with the aid of some important publications of reference genome sequence assemblies for *G. raimondii*, and *G.arboreum* (Wang *et al.*, 2012; Paterson *et al.*, 2012; Li *et al.*, 2014) and the improvement of next-generation sequencing technologies, a great number of SNP markers have been identified and gathered for development of a high-throughput genotyping assay (Zhu *et al.*, 2014; Hulse-Kemp *et al.*, 2015). An international consortium developed the highly multiplexed SNP assay for cotton based on the Infinium-II SNP genotyping method of Illumina (Steemers *et al.*, 2006). The CottonSNP63k array was designed to detect up to 70,000 SNP markers,

50,000 of them designed from SNPs detected at intraspecific levels within *G. hirsutum* L., and the remaing 20,000 assays were designed to detect SNPs interspecifically, between *G. hirsutum* L. versus other *Gossypium* species (Hulse-Kemp *et al.*, 2015). Both sets included a mixture of gene-based (transcriptome-based) and genome-based (gDNA-based) sequence comparisons. The Illumina Infinium-II technology utilizes bead-based approach with two-fluorophores to express the relative amount of signal and signal intensities of SNP makers. Fluorescent signals from the array are assigned to specific SNP markers according to the "manifest file", and then assigned to allelic genotypes (e.g., A, B, or H) using the "cluster file" and Illumina's Genome Studio program.

High-throughput SNP genotyping arrays have been well established for many economic important crops and in some cases applied in cultivar breeding or other genomic research (Ganal *et al.*, 2011; Hamilton *et al.*, 2011; Song *et al.*, 2013; Truco *et al.*, 2013; Bianco *et al.*, 2014; Chen *et al.*, 2014; Dalton-Morgan *et al.*, 2014; Tinker *et al.*, 2014; Wang *et al.*, 2014). Few studies have been reported in cotton based on high-throughput genotyping technologies. In my thesis research, the applicability of the high-throughput SNP genotyping technology has been examined for [1] characterization of an interspecific chromosome-specific recombinant inbred line populations that involves germplasm introgressed from *G. barbadense* (L.) into Upland cotton, *G. hirsutum* L., and its use for QTL analysis of several fiber yield, quality traits, and Fusarium wilt race 4 resistance, and also [2] the construction of an interspecific linkage map between the

domesticated Upland cotton species and a wild AD tetraploid species from Brazil, *G. mustelinum* (Miers ex Watt).

CHAPTER II

HIGH-DENSITY SNP-BASED MAPPING AND MULTI-TRAIT QTL ANALYSIS OF ISOGENIC CHROMOSOME SPECIFIC CS-B17 RILS IN UPLAND COTTON (*GOSSYPIUM HIRSUTUM* L.)

Introduction

Cotton (*Gossypium spp.*) has been widely used in the manufacture of apparel, house furnishing, and industrial products, and also became one of the most important economic crops in the world. Nowadays, almost 80 countries grow cotton on 30.3 million hectares (USDA-FAS, https://apps.fas.usda.gov/psdonline/app/index.html - /app/downloads) and produce 105 million 480-lb bales all over the world in 2016. India, China, and the United State are the three largest cotton production countries and together supply two-third of cotton in the world. In 2016, the 17 southern states in the US provided 17 million bales of cotton, which was approximately worth about $5.6 billion, and led to overall economic revenue of approximately $30 billion (National Agricultural Statistics Service, https://www.nass.usda.gov). Cotton production by the United States and the whole world in 2017 are anticipated to be 19.2 million and 114.7 million bales, respectively. Surprisingly, contemporary cotton production relies on one allotetraploid cotton species, *Gossypium hirsutum* L. $[AD]_1$ (2n = 4x = 52), which dominates almost 95% cotton production of the world. Another tetraploid species, *G. barbadense* L. $[AD]_2$, accounts for the most of the remaining production, and only few region in Asia

grow A-genome diploid species (2n = 2x = 26), *G. arboreum* L. and *G. herbaceum* L. (Lee, 1984; Campbell *et al.*, 2010; Percy *et al.*, 2014).

Observing that fiber yield growth has slowed and yield variation between years has increased, several studies have argued that the narrow genetic basis utilized in breeding of US cultivated cotton, *G. hirsutum*, has been one of the major reasons for the fiber yield plateau in the last few decades (May, Bowman and Calhoun, 1995; Meredith, Jr., 2000; Paterson *et al.*, 2004). Despite the wide genetic diversity present among the 52 currently recognized species in the *Gossypium* genus, including 45 diploid species (2n = 2x =26) and 7 allotetraploid species (2n = 4x = 52) (Hutchinson, Silow and Stephens, 1947; Saunders, 1961; Fryxell, Craven and Stewart, 1992; Cronn et al., 2002; Wendel and Grover, 2015), almost all new cultivars have been developed by reselection within released lines or germplasm, mainly *G. hirsutum* species, and very few new cultivars have resulted from the infusion of interspecific germplasm (May, Bowman and Calhoun, 1995; Esbroeck and Bowman, 1998). Because of differences in the ploidy levels, meiotic affinity and chromosomal structure, introgression from diploid species is expected to be difficult and so diploid cannot be considered as a primary genetic resource for Upland cotton. On the other hand, introgression from different tetraploid species seems to be rational approach since they have identical ploidy level and were separated from the same ancestor around 1 to 2 million years ago (mya) (Wendel, 1989; Wendel and Cronn, 2003). However, the genetic differences that arose during speciation of the tetraploid cottons lead to genetic incompatibilities, sterility, weakness, and distorted segregation among hybrids, which majorly hinders the utilization of

interspecific introgression (Stephens, 1949; McKenzie, 1970; Reinisch *et al.*, 1994). The exact mechanisms that degrade opportunities for successful interspecific germplasm introgression and their relative importance are not yet entirely clear, but several plausible causes have been observed. For example, zygotic selection can arise from lethal chlorophyll deficiencies that can be readily observed among seedlings, due to segregation duplicate recessive complementary genes in F2 interspecific hybrids. Other factors could affect traits prior to seed formation and/or traits that are not so obvious – for example, high degrees of sterility have been observed among viable interspecific F2 plants, due duplicate recessive complementary genes that undermine meiotic asynapsis of homologs. Altered distributions and suppressed rates of recombination and pollen sterility can be caused by sequence differences and chromosomal structure differences between tetraploid species, including heterozygosity for translocations, inversions, or deletions. Vigor and fertility of certain interspecific hybrids can be reduced by factors causing hybrid lethality or weakness, while odd ploidy levels and late reproductive maturity (juvenility and/or photoperiodicity) can complicate usage of hybrids (Pundir, 1972; Percy and Kohel, 1999; Zhang, Percy and McCarty, 2014).

Another challenge during interspecific hybridization is the inadvertent loss of desired genes during the long process of background recovery, especially in conventional breeding schemes. Backcrossing is required to recover the vast majority of the domesticated species' genotype, without which all genetic products will be ill-adapted to agricultural use. In interspecific hybrids, the numbers of homologous recombination events may be reduced overall and/or in specific regions, plus population

sizes are typically small, so favorable alien gene(s) may be lost during the selection, or unwanted gene(s) may be hard to remove from the breeding population, due to "linkage drag" with selected loci.

With the appearance of molecular markers, information of the genetic architecture for breeding materials can be obtained. Breeders are able to retain the desired gene(s) and get rid of unfavorable gene(s) using phenotypic evaluations and genotypic information for the breeding purposes in a more efficient manner, which is also known as marker-assisted selection (MAS) (Francia *et al.*, 2005). For multi-genic traits, quantitative trait locus (QTL) analysis has been widely used to determine the association between specific loci and traits of interest, and then further been applied into MAS. In cotton, QTL analyses have been conducted for many traits, such as yield, fiber-quality relative traits, and biotic/abiotic resistance, has been studied for a while and achieved some successful results (Wang *et al.*, 2008, 2015; Yu *et al.*, 2014). Nonetheless, few of the studies compares results with others because of the lack of "common" markers and insufficient knowledge of marker synteny between linkage maps (Said *et al.*, 2015). Lacape *et al.* compared QTL results from three backcross populations, BC_1, BC_2, BC_2S_1 from the *G.h.* x *G.b.*, and observed only 20 % of the QTLs repeatedly showed up in 2 or 3 data sets, and 30% of them agree with at least one previous studies for chromosomal position and parental origins (Lacape *et al.*, 2005). Rong *et al.* attempted to investigate the similarity of 405 QTLs from ten different *G.h.* x *G.b.* populations based on their positions on a reference map, and observed around 10% of the total QTL were corresponded (Rong *et al.*, 2007). Both studies concluded that the

low congruence of QTLs from different studies was caused by the variation from environments, genetic background, interactions, and others. Similar situations in the QTL studies for other crops, the additive effects of individual QTLs would sometimes be obscured by other effects, such as the presence of a major QTL, epistasis, interactions with non-allelic genes and/or environmental effects (Holland, 2001; Liao *et al.*, 2001; Li *et al.*, 2003; Charcosset *et al.*, 2004; Steele *et al.*, 2006). Markers dispersed across entire genome and multiple experiment replications are absolutely necessary to deduce a comprehensive understanding of the genetic effects, i.e., to conduct a statistical analysis capable of separating influence factors and uncovering important and minor QTLs. Unfortunately, the genetic interactions are too complicated to investigate, and not surprisingly, the overall costs for such experiments can be very high, and may render them unaffordable.

Chromosome substitution was developed as a genetic and breeding approach that reduces the genetic incompatibility during interspecific introgression, and was first documented in wheat (*Triticum aestivum*, 2n = 42) (Knott, 1987). The same concept was applied in cotton (Kohel, Endrizzi and White, 1977). A set of seventeen chromosome substituted lines of Upland cotton were formally released, each disomic for one *G. barbadense* chromosome or chromosome arm-like segment (Stelly *et al.*, 2005). The procedure used to develop these lines involved [1] creating isogenic versions of Upland cotton inbred TM-1 that were monosomic or "monotelodisomic" (acrocentric), [2] for each different chromosome substitution, crossing the donor, *G. barbadense* line "3-79", to a specific TM-1 hypoaneuploid, [3] recovery of the corresponding F1

hypoaneuploid hybrid (hemizygous for the alien chromosomes), [4] backcrossing it to the same type of TM-1 hypoaneuploid, [5] repeating steps #3 and #4 to reach the BC5F1, and then "selfing" (self-pollinating) the identified BC5F1 hypoaneuploid hybrid, [6] identifying a disomic BC5S1 individual and [7] selfing that individual (euploid, and true breeding for the alien chromosome pair) to increase seed supplies (**Figure 1**). Ideally, each CS-B lines contains just one homologous pair of chromosomes or chromosomal segments from *G. barbadense*, and very minor amounts of inadvertently retained small donor segments elsewhere in the genome, so the chromosome substitution (CS) lines are nearly isogenic to recurrent parent, TM-1, as well as to each other (Saha *et al.*, 2004, 2006; Saha, Stelly and Raska, 2011).

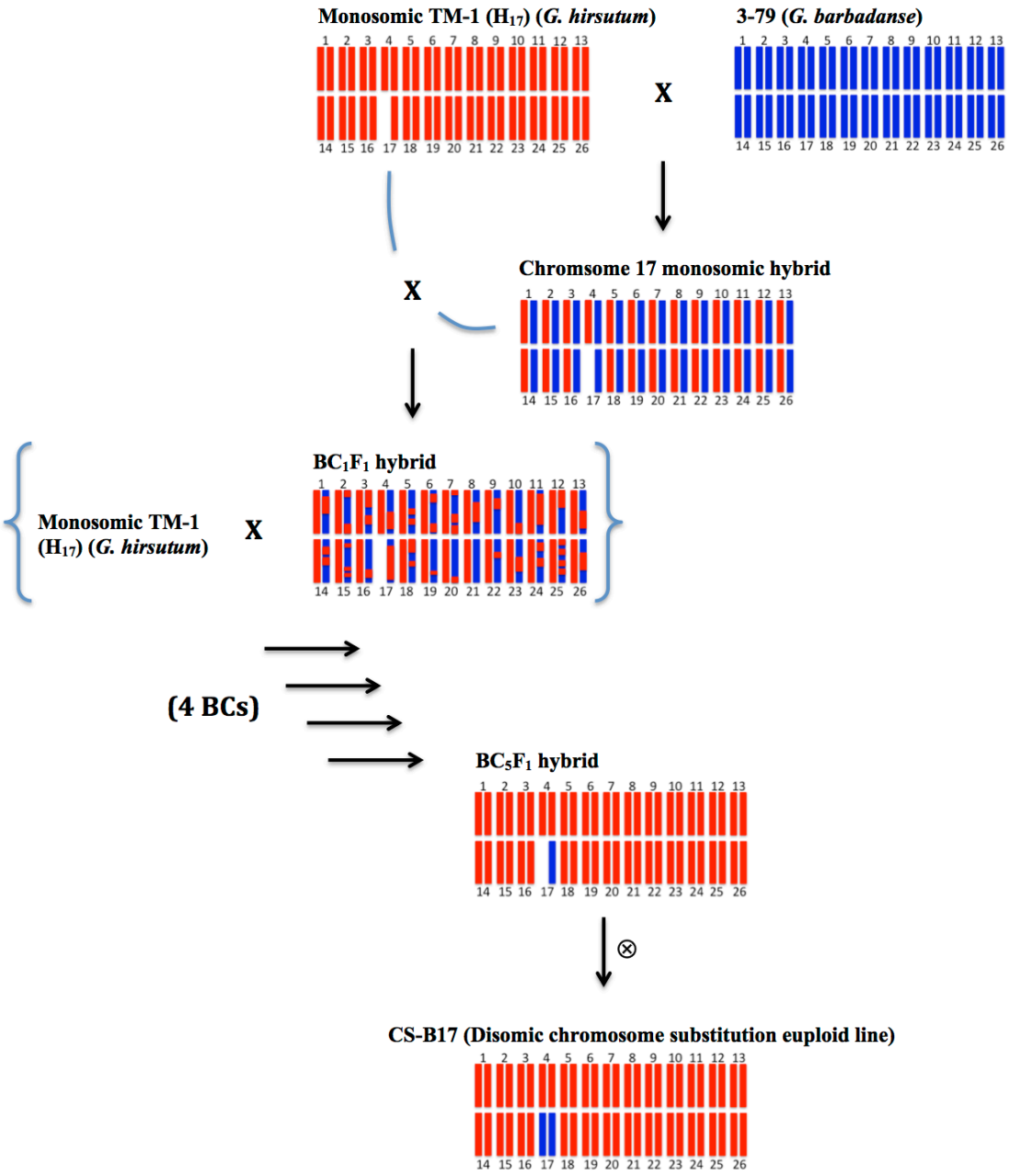


Figure 1. Diagram of chromosome substitution line (CS line) development. The designation of the CS-B17 indicates the chromosome 17 is substituted from the donor parent, cultivar 3-79, which is *G. barbadense*.

Multiple methods have been used to detect and dissect the genetic effects for many fiber quality and yield related traits associated with cotton CS lines (Saha *et al.*, 2004; Jenkins *et al.*, 2007). These include quantitative comparisons of the lines, per se, and quantitative partitioning by analysis of various type of families, e.g., F2, F3, sets of intercrosses and top-crosses (Saha *et al.*, 2010b). Many of these analyses were aided by the AD and related statistical models (Saha *et al.*, 2011) to separate variation components and successfully detect significant effects to fiber traits, such as high additive effect for lint percentage (54%), and moderate to high dominance by environment interaction effect for boll weight (57%), lint yield (34%), and seed cotton yield (25%) (Saha *et al.*, 2010b).

The usefulness of chromosome substitution lines can be greatly extended by using them as parents to create chromosome-specific recombination inbred lines (CS-RILs), which can be used for QTL analysis, further introgression and development of improved parents for breeding. CS-RILs are quasi-isogenic to each other, and segregation among them is limited essentially to just one chromosome pair, which significantly simplifies and thus improves genetic analysis of the respective cotton CS line parent (Rousset *et al.*, 2001; Luan *et al.*, 2009).

CS-B17 is one of the *G. barbadense* CS lines jointly developed by the Stelly Laboratory at Texas A&M AgriLife Research and characterized by the Saha-Jenkins-McCarty team at the USDA-ARS at Starkville, Mississippi. The line was associated with significant affects on a number of important traits, e.g., lint percentage, micronaire, and FOV4 susceptibility (Saha *et al.*, 2010a; Ulloa *et al.*, 2016a), and was selected as a

parent to create a small population of ~50 chromosome-specific RILs, i.e., CS-B17 RILs, that would be increased and then quantitatively analyzed for traits of interest. These CS-RILs were subjected to extensive assessment in replicated multi-environment experiments in Mississippi and South Carolina for plant and fiber traits, as well as in California for FOV resistance. However, previous mapping efforts to quantitatively dissect the RIL population traits with SSRs by Saha et al. (unpublished), and then by modest numbers (~100) SNPs alone, and the combination of SSR and SNP data, failed to yield internally congruent linkage mapping results and raised serious concerns about the accuracy of the data and/or population (Hulse-Kemp and Stelly, personal communication). This failure precluded reliable quantitative genetic dissection and QTL analysis, and suggested that additional data were needed to sleuth and understand the germplasm and data, i.e., before a robust and reliable quantitative analysis could be completed. These inferences elevated the prospective value of high-quality, high-density genotyping, and justified the necessary investments for analysis using the CottonSNP63K array. Although the array was designed for whole-genome analysis, and thus proportionately quite expensive for single-chromosome analysis, the highly reliable genotype calls and high densities of chromosome-17 marker were expected to markedly enhance the quantitative genetic dissection and QTL analysis. Moreover, the detailed data were expected to be especially important in this case (given the problems noted above), for validation of the genomic content of each CS-RIL, accurate detection of recombination events, and accurate definition of the chromosome-17 segments in each CS-B17 RIL.

The CS-B17 RILs were also sent to California for *Fusarium oxysporum* f. sp. *vasinfectum* Race 4 (FOV4) resistance performance testing, from which significant results were also obtained. Fusarium wilt is a serious disease to cotton in the US and other cotton production countries, and often accompanies with other diseases to cause huge cotton yield loss, e.g. root-knot nematode (RKN) and verticillium wilt (Moricca *et al.*, 1998; Ulloa *et al.*, 2011). This pathogen is a soil-borne fungus, which can survive in a field for many years in dormant status in soil or plant debris, or it can sustain on other plant species but not virulent to them, such as weeds. Therefore, once FOV infects a cotton field, it is difficult to eradicate the fungus.

FOV usually infects through the wounds on the roots of cotton, spreads along the xylem, where it causes phenolic compounds to accumulate in the vascular tissue, and leads to the normal Fusarium wilt symptoms of vascular browning, or discoloration, whole-plant wilting, and plant death eventually (Dowd, Wilson and McFadden, 2004; Hall, Heath and Guest, 2013). In California, two major FOV isolates infecting cotton have been identified as: FOV race 1 (VCG 0111), and FOV race 4 (VCG 0114) (Bell *et al.*, 2017). FOV race 1 (VCG 0111) belongs to vascular-competent isolates, which will cause severe disease damage in cotton only if root-knot nematode is present in the field (Jorgenson *et al.*, 1978; Garber *et al.*, 1979). Thus, breeding for field resistance to FOV race 1 (VCG 0111) can be also achieved by resistance to nematodes (Wang and Roberts, 2006). On the contrary, race 4 (VCG 0114) belongs to root-rot pathotype because it causes extensive root damage and the whole-plant wilting regardless of the presence of the nematode (Kim, Hutmacher and Davis, 2005). The breeding against race 4 (VCG

0114) depends on the genetic resources that can resist the infection of the pathogen. Since 2005, when the occurrence of FOV race 4 (VCG 0114) in California and USA was first reported, this pathogen immediately became a growing threat to cotton production and attracted scientists' attention of resistance development as well as the prevention of pathogen spread outside the California (Kim, Hutmacher and Davis, 2005; Cianchetta *et al.*, 2015).

Therefore, the primary goal of this study was to conduct a QTL analysis research for fiber traits and FOV 4 resistance on an interspecific chromosome-17 substitution recombinant inbred line (CS-B17-RIL) population using high-throughput SNP genotyping. The effectiveness of using chromosome substitution material to remove epistasis and genic interactions from non-target chromosomes was evaluated from the perspective of increasing detection power and stability of the QTLs additive effects on target chromosome between experiments. Previously, the CottonSNP63K array was used to mapp 500 markers to a chromosome-17 linkage group in the published *G.h.* x *G.b.* map (Hulse-Kemp *et al.*, 2015). These markers were used to construct a high-density linkage map for the QTL analysis in fiber traits and FOV 4 resistance, and the detected QTLs in this study were identified by the associations with highly linked SNP markers. Results from the QTL analyses were compared with other QTL studies using same parental lines, TM-1 x 3-79, or the same species combination, *G.h.* x *G.b*. for their positions on *G. hirsutum* physical map (Saski *et al.*, 2017, in press), the relative amounts of phenotypic variation accounted for by the QTL(s), and the parental source of each favorable allele.

Material and Methods

Plant materials

Through the cooperation between TAMU and USDA-ARS research team at Mississippi State, the CS-B17-RIL population was established by selfing the F_2 plants to F_6 generations via single seed decent (SSD) method from the cross between TM-1 (*G. hirsutum* L.) and its isogenic line, CS-B17, with chromosome-17 substituted from doubled haploid inbred line, 3-79 (*G. barbadense* L.) (**Figure 2**). In 2004, the TM-1/CS-B17 F_2 seeds were generated at Tecoman, Mexico, and fifty individual F_2 plants were inbred through SSD method for four generations in the greenhouse at USDA/ARS, Mississippi State. The F_6 lines were grown at Tecoman, Mexico to increase F_7 seeds to develop a 50 lines CS-B17-RILs population for further research. Theoretically, each CS-B17-RIL line is homozygous for a unique array of parental segments within chromosome 17, and is homozygous for TM-1 genetic alleles for the other 25 chromosome pairs (**Figure 2**). Given that CS-B17 RILs are nearly isogenic for 25 of the 26 chromosome pairs, most heritable phenotypic differences among them are attributable to variations in chromosome 17 composition. Opportunities for inter-locus genetic interactions are also be greatly reduced, rendering the chromosome-17 QTL additive effects much easier to detect. More detailed procedures regarding the development of CS lines and the derived recombinant inbred lines have been described (Stelly *et al.*, 2005; Saha, Stelly and Raska, 2011).

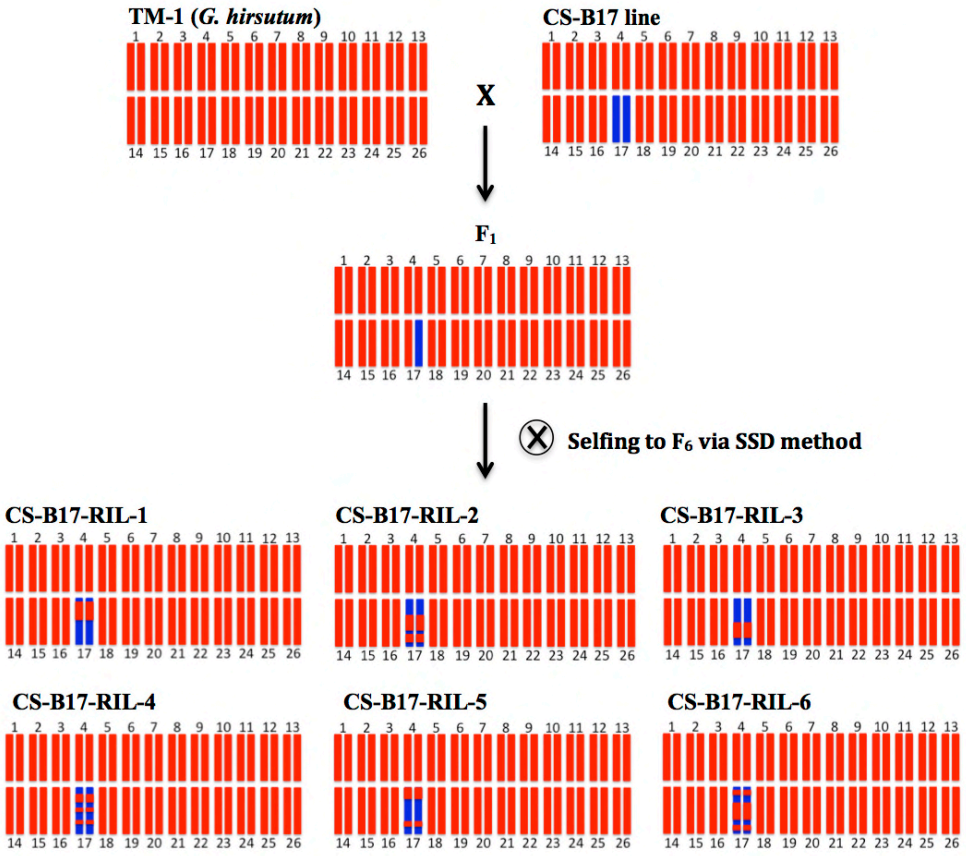


Figure 2. Diagram of CS-B17-RILs population development. Differences among RILs of the CS-B17-RILs population are relegated to chromosome 17 only, because all other chromosomes of the lines are expectedly homozygous and isogenic to the recurrent parent, TM-1.

Phenotyping

Fiber traits

The 50 lines of CS-B17-RILs population were field-tested in 2008 at two sites at Mississippi State (33.4° N, 88.8°W), and two sites in 2009, one at Mississippi State and one at Florence, South Carolina State (34.1°N, 79.4° W); these four environments are denoted as ST8.18, ST8.19, ST8.09, and TC17L respectively. In each environment, the

CS-B17-RILs population was grown under the randomized complete block design (RCBD) with 4 replications. Each entry was planted in a 12-m single-row plot with 97-cm row spacing and 10-cm spacing between plants in rows; a total of 120 plants were grown in a plot. In the experiments at Mississippi State, one row was skipped after two rows were planted, but no rows were skipped in the experiment at Florence, SC.

In every plot, a 25-boll sample of the first positions bolls at the middle nodes of the plants was hand-harvested for the fiber properties. Seedcotton samples were ginned using a 10-saw laboratory gin, and the lint percentage of each samples can be determined from absolute weights, i.e., 100% x (lint weight / seedcotton weight); the ginned lint samples were submitted to the Cotton Incorporated laboratory for HVI fiber measurements. All plots were later harvested by a commercial cotton picker to evaluate the lint yield and seed cotton yield. In this study, three agronomic traits and seven fiber quality traits were analyzed, including boll size (g), lint yield (LYLDHA, kg ha^{-1}), seedcotton yield (YLDHA, kg ha^{-1}), lint percentage (%), upper half mean length (UHM, mm), uniformity (UI, %), strength (kN m kg^{-1}), elongation (ELO, %), micronaire (MIC), degree of reflectance (Rd), and degree of yellowness (+b) (Saha *et al.*, 2017).

Fungus inoculum

An identified *Fusarium oxysporum* f.sp. *vasinfectum* (FOV) race 4 isolate from an infested field in the San Joaquin Valley was used in the greenhouse evaluations in 2016 at the University of California, Riverside (UCR), CA and at the University of California Kearney (UCK) Research and Extension Center (Parlier, CA, USA) (Kim,

Hutmacher and Davis, 2005; Ulloa *et al.*, 2006). Single spore cultures were stored on the filter paper at -20°C for inoculum preparation. Growing in 9 cm diameter Petri dishes containing 20 ml of potato dextrose agar (PDA) with 3 mM of streptomycin, the isolate was cultured at room temperature for 1 to 2 weeks. Then, water was poured in the plate, and conidia were scraped using a bacterial loop to dislodge spores/conidia into water. The spore suspension was filtered through four layers of cheesecloth to remove hyphae. The filtered suspension was quantified using hemocytometer for spore counts and adjusted to the final desired inoculum concentration.

Fusarium wilt race 4 greenhouse assay

A root-cut dip method, described by Smith et al. and modified by Ulloa et al., was used to inoculate the CS-B17-RILs population in both greenhouse assays (Smith *et al.*, 1981; Ulloa *et al.*, 2011, 2013). Seed were germinated in composite medium of vermiculite and peat moss. At 3 weeks, seedlings were removed from the medium, their roots cleaned, rinsed gently with water, and then dipped into the FOV4 inoculum with 1 x 106 conidia per ml for three minutes (UCR and UCK). In the UCR experimental site, the seedlings were transplanted in mix #2 soil (Baker, 1957) in pots with 2 to 5 plants per pot (subsamples), and each plot was used as a replication. The assay used a randomized complete block design (RCBD) with four replications as experimental design. From 8 to 20 plants per entry were examined for FOV4 infection. In the UCK experimental site, the seedlings were transplanted in steam-sterilized U.C. Mix #2 (Baker, 1957) soil in 6 x 15 cm (500 ml) box pots with one plant per pot (Ulloa *et al.*,

2009), following a complete randomized design (CRD) with 6 plant-replications per entry. Individual plants were evaluated 24 days after inoculation (dai).

Disease index (DI), plant height (PH, cm), shoots weight (SW, g), and vascular discoloration length (VDL, cm) were measured in the UCR greenhouse assay. VDL was measured by cutting the stem longitudinally and evaluating the discoloration part, and the ratio of discoloration length to total length could be calculated (VDL/PH*100). The disease index of the leaves was based on the following scale: 0 = no symptoms; 1 = epinasty of leaves; 2 = 1–30 % of leaves chlorotic and wilted; 3 = 31–80 % of leaves chlorotic and wilted; 4 = 81–100 % of leaves chlorotic and wilted; and 5 = plant death (Ulloa *et al.*, 2011, 2016a). In the UCK greenhouse assay, height (cm), number of nodes, average internode length, foliar disease index (FDI), and vascular stem and root staining (VRS) index were measured for the RILs. The scale of FDI in UCK was same as the scale of DI in UCR, and the scale for VRS index was based on: 0 = no any symptom, 1 = light spotty areas or thin line, 2 = more continuous than 1, but light colored staining covering an area between one quarter and one half of the stem cross-section, 3 = moderate brown/black staining evident in a band encircling most of the stem-root cross-section, 4 = brown/black staining evident across most vascular tissue in stem cross-section, and 5 = plant severely damaged or plant death with staining evident throughout the root tissue. VRS was measured by cutting the stem from node 0 to the end of the root longitudinally and evaluating the discoloration part (Ulloa *et al.*, 2009, 2013, 2016b).

In the UCR assay, 49 RILs were successfully germinated for the test for the 2-5 plants RCBD with 4 repeats experiment, without CS-B17-RIL 59. Similar situation in

the UCK greenhouse FOV4 assay, only 44 RILs were germinated enough for the 6 plant-replications CRD experiment, excluding CS-B17-RIL 59, 60, 64, 77, 88, and 98.

Phenotypic data analysis

Statistical analyses were conducted for each trait by environment separately for their sample mean, standard deviation, minimum, maximum, coefficient of variance, and the one-way ANOVA, including entry effect (genotype effect) and blocking effect (not shown). Broad-sense heritability of each trait within each environment, i.e., the ratio of the genetic variance to the total variance, was calculated from their ANOVA statistics. Pearson correlation analysis was also carried out to investigate the association between traits within environment. PROC GLM function in SAS/STAT® (SAS, ver. 9.3, SAS Institute, Cary, NC, USA) and excel 2013 software were mainly used for the one-way ANOVA and the above statistical analysis. Basic tests for normality and homogeneity of variance were also performed by using SAS/STAT® and R software.

Genotyping

The CS-B17-RILs population was grown in the field on FnB road, College Station, TX in 2012 and New Beasley lab green house in 2016 for the DNA sample preparation. Fifty CS-B17 RILs were grown well and their young, folded leaf tissue were sampled for DNA extraction using NucleoSpin® Plant II genomic DNA extraction kit for plant and fungi (Macherey-Nagel, Duren, Germany). A Nanodrop Spectrophotometer (Thermo Fisher Scientific, Waltham, USA) was used to determine

the DNA concentrations. DNA concentrations were standardized at 50 ng/µl, and the two parents, TM-1 and 3-79, their F1 hybrid and the 50 RILs were genotyped using the CottonSNP63K array at Texas A&M University according to Illumina protocols. After the single-base extension, the chip was scanned by Illumina iScan to generate the image files, which were then saved in GenomeStudio software to used to make a genotype call of each SNP characterized in the cluster file for tetraploid cotton genotyping (Hulse-Kemp *et al.*, 2015) (available at http://www.cottongen.org/node/ add/cotton-cluster-file-request). Genotyping data of our RIL population for the 63,058 SNP markers were transformed into "ABH" format, and only makers that expressed opposite homozygous genotype calls between two parents were utilized in further linkage mapping as well as QTL mapping analysis.

Linkage mapping analysis

With segregation in our CS-B17-RILs population to be analyzed only for chromosome 17, some minor adjustments for linkage mapping procedure were necessary. Instead of grouping markers based on recombination fraction and LOD score, the *G. barbadense* x *G. hirsutum* map (Hulse-Kemp *et al.*, 2015) was used as reference to select the markers (500) located in the c17 (chromosome-17) linkage group. Subsequently, the genotype data with a set of the 500 SNP markers were mapped with R/Onemap package (Margarido, Souza and Garcia, 2007) using recombination counting and ordering algorithms (RECORD) (Hans Van Os, Piet Stam, 2005) and Kosambi map function with multiple replications for obtaining the best result (Kosambi, 1944). Since the mapping

analysis relied heavily on genotypic data, co-segregated markers that had identical genotypic patterns across the mapping population would be placed into common bin groups and mapped to the same positions. Therefore, a small set of markers comprising of one selected marker from each bin group was used for analysis in order to increase computational efficiency. After initial map was created, visualized genotype examination and crossover counts were carried out on the ABH data file to remove questionable markers in Microsoft Excel software; then we could conduct the linkage mapping analysis again for the final map. Linkage disequilibrium were plotted to examine the marker order of the final map by using the CheckMatrix software (http://www.atgc.org/XLinkage/); the final map was drawn via MapChart software (Voorrips, 2002).

QTL analysis

Once the final linkage map for chromosome 17 was created, QTL analysis was performed using R/qtl package (http://www.rqtl.org/) for every trait-by-environment. Initial scanning for the first QTL was conducted using the simple interval mapping (SIM) method, in conjunction with three computational methods, maximum likelihood via EM algorithm (Dempster, Laird and Rubin, 1977), Haley-Knott regression (Haley and Knott, 1992), and multiple imputation (Churchill and Harbor, 2001). A composite interval mapping (CIM) strategy was performed to search the additional QTL while assuming the first detected QTL as covariate (Zeng, 1993). A significance threshold for LOD score was generated from the distribution of the genome-wide maximum LOD score of 1,000

permutation tests under the null hypothesis of no QTL presented. In every permutation test, the phenotypic data were randomly paired with genotypic data, and the QTL mapping analysis was performed to derive the maximum LOD score. From the 1,000 permutation tests, the 95th, 99th, and 99.9th percentiles of all maximum LOD values were considered as thresholds in this study. Potential QTLs were identified only if their LOD scores were above the thresholds. Information included QTL position on linkage map, 1.5-LOD support intervals, the amount of phenotypic variation explained by the QTL (R^2), and additive effect. The QTL analysis procedure was illustrated in detail using the experiment UHM_ST8.19 data in the result section. The temporary nomenclature of QTL in this study followed the rule: q + trait abbreviation + environment name + serial number of marker associated with the traits in environment **(CottonGen, 2010) (https://www.cottongen.org/help/nomenclature_qtl).** For example, qUHM-ST8.18-2 represented the second detected QTL linked to UHM in environment ST8.18.

QTLs comparison with published results

Published QTLs on chromosome 17 from other publications were used for comparison with our analysis in terms of position on reference genome and estimated position on genetic map, additive effect, and R^2. The sequence of inferred markers associated to published QTLs, from supplemental material or from Cottongen website (https://www.cottongen.org/search/markers), were used for Basic Local Alignment Search Tool (BLAST) via CottonFGD website (Zhu *et al.*, 2017) on JGI *G.hirsutum* genome database (Saski *et al.*, 2017, in press). Then, we searched the two nearest SNP

markers flanking the QTL-associated markers based on their sequence positions on reference genome. Given knowledge about positions of the two flanking SNP markers in the genome assembly, a common reference, the position of the respective QTL(s) was relatable to our genetic map. With the information of knowing the two flanking SNP markers, the position of the published QTLs could now be estimated in our genetic map. If the locations of our and previously reported QTLs were different, it can be inferred that the QTLs were different, whereas if they were col-localized, we inferred that our QTL was likely identical to the reported QTL.

Results

Phenotypic data analysis

Mean, standard deviation, minimum value, maximum value, p-value of the entry effect from ANOVA, coefficient of variance, and broad-sense heritability for traits are listed in **Table 1** and **Table 2**. ANOVA was conducted according to RCBD for most experiments and CRD for the UCK FOV4 assay. Probability values for treatment effects (genotypes) were very low (p-value < 0.0001) in most experiments, which indicated the significant difference existed between CS-B17-RIL lines for most traits and locations, except for the lint yield (LYLDHA) on ST8.19, and seedcotton yield (YLDHA) in experiments ST8.19 and TC17L. The coefficient of variance (CV) for LYLDHA, YLDHA, plant height (PH), shoot weight (SW), vascular discoloration length (VDL), foliar disease index (FDI) and vascular stem and root staining (VRS) index were high from 57% to 26%, which implies these traits were greatly influenced by environmental

effects. The CV% for disease index (DI), and height, node, inter node in UCK had moderate CV were moderate, around 12 to 20%, where as the rest fiber traits had low CVs less than 10% - indicating good reliability for their phenotypic data. Statistically, the FOV4 resistance traits were more affected by other factors than the fiber traits. Lint% had the highest broad-sense heritability, $H^2 = 0.828$, averaging across four environments, followed by MIC with an average heritability $H^2 = 0.702$. UHM also exhibited consistent heritability across environments, with a mean of value of 0.449. Heritability of UI, strength, Rd, and +b were low, but similar in different experiments. Heritability of boll size and ELO were moderately high but varied more across environments.

For FOV4 resistance experiments, the p-values of all measured traits from the ANOVA were extremely small, leading to strong rejection of the null hypotheses and thus claiming of significant genotype effects on these CS-B17-RIL traits. Among those traits, shoot weight had the highest CV, 57.43%, and others had CV between 12.39% to 30.74%, which implied resistance assay was more sensitive to environmental effect, compared to fiber traits experiment. The heritability of traits in FOV4 assay was averagely higher than fiber trait experiments. In the UCR experiment, plant height had the highest H^2, 0.708, followed by DI with 0.685, and SW for 0.638. In the UCK experiment, VRS had the highest H^2, 0.53, followed by height with 0.511, and FDI for 0.482.

Correlation analysis

Correlations among traits measured in the fiber-trait experiments were analyzed separately **(Table 3)** from traits measured in the FOV4-resistance experiments (**Table 4)**. Among fiber traits, boll size was positively correlated with lint yield and seedcotton yield in environment ST8.09 and TC17L, and the two yield traits (seedcotton and lint) shared the highest correlation coefficient, 0.99, among all environments. For fiber quality, MIC and UI were the traits that having most positive correlation with others. MIC was significantly and positively correlated to boll size, lint%, uniformity index, and strength in all the environments, where they were investigated, and MIC and lint% were the second highest correlated traits. Similarly, UI was also positively, correlated to boll size, lint%, UHM, strength, and MIC in all the three environments, except for the environment ST8.09 with strength. Among all the traits, ELO was one of the few traits negatively correlated with others. In environment ST8.18 and ST8.19, ELO was negatively correlated with lint%, UHM, and MIC, and two yield traits were significantly negatively correlated with ELO as well.

In FOV4 resistance experiments, the disease index (DI), vascular discoloration length/plant height (VDL/PH) ratio, foliar disease index (FDI), and vascular stem and root staining (VRS) index were positively correlated with each other. Among them, DI and VDL/PH ratio had the highest correlation coefficient of 0.9489, and then followed by the coefficient of FDI versus VRS, for which the coefficient was 0.9049; the coefficient of the rest combinations ranged between 0.7951 and 0.5997, which indicated the resistance index methods were quite concordant in both UCR and UCK FOV4 assay.

Plant height (PH) and seedling weight (SW) in the UCR experiment were strongly correlated (r = 0.9265), and both traits were significantly and negatively correlated to the previous mentioned resistance index traits. DI and SW, VDL/PH ratio and SW, and DI and PH had the highest three negative correlations, -0.939, -0.9251, and -0.8536 respectively. PH and SW were also negatively correlated to FDI and VRS index with the coefficients ranging from -0.7890 to -0.6812. VDL was moderately shared positive correlations with SW and PH, and negative relationships with DI, FDI, VRS index, but there was no strongly relationship to the VDL/PH ratio. Height, number of nodes, and average internode length in UCK were correlated with each other but not most of the above traits.

Table 1. Fiber traits phenotypic data summary of 50 CS-B17-RILs.

Traits [a]	*Environments* [b]	*Mean*	*SD*	*Min*	*Max*
Boll size	*ST8.18*	4.57	0.459	3.26	5.54
	ST8.19	5.13	0.402	3.76	6.22
	ST8.09 [c]	5.04	0.551	2.29	6.47
	TC17L	6.02	0.492	4.47	7.14
Lint percentage (Lint%)	*ST8.18*	33.28	1.814	29.26	38.51
	ST8.19	32.63	1.892	28.48	37.13
	ST8.09 [c]	31.19	1.822	27.78	35.37
	TC17L	33.54	2.008	29.52	38.73
Lint yield (LYLDHA)	*ST8.18*	--	--	--	--
	ST8.19	328.12	168.036	47.05	856.42
	ST8.09	226.52	105.487	40.79	634.69
	TC17L	1279.54	361.323	419.06	2701.03
Seedcotton yield (YLDHA)	*ST8.18*	--	--	--	--
	ST8.19	1006.40	513.288	150.36	2584.90
	ST8.09	726.74	336.697	121.31	2143.86
	TC17L	3807.68	1025.00	1189.44	7830.48
Upper half mean length (UHM)	*ST8.18*	27.89	0.709	26.16	29.97
	ST8.19	27.76	0.660	25.65	29.72
	ST8.09 [c]	28.55	0.663	26.92	29.97
	TC17L	--	--	--	--
Uniformity index (UI)	*ST8.18*	82.93	0.884	80.30	85.0
	ST8.19	83.70	0.746	81.20	85.4
	ST8.09 [c]	83.57	0.818	81.50	85.6
	TC17L	--	--	--	--
Strength (ST)	*ST8.18*	273.92	8.750	252.99	317.71
	ST8.19	271.39	8.522	249.07	297.12
	ST8.09 [c]	275.45	12.201	251.03	310.85
	TC17L	--	--	--	--
Elongation (ELO)	*ST8.18*	6.28	0.511	4.80	7.70
	ST8.19	6.62	0.577	5.10	8.70
	ST8.09 [c]	7.23	0.385	6.10	8.10
	TC17L	--	--	--	--
Micronaire (MIC)	*ST8.18*	4.24	0.438	3.42	5.35
	ST8.19	4.46	0.482	3.64	5.78
	ST8.09 [c]	3.72	0.402	2.90	4.60
	TC17L	--	--	--	--
Reflectance (Rd)	*ST8.09* [c]	76.06	33.051	70.10	79.80
Yellowness (+b)	*ST8.09* [cD]	9.03	3.952	7.70	11.00

[a] : Boll size (g), lint % (%), LYLDHA (kg ha^{-1}), YLDHA (kg ha^{-1}), UHM (mm), UI (%), ST (kN m kg-1), ELO (%).
[b] : "ST" indicate experiments at Mississippi State and "TC" represent experiment in South Carolina. First two experiments conducted in 2008, and later two conducted in 2009.
[c] : RCBD with 3 replication in the experiments.

Table 1. Continued.

Traits [a]	*Environments* [b]	*Entry p-value*	*CV (%)*	H^2
Boll size	*ST8.18*	<0.0001	7.70	0.411
	ST8.19	<0.0001	5.12	0.577
	ST8.09 [c]	<0.0001	7.31	0.560
	TC17L	<0.0001	5.11	0.610
Lint percentage (Lint%)	*ST8.18*	<0.0001	2.45	0.800
	ST8.19	<0.0001	2.59	0.803
	ST8.09 [c]	<0.0001	2.30	0.847
	TC17L	<0.0001	2.22	0.862
Lint yield (LYLDHA)	*ST8.18*	--	--	--
	ST8.19	0.2638	49.77	0.035
	ST8.09	0.0019	42.30	0.182
	TC17L	0.0278	24.44	0.117
Seedcotton yield (YLDHA)	*ST8.18*	--	--	--
	ST8.19	0.2939	46.69	0.030
	ST8.09	0.0074	42.85	0.151
	TC17L	0.3754	24.32	0.016
Upper half mean length (UHM)	*ST8.18*	<0.0001	1.85	0.444
	ST8.19	<0.0001	1.79	0.423
	ST8.09 [c]	<0.0001	1.65	0.480
	TC17L	--	--	--
Uniformity index (UI)	*ST8.18*	<0.0001	0.90	0.258
	ST8.19	0.0320	0.84	0.113
	ST8.09 [c]	0.0002	0.81	0.312
	TC17L	--	--	--
Strength (ST)	*ST8.18*	0.0001	2.79	0.232
	ST8.19	<0.0001	2.74	0.250
	ST8.09 [c]	0.0067	3.71	0.212
	TC17L	--	--	--
Elongation (ELO)	*ST8.18*	<0.0001	6.27	0.382
	ST8.19	<0.0001	5.54	0.583
	ST8.09 [c]	0.0003	4.40	0.237
	TC17L	--	--	--
Micronaire (MIC)	*ST8.18*	<0.0001	6.45	0.615
	ST8.19	<0.0001	5.14	0.777
	ST8.09 [c]	<0.0001	5.78	0.715
	TC17L	--	--	--
Reflectance (Rd)	*ST8.09* [c]	0.0002	1.86	0.304
Yellowness (+b)	*ST8.09* [c]	0.0003	5.25	0.298

[a] : Boll size (g), lint % (%), LYLDHA (kg ha^{-1}), YLDHA (kg ha^{-1}), UHM (mm), UI (%), ST (kN m kg-1), ELO (%).
[b] : "ST" indicate experiments at Mississippi State and "TC" represent experiment in South Carolina. First two experiments conducted in 2008, and later two conducted in 2009.
[c] : RCBD with 3 replication in the experiments.

Table 2. Fusarium wilt race 4 resistance phenotypic data summary of CS-B17-RILs.

Traits [a]	*Environments* [bc]	*Mean*	*SD*	*Min*	*Max*
Disease index (DI)	*UCR*	4.04	0.885	0.80	5.00
Plant height (PH)	*UCR*	16.13	8.034	5.94	38.06
Shoot weight (SW)	*UCR*	2.11	2.003	0.11	12.29
Vascular discoloration length (VDL)	*UCR*	10.20	3.300	4.66	22.78
Height	*UCK*	5.95	1.330	2.50	9.00
Number of nodes	*UCK*	3.02	0.658	1.00	4.00
Average internode length	*UCK*	2.04	0.555	0.63	3.90
Foliar disease index (FDI)	*UCK*	2.87	1.099	0.00	5.00
Vascular stem and root staining index (VRS)	*UCK*	2.63	1.173	1.00	5.00

[a] : PH (cm), SW (g), VDL (cm), height (cm), average internode length (cm).
[b] : 49 RILs under RCBD with 4 repeats in UCR and 44 RILs under CRD with 6 repeats in UCK.
[c] : University of California, Riverside (UCR); University of California Kearney (UCK) Research and Extension Center (Parlier, CA, USA)

Table 2. Continued.

Traits [a]	*Environments* [bc]	*Entry p-value*	*CV (%)*	H^2
Disease index (DI)	*UCR*	<0.0001	12.39	0.685
Plant height (PH)	*UCR*	<0.0001	27.13	0.708
Shoot weight (SW)	*UCR*	<0.0001	57.43	0.638
Vascular discoloration length (VDL)	*UCR*	<0.0001	23.33	0.487
Height	*UCK*	<0.0001	15.72	0.511
Number of nodes	*UCK*	<0.0001	17.46	0.364
Average internode length	*UCK*	<0.0001	20.83	0.417
Foliar disease index (FDI)	*UCK*	<0.0001	27.68	0.482
Vascular stem and root staining index (VRS)	*UCK*	<0.0001	30.74	0.530

[a] : PH (cm), SW (g), VDL (cm), height (cm), average internode length (cm).
[b] : 49 RILs under RCBD with 4 repeats in UCR and 44 RILs under CRD with 6 repeats in UCK.
[c] : University of California, Riverside (UCR); University of California Kearney (UCK) Research and Extension Center (Parlier, CA, USA)

Table 3. Pearson correlation coefficients between fiber traits for CS-B17-RILs.

Traits	*Env.*	*Boll size*	*Lint%*	*LYLD HA*	*YLD HA*	*UHM*	*UI*	*ST*	*ELO*	*MIC*	*Rd*
Boll size	*ST8.18*	1									
	ST8.19	1									
	ST8.09	1									
	TC17L	1									
Lint%	*ST8.18*	0.21*	1								
	ST8.19	0.14	1								
	ST8.09	0.11	1								
	TC17L	0.06	1								
LYLD HA	*ST8.18*	--	--								
	ST8.19	0.03	0.03	1							
	ST8.09	0.35*	0.18*	1							
	TC17L	0.22*	0.29*	1							
YLDHA	*ST8.18*	--	--	--							
	ST8.19	0.01	-0.06	0.99*	1						
	ST8.09	0.32*	0.07	0.99*	1						
	TC17L	0.22*	0.11	0.98*	1						
UHM	*ST8.18*	0.16*	0.17*	--	--	1					
	ST8.19	0.03	0.08	0.20*	0.19*	1					
	ST8.09	0.13	0.13	0.18*	0.16	1					
	TC17L	--	--	--	--						
UI	*ST8.18*	0.41*	0.33*	--	--	0.51*	1				
	ST8.19	0.26*	0.19*	0.08	0.06	0.29*	1				
	ST8.09	0.33*	0.29*	0.19*	0.16*	0.59*	1				
	TC17L	--	--	--	--	--	--				
Strength	*ST8.18*	0.09	0.24*	--	--	0.37*	0.31*	1			
	ST8.19	0.21*	0.24*	-0.02	-0.05	0.33*	0.21*	1			
	ST8.09	0.13	0.14	0.09	0.07	-0.13	0.14	1			
	TC17L	--	--	--	--	--	--	--			
ELO	*ST8.18*	-0.08	-0.39*	--	--	-0.15*	-0.06	-0.03	1		
	ST8.19	0.03	-0.45*	-0.23*	-0.19*	-0.20*	-0.04	<0.01	1		
	ST8.09	0.18*	-0.13	0.05	0.07	-0.11	0.13	0.40	1		
	TC17L	--	--	--	--	--	--	--	--		
MIC	*ST8.18*	0.48*	0.69*	--	--	0.12	0.36*	0.21*	-0.21*	1	
	ST8.19	0.36*	0.72*	0.05	-0.02	-0.03	0.31*	0.28*	-0.27*	1	
	ST8.09	0.37*	0.67*	0.17*	0.08	0.08	0.39*	0.26*	0.09	1	
	TC17L	--	--	--	--	--	--	--	--	--	
Rd	*ST8.09*	0.17	-0.31*	0.033	0.08	0.26*	0.13	-0.14	0.01	-0.32*	1
+b	*ST8.09*	-0.04	0.15	-0.13	-0.16*	-0.29*	-0.13	0.13	0.05	0.09	-0.57*

* : p-value of the coefficients are smaller than 0.05, which indicating the variables are significantly correlated.
-- : Indicates missing data.

Table 4. Pearson correlation coefficients between traits in FOV4 resistance assay.

Trait / Site	*DI_ UCR*	*PH_ UCR*	*SW_ UCR*	*VDL_ UCR*	*VDL/PH_UCR*	*Height_ UCK*	*Node_ UCK*	*Internode length_ UCK*	*FDI_ UCK*	*VRS_ UCK*
DI_UCR	1.00									
PH_UCR	-0.85*	1.00								
SW_UCR	-0.94*	0.93*	1.00							
VDL_ UCR	-0.33*	0.72*	0.49*	1.00						
VDL/PH_ UCR	0.95*	-0.85*	-0.93*	-0.28	1.00					
Height_ UCK	-0.22	0.15	0.10	0.04	-0.16	1.00				
Node_ UCK	-0.30	0.34*	0.34*	0.31*	-0.26	0.38*	1.00			
Internode length_ UCK	0.05	-0.14	-0.17	-0.21	0.08	0.61*	-0.49*	1.00		
FDI_ UCK	0.80*	-0.79*	-0.79*	-0.48*	0.74*	-0.37*	-0.51*	0.08	1.00	
VRS_UCK	0.68*	-0.72*	-0.68*	-0.57*	0.60*	-0.36*	-0.50*	0.087	0.90*	1.00

* : p-value of the coefficients are smaller than 0.05, which indicating the variables are significantly correlated.

Genotypic analysis of CS-B17-RILs

The CS-B17-RILs was genotyped using the CottonSNP63K array, and the genotype data were transformed into the "ABH" format for mapping analysis, where "A" is homozygous for the TM-1 allele, "B" is homozygous for the 3-79 allele and "H" is heterozygous. Functional markers, showing polymorphism between parents and having over 95% success genotype calls, were retained and coordinated with *G. hirsutum* x *G. barbadense* linkage map information (Hulse-Kemp *et al.*, 2015) for preliminary examination of our population. Categorized according to chromosome number from previous mapping results, SNP genotype calls for the entire population were summed up to calculate the allele proportion by chromosomes (**Figure 3**). As expected, segregation was mainly relegated to chromosome 17, within which SNPs averaged 52% of genotype A, 45% of B, and 2% of H, and closely matched the 1:1 allele frequency expectation for these CS-RILs. Interestingly, chromosomes 3 and 23 also possessed appreciable amounts of genotype B, 4.1% and 12.4%, respectively, reflecting pieces of Gb 3-79 chromosomes 3 and 23 detected previously in CS-B17 by CottonSNP63K analysis. The observation of genotype B for entire population across genome, 2.16%, closely approximated the expected percentage of 1.36%; it is likely that the responsible *G. barbadense* segments are not in chromosome 17 were probably residual carryover from the backcross-based development of CS-B17 line. The presence of sizable segments of *G. barbadense* chromosomes 3 and 23 in CS-B17 might have arisen by chance, or it may be that they resulted from segmental deficiencies in chromosomes 3 and 23 of the monosomic *G. hirsutum* cytogenetic stock that was used as recurrent parent in

development of the CS-B17 line. This possibility can easily be tested by dosage analysis of the H17 *G. hirsutum* parent in the future.

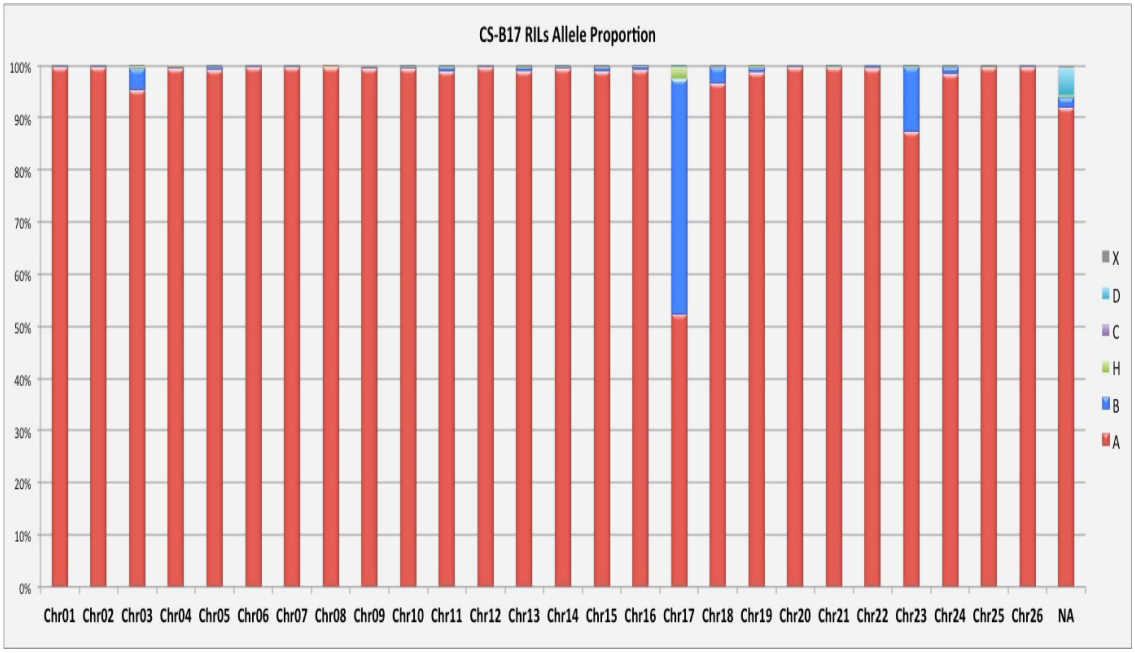


Figure 3. Allele proportions of SNP loci across all 26 chromosome pairs of 50 CS-B17-RILs. "A" genotype represents the allele derived from the *G. hirsutum*, TM-1, parent 1in the study, and "B" represents the allele from 3-79, *G. barbadense*. "H" indicates the heterozygous loci. "C", and "D" are used for dominant marker types. "X" represents missing genotype call.

Linkage map construction

Exactly 500 SNP markers previously mapped to chromosome 17 were selected for constructing a novel linkage map for 50 lines of CS-B17-RILs population. Using the order of loci in our initial map, the number of chromosome-17 crossovers in each line was calculated (**Figure 4**) to provide a means to evaluate the quality of the linkage map.

The overall average was 4.3 recombination events per RIL, but among RILs in the population, CS-B17-RIL60 was an "outlier" in that it possessed a uniquely high number of putative crossovers in chromosome 17 (**Figure 5**). Moreover, CS-B17-RIL60 was also unique among the RILs in that it contained several *G. barbadense* segments scattered across entire genome, i.e., in chromosomes other than 17. To verify that the exceptional characteristics CS-B17-RIL60 were not do a spurious single-seed event, a DNA sample was extracted from another individual CS-B17-RIL60 plant that was grown from backup bulk seed, and then genotyped to further assess the possibility of contamination during wet lab work and to verify the abnormality of CS-B17-RIL60. Comparison of allele component density between CS-B17-RIL60 and other two CS-B17-RILs clearly demonstrated the *G. barbadense* genetic germplasm was unintentionally involved into CS-B17-RIL60 development, i.e., beyond what CS-B17 could have provided (**Figure 6**). The step at which CS-B17-RIL60 development was contaminated remains uncertain at this time. Given the contamination, CS-B17-RIL60 was removed for linkage mapping and QTL analysis. Based on the remaining 49 lines, 500 SNP markers were categorized into 84 bin groups, and the final linkage map for chromosome 17 spanned 85 cM with average interval 1.02 cM per bin. The largest gap between markers was 3.95 cM, which occurred between the positions of 55.64 cM and 59.59 cM on the chromosome-17 genetic map. Detailed mapping information of the SNP markers can be found in **Supplemental Table S1**.

CheckMatrix 2D software was applied to examine the marker order of the final linkage map (**Figure 7**). Markers having small recombination fractions were denoted in

red color, indicating high likelihood of linkage disequilibrium; on the other hand, cool color was used for linkage equilibrium, i.e., that markers tended not to link together. In **Figure 7**, the red color along the diagonal of the plot illustrated the deduced order of markers was internally congruent. In the plot, it can be observed that markers in the same bin group have identical genotype patterns across population, and result in a red square in the heatmap. The size of each red square denotes the number of markers in the respective bin group. The clear intersection between two major red squares corresponded to the largest gap of our linkage map, 3.95 cM; the same phenomenon was observed at the same position of the *G. hirsutum* x *G. barbadense* interspecific map (Hulse-Kemp *et al.*, 2015). High recombination areas can result in analytical fragmentation a linkage group and lead to multiple linkage groups, especially when the number of molecular markers available for linkage mapping is too low (Zhang *et al.*, 2012; Yu *et al.*, 2013; Shang *et al.*, 2016). More SNP markers should be developed in that region to increase its linkage map resolution.

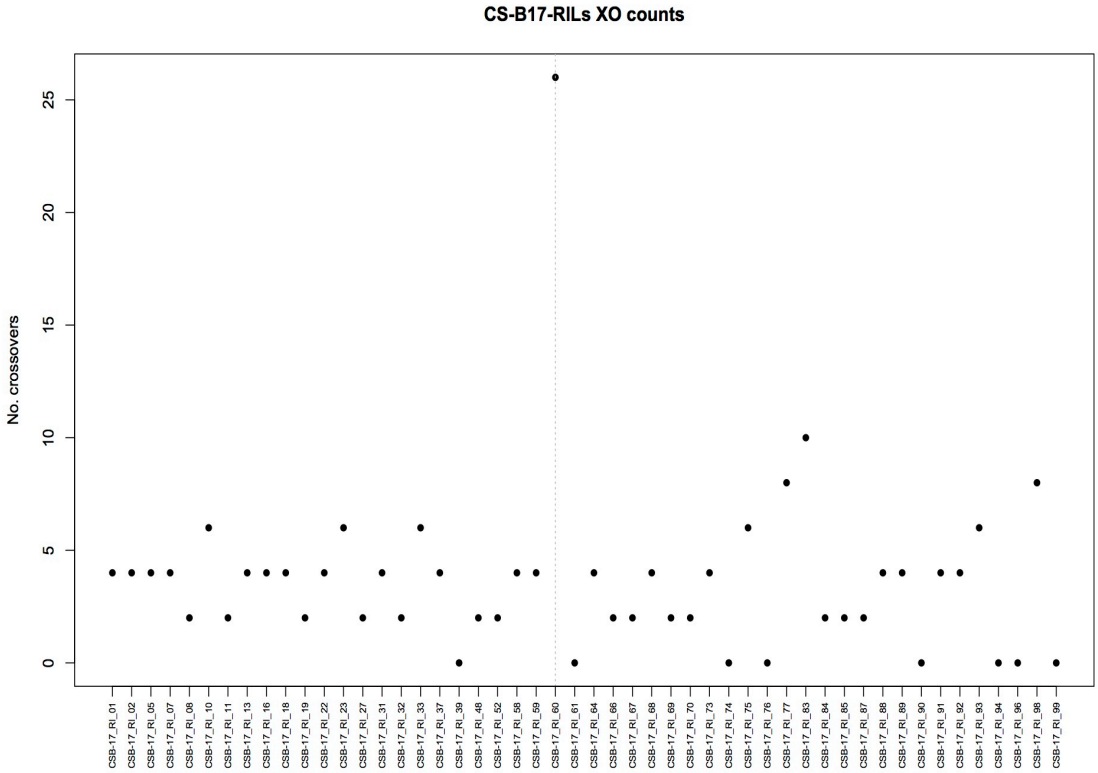


Figure 4. Crossover counts for CS-B17-RILs based on the order of SNP loci inferred from the initial linkage analysis. The number of inferred crossovers was exceptionally high for CS-B17-RIL60.

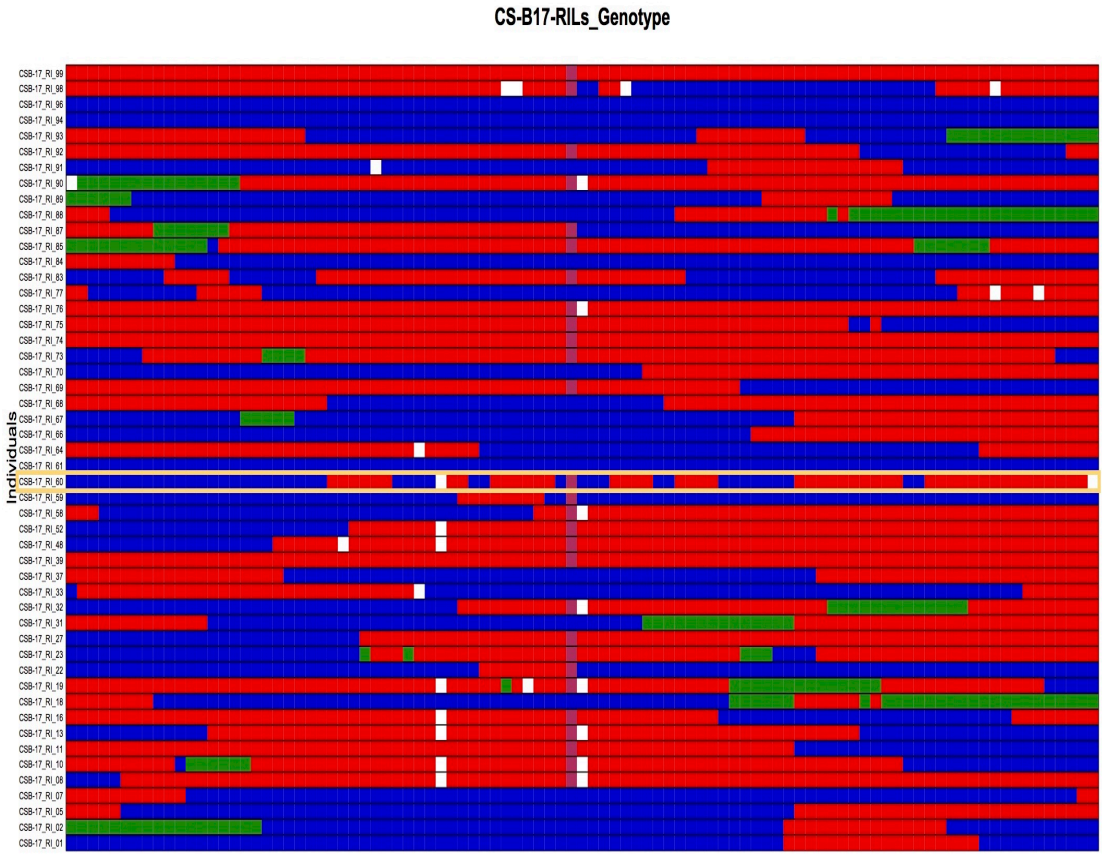


Figure 5. Genotype visualization for CS-B17-RILs. Markers were ordered along the X-axis according to the initial linkage mapping results, and the genotype of each RIL was represented as horizontal line. Red indicates homozygous genotype A, i.e., the *G. hirsutum* parent; blue denotes homozygous genotype B, from the *G. barbadense* parent; green denotes heterozygous genotype H. Maroon color was used to represent the genotype D, where genotype A and genotype H were distinguishable, i.e., when the *G. barbadense* marker was dominant. The suspect line CS-B17-RIL60 is highlighted in orange rectangular.

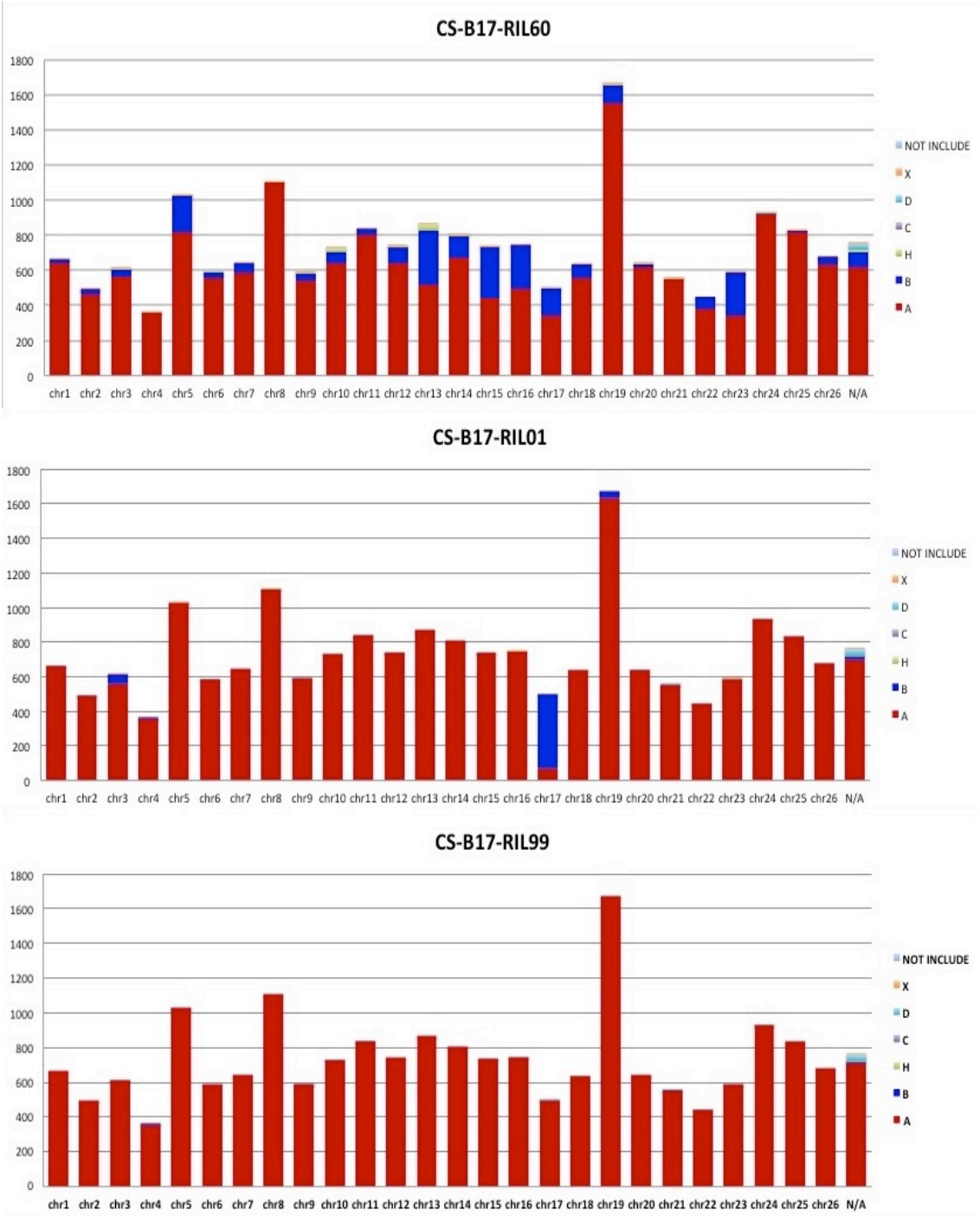


Figure 6. Allele component density of CS-B17-RIL60, CS-B17-RIL01, and CS-B17-RIL99. In CS-B17-RIL60, the *G. barbadense* alleles were distributed across entire genome, not mainly in chromosome 17; the distribution in the CS-B17-RIL60 genome differed markedly from distributions in genomes of other RILs, such as CS-B17-RIL01 and CS-B17-RIL99.

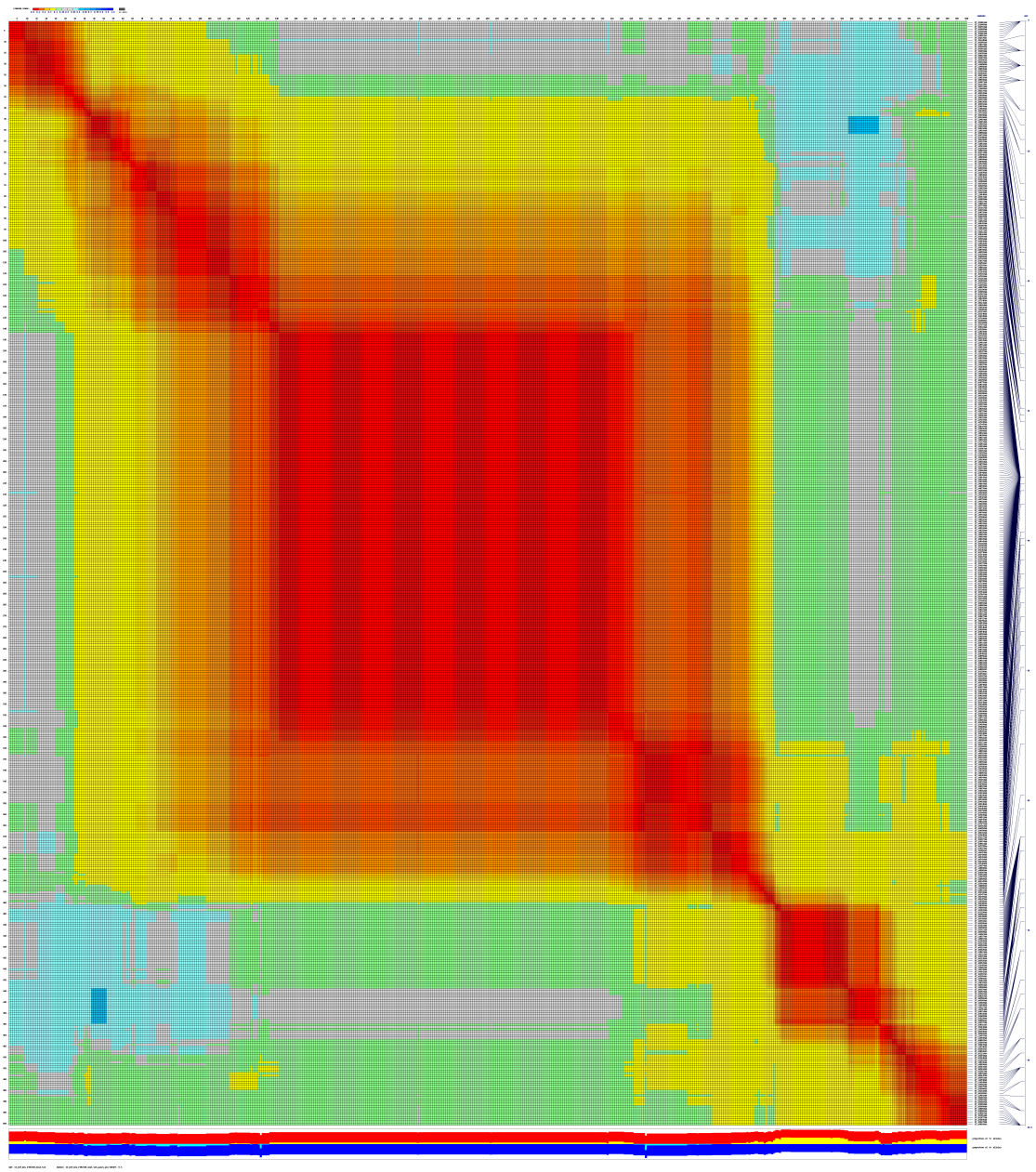


Figure 7. Linkage disequilibrium and recombination plot for chromosome 17 using CheckMatrix software. Warm colors, e.g., red, represent higher likelihoods of two markers are linked together; on the contrary, cool colors, e.g., yellow, represent markers that are less linked.

QTL mapping analysis

Details of the QTL analysis are illustrated using experiment UHM_ST8.19 (**Figure 8**). The three methods of QTL analysis - maximum likelihood via EM algorithm, Haley-Knott regression, and multiple imputation method - yielded very similar results in every experiment because of the high density and high rates of successful genotype calling from the CottonSNP63K array. LOD scores were nearly identical, such that the graphic plot comparing them, their respective lines were almost overlapping (**Figure 8.1**). The methods used for subsequent analysis are comparatively exemplified relative to results with the Haley-Knott regression method. SIM and CIM scanning detected 31 QTLs with significantly affecting the various traits analyzed in the fiber trait and FOV4 resistance experiments (**Table 5, Figure 9**). For most traits, at least one QTL was detected in each environment, but there was only one QTL found for boll size, LYLDHA, and Rd in all the testing environments. None was found for YLDHA, the degree of yellowness. In the UCK FOV4 experiment, no significant QTL was detected for plant height, node number or internode length.

For lint%, a QTL was detected in each of the four environments, qLP-ST8.18-1, qLP-ST8.19-1, qLP-ST8.09-1, and qLP-TC17L-1, were highly significant and consistent. Their LOD scores exceeded 0.1% probability threshold by more than 1.5 LOD score. Moreover, the four QTLs were at the exactly same position on the linkage map, 40.35 cM, with a very narrow 1.5-LOD support interval, and explained a similar proportion (65-75%) of the phenotypic variance in each experiment, respectively.

Additive effects of 4 QTL were also similar and ranged from 1.5% to 1.3% increase in lint percent due to the genetic contribution from the *G. hirsutum* parent.

For MIC, one QTL was detected in each the three test environments it was evaluated, and all were very highly significant. The qMIC-sT8.18-1 and qMIC-sT8.09-1 QTLs were located at essentially identical positions, and qMIC-sT8.19-1 was only 3 cM apart from the previous two. The R^2 of three QTL were similar, 59.26%, 63.58%, and 68.52% of phenotypic variation respectively, and exhibited narrow support intervals and strong additive effects. Parental SNP associations indicated that the MIC-increasing "allele" at this QTL came from the *G. hirsutum* background.

The QTLs for UHM ST8.18 and ST8.19 were highly significant and consistent. Moreover, two QTLs were detected in each of these two environments. The first (strongest) detected QTL in each environment, namely qUHM-ST8.18-1 and qUHM-ST8.19-1, was located similarly at the very end of linkage group, 81.86 cM and 83.89 cM, and explained 56.72% and 43.72% of variation, respectively. The second QTL in each experiment, qUHM-ST8.18-2 and qUHM-sT8.19-2, mapped similarly to the middle of the chromosome, i.e., 44.77 cM and 51.42 cM, respectively. The phenotypic variance explained by the first and second QTLs rose up to 72.76% and 58.01% in ST8.18 and ST8.19, i.e., roughly 16% higher than by the first-detected QTL. In experiment ST8.09, only one QTL, qUHM-ST8.09-1, was detected, at 69.34 cM of our linkage map, which was just in the middle of the two QTLs in the first two environments. The 1.5-LOD support interval of qUHM-ST8.09-1was obviously wider, 40 cM, and the R^2 was lower,

24.81% of the phenotypic variation. Parental SNP associations indicated that the UHM-increasing "alleles" at all five QTLs came from the *G. hirsutum* background.

Since the QTLs of lint%, MIC, and UHM in this study were extremely similar in terms of the significance, position on linkage map, additive effect, and the R^2, the QTLs detected in different environments might be the same. That is, the effects of chromosome-17 QTLs on lint%, MIC, and UHM were consistently detectable across environments.

Compared to the QTLs affecting lint%, MIC, and UHM, analysis results for UI, strength, and ELO varied more, and were less significant in different environments. For UI, one QTL was detected in each Mississippi environment; qUI-ST8.18-1, qUI-ST8.19-1, and qUI-ST8.09-1 were positioned at 39.83 cM, 40.35 cM, and 40.35 cM of the genetic map with 1.5-LOD support interval spanning 10 cM - 21 cM. The qUI-ST8.18-1 QTL showed the most significance, which over the 0.1% threshold by more than 1.5 LOD score, while the other two QTL were just above the 1% threshold but below the 0.1% threshold. Phenotypic variance explained by QTLs also differed from 28% to 41%. As for all of the traits above, *G. hirsutum* contributed the QTL "allele" that raised CS-B17 RIL phenotypic values for fiber uniformity index.

For ELO, significant QTL effects were detected in experiments ST8.18 and ST8.19 but not ST8.09. The positions of both QTLs on the genetic map were the same (36.15 cM) and exhibited a similar level of significance, over 1% threshold but less than 0.1% threshold. qELO-ST8.18-1 encompassed a 28.32 cM interval and accounted for 24.50% of phenotypic variation. QTL qELO-ST8.19-1 encompassed an even longer

interval, 40.11 cM, and explained 27.84% of variation. However, for this trait, the favorable QTL "allele" was from *G. barbadense* donor parent, instead of *G. hirsutum*.

For fiber strength, four QTLs were detected in three environments with different levels of significance. The positions of three QTL were near each other, namely qST-ST8.18-1, qST-ST8.19-1, and qST-ST8.09-1, at chromosome-17 map positions 41.45 cM, 40.35 cM, and 35.11cM, whereas the fourth QTL, qST-ST8.09-2, was located far away, at 76.05 cM of the linkage map. Corresponding to the significance levels, the 1.5-LOD support interval was narrower if QTL was strongly significant: qST-ST8.09-2 had the interval of 10 cM, 14 cM for qST-ST8.18-1, 26cM for qST-ST8.19-1, and 54 cM for qST-ST8.09-1. QTL qST-ST8.18-1 explained 33.61% of phenotypic variation, whereas qST-ST8.19-1 accounted for 27.69 %, and 22% was explained by qST-ST8.09-1 only but 48% was explained when combining qST-ST8.09-1 and qST-ST8.09-2. An interesting fact was that the ST-increasing "alleles" at qST-ST8.18-1, qST-ST8.19-1, and qST-ST8.09-1 came from *G. hirsutum,* whereas the ST-increasing "allele" at qST-ST8.09-2 was from *G. barbadense*. Moreover, significance of qST-ST8.09-1, just above 5% threshold, might be offset by the qST-ST8.09-2, since the contributions of the two QTLs were opposite. Effects from two QTLs are readily observed from the phenotype means of genotype groups (**Figure 10**).

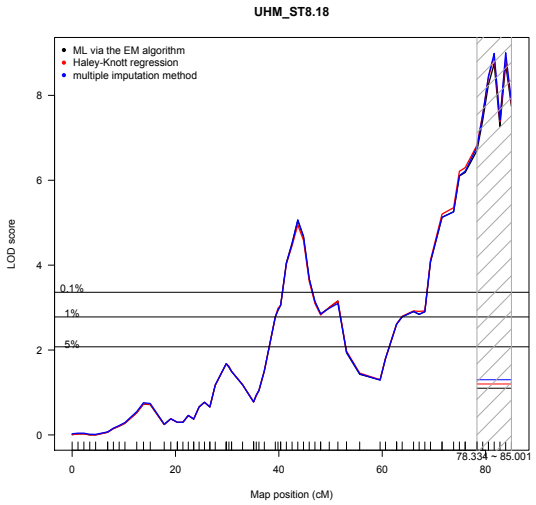
Interaction effects between QTLs in the same experiment were evaluated by 2-dimensional scans, using the *scantwo* function in R/qtl package. No interaction effect was detected among QTLs in this study. QTLs for boll size and Rd showed up in one field only in Mississippi State, but not in the other two experiments in MS. qBS-ST8.18-

1 and qRd-ST8.09-1 might reflect environmental effects or other interactions, instead of the additive genetic effect we mainly searching for. A QTL affecting LYLDHA was detected in the field in South Carolina State, and explained 32.07% of phenotypic variation; however, no other experiments were carried out in South Carolina for validation or assessment of location-relevant QTL stability. Besides, cotton yield might easily be influenced by environmental factors, huge variation was observed between experiments in different locations. For example, the yield mean in experiment in SC was roughly three times more than the average in experiments in MS (**Table 1**). Given the complexity of yield determination and inconsistent results, more research would be needed to verify the authenticity of qLYLDHA-TC17L-1.

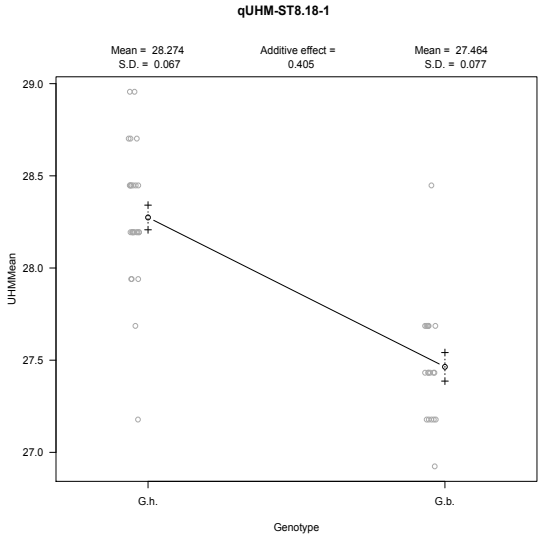
Fusarium wilt race 4 resistance experiments in UCR and UCK led to detection of 7 QTLs - DI (disease index), PH (plant height), SW (seedling weight), VDL (vascular discoloration length), VDL/PH (vascular discoloration length / plant height), FDI (foliar disease index), and VRS (vascular stem and root staining) index. All QTL mapped to the same position on the genetic map, 17.82 cM. The 1.5 LOD support intervals for these 7 QTLs were very similar, and ranged between 15.73 cM 7.26 cM. The proportions of phenotypic variation affected by QTLs were quite different: qPH-UCR-1 explained the most among the seven, 58.23% of phenotypic variation in plant height in UCR; qSW-UCR-1 and qVDL/PH-UCR-1 accounted for 53.99% and 52.62% of phenotypic variation in shoot weight and the ratio of vascular discoloration length over plant height.qDI-UCR-1, qFDI-UCK-1, and qVRS-UCK-1 explained 44.69%, 40.85%, and 37.48% of variation respectively, and only 24.46% of variation was explained by

qVDL-UCR-1. According to the sign of the additive effect for each QTL, *G. hirsutum* provided higher value in PH, SW, and VDL; and concordantly, the DI, VDL/PH, FDI, and VRS index would increase, which meant more susceptible to FOV4, when the QTL allele transmitted from *G. barbadense* parent. In other words, all seven QTLs drove to the same conclusion that *G. hirsutum* provide more resistance than *G. barbadense* at this QTL.

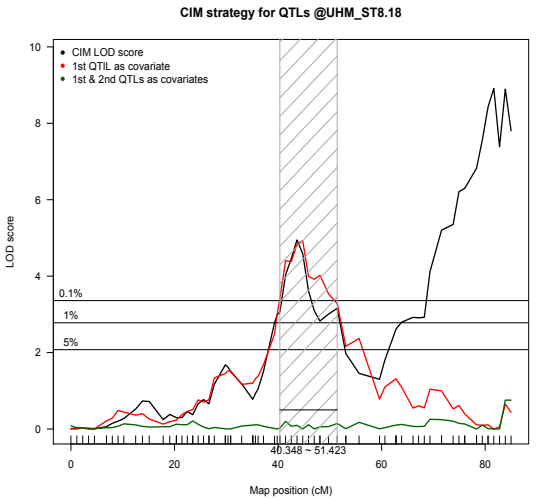
8.1 LOD score plot from SIM method



8.2 Marker regression for the 1st QTL



8.3 CIM strategy for addition QTL scan



8.4 Marker regression for the 2nd QTL

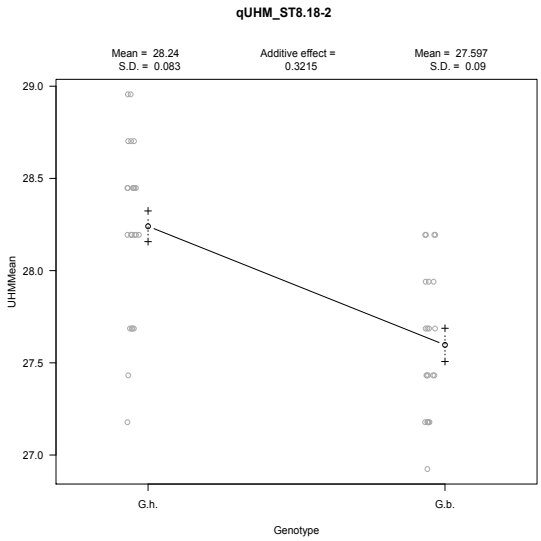
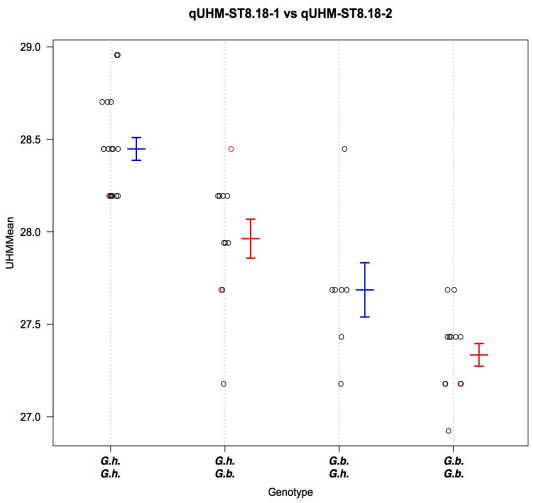


Figure 8. QTL mapping analysis procedure demonstration for experiments. (8.1) Simple interval mapping (SIM) was performed to search for the 1st QTL based on the LOD score using three methods: maximum likelihood estimation (navy blue), Haley-Knott regression (red), and multiple imputations (royal blue). **(8.2)** Phenotype means for genotype groups were plotted at the inferred QTL, and its additive effect was estimated by the half of the difference between the phenotypic means of two genotype groups. **(8.3)** Composite interval mapping was utilized to search for additional QTLs while using the first-detected QTL as covariate. The black line indicates the LOD score from SIM; the red line represented the LOD score from CIM while using the 1st QTL as covariate,

and the green line expresses the LOD score from CIM while using the 1st and 2nd QTLs as covariates. **(8.4)** Phenotype means of inferred QTL-genotype groups were plotted, and the additive effect of the QTL was calculated. **(8.5)** Phenotype means plotted within each genotypic group involving combinations of at two inferred QTLs. **(8.6)** Two-dimensional scans for QTLs interactions, shown as color-encoded graphic display. Lower right triangle: the likelihood estimates between 2 QTLs plus interaction model versus the null QTL model, using the index on the right side of the color scale bar. Upper left triangle: evaluation of full model, including interaction effects between 2 QTLs, versus reduced model with 2 additive QTL effects, using the index on the left side of the color scale bar.

8.5 Interaction evaluation for QTLs



8.6 QTLs interaction scanning

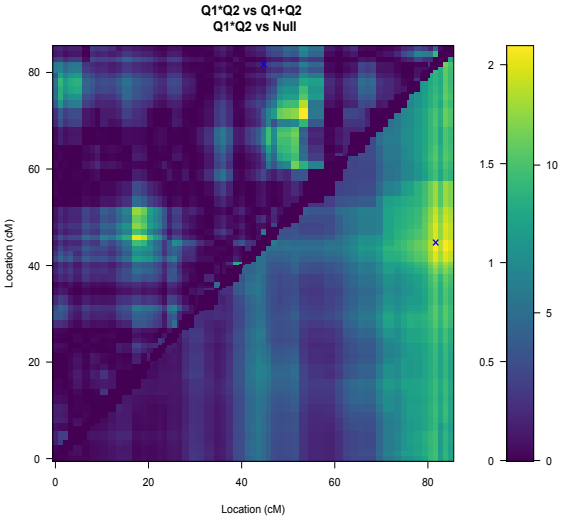


Figure 8. Continued.

Table 5. QTL results summary.

QTLs	*LOD score*	*Position (cM)*	*1.5-LOD support interval*	*Nearest Marker*	*Additive effect* [a]	*R^2 (%)* [b]
qBS-ST8.18-1	3.47**	39.32	27.69~42.56	i53710Gb	0.17	27.84%
qLP-ST8.18-1	13.15****	40.35	39.32~41.45	i24956Gh	1.30	70.95%
qLP-ST8.19-1	11.72****	40.35	39.32~41.45	i24956Gh	1.37	66.75%
qLP-ST8.09-1	14.24****	40.35	39.32~41.45	i24956Gh	1.42	73.77%
qLP-TC17L-1	11.87****	40.35	39.32~41.45	i24956Gh	1.50	67.23%
qLYLDHA-TC17L-1	4.13***	40.35	27.69~49.75	i24956Gh	109.6	32.07%
qUHM-ST8.18-1	8.75****	81.66	78.34~85.00	i61202Gt	0.41	56.72%
qUHM-ST8.18-2	4.66***	44.77	40.35~51.42	i52934Gb	0.32	72.76%
qUHM-ST8.19-1	6.31****	83.89	79.45~85.00	i03664Gh	0.35	43.72%
qUHM-ST8.19-2	3.44**	51.42	39.83~55.64	i03477Gh	0.27	58.01%
qUHM-ST8.09-1	2.99**	69.34	41.45~85.00	i60921Gt	0.27	24.81%
qUI-ST8.18-1	5.54****	39.83	37.19~46.98	i53710Gb	0.37	41.20%
qUI-ST8.19-1	3.93**	40.35	27.69~48.09	i24956Gh	0.24	30.86%
qUI-ST8.09-1	3.52**	40.35	36.15~51.42	i24956Gh	0.31	28.19%
qST-ST8.18-1	4.33***	41.45	32.98~46.98	i44474Gh	3.01	33.61%
qST-ST8.19-1	3.45**	40.35	27.69~53.09	i24956Gh	2.75	27.69%
qST-ST8.09-1	2.72*	35.11	30.85~85.00	i43236Gh	3.50	22.53%
qST-ST8.09-2	2.38***	76.05	73.81~83.89	i60936Gt	-3.35	48.05%
qELO-ST8.18-1	2.99**	36.15	21.43~49.75	i54458Gb	-0.17	24.50%
qELO-ST8.19-1	3.47**	36.15	7.98~48.09	i54458Gb	-0.23	27.84%
qMIC-ST8.18-1	9.55****	37.19	35.11~41.45	i28727Gh	0.29	59.26%
qMIC-ST8.19-1	10.74****	37.19	35.11~41.45	i28727Gh	0.35	63.58%
qMIC-ST8.09-1	12.3****	40.35	39.32~41.45	i24956Gh	0.30	68.52%
qRd-ST8.09-1	3.74***	32.98	20.37~49.75	i28538Gh	-0.68	29.63%
qDI-UCR-1	7.56****	17.82	12.50~23.52	i55318Gb	-0.58	44.69%
qPH-UCR-1	9.09****	17.82	15.12~22.48	i55318Gb	5.56	58.23%
qSW-UCR-1	8.55****	17.82	15.12~22.48	i55318Gb	1.31	53.99%
qVDL-UCR-1	2.87**	17.82	15.12~30.85	i55318Gb	1.28	24.46%
qVDL/PH-UCR-1	9.25****	17.82	12.50~22.48	i55318Gb	-15.85	52.62%
qFDI-UCK-1	5.24***	17.82	15.12~22.48	i55318Gb	-0.56	40.85%
qVRS-UCK-1	4.43**	17.82	15.12~22.48	i55318Gb	-0.57	37.48%

*: LOD score above the 5% threshold; **: LOD score above the 1% threshold;
: LOD score above the 0.1% threshold; *: LOD score above the 0.1% threshold + 1.5 score more.

[a]: Positive sign indicated ***G. hirsutum*** allele contributed higher value for the trait, and negative sign represented ***G. barbadense*** allele contributed higher value.

[b]: If the inferred QTL was the second QTL in the experiment, the variance explained by both 1st and 2nd QTLs was displayed in R^2 column.

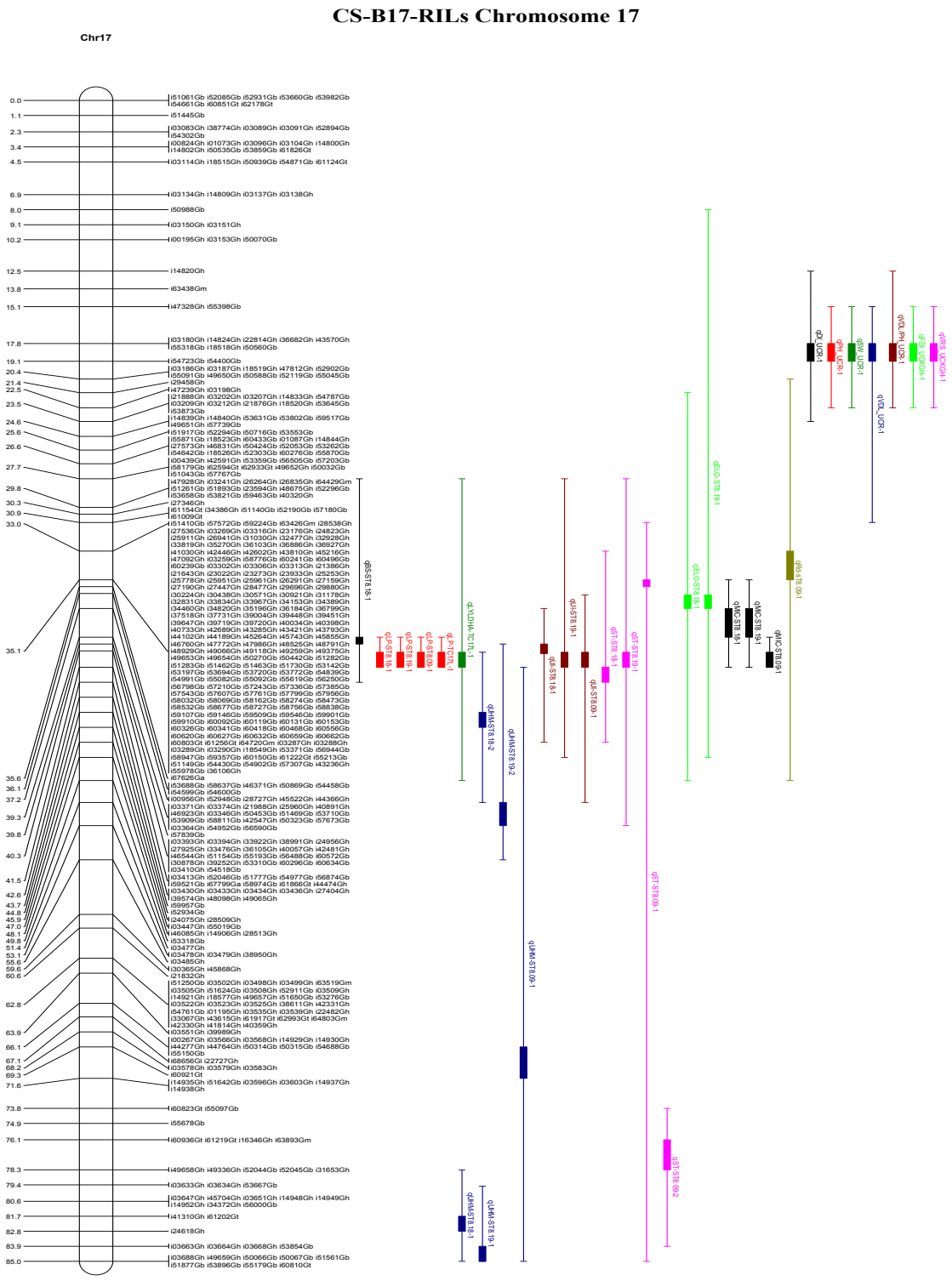


Figure 9. QTL mapping results on chromosome 17.

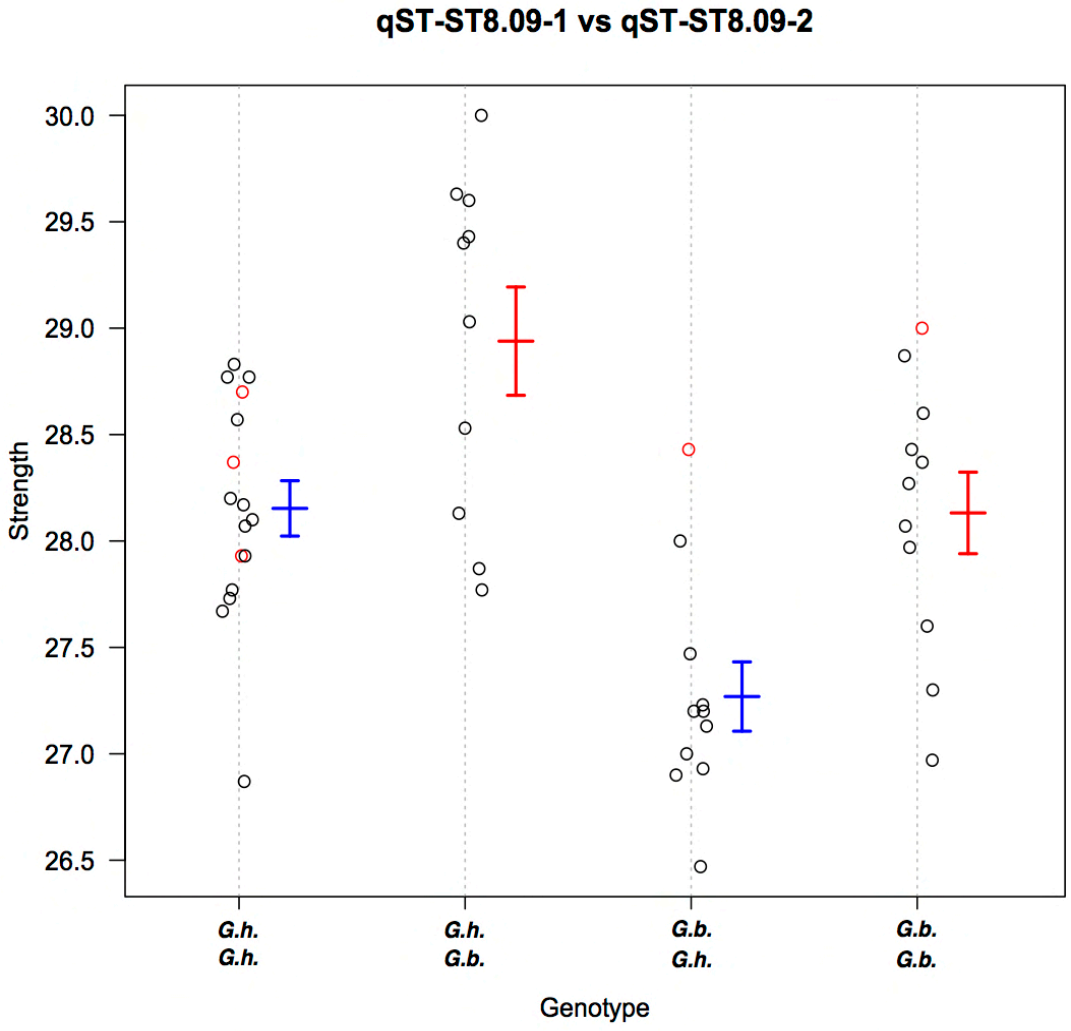


Figure 10. Phenotype mean by genotypic groups on two inferred fiber strength QTLs at ST8.09. In each genotype combination on the X-axis, the upper genotype was qST-ST8.09-1, and the lower was qST-ST8.09-2. *G. hirsutum* allele at qST-ST8.09-1 and the *G. barbadense* allele at ST-ST8.09-2 increased the fiber strength. Therefore, groups with the *G.h.* allele at qST-ST8.09-1 and the *G.b.* allele at qST-ST8.09-2 had the highest phenotypic mean, and the opposite combination had the lowest mean of fiber strength.

Discussion

QTLs comparison

Two major parents, TM-1 and 3-79, are considered as one of the representative lines for domesticated non-phototperiodic *G. hirsutum* and non-photoperiodic forms of *G. barbadense*, and they have been involved for various cotton researches for long time, definitely including fiber traits and pathogen resistance studies (Said *et al.*, 2013). However, previous QTL research that used TM-1 and 3-79 as parental lines for fiber traits and FOV4 resistance seldom reported findings on chromosome 17, compared to other chromosomes (Kohel *et al.*, 2001; Park *et al.*, 2005; Frelichowski *et al.*, 2006; Said *et al.*, 2013; Yu *et al.*, 2014; M. Ulloa *et al.*, 2016a). Only two markers were presumably associated to FOV4 resistance in the early study of CS-B lines evaluation (Ulloa *et al.*, 2013; Ulloa *et al.*, 2016a), and few markers related to fitness with little significance or lack of stability across environment in the studies of using RIL population from TM-1 crossed 3-79 (Frelichowski *et al.*, 2006; Yu *et al.*, 2014). Studies using *G. hirsutum* x *G. barbadense* and *G. hirsutum* x *G. hirsutum* as mapping population were further reviewed, but still not many QTLs have been reported on chromosome 17 (Lopez-Lavalle *et al.*, 2012; P. Wang *et al.*, 2012; Sun *et al.*, 2012; Zhang *et al.*, 2012, 2017; Yu *et al.*, 2013; Wang *et al.*, 2015; Li *et al.*, 2016). QTLs from two recent studies using same interspecific combination were chosen for comparison: one study involved mapping QTL for lint percentage among three different types of backcross populations between *G.hirsutum* and *G. barbadense*, and the other one used chromosome segment substitution lines of *G. barbadense* into *G.hirsutum* for

double-cross hybrid populations for fiber quality QTL mapping (Shi *et al.*, 2015; Zhai *et al.*, 2016). Among all the traits we investigated, lint%, MIC, UHM, and FOV4 resistance had high heritability in this study were selected for comparison because of their high stability and great significance across environments, which probably were the consequence of detecting true genetic effect rather than inflation by environmental effect. Same concept applied in previous study, markers that were identified for the same trait in more than two mapping population were chosen for comparison. Only four lint% QTLs, two MIC QTL, and one FOV4 resistance QTL were qualified. The detailed information of those selected QTLs from previous studies, their position on reference genome, and the information of the flanking SNP markers were all included in **Table 6**.

QTLs from previous studies were identified by SSR markers; therefore, the comparison of the QTL localization between studies must be through the *G. hirsutum* reference genome. The primer sequences were blasted on the JGI *G.hirsutum* reference genome database to determine their positions in the physical map. Among them, qFM-17-7 and qLP-17-7 were both identified by marker NAU2909, and it was aligned to chromosome A02 of the reference genome assembly. The other QTL-associated SSR markers were all detected in chromosome D03 of the reference genome, and their flanking SNP markers were used to infer their plausible positions on our genetic map. According to the position on JGI *G. hirsutum* reference genome, Fov4-C14$_1$ was estimated to be located between markers i18515Gh and i50560Gb, and its position on genetic map was 17.82 cM, which is exactly the same as our FOV4 resistance QTL. The qFM-17-8 QTL identified by NAU2325 SSR marker was surrounded by i49652Gh and

i26264Gh, and its plausible region was deduced within the interval from 27.69 cM to 29.82 cM. It was 7 to 12 cM away from our QTLs, qMIC-ST8.18-1, qMIC-ST8.19-1, and qMIC-ST8.09-1 (**Table 5**). The qLP-17-8 QTL was also identified by marker NAU2325, and there was 10.53 to 12.66 cM distance difference from our Lint% QTL, 40.35cM on genetic map. QTLs qLP-C17-1 and qLP-C17-2 were identified by SSR markers NAU6542 and JESPR195, and their estimated positions on linkage map were 37.19 cM and 39.32 cM respectively, only 3 and 1 cM far from our Lint% QTL. QTL qLP-C17-2 was just located at the lower limit of 1.5 LOD support interval for the Lint% QTL, 39.32 cM to 41.45 cM.

The sign of the additive effects of all QTLs we selected from previous studies indicated the contribution parent for the favorable traits: positive sign represents that *G. hirsutum* allele increases the phenotype; on the contrary, the negative sign of the additive effect indicates *G. barbadense* allele increases the phenotype value. Among the QTLs from early studies, only QTL Fov4-C14$_1$ from RIL mapping population had negative sign, but all the rest QTL had positive sign of additive effect, suggesting the phenotypic value increased by having *G. barbadense* allele on the locus. For lint% and MIC, superior allele was concordantly provided from *G. hirsutum* in this study and previous researches on chromosome 17. For FOV4 resistance, Fov4-C14$_1$ from RIL population, which used same parental lines as this study, was detected with resistance provided from TM-1, *G. hirsutum,* as concluded here. In two other experiments based on F_2 population that involved other parental lines, 3-79 was also reported to carry FOV4 susceptible allele; however, an FOV4 resistance with stronger effect was detected in the *G.*

barbadense cultivar, Pima-S6. A plausible explanation is that either the TM-1 allele exerted only a minor effect and the Pima-S6 allele exerted a major effect, or the strong effect from Pima-S6 masked the nearby TM-1 resistance effect (Ulloa *et al.*, 2013; M. Ulloa *et al.*, 2016a).

When compared to similar studies with different family structure, the mapping results from this study are far more consistent and increased the statistical authority of the detection for FOV4 resistance and lint% QTL. For MIC, there was roughly 12 cM distance difference between QTL in this project and previous research, more experiments cooperated with high-density linkage map are required to verify and determine their actual position (Zhai *et al.*, 2016). However, the difference in distance is small enough to emphasis the true genetic additive effect of QTL in MIC on chromosome 17.

Table 6. Fiber quality and FOV4 resistance QTLs from previous research.

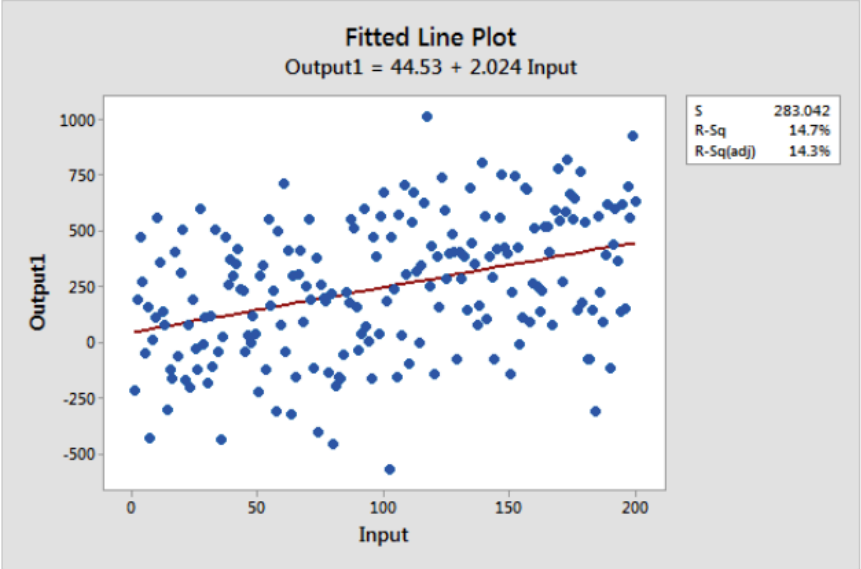
QTLs	*Mapping Popu.*	*Inferred Marker*	*Traits*	*LOD*	*R^2 (%)*	*Add. effect*[a]	*Ref.*	*Chr.*[b]	*Pos.*[c]	*L Marker*[d]	*LM Pos.*[c]	*LM Pos.*[e]	*R Marker*[d]	*RM Pos.*[c]	—
Fov4-C14-1	RIL	MUSS354	FOV4	3.00	18.0	-0.36	(Ulloa *et al., 2013*)	D03	2425986	i18518Gh	2405152	17.82	i50560Gb	2428011	17.82
	F_2		FOV4	5.00	80.0	1.58	(Ulloa *et al., 2013*)								
	F_2		FOV4	11.0	76.0	1.10	(Ulloa *et al., 2013*)								
qFM-17-7	F_2	NAU2909	MIC	22.65	11.69	0.318	(Zhai *et al.* 2016)	A02	62588579						
	$F_{2:3}$		MIC	2.79	5.12	0.101	(Zhai *et al.* 2016)								
qFM-17-8	F_2	NAU2325	MIC	12.8	6.78	0.138	(Zhai *et al.* 2016)	D03	6796228	i49652Gh	6678911	27.69	i26264Gh	6980278	29.82
	$F_{2:3}$		MIC	4.04	7.67	0.118	(Zhai *et al.* 2016)								
qLP-17-7	F_1 (Double-cross)	NAU2909	LP	4.25	7.10	0.93	(Zhai *et al.* 2016)	A02	62588579						
	F_2		LP	26.38	13.5	2.45	(Zhai *et al.* 2016)								
	$F_{2:3}$		LP	2.95	5.28	0.70	(Zhai *et al.* 2016)								
qLP-17-8	F_1 (Double-cross)	NAU2325	LP	2.37	4.40	0.66	(Zhai *et al.* 2016)	D03	6796228	i49652Gh	6678911	27.69	i26264Gh	6980278	29.82
	F_2		LP	15.77	8.30	1.13	(Zhai *et al.* 2016)								
	$F_{2:3}$		LP	6.19	11.57	0.99	(Zhai *et al.* 2016)								
qLP-C17-1	BC_1F_1	NAU6542	LP	10.81	27.51	3.77	(Shi *et al.* 2015)	D03	36488724	i00956Gh	36481017	37.19	i45522Gh	36518545	37.19
	BC_2F_1		LP	5.80	12.44	1.56	(Shi *et al.* 2015)								
	BC_2F_1-NY		LP	3.65	8.61	1.39	(Shi *et al.* 2015)								
qLP-C17-2	BC_1F_1	JESPR195	LP	11.7	26.4	3.70	(Shi *et al.* 2015)	D03	39530706	i03364Gh	39487016	39.32	i56590Gh	39595428	39.32
	BC_1F_1		LP	5.74	12.37	1.53	(Shi *et al.* 2015)								
	BC_2F_1-NY		LP	3.99	8.69	1.40	(Shi *et al.* 2015)								

a: Positive sign indicated *G.hirsutum* increase trait performance, and negative sign indicated *G. barbadense* increased trait performance. b: Chromosome number of the JGI *G. hirsutum* reference genome database where marker sequences were aligned. c: The start position of the marker sequences on the JGI *G. hirsutum* reference genome database. d: SNP chip-based flanking markers of the QTLs inferred markers based on position on reference genome database. e: Linkage map position of SNP markers on chromosome 17.

Another interesting and noteworthy fact was that the phenotypic variance explained by our QTLs (**Table 5**) were much higher than the QTLs detected in normal mapping populations (**Table 6**) (Frelichowski *et al.*, 2006; Lopez-Lavalle *et al.*, 2012; Yu *et al.*, 2014; Zhang *et al.*, 2017). The highest R^2 of a single QTL we detected was 73.77% by qLP-ST8.09-1 and the lowest was 22.53% from qST-ST8.09-1. In QTL analysis, R^2 serves as an indicator to account for the percentage of variation contributed on specific locus (QTL) in the overall phenotypic variation. It can also be used to describe the fitness status of observations to their regression line from the point of view in statistics. **Figure 11** serves to illustrate the difference between high and low R^2, which can facilitate better insight for our analysis. The two datasets depicted in **Figure 11.1** and **Figure 11.2** possessed similar trend, with slope 2.204 in **Figure 11.1** and slope 2.134 in **Figure 11.2**, but differ markedly in R^2, i.e., the degree to which the overall variance among data points in a distribution is accounted for by the respective regression line, i.e., 14.7% versus 86.5%, respectively (Frost, 2014). Data in **Figure 11.1** are clearly scattered farther away from the regression line, whereas **Figure 11.2** they are very close to the regression line. Thus, in a high-R^2 data set, the predictor variable largely determines the response variable, but, in low R^2 data set, the predictor variable in the model explains little of variation of the response variable. Analogously, the higher the R^2 of the QTL(s), the larger proportion of the phenotype performance depends on the additive effect of the QTL(s) in the regression model. From **Table 5**, the QTLs affecting lint%, MIC, UHM, and FOV4 resistance had very high R^2 and also exhibited incredible stability in different environments. This reflects the huge success of removing most

genic interactions and GxE interactions, those involving loci outside chromosome 17, especially for the fiber traits like length and micronaire, which are heavily influenced by environmental effect interactions (Snider *et al.*, 2013). While the CS-RIL strategy precludes opportunities to explore the major QTL effects and epistatsis involving other chromosomes, the utilization of chromosome substitution inbred lines greatly improves the power and accuracy of QTL additive effect detection for the specific chromosome being analyzed by greatly limiting genetic background "noise", which has been considered as one of the major reason for the low level congruence of QTL localization between experiments (Tanksley and Nelson, 1996; Nagata *et al.*, 2015; Wang *et al.*, 2017).

11.1 Data set with low R^2



11.2 Data set with high R^2

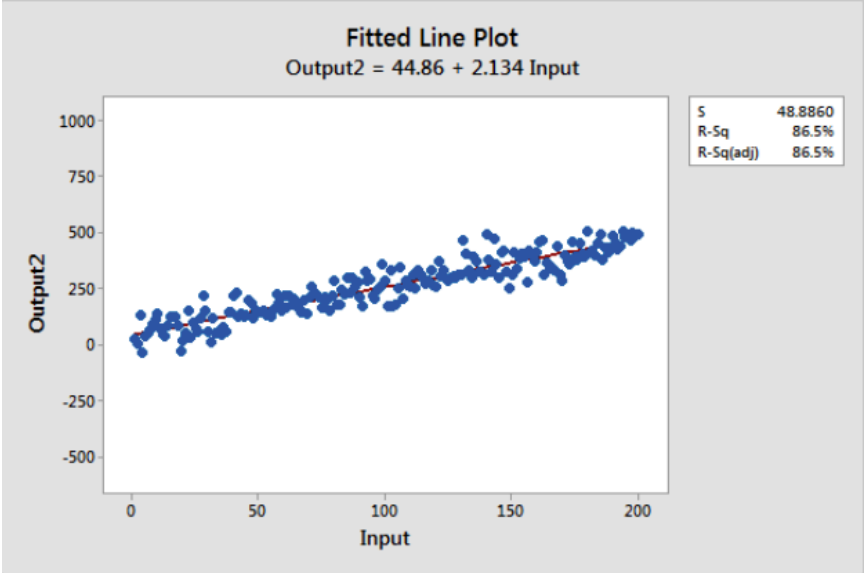


Figure 11. Demonstration of high- and low-R data sets in regression model reprinted from Frost, with permission from Minitab, copyright 2014. (11.1) Data from the low-R^2 data set were scattered much farther from the regression line. (11.2) Data points of the high-R^2 data set occur much closer to the fitted regression line.

One of the main advantages of chromosome substitution (CS) lines, chromosome substitution recombinant inbred lines (CS-RILs), and chromosome segment substitution lines (CSSLs) is that they are nearly isogenic to recurrent backcross parent, and thus nearly isogenic to each other, as well. Collectively, they can provide a powerful platform for integrated genetic/genomics/breeding research. The homogeneity is accentuated if the recurrent parent is highly inbred and homozygous, e.g., as for Upland inbred TM-1, which was derived from a commercial cultivar. These breeding strategies allow for new genetic variation to be introgressed from alien species into elite germplasm of a cultivated species such as Upland cotton while avoiding the severe interspecific hybrid breakdown during the breeding process. The more distant the donor line/species is from Upland cotton, the more helpful and valuable these backcross-mediated introgression approaches tend to be, as heavy phenotypic penalties tend to result from any early-generation inbreeding. *G. hirsutum* is known for its yield potential, and *G. barbadense*, on the contrary, is utilized mainly for its superior fiber quality. A long-standing challenge for cotton breeding has been that fiber yield and fiber quality are negatively correlated most of the time (Wang *et al.*, 2015; Li *et al.*, 2016; Zhai *et al.*, 2016; Zhang *et al.*, 2016). Despite the fact that there would be little interest in using deleterious *G. barbadense* QTL "alleles" for breeding, the information is extremely valuable for background *negative selection* against *G. barbadense* on chromosome 17 in Upland cotton breeding in order to prevent the introgression of undesired allele. Furthermore, knowledge of the location of a QTL is important, to some degree irrespective of the degree to which the QTL controls a trait, and which the alien

allele is beneficial or deleterious. Not uncommonly, the location is important, per se, because in other parental species combinations, the magnitude and direction of the QTL effects can be altogether different.

An interesting point was that the two QTLs for strength observed in ST8.09 were derived from the two opposing parents; the beneficial "allele" for qST-ST8.09-1 came from the *G. hirsutum* recurrent parent, whereas the beneficial "allele" for qST-ST8.09-2 came from the *G. barbadense* donor parent 3-79. Neutralization of the effects from two opposite genetic origin QTLs occurred when using SIM method, and LOD scores for both QTLs were just slightly over the 5% threshold, Figure 12. However, when the first detected QTL was assumed as covariate, the LOD score of the second QTL highly increased greatly, beyond the 0.1% threshold, **Figure 12**. Out of the three experiments in MS, two QTLs for strength were detected only in experiment ST8.09, so more investigations are needed to verify the QTL analysis result. The concept of effect neutralization from two nearby QTLs with opposing effects might partially explain the fact that not many fiber traits QTL have been reported on chromosome 17 from the cross of TM-1 x 3-79 or *G. hirsutum* x *G. barbadense*. However, generally speaking, our results seem to indicate that most fiber QTLs on chromosome 17 involve favorable *G. hirsutum* alleles and have only minor effects. Therefore, in previous studies using normal genome-wide mapping populations like F_2, BC_1F_1, or RIL, chromosome-17 fiber quality QTLs have rarely been detected, inferably because effects of most of the beneficial *G. hirsutum* chromosome-17 fiber QTLs are camouflaged by relatively strong beneficial *G. barbadense* QTLs on other chromosomes. The present study shows,

however, that chromosome-17 QTLs can be exposed by removing most of extraneous genic effects, epistasis, and GxE interactions.

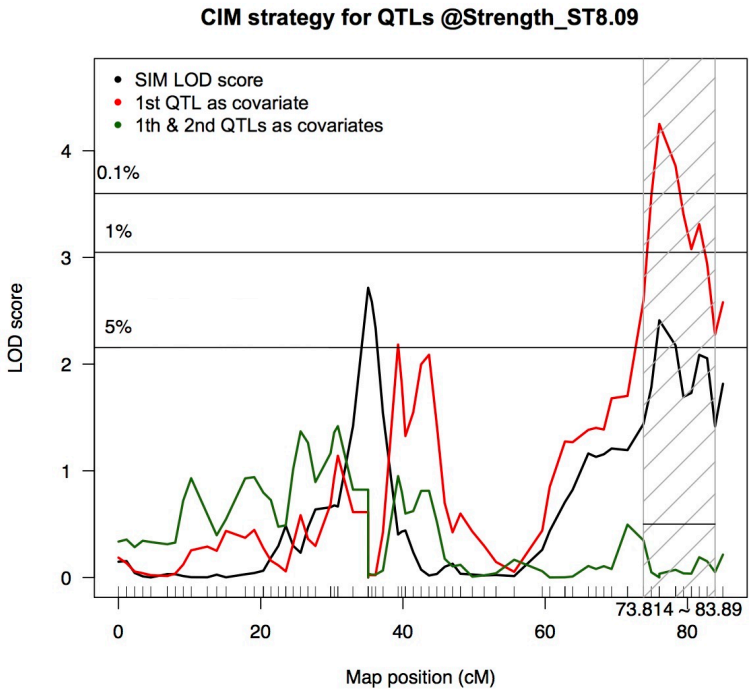


Figure 12. CIM strategy scanning for two opposite genetic background QTLs. Black line: LOD score from the initial QTL scanning using the SIM method revealed a strong QTL in the middle of chromosome 17. Red line: CIM method reveals an additional QTL while using the first QTL as a covariate, resulting in a higher LOD score for a second (end-chromosome) QTL, which exceeded the 0.1% probability threshold. Green line: Same strategy, attempting to detect a significant third QTL while using the two previously identified QTLs as covariates, but none were detected.

CS-B17-RIL60

Initially, the primary impetus for applying the CottonSNP63K to the CS-B17 RIL population was to increase the density and reliability of genotypic calls. Prior

efforts with SSRs alone, with low-density simplex KASP assays alone, and with SSR and SNP data combined, had not resolved what seemed to questionable linkage mapping results and highly erratic QTL LOD score plots. Neither seemed tenable, and a more powerful approach was sought using the recently developed high-density SNP-based genotyping platform. However, results show that the high-density genome-wide genotyping did more much more than increase the linkage map resolution. Using the order of hundreds of c17 loci established by the first round of linkage mapping, we examined the parental contribution to each pair of chromosome 17 of each CS-B17 RIL. CS-B17-RIL60 was immediately flagged as suspicious, because it exhibited exceptional numbers of recombinant events, far higher than any of the other CS-B17 RILs (**Figure 5 & 6**). Further analysis on a genome-wide scale using non-c17 SNPs revealed large numbers of large, unexpected *G. barbadense* segments scattered across the entire genome of CS-B17-RIL60, an observation that also was unique to this CS-B17 RIL. The finding was confirmed by extracting and CottonSNP63K-genotyping another DNA sample from a different individual plant of CS-B17-RIL60 line. Concordance of genotypes across the two samples eliminated hypothetical explanations of contamination from other DNA sample during the wet lab process. It is reasonable to speculate that the lineage of CS-B17-RIL60 was contaminated by open pollen while selfing. The overall amount of non-TM1 loci in the CS-B17-RIL60 genome (~15%) and their low heterozygosity levels, suggest that the contamination (cross pollination) may have occurred in a relatively early generation during its development.

Knowing the flawed composition of CS-B17-RIL60, its removal from the analysis improved the accuracy of both linkage mapping and QTL analysis. The total length of the linkage map and the average interval between markers were thus reduced. More importantly, the LOD score plot from QTL analysis was converted from a highly suspicious abnormal shape with excessive vertical zig-zag, i.e., high-low LOD oscillations, within narrow map distances, to one with a shape far more typical of LOD plots. The influence of including versus excluding CS-B17-RIL60 lines is exemplified by using lint% and UHM in experiment ST8.09. **Table 6** documents the phenotype data of CS-B17-RIL60, the mapping population average, mapping population standard deviation, and the rank of CS-B17-RIL60 among the population for each trait. Out of the 50 lines, CS-B17-RIL60 ranks at the extreme place in many traits, such as boll size, lint%, lint yield, seedcotton yield, uniformity, elongation, and micronaire, which further confirmed the abnormality of this line relative the other CS-B17 RILs. When including CS-B17-RIL60, the total length of linkage map for chromosome 17 extended to 107.3 cM with average interval of 1.08 cM, and the largest gap became 6.58 cM long (**Figure 13**). The effects of CS-B17-RIL60 on QTL analysis is illustrated using lint% and UHM as examples; CS-B17-RIL60 ranks first and twenty-eighth respectively (**Figure 13**). Therefore, the influence caused by CS-B17-RIL60 is more severe in lint% than UHM. From **Figure 13.1**, the zigzag shape can be observed, especially near the highest peak of the plot. Comparison of the LOD line with CS-B17-RIL60 genotype revealed that the dramatic dips in the score occur at map positions where CS-B17-RIL60 has *G. barbadense* allele at the locus. Whereas higher lint% was generally associated with the

TM-1 SNPs near the TM-1 QTL Lint%_ST8.09 “allele”, the opposite effect was imposed by CS-B17-RIL60, because its extra high lint% values (highest among the CS-B17 RILs) were uniquely associated the *G. barbadense* Lint%_ST8.09 “alleles”. Thus, only at the loci where CS-B17-RIL60 exhibited a *G. barbadense* Lint%_ST8.09 “allele”, the phenotype differences between genotypes was markedly reduced, so the likelihood of having QTL was also decreased, which resulted the zigzag shape of the plot LOD scores relative to map position.

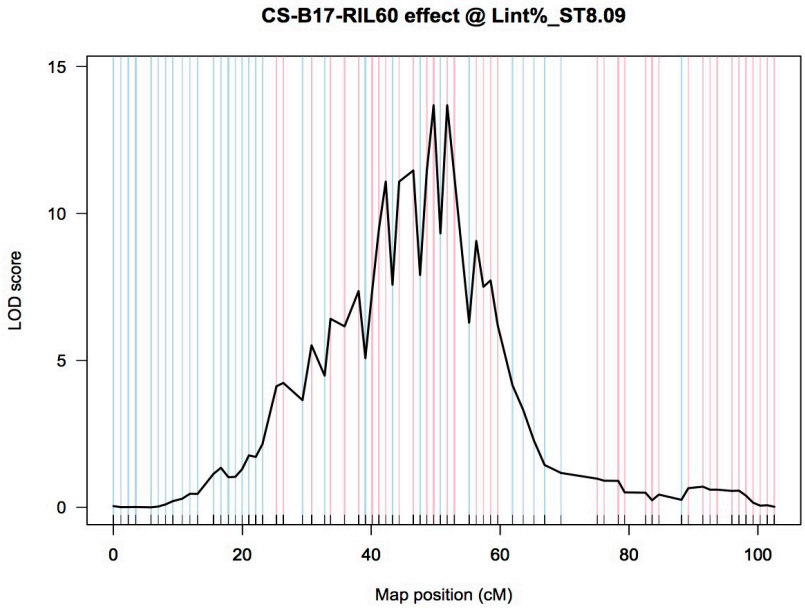
The rationale above can be extended to the other traits, provided ample consideration is given to the relative performance of CS-B17-RIL60 versus other CS-B17 RILs. For example, the UHM phenotype of CS-B17-RIL60 was intermediate, therefore, the phenotypic mean difference between genotype groups did not change too much at the loci where CS-B17-RIL60 has *G. barbadense* allele, so LOD scores were little affected and the plot of LOD scores did not exhibit a such an extreme zigzag shape (**Figure 13.2)**. Fewer dramatic high-low LOD oscillations are observed and the oscillations are not correlated to the genotype of CS-B17-RIL60 anymore. These findings indicate that development of accurate high-throughput genotyping technology in cotton not only increase the accuracy of QTL analysis by greater abundance and density of polymorphic markers, but also by empowering thorough genome-wide genomic examinations of the research plant materials.

Table 7. CS-B17-RIL60 and population phenotype statistics in experiment ST8.09

Traits	*CS-B17-RIL60 mean*	*Population average*	*Population S.D.*	*Phenotypic Rank*[a]
Boll size	3.14	5.04	0.471	50
Lint %	34.92	31.23	1.727	1
YLDHA	370	728.63	205.470	49
LYLDHA	129	227.43	66.379	48
UHM	28.448	28.562	0.54393	28
UI	82.57	83.59	0.598	49
Strength	273.879	275.602	7.9511	29
ELO	6.33	7.23	0.255	50
MIC	3.23	3.73	0.363	48
Rd	75.73	76.04	1.248	33
+b	8.77	9.04	0.412	39

[a] : Ranking from high to low value.

13.1 LOD score plot of lint% in experiment ST8.09.



13.2 LOD score plot of UHM in experiment ST8.09.

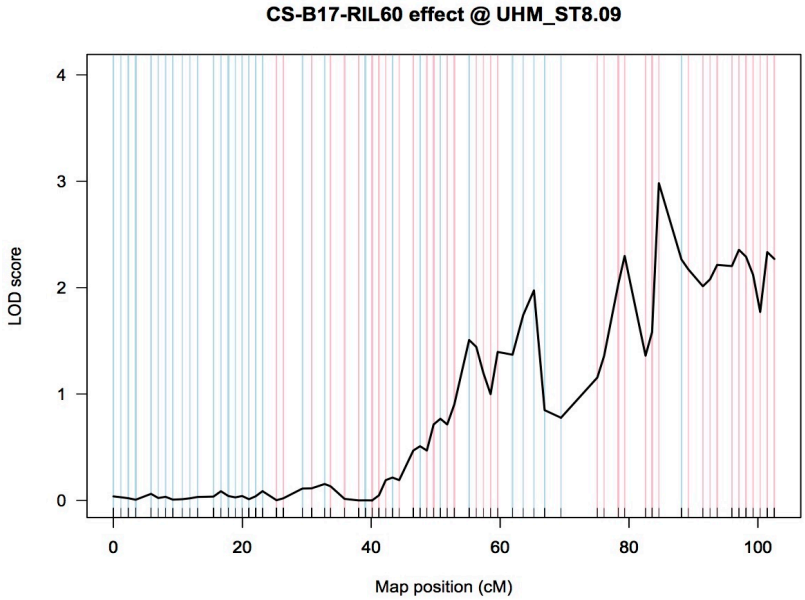


Figure 13. CS-B17-RIL60 effect on QTL analysis in lint% and UHM. Black lines: LOD scores for Lint% (Figure 13.1) and UHM (Figure 13.2) based on SIM. Color of vertical lines indicates the genotype call (parental assignment) at the each SNP locus of CS-B17-RIL60. Genotype A, allele from *G. hirsutum*, was identified as pink line, and Genotype B, allele from *G. barbadense*, was identified using light blue color.

CHAPTER III

CONSTRUCTION OF AN INTERSPECIFIC LINKAGE MAP BETWEEN UPLAND COTTON (*GOSSYPIUM HIRSUTUM* L. $(AD)_1$) AND *GOSSYPIUM MUSTELINUM* MIERS EX WATT $(AD)_4$

Introduction

Seven tetraploid species and forty-five diploid species of cotton have been identified in *Gossypium* genus, but only two domesticated tetraploids, *G. hirsutum and G. barbadense*, are used extensively in research and production due to their spinnable fiber (Wendel and Grover, 2015). Several factors have contributed to heavy reliance on a narrow gene pool for cotton production and elite cultivar breeding. Serious genetic incompatibilities cause varying degrees of sterility, distorted segregation and weakness tend to occur in interspecific hybrids; these have tended to dissuade breeders from widespread use of inter-specific research and breeding (Stephens, 1949; McKenzie, 1970; Reinisch *et al.*, 1994). Even worse, the extreme difficulty in recovering an agriculturally elite product involving interspecific germplasm has led to a heavy reliance on the closely related elite lines for genetic improvement of Upland cotton; this has greatly narrowed its genetic diversity pool, and made it difficult to create new allelic combinations that substantially improve performance, thus the rate of trait progress of cotton cultivars has slowed (May, Bowman and Calhoun, 1995; Paterson *et al.*, 2004). To sustain long-term improvement, new germplasm is essential and critical. Several technologies seem poised to help, especially high-throughput phenomics, genomics and

related biotechnologies, including mutagenesis, RNAi, gene editing. New molecular technologies that lead to and enable marker-assisted selection and genome selection hold great promise for improving research on and methods for exotic *Gossypium* germplasm usage to increase Upland cotton diversity and sustainability of cotton as a natural fiber crop.

Introgression of superior fiber quality from *G. barbadense* germplasm into Upland cotton has been investigated for several decades, but few studies have focused on the other wild tetraploid cotton species. Among them, *G. tomentosum* and *G. darwinii* recently attracted people's attention since they are closely related to domesticated species, *G. hirsutum* and *G. barbadense* respectively (Grover *et al.*, 2015). High-density linkage maps of them crossed by *G. hirsutum* have been constructed (Hou *et al.*, 2013; Chen *et al.*, 2015), and several valuable QTL related to fiber quality, like elongation, uniformity, strength, have been identified and been used into Upland cotton improvement (Zhang *et al.*, 2011; B. Wang *et al.*, 2012). As the earliest phylogenetic branch alone among 52-chromosome AD-genome cotton species, *G. mustelinum* was the first to evolutionarily separate from all six other extant tetraploid species, and it is the AD-genome species most genetically divergent from *G. hirsutum*. Although *G. mustelinum* clearly harbors potentially valuable traits, e.g., pest-resistance chemistry (Khan, Stewart and Murphy, 1999), it has be subjected to far less research than some other AD species, perhaps for reasons mentioned above (Wendel, Rowley and Stewart, 1994).

First documented by Watt (1907), *G. mustelinum* is known to occur naturally in the small area of northeast Brazil, where faces an extreme arid environment (Wendel and Grover, 2015). Including recent reports, several populations have been discovered, typically growing along waterways (Alves *et al.*, 2013; Menezes *et al.*, 2014). Given its natural environment, its reasonable to deduce that drought tolerance traits might be valuable derivatives from introgression into Upland cotton. Wang *et al*. (2016) constructed a *G. hirsutum* x *G. mustelinum* SSR and RFLP linkage map with 1055 loci across 26 chromosomes and 5595 cM in length. Cooperated with the mapping information, QTL analyses on the backcross populations were further performed, and several QTLs that increased fiber elongation, upper half mean length, and uniformity were detected and actually beneficial from *G. mustelinum* allele (Wang *et al.*, 2016, 2017). One QTL, which *G. mustelinum* allele decreased the short fiber content, was also identified in the research. Despite *G. mustelinum* does not possess spinnable fiber, it still can be used as valuable genetic resources for Upland cotton genetic improvement (Gardunia, 2006; Xu, 2014).

Linkage maps are a fundamental resource and can foster many sorts of contemporary research, including germplasm characterization, genomic analysis, and diversity analysis. To date, a robust and high-density linkage map of SNPs between the wild Brazilian 52-chromosome cotton species *G. mustelinum* and *G. hirsutum* is still lacking but greatly needed to conduct in-house and other global breeding efforts involving the germplasm resource. Therefore, the main goal of this study is to create such a linkage map on a BC_1F_1 population using high-throughput genotyping, the

CottonSNP63K Array. Additionally, minor goals will be to relate the new map to important resources are now available that were not available at the outset of this effort, including an independently created SSR-based map that also involves *G. mustelinum* and is fairly extensive, a high-density inter-specific *G. hirsutum* - *G. barbadense* SNP-based map, and a latest released version of the Upland cotton genome sequence assembly (Hulse-Kemp *et al.*, 2015; Wang *et al.*, 2016; Saski *et al.*, 2017, in press).

Material and Methods

Plant materials

The *G. hirsutum* inbred "TM-1" was crossed as the ovule parent with a *G. mustelinum* tree grown for multiple years in a soil "pit" in the Cotton Cytogenetics greenhouse on the Texas A&M University campus, when it flowered briefly in 1988 to generate F1 seeds in the greenhouse of new Beasley lab. Three different F1 plants were grown in 2002 and reciprocally backcrossed to recurrent parent, TM-1, to generate BC_1F_1 seeds in the same greenhouse. A total of 59 BC_1F_1 seeds were randomly selected and grown in the field on FnB road, College Station, TX in 2012 and 2014. Very young unfurled leaves of each plant were place into 2 ml tubes for DNA extraction using NucleoSpin® Plant II genomic DNA extraction kit for plant and fungi (Macherey-Nagel, Duren, Germany).

Genotyping

DNA concentrations of the 59 BC_1F_1 plants, two parents, and F_1, were determined via Nanodrop Spectrophotometer (Thermo Fisher Scientific, Waltham, USA) and standardized at 50 ng/µl for genotyping using the CottonSNP63K array at Texas A&M University according to Illumina protocols. After the single-base extension, the chip was scanned by Illumina iScan to generate the image files, which were then saved in GenomeStudio software to determine the genotype call of each SNP according to the cluster file developed for tetraploid cotton genotyping (Hulse-Kemp et al., 2015) (available at http://www.cottongen.org/node/add/cotton-cluster-file-request). Genotype data of the 63,058 SNP markers across 59 individuals were transformed into "ABH" format based on the genotype call of two parents and F_1 plant. Only co-dominant makers that expressed different homozygous genotype calls between two parents and heterozygous call in F_1 plant were retained.

Linkage mapping analysis

After filtering out markers that had genotype calls less than 95% across 59 individuals, only 15,914 markers remained for the linkage mapping analysis. The 15,914 markers were then categorized into groups based on their genotype pattern across mapping population. To increase the computational efficiency of the linkage analysis, only one randomly selected marker from each marker-pattern group was used. Since the inclusion of missing data commonly affects mapping results, e.g. by causing mis-grouping of markers into bins or even the reversal of neighboring markers, the function

"FindDup" of the "R" program package "qtl" was applied for marker categorization; the resulting missing data were considered as genotypes, as a means to avoid subsequent complications and erroneous inferences during further steps (Waghmare *et al.*, 2005). A total of 3,032 marker bins were created and one marker from every bin was selected for mapping using the "Onemap" package (Margarido, Souza and Garcia, 2007) in R.

A LOD score of 6 was chosen as the minimum grouping criteria for dividing markers into 26 major linkage groups. The *G. barbadense* x *G. hirsutum* genetic map of CottonSNP63K SNP markers (Hulse-Kemp *et al.*, 2015) was used as preliminary information to associate linkage groups with chromosomes. Third, within each linkage group, all markers were included for the first linkage analysis using three different ordering algorithms: recombination counting and ordering algorithms (RECORD) (Hans Van Os, Piet Stam, 2005), rapid chain delineation (RCD) (Doerge and Weir, 1994), and unidirectional growth (UG) (Tan and Fu, 2006). Interval distances between markers were estimated using Kosambi map function (Kosambi, 1944). Linkage map construction has been recognized as a "traveling salesman problem" (TSP), i.e., the exact answer can only be found if all the possible combinations were listed and compared (Hoffman, Padberg and Rinaldi, 2013). Current ordering algorithms merely lead to plausible approximations to the correct answer, and it is necessary to repeatedly run a mapping analysis to obtain the closest answer. The initial mapping analysis was repeated thirty times, and only the result having highest likelihood score was retained.

Among the three algorithms used in this project, the RECORD method generated results with the highest likelihood score, shortest length and best stability, so only

RECORD was applied in the further steps (**Figure 14**). Linkage maps and recombination fraction plots (**Figure 15** & **16**), served as references to guide subsequent marker elimination in subsequent procedures of the analysis, specifically during the fifth step. Fourth, by progressive increasing the LOD score used as a threshold for marker grouping, it was possible to detect a LOD level for which a linkage group remained intact, but above which few markers would be split away from the linkage group. The highest LOD score associated with each intact linkage group was recorded. Fifth, the markers that split out from a linkage group at higher LOD scores were then examined to determine whether they should be omitted, kept, or placed in the a different linkage groups; factors in this decision included their distance from adjacent markers in the linkage map, the recombination fraction with other markers -- using plots by the "plotRF" function, as well as their genotype calls.

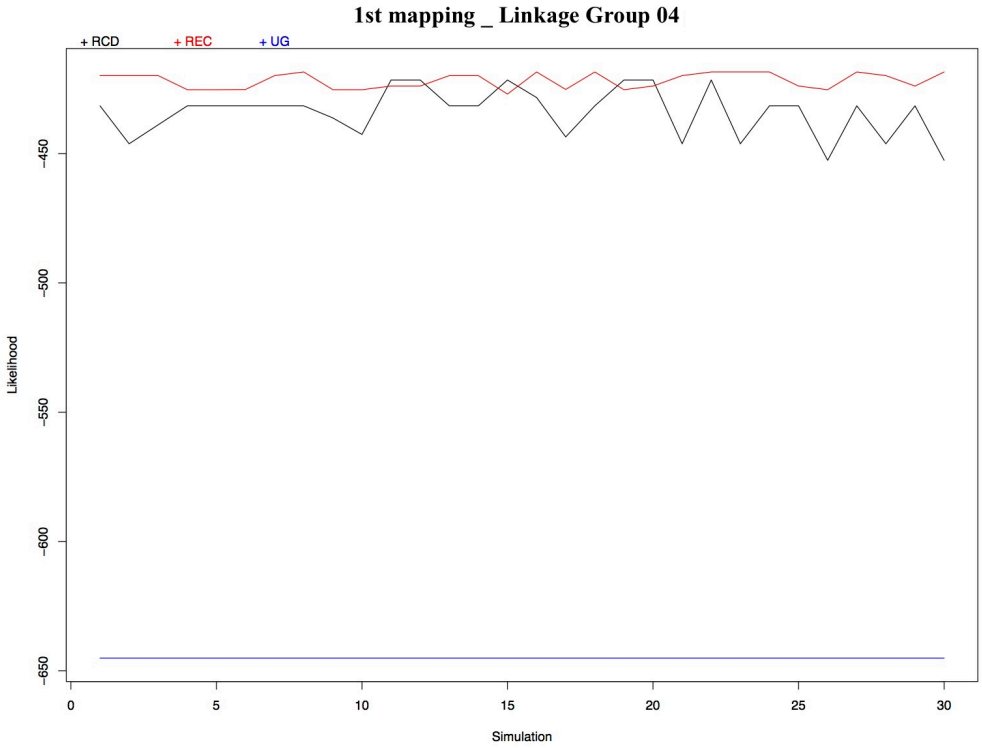


Figure 14. Likelihood scores after each of 30 cycles from three algorithms on linkage group 04. Black line: RCD method. Red line: RECORD method. Blue line: UG algorithm.

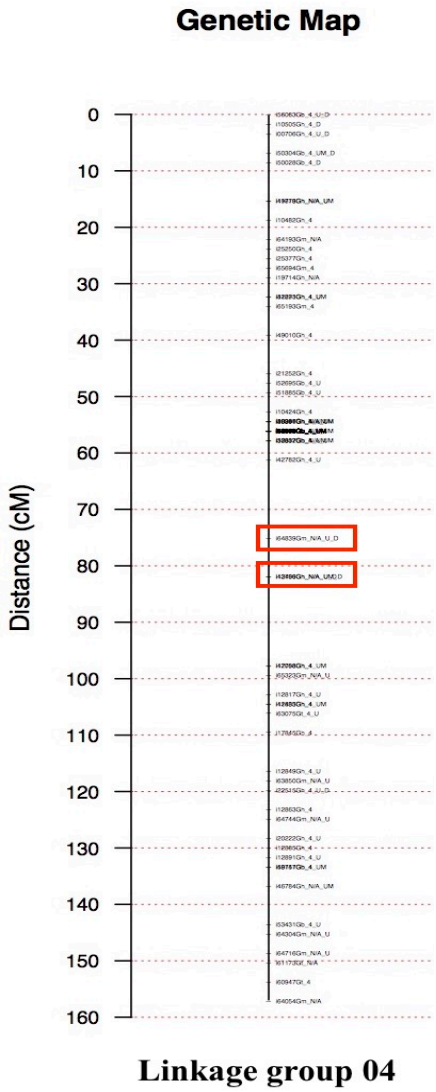


Figure 15. Example of initial mapping results from RECORD for linkage group 04. Red boxes denote two suspicious markers, which were far apart from their adjacent markers, suggesting that they might not belong to this linkage group.

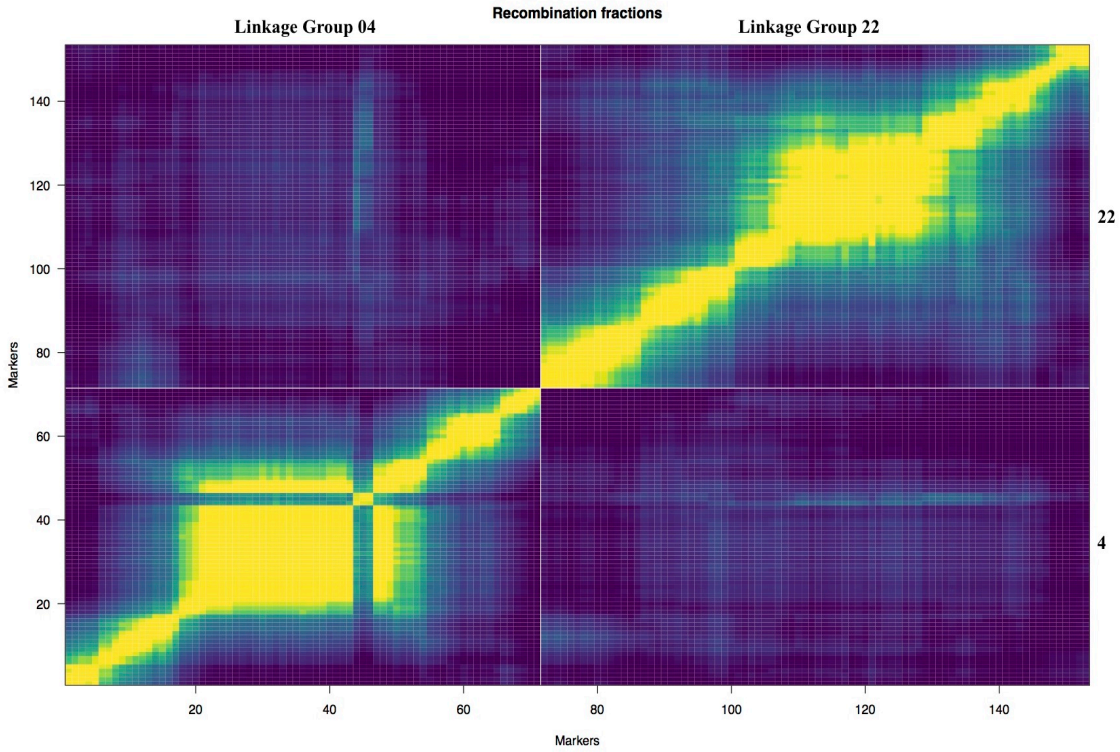


Figure 16. Recombination fraction plots, illustrated for two chromosomes (4 and 22). Light color indicates high likelihood of linkage, whereas dark color indicates low likelihood of linkage. Three markers in the middle of linkage group 04 showed tight linkage to each other, but high recombination rates with other markers in linkage group 04. Moreover, the same loci had slight association with markers in linkage group 22.

After eliminating suspicious markers from the previous step, the remaining 2,951 bin markers were mapped again using RECORD for ordering and Kosambi function for estimating distance; the analysis was repeated 100 times. Only the mapping results with best highest likelihood score were retained. As described for the initial mapping results, results from the second round of mapping were analyzed similarly: the linkage map, recombination fraction plot, and raw genotype data were generated to verify no additional markers should be removed.

Missing data often affects linkage mapping analysis and QTL analysis (Jiang and Zeng, 1997; Waghmare *et al.*, 2005). However, the effects depend on context. In general, two markers would be closely arranged together with distance less than 0.01 cM if a missing data did not occur at the plausible recombinant event breakpoint. Conversely, if a missing data point happened to occur at a seeming crossing over point two markers would be separated as two loci (**Supplemental Figure S2**). To differentiate the types of missing data above and to ameliorate the quality of the genetic map influenced by missing data, a more conservative approach was chosen in this project, where instances of missing data were considered as a genotype when categorizing markers into bin groups at the initial step. A negative ramification inherent associated with this approach was that as more markers were selected for analysis with limited availability of iterations, it became more likely to falsely separate markers that shared an otherwise identical genotype pattern. However, this few iterated mapping result could still aid to determine the influence caused by a missing data on mapping analysis. Markers with missing data that does not affect the analysis would very close to one of their adjacent bins, which suggests that the markers should be combined into the bin group. As the number of bin group decreases, the availability of iterations can increase, and a better quality of genetic map can be obtained. The above processes can be repeated multiple times to refine the genetic map when dealing with missing genotypic data.

Results from the second mapping with RECORD (100 repeats) was applied to the entire SNP marker set according to the bin group relationships established in step 1.

The dot plots (**Figure 17**) were made by aligning sequences of all linkage map markers to the JGI *G. hirsutum* physical map and sequence assembly (Saski *et al.*, 2017, in press). Heat maps (**Figure 18**) for linkage disequilibrium were plotted from the CheckMatrix software (http://www.atgc.org/XLinkage/). Dot plots and heat map were used for marker order examination within each linkage group. The example in **Figure 17** exemplifies situations where markers with a similar genotype pattern across population could be placed too far apart, 72 cM and 82 cM, simply because of the order that markers were entered into analysis. To solve this problem, similar mapping procedures were conducted again: step 1 bin grouping, step 2 ordering, and step 3 examination. Since some markers were placed at extremely close positions based on the second-round mapping result, we were able to re-grouped the 2nd set of 2,951 bins into the 3rd set of 1,806 bins. A set of 1,806 bin-specific markers was then used for a third round analysis, including 1,000 repeated analyses with RECORD, while using Kosambi function. Repetitive procedures could be conducted again with higher number of repeats and equal or smaller number of bin groups if missing data were determined not to affect the analysis until linkage map results became fine and stable. Once the final map was achieved, high quality linkage map for 26 groups were drawn by using MapChart software (Voorrips, 2002). The finalized mapping results are listed in the **Supplemental Table S3.**

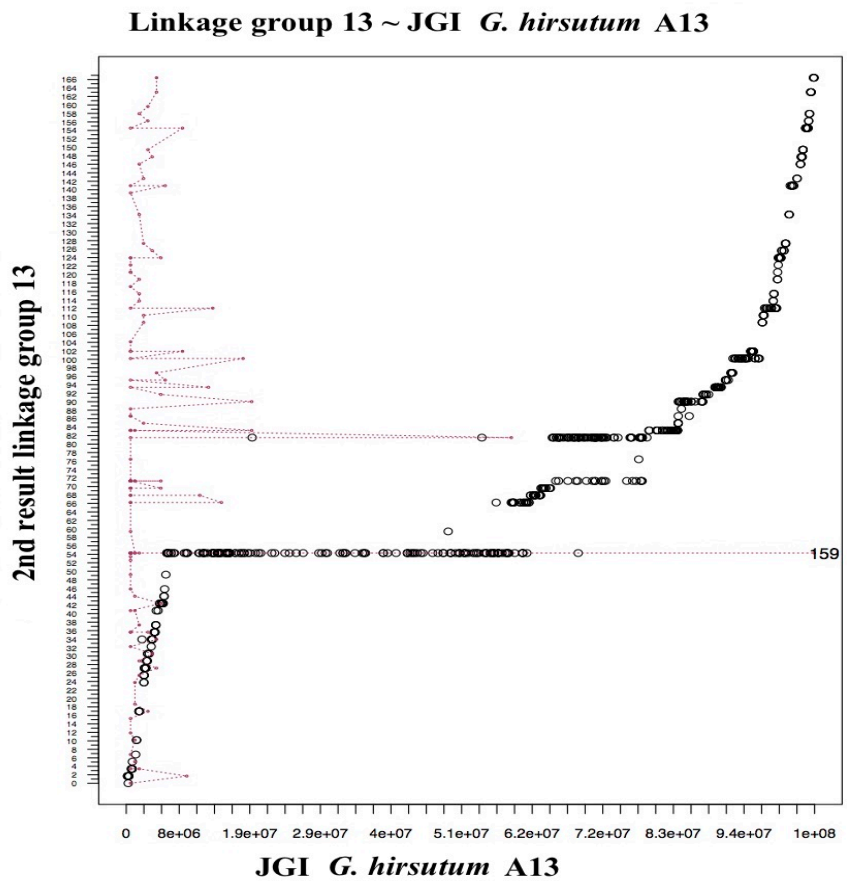


Figure 17. Dot plot of marker positions on linkage group 13 associated with physical map positions in the *G. hirsutum* genome assembly (Saski *et al.*, 2017, in press) posted at JGI, based on sequence alignments. Markers in the middle segment were linkage mapped to different positions than markers above them, but they have similar positions on physical map (i.e., in middle group versus upper group). The clear separation on linkage map was caused by missing data and the mapping procedure setting. The maroon line indicates the numbers of markers in each bin group.

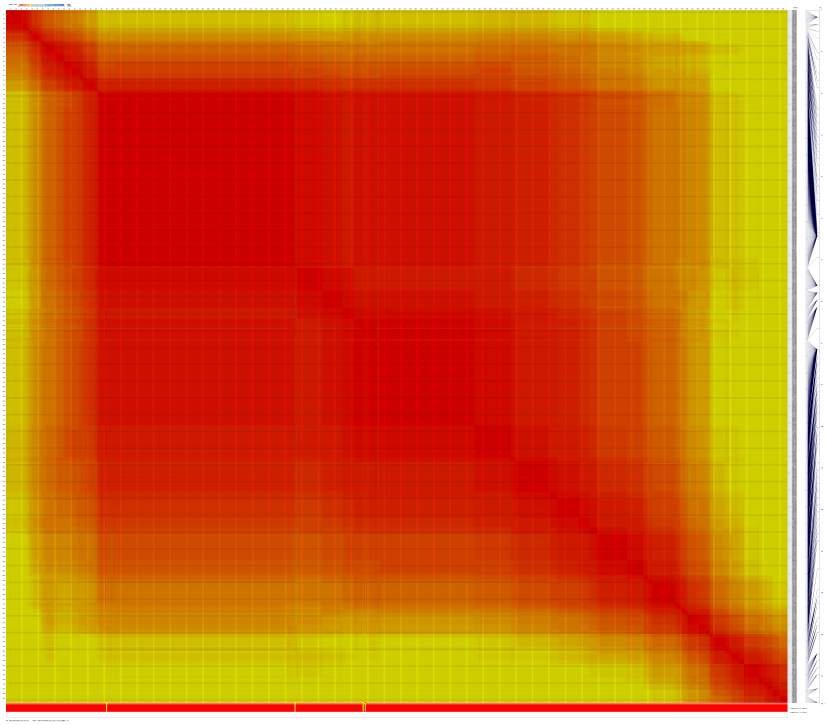


Figure 18. 2D heat map of linkage group 13 using CheckMatrix software for linkage disequilibrium examination. Red denotes markers linked together, whereas yellow denotes absence of linkage.

Synteny analyses

Sequences of all SNP markers were aligned to the *G. hirsutum* (AD_1) reference genome (Saski *et al.*, 2017, in press) using BLAST+ command line under UNIX system with default parameter settings (Altschul *et al.*, 1990; Camacho *et al.*, 2009). The BLAST hit with highest BIT score and lowest e-value was selected for each marker. *G. hirsutum* x *G. mustelinum* linkage map positions against *G. hirsutum* (AD_1) reference genome alignment were plotted using normal default plotting function in R software. Detailed alignment information between linkage groups and reference genome were depicted by using the R package "RCircos."

Linkage map comparisons

The final version of our high-density *G. hirsutum* x *G. mustelinum* Cotton64K-based SNP linkage map was aligned to the recently reported high-density Cotton64K-based *G. hirsutum* x *G. barbadense* interspecific linkage map (Hulse-Kemp *et al.*, 2015) based on common loci, and then compared in terms of map length, bin groups, average interval, marker distribution, and marker correlation between maps. An alignment was also made between our linkage map and a recently reported SSR *G. hirsutum* x *G. mustelinum* linkage map (Wang *et al.*, 2016), based on relative alignments to the recently released cotton genome assembly (Saski et al. 2017, in press), which was used as a common reference for synteny analysis and relative marker ordering. Basic linkage map characteristics were then compared.

Results

Linkage map construction

A new linkage map was developed by CottonSNP63K-based genotyping of a population of 59 BC_1F_1 individuals from the cross between *G. hirsutum* inbred "TM-1" and *G. mustelinum*, the parents and F1 hybrid. A total of 15,825 SNP markers were mapped and represented by 1,776 bins, re-grouped from the 1,806 bins after the third around mapping analysis, with overall length of 4193.82 cM (**Figure 19**). In average, the interval between bins was 2.39 cM with approximately 8.9 markers per bin. According to the origins of the 70,000 SNPs used to populate the CottonSNP63K array (Hulse-Kemp et al., 2015), the 15,825 SNP markers mapped here included 9,939 SNP

markers (62%) came from the CottonSNP63K *G. hirsutum* intraspecific data set (50,000 SNPs), 1,841 SNP markers (11%) contributed from the CottonSNP63K *G. mustelinum*-derived set of 4,758 SNPs, and the remaining 4,045 markers were from other cotton species used to populate the CottonSNP63K, mostly from the *G. barbadense*-derived set of 5,233 SNPs (**Figure 20**).

The linkage groups containing the most markers correspond to chromosomes 19, 8, and 5. Maps of these linkage groups were also the three longest, i.e., most recombinant, with 247, 214 and 240 cM, respectively. However, SNPs/cM ranged from 2.37 to 5.42, so the marker counts were not always associated with map lengths. For example, nearly 800 markers were mapped on both chromosome 24 and 13, but the map lengths of these two linkage groups were merely 166 cM and 146 cM respectively. Conversely, chromosome 11 and 21 had the next longest map length after the three, 200 cM and 197 cM, but only 680 and 476 markers were mapped on them respectively.

No significant difference between two sub-genomes was found in terms of marker distribution, average interval, markers per bin, and so on. A total of 7,764 markers were mapped to the A subgenome, with average map length of 165.8 cM per linkage group, an average interval of 2.45 cM between bins, and 8.7 markers per bin. For comparison, 8,061 markers were mapped to the D subgenome, and the average linkage group length was about 6% shorter on average, 156.8 cM. In the D-subgenome, the average interval between bins was 2.33 cM and 9.1 markers were found per bin.

Among the 15,825 SNP markers mapped in this analysis of BC1F1, 659 markers (4.16%) showed significant segregation distortion, p-value < 0.05 in chi-square test, and

199 of them were mapped on A sub-genome and 460 on D sub-genome (**Figure 19**). The markers that underwent significantly distorted segregation were non-randomly distributed among chromosomes, e.g., 225 (37%) were in chromosome 26, 120 (16.8%) in chromosome 14, and 79 (15.7%) in chromosome 09. Segregation distortion also occurred for markers in chromosome 24, 08, 16, 10, and 11. The distortions favored the wild species parent, in that 194 SNP markers were skewed toward *G. hirsutum* and the recovery of homozygous backcross products; these loci were in chromosomes 24, 08, 16, 10, and 11, with average A/H genotype ratio of 1.81, 1.85, 1.83, 1.99, and 1.54 respectively, versus the expectation ratio of 1. On the other hand, 465 markers (70.6%) in chromosomes 26, 14, 09, or 05 tended to remain heterozygous, with average A/H ratios of 0.47, 0.54, 0.52, and 0.44, versus the expectation ratio of 1. However, the effect and reason of segregation distortion are still unclear. Further research is needed in the future.

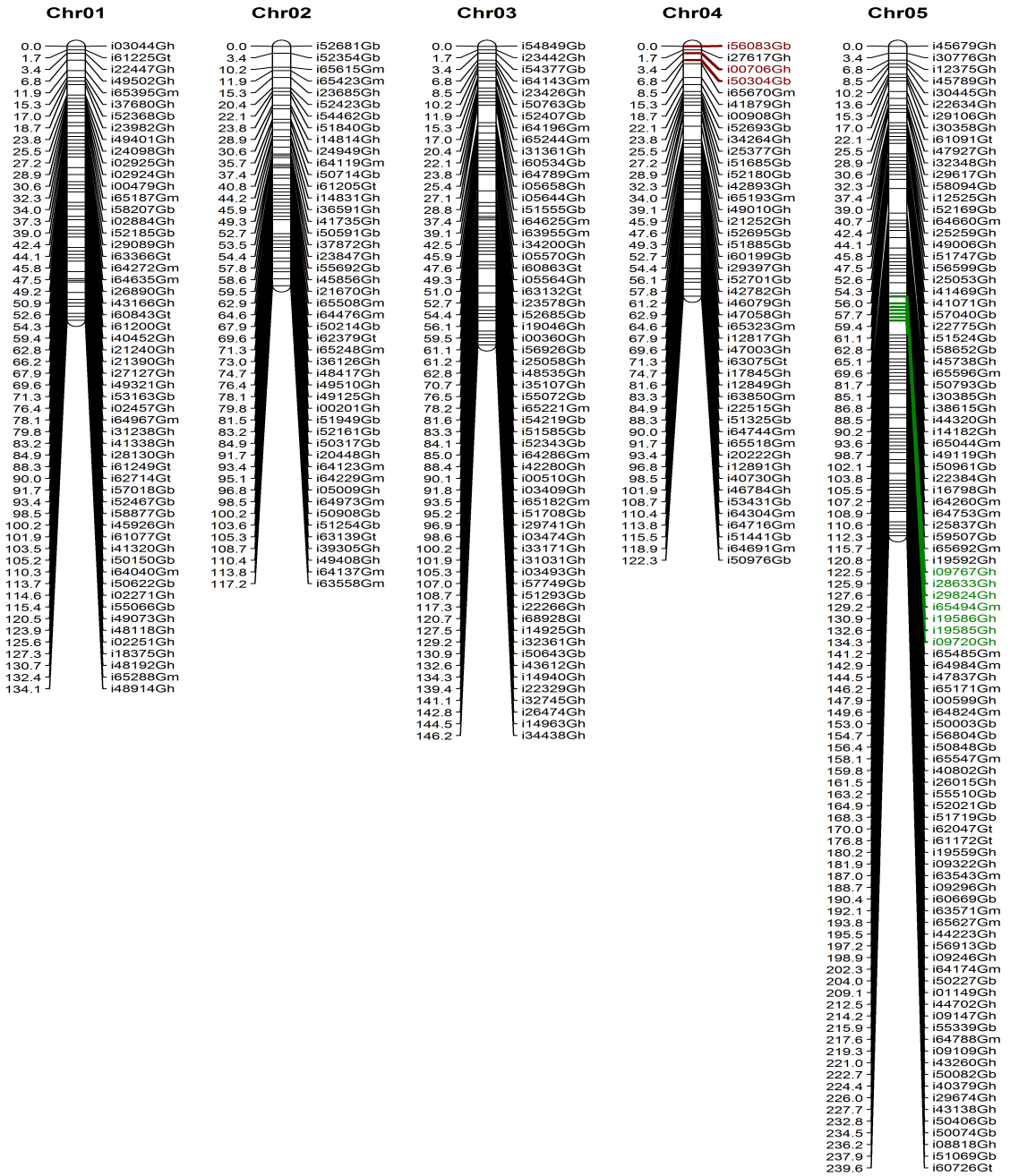


Figure 19. Linkage maps of 26 chromosomes based on linkage analysis of 59 BC1F1 from *Gossypium hirsutum* 'TM-1' x (*G. hirsutum* 'TM-1' x *G. mustelinum*). The right column of text list CottonSNP63K marker identifiers, one per bin group and collectively representing all 1,776 loci (marker bins) across the entire genome. The left column of numbers indicates the calculated map position in centiMorgan (cM). Markers showing segregation distortions are colored: loci favoring transmission/recovery of the *G. hirsutum* allele are maroon-colored, whereas those favoring the *G. mustelinum* allele are green-colored.

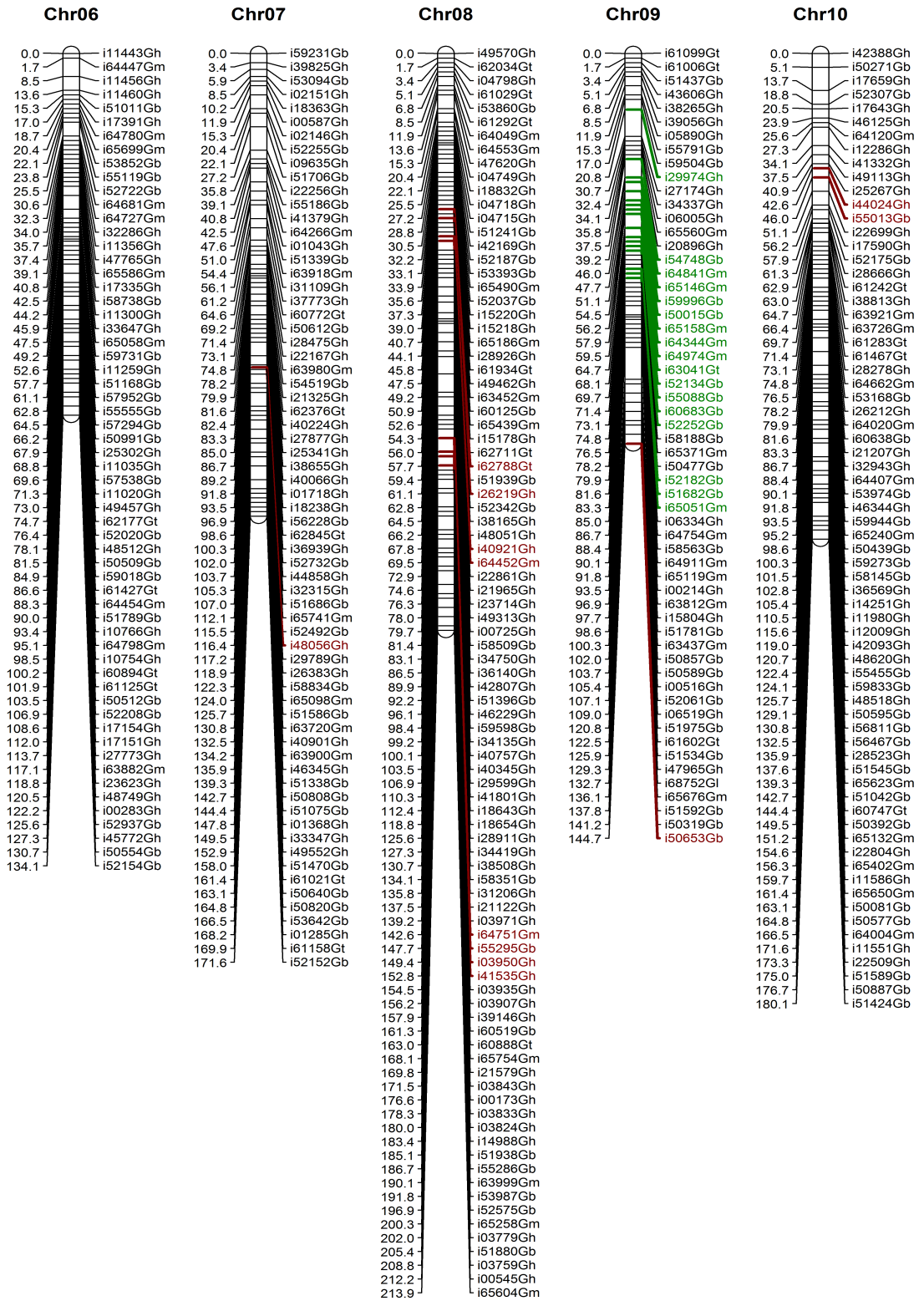


Figure 19. Continued.

Figured 19. Continued.

Figure 19. Continued.

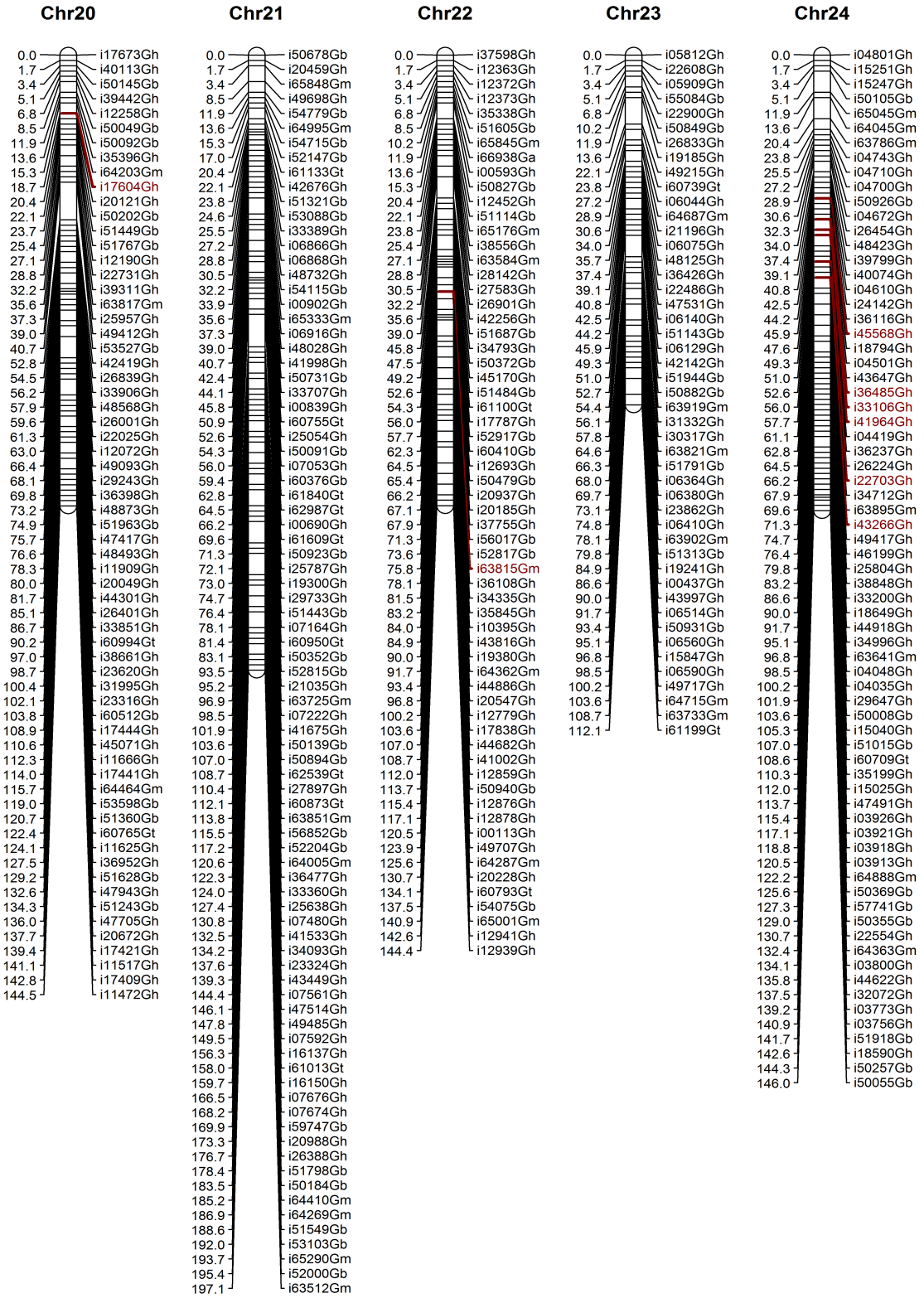


Figure 19. Continued.

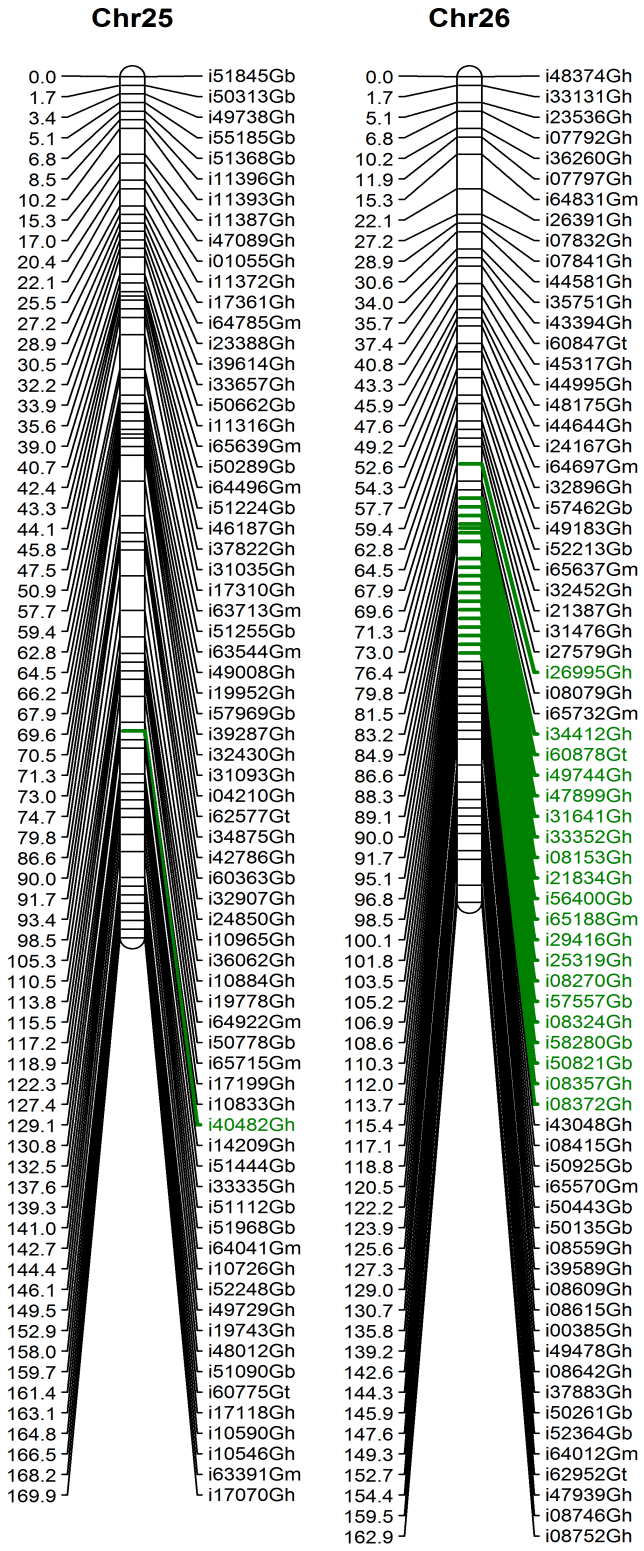


Figure 19. Continued.

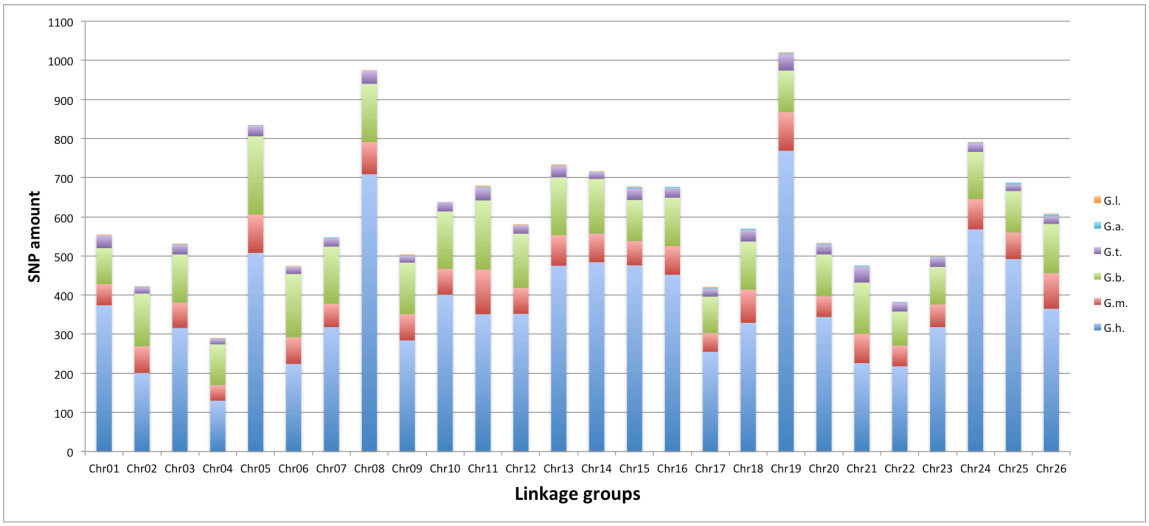


Figure 20. Numbers of CottonSNP63K markers, by source, that were mapped to the 26 linkage groups in the *Gossypium hirsutum* 'TM-1' x (*G. hirsutum* 'TM-1' x *G. mustelinum*) BC1F1 family. This bar graph depicts chromosomal distributions of mapped markers from the CottonSNP63K array, based on calculations dependent on information by Hulse-Kemp et al. (2015) about species origins of CottonSNP63K SNPs: Blue - *G. hirsutum* set. Red - *G. mustelinum.* Green - *G. barbadense.* Purple - *G. tomentosum*. Light blue - *G. armourianum.* Orange - *G. longicalyx.*

Synteny analysis

Primer sequences of all 70,000 SNP markers on the CottonSNP63K array were aligned to the *G. hirsutum* reference genome database at JGI using BLAST+ command function under UNIX system, and only the best hit per marker was used for syntenic analysis. Among markers mapped in the linkage maps, 15,500 (97.95%) of them were detected on chromosomes of the reference genome database, while 301 (1.9%) were aligned to scaffolds, and 24 (0.15%) marker sequences were not detected. A dot plot shows marker positions in both the linkage maps and reference genome assembly, including 5,947 markers in A sub-genome versus 9,553 markers in the D sub-genome.

The dot plot displayed high collinearity between linkage groups and reference genome database along the diagonal (**Figure 21**). However, it also revealed significant homeologyl relationships between the A and D sub-genomes, as well as ancestral translocations between chromosome 2, 3, and 4, 5 in the A subgenome, relative to the D-subgenome. The R package "RCircos" was applied to depict detailed alignments of SNP markers by linkage groups (**Figure 22**). High collinearity between linkage groups and chromosomes from reference genome was evident on circular alignment plots. Most markers aligned toward the corresponding chromosomes, but some markers were detected on the homeologous chromosome of the corresponding ones, and a few markers were found in assemblies of other chromosomes. The relationship to homeologous sequences in the genome assembly occurred more often in linkage groups of A sub-genome than groups of D sub-genome.

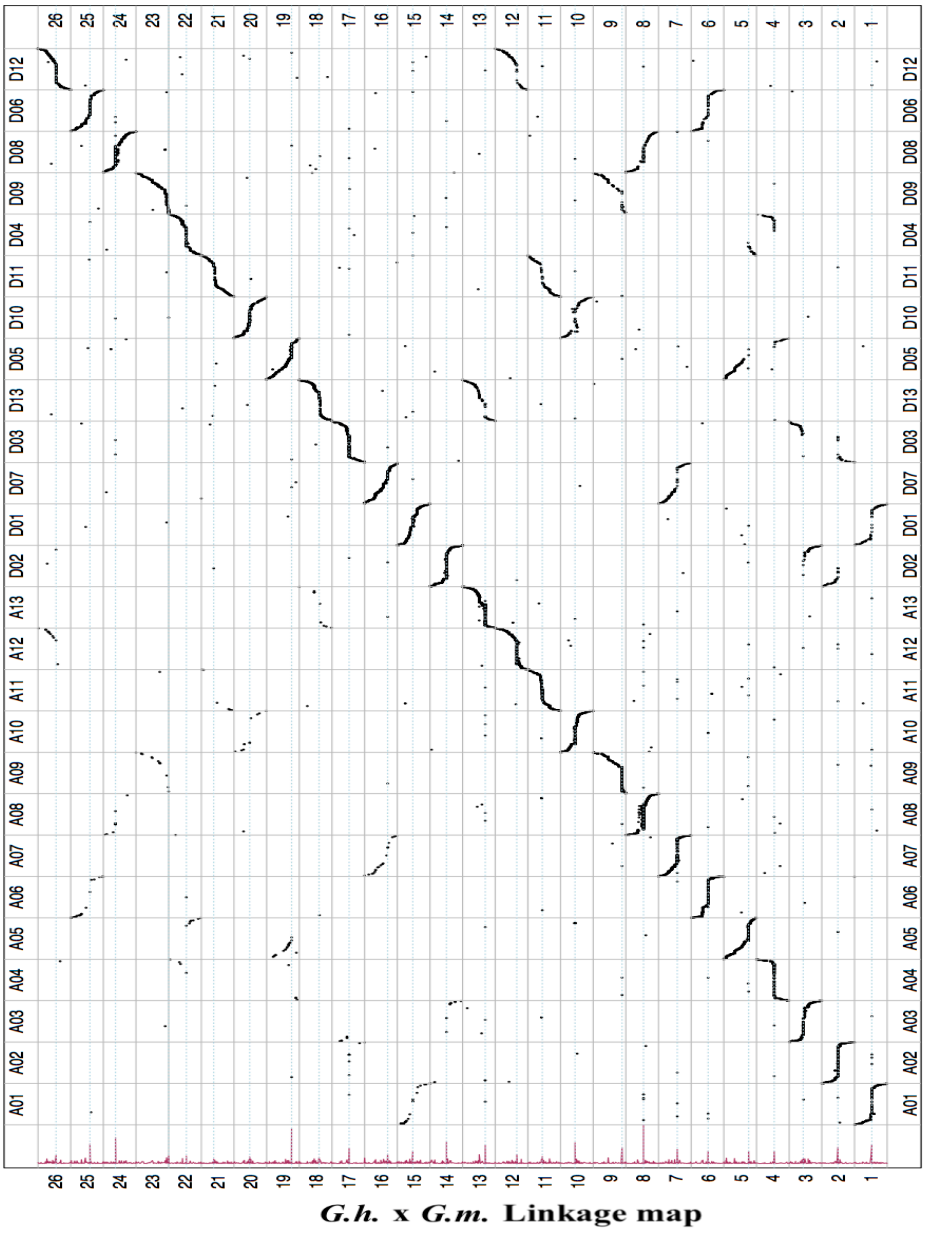


Figure 21. Dot plot showing syntenic relationships deduced from sequence alignments of linkage map SNP loci to the most recent public *G. hirsutum* 'TM-1' reference genome (Saski et al., 2017, in press). Linkage groups are numbered according to chromosome identities and chromosome scaffolds are numbered according to chromosome or segmental homeology relationships noted in the sequence assembly (Saski et al., 2017, in press). The maroon line along the linkage map (Y-axis) indicates the relative numbers of mapped CottonSNP63K SNPs within each recombinationally defined marker bin of the BC1F1 mapping population from *Gossypium hirsutum* 'TM-1' x (*G. hirsutum* 'TM-1' x *G. mustelinum*).

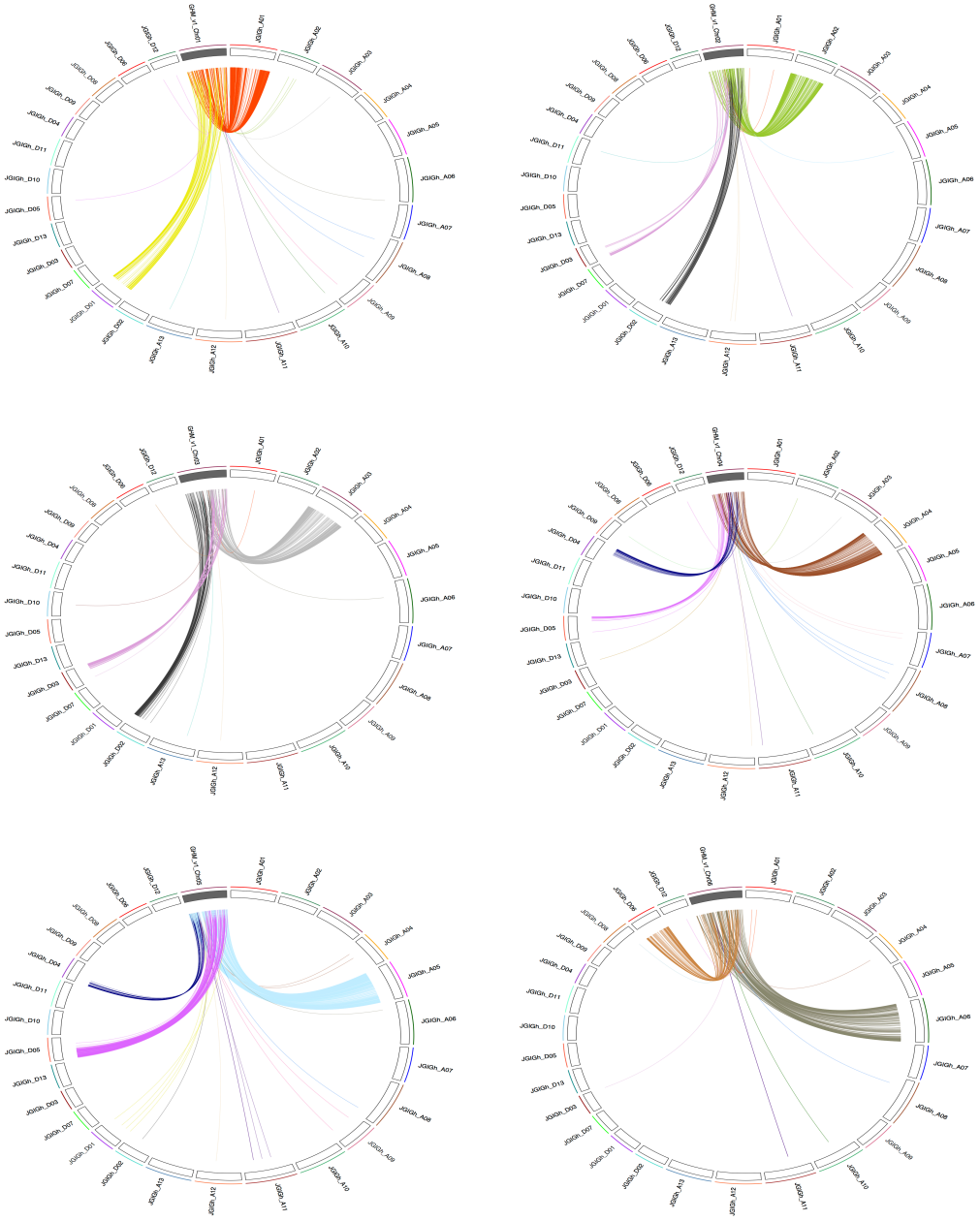


Figure 22. R Circos plots displaying syntenic relationships between linkage map SNP loci to the most recent public *G. hirsutum* 'TM-1' reference genome (Saski et al., 2017, in press). Syntenic relationships were deduced from sequence alignments to the genome assembly scaffolds. Each circle represents the alignment of CottonSNP63K markers mapped to one linkage group, demarcated as a dark grey block of each plot (near the 12-o'clock position), versus the 26 scaffolds in the reference genome assembly (26 white blocks). Different colors denote alignments to different chromosomes of the sequence assembly.

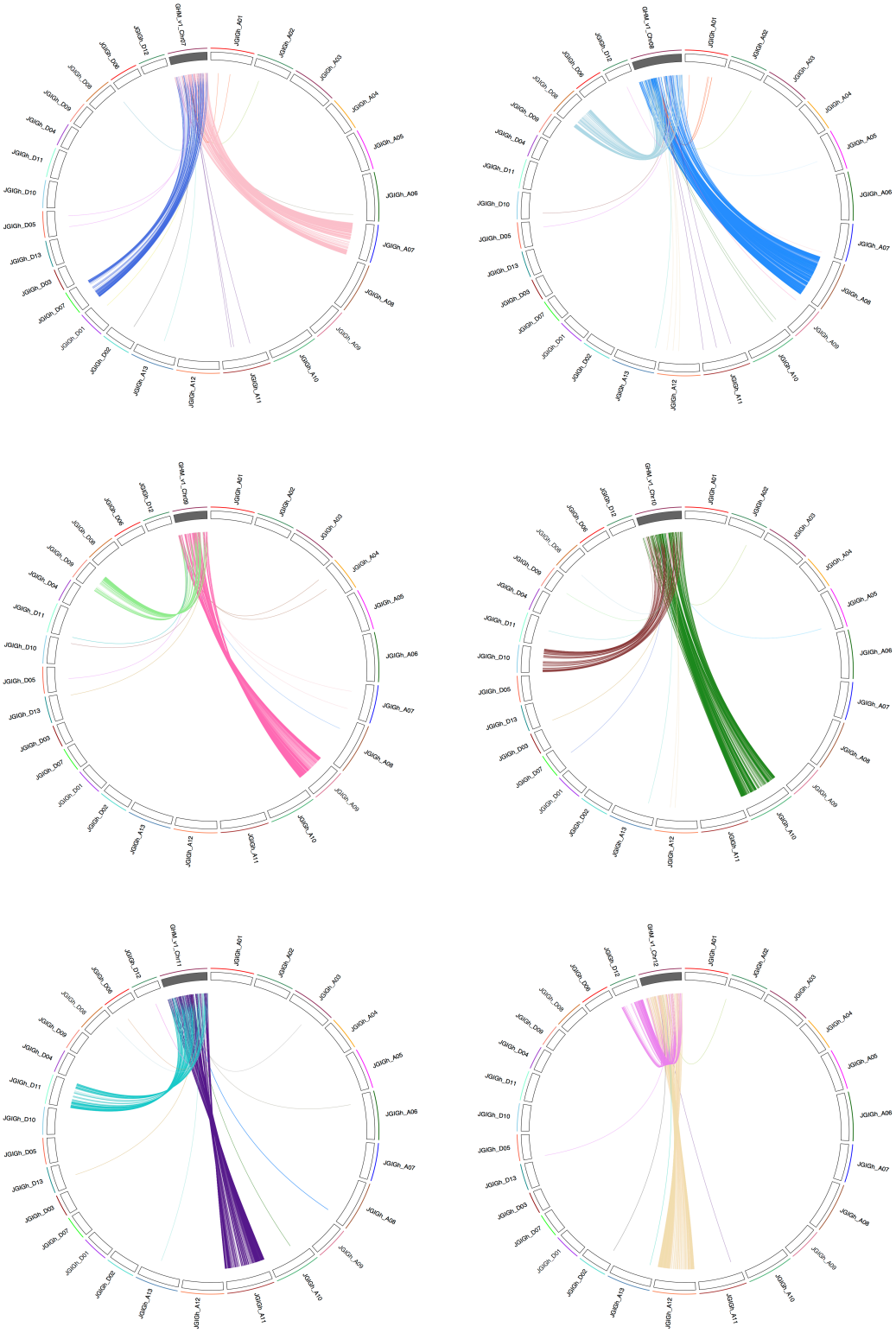


Figure 22. Continued.

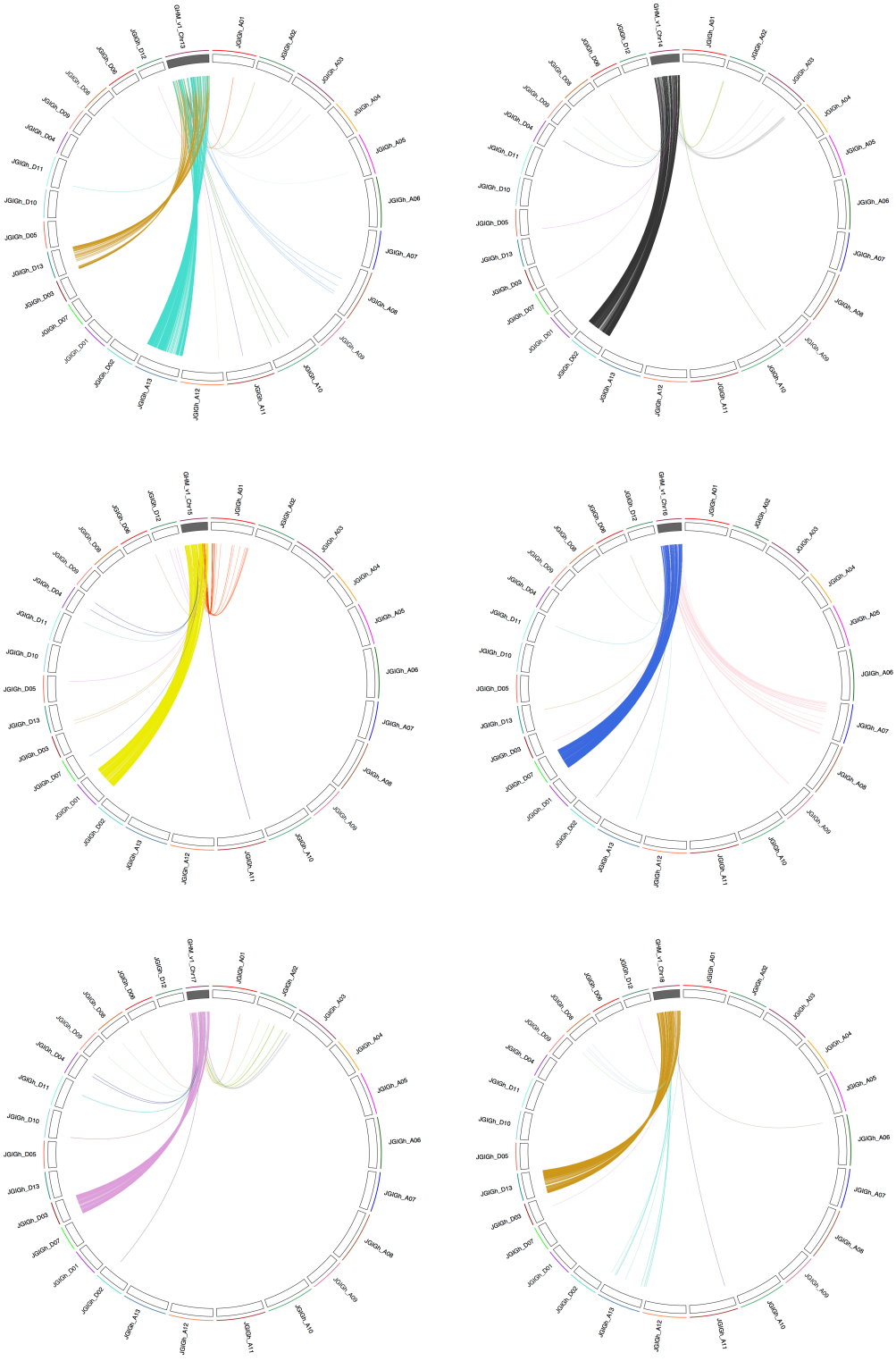


Figure 22. Continued.

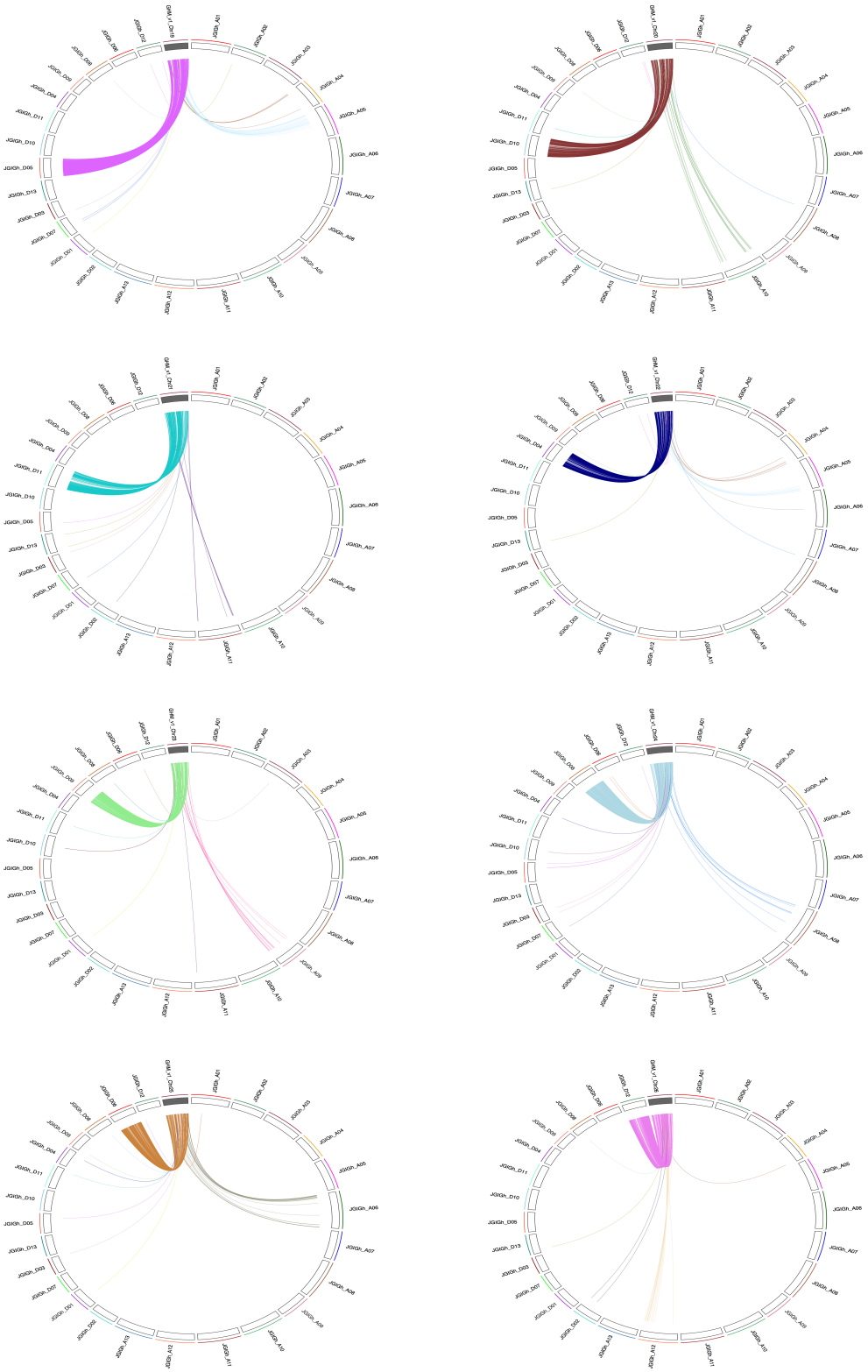


Figure 22. Continued.

Comparison to Other Linkage Maps

The high-density *G. hirsutum* x *G. barbadense* (*G.h.* x *G.b.*) linkage map of (Hulse-Kemp *et al.*, 2015), was also constructed by genotyping SNP markers on the CottonSNP63K array. Thus, the use of common markers facilitated alignments and comparisons with the *G. hirsutum* x *G. mustelinum* linkage map developed as part of this study. A total of 15,825 markers were mapped using the *G. hirsutum* x *G. mustelinum* (*G.h. x G.m*) population, and these spanned 4,193.82 cM. In comparison, the *G.h.* x *G.b.* map included 19,191 SNP markers and covered 3854.3 cM. Among markers from two maps, 58% of the markers were in common, while 13% of markers were unique to the *G.h.* x *G.m.* map, and 28% of markers were unique to the *G.h.* x *G.b.*, except chromosome 19. For that chromosome, 48% of markers were common to both maps, and the species-specific proportions of unique markers were more lopsided, i.e., 7% and 48% respectively, in the *G.h.* x *G.m.* and *G.h.* x *G.b.* maps.

Fewer recombination events contributed to the *G.h.* x *G.m.* mapping population than to the *G.h.* x *G.b.* mapping population, and the resulting linkage map consisted of fewer recombination bins, 1,776, when compared to the *G.h.* x *G.b.* genetic map that comprises 4,220 bins. The average size of recombinationally defined marker bins was 2.39 cM in the *G.h.* x *G.m.* map, versus 0.92 cM in the *G.h.* x *G.b.* map, and the average number of SNP markers per bin was 8.9 for *G.h.* x *G.m.* map versus 4.5 in the *G.h.* x *G.b.* map. While the order of common markers was largely conserved across the two maps, about 8% of the common markers had seemingly inverted order (Figure 23). In many cases, markers that had been recombinationally separated into different *G.h x G.b.* map

bins were co-localized to a *G.h x G.m.* map bin (see example marked with blue arrow in **Figure 23)**, whereas very few markers exhibited the reverse relationship (see example marked with red arrow in **Figure 23)** (**Supplemental Figure S4**).

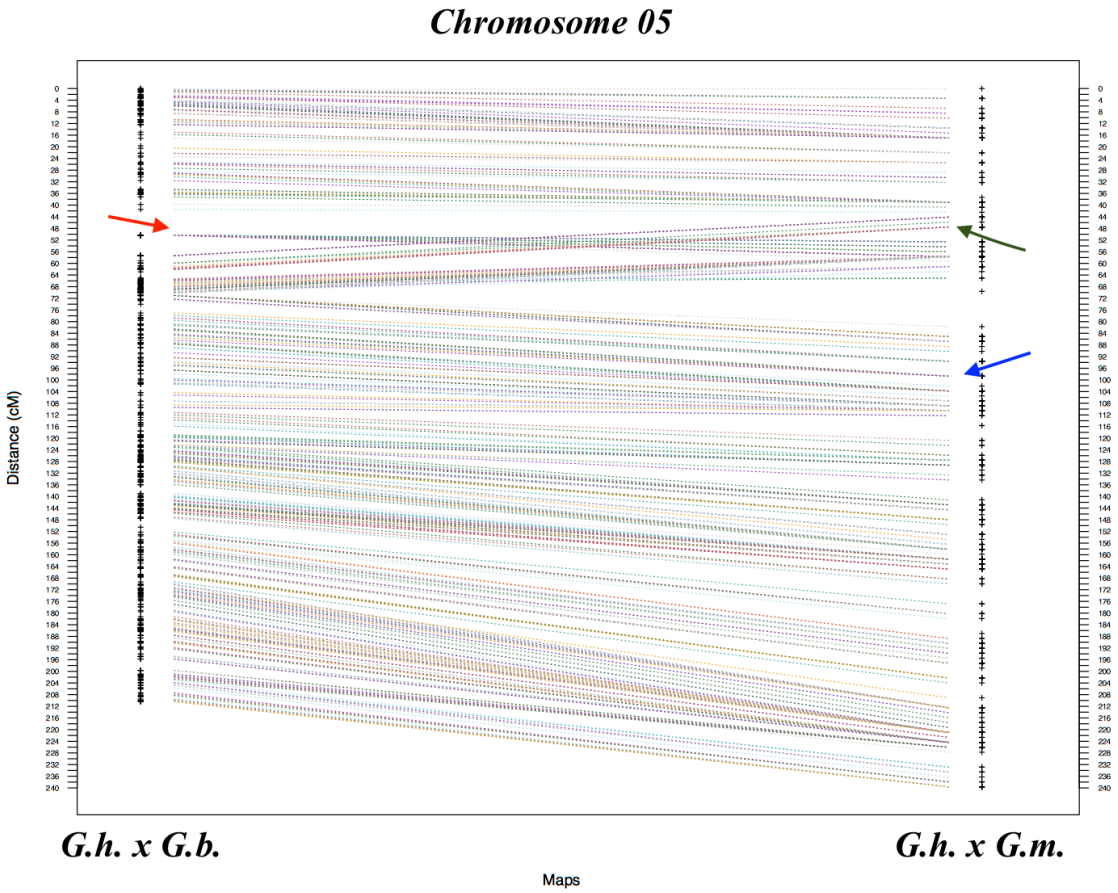


Figure 23. Alignment of *G.h. x G.b.* and *G.h. x G.m.* maps based on common CottonSNP63K SNP markers. Green arrow - example where the order of loci was inverted. Blue arrow - example where markers that were clustered into *G.h x G.m.* map one bin, but recombinationally separated into different *G.h x G.b.* map bins. Red arrow - example of reverse relationship, i.e., where loci are separated into different *G.h x G.m.* map bins, but recombinationally clustered into one *G.h x G.b.* map bin.

Using sequence alignments to a common reference - the genome sequence assembly of (Saski et al., 2017, in press), it was possible to align and compare our SNP-based map with the SSR-based *G.h.* x *G.m.* linkage map (Wang *et al.*, 2016). The total length of the SNP-based map was 1,400 cM shorter than the SSR-based map, and the average bin size was narrowed by half, from 5.3 to 2.4 cM. In both maps, there was a slight excess of markers mapping to the D subgenome, but the percentage was a bit higher for SSRs (6.9%) versus SNPs (3.8%). Marker order was compared by their position on reference genome and showed high colinearity (**Supplemental Figure S5**).

Discussion

Recent development of a high-density molecular marker array for high-quality genotyping of cotton has greatly improved the efficiency, capability and accuracy for molecular and genetic research in cotton (Hulse-Kemp *et al.*, 2015). While the *G. hirsutum* - *G. mustelinum* mapping population used in this study is smaller than usual, the interspecific genetic map that was developed in this study is the most saturated map so far, and offers considerable insight and utility to the community. For breeding purposes, the map, its markers, and information that can be linked to it informatically, all serve as a fundamental resources for QTL analysis, genetic dissection and many forms of marker-assisted selection. These are especially important to the use of interspecific materials for superior traits introgression into Upland cotton and the subsequent derivation of agriculturally palatable genetic products. To overcome serious genetic incompatibilities when utilizing an alien genetic resource, such as wild species *G.*

mustelinum, backcrossing to create near-isogenic materials is usually considered as an appropriate approach. Special mapping populations like chromosome segments substitution lines or advanced backcross inbred lines are heavily rely on the linkage mapping information from other studies to determine the general genetic structure of the population, to identify segments of interests for breeding, and to mount applied marker-assisted selection (P. Wang *et al.*, 2012; Yu *et al.*, 2013; Zhai *et al.*, 2016). Similar concept of research has been carried out on other germplasm by several other students in our research group at the Stelly Lab. Analogous considerations extend to large-scale germplasm analysis and development of genome selection. This marks a new stage in interspecific breeding in cotton. For example, previous students in the Stelly lab have conducted large-scale germplasm backcross introgression without the benefit of genome-wide applicability of marker-assisted methods. Thus, phenotypic analysis was generally not integrated with genome-wide marker analysis, because marker analysis was relegated to small numbers of loci near individual targets of high scientific, genetic and/or agricultural value. Many of these constraints are reduced or eliminated for *G. mustelinum* and other species by advent of the CottonSNP63K, i.e., recombination frequency analysis in random mating populations and QTL analysis in fiber traits among backcross populations and (Gardunia, 2006), and transmission rate analysis between *G. hirsutum* and *G. mustelinum* among different backcross generations in different environments (Xu, 2014).

Comparisons between interspecific genetic maps can be used as an approach for studying chromosome structural differences, e.g., relative to speciation, in addition to a

number of other methods that range from simple karyotyping to whole-genome sequencing (Chen *et al.*, 2006; Kirkpatrick, 2010; Mano, Omori and Takeda, 2012). A comparison of the *G.h.* x *G.b.* map versus the *G.h. x G.m.* map reveals a major difference in numbers of recombination bins, i.e., 4,220 versus 1776, respectively. Likely explanations include the differences in structures and the sizes of the respective mapping populations: 118 F_2 individuals for the *G.h.* x *G.b.* map and 59 BC_1F_1 plants for the *G.h.* x *G.m.* map. Thus, the *G.h.* x *G.b.* map, was based on 236 meiotic events in 118 microsporocytes and 118 megasporocytes, and the *G.h.* x *G.m.* map was based on 59 meiotic events in 11 microsporocytes and 48 megasporocytes. The pathway of transmitting recombination to the next generation is not necessarily but can be a significant factor in transmission, and so could affect the genotypic composition of the progeny, the perceived recombination intensities and distributions. F_2 populations can be influenced by gene-based biases that affect female and male gamete viability or functionality, as well as zygotic homozygosity; BC_1F_1 populations are more likely to be differentially affected by only the male or the female gamete, not both (Rooney and Stelly, 1991; Xu, 2014). Polygenic relationship between two parental species might also play a role for the difference in recombination bins (Grover *et al.*, 2015; Wendel and Grover, 2015). Because of the huge difference in numbers of recombination bins between two maps, it was not surprising to observe that many markers placed in different bins in the *G.h.* x *G.b.* map were co-segregated into same bin in the *G.h.* x *G.m.* map. More informative comparisons between two maps were the average widths of individual bins in the two maps, as well as difference in marker order. The average

interval width between bins in the *G.h.* x *G.b.* map, 0.92 cM, versus the average interval of 2.39 cM in the *G.h.* x *G.m.* map. On a per bivalent basis, the recombination rate was higher between *G. hirsutum* and *G. mustelinum.* Cases were observed where two adjacent bins spanned a longer distance in the *G.h.* x *G.b.* map than they did in the *G.h.* x *G.m.* map (**Figure 23**). Detail statistical analysis about the interval distance difference between two maps and the inverted bin order on the genetic maps will be conducted in the future.

Segregation distortion has been widely observed in many studies in both intra-specific and inter-specific mapping populations, and in this project as well (Yu *et al.*, 2013; Zhiyuan *et al.*, 2014; Hulse-Kemp *et al.*, 2015; Wang *et al.*, 2015). More interestingly, these markers were distributed in the particular chromosomes only as well as with the tendency to either genotype, homozygous genotype "A" or heterozygous genotype "H". No mixed types of segregation distorted markers was observed in the same chromosome. Some well-studied cases to explain the segregation distortion were the combination of complementary recessive genes, i.e., chlorophyll deficiency, asynapsis, and sterility (Zhang, Percy and McCarty, 2014). Specific gamete or combination of gametes would lead to abnormal plant form or even lethal traits, which resulted in the selection toward zygotes/gametes. Additionally, recombination rate on particular regions in chromosomes, such as hotspot or cold-spot, also influences the allelic transmission rate to the offspring. More research is necessary regarding the segregation distortion area of this genetic map.

Linkage maps consist of recombination distances in centiMorgan between "markers", based on the frequencies of observed of recombination due to crossovers (homologous recombination). "Markers" can take many forms, from phenotypic traits to DNA sequences, even single bases. All are the informative tools for genetic research and serve as mutual references for each other. A common application of linkage maps is to help detect and correct misplaced segments in physical and/or sequence contig assemblies. Genome assembly involves progressive assembly of segments at many genome locations into larger contigs, placement along scaffolds and chromosome; many aspects of this process are based on sequence alignments and similarities. However, repetitive sequence around the entire genome in plants and the complex association between chromosomes due to polyploidization and speciation often deepen the difficulty and lead to some inaccuracy of genome assemblies. Sometimes even published assemblies are extensively flawed. Therefore, the recombination fraction between molecular markers from linkage maps can help to determine the correct locations and order of individual loci or groups of loci, and both can help the assembly process; this tends to be especially important when encountering ambiguous situation involving similar sequences in different segments. Such segments tend to be more common in polyploids, paleopolyploids and species with high content of dispersed repeats.

Analysis of collinearity between linkage maps and physical maps, e.g., **Figure 21**, revealed high congruity indicated by the heavy presence of dots along the diagonal, which indicates that order of markers in one linkage group are associated to the order of corresponding marker-associated sequences in the physical map or sequence assembly

for the respective chromosome. In **Figure 21,** the *G.h.* x *G.m.* map is shown and most loci are well ordered relative to the *G. hirsutum* reference genome, accounting for 13,211 markers out of 15,815, i.e., 83.53%. When looking at the upper left and lower right triangles that compare A-D and D-A homeologous relationships across the two maps (linkage versus physical), 1,996 (12.62%) of the rest markers are instead sequence-aligned to assemblies for the homeologous chromosomes, i.e., not the linkage-mapped chromosome. For example, some markers from linkage group 01 were sequence-aligned to the sequence assembly for its homeolog, D01, of the reference genome, rather than its homolog, A01, where expected (**Supplemental Figure S6**). In dot plots, such aberrant markers reveal well known homeologous relationships, as well as ancestral A-genome translocation events, e.g., those involving ancestral progenitors of chromosome 2 and 3, as well as chromosome 4 and 5 (Desai *et al.*, 2006). For example, in the second row, most markers from linkage group 2 are detected in A02 chromosome, but one half of rest are located in D02 and the other half in D03 chromosome because of the translocation between chromosome 2 and 3. Similar observations were reported previously for *G.h.* x *G.b.* map (Hulse-Kemp *et al.*, 2015), translocation events involving the complete arms of the chromosomes when examining the markers on homeologous chromosomes of those (**Figure 24**) (Blenda *et al.*, 2012). The findings indicate that the ancestral translocation involved breaks in the centromere or nearby pericentromeric heterochromatin, and therefore involved complete arms. Analogous phenomena are also observed in linkage groups 3, 4, and 5.

If we compare the two triangles, markers in upper left triangle are obviously fewer in number than the markers in lower right triangle, which implies the D sub-genome of reference genome assembly is better than the A sub-genome assembly. A very plausible explanation relates to the relative quality of related diploid genome assemblies available as references to facilitate assembly of the polyploid AD genome of *G. hirsutum*. The D5 genome assembly is more solid and robust than the draft A2 genome assembly (Paterson *et al.*, 2012; Li *et al.*, 2014). The tetraploid cotton genome database therefore had only one extensively useful reference diploid genome, i.e., that of the D5 genome database, only. It is highly plausible that statistical sequence alignment procedures found no diploid A-genome match for many small segments that supposed to be placed in A sub-genome but did find a match in the diploid D-genome references, and as a consequence might have misassembled the affected A-genome sequences into D sub-genome instead. The linkage map information in this study can therefore help in the correction of segments that were misplaced.

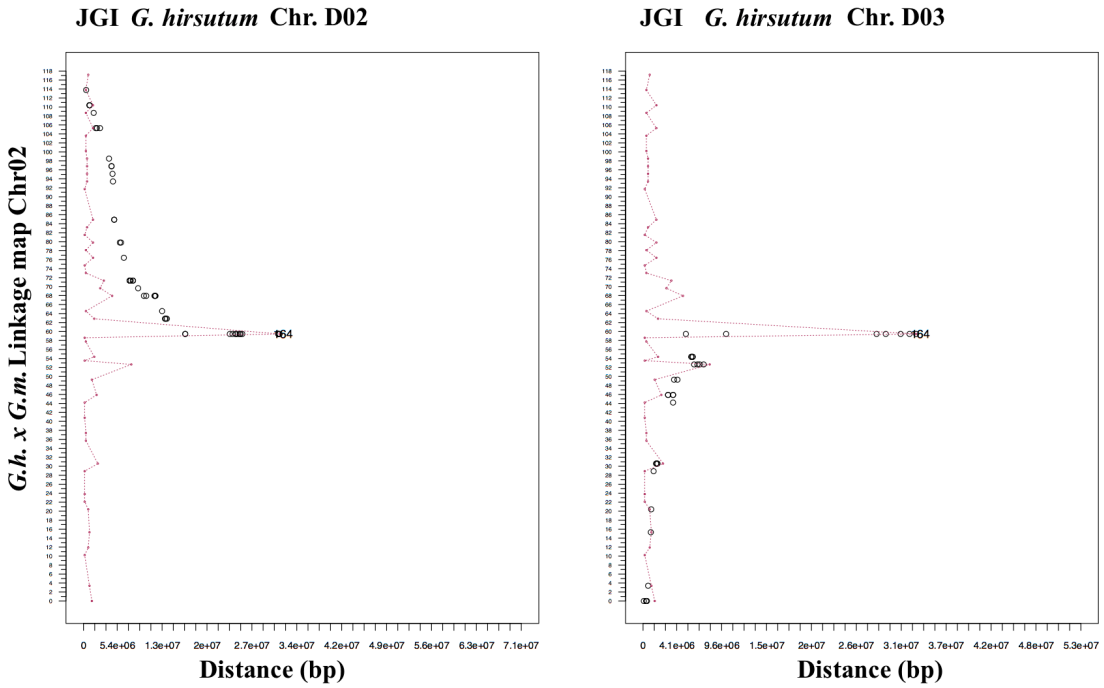


Figure 24. Putative arm-specificity of ancestral A-genome translocations demonstrated by alignment of linkage mapped marker sequences to physically mapped sequences of homeologous D-subgenome chromosomes, illustrated for linkage group 02. Maroon line (left side of each plot): numbers of markers per recombinationally defined marker bin in linkage group 02. The peak (164 co-segregated markers) is thought to be the peri-centromeric region of the chromosome. All markers associated with the physical assembly of chromosome D02 were linkage mapped to the upper part of the linkage group 02 (one arm), whereas all markers associated with the physical assembly of chromosome D03 were only arranged on the lower part of the linkage group (opposite arm).

GenTrain score

Dot plots provided a basis for comparison between the *G.h. x G.m.* linkage map versus *G. hirsutum* genome assembly. Markers that concordantly detected loci in linkage groups and chromosomally associated physical sequence assemblies were referred as "Type I" markers, whereas those detected on the homeologous chromosomes

of their linkage groups were designated as "Type II" markers. As part of an effort to dissect the difference between two types of markers, detailed characteristics of the marker genotyping, including the cluster file, fluorophore signals intensity and ratio, and genotype call reliability, were examined for some markers through the SNP Graph in GenomeStudio Genotyping Module (V 1.9.4, Illumina, Inc.) (**Figure 25**). However, it would have been extremely time-consuming for examining every SNP graph in detail given such large numbers of markers. For convenience to users, Illumina provides a simple indexed value, the "GenTrain score", that summarizes multiple characteristics of markers for speedy assessment decision-making. For example, in development of earlier CottonSNP63-based maps, the GenTrain score was utilized to classify markers into four different patterns (Hulse-Kemp *et al.*, 2015). Scores higher than 0.6 represents normal co-dominant markers that perform like diploid with three clear and separated clusters for genotype AA, AB, and BB, where the two homozygous clusters locate at 0 and 1 on the X axis. GenTrain scores between 0.3 to 0.59 in cotton often involve a second pattern of markers involving two loci, usually from homeologous chromosomes, but only one of the two loci is polymorphic, so genotypes will be AAAA, AAAB, AABB. Though closer, such clusters are still usually statistically readily separable but with one of the homozygous classes located near 0.5 on the X axis, instead of at an extreme (0 or 1). The third and fourth pattern of markers have GenTrain score between 0.21 to 0.29, and less than 0.2 respectively, and in these cases, multiple loci may be detected with just one locus showing polymorphic, where the other loci are monomorphic and these contribute differentially to signal for one allele, causing skewness in the overall signal distributions.

However, low intensity of fluorophore signals of samples can also lead to low GenTrain scores, as exemplified by the lower right graph in **Figure 25**. Summarily, high GenTrain scores generally indicate high specificity and reliability of markers, well-defined cluster files and accurate genotype calling. However, many factors may also affect the score, especially lower scores. Because they have relatively higher rates of sequence redundancy, polyploids invariably exhibit relatively higher numbers of GenTrain scores below 0.6, for them the informativeness and utility of GenTrain scores tends to be lower than for diploid species.

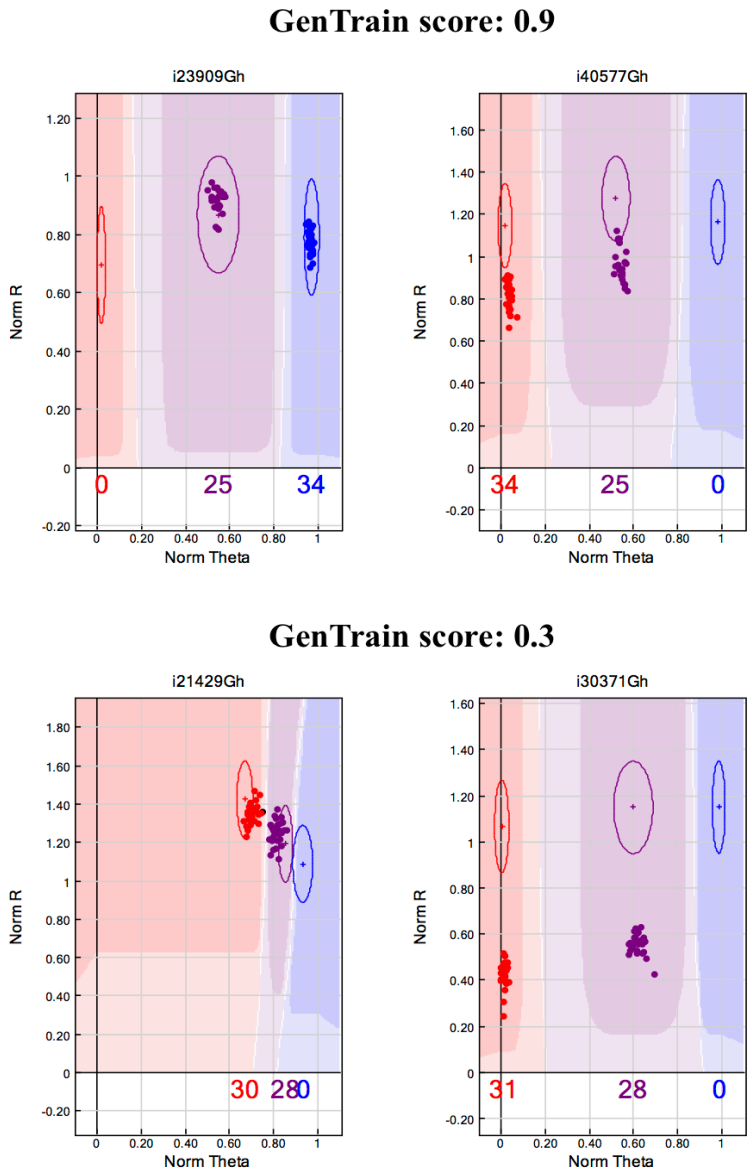


Figure 25. Examples of polar coordinates SNP graphs in GenomeStudio. The X axis represents normalized theta, the angle deviation from pure A signal. 0 indicates pure fluorophore-A signal and 1 indicates pure fluorophore-B signal. The Y axis represents the normalized fluorophore signal intensity. In each figure, the three ellipses and their surrounding shaded regions are the cluster files for each of the two homozygous genotypes and one heterozygous genotype. For each genotype call (dot), the shorter distances between the center of the ellipse and sample's dot indicates higher reliability. The upper two graphs are the examples for SNP marker having 0.9 GenTrain score in this study, whereas the lower two graphs exemplify SNP markers having 0.3 GenTrain score, one due to proximity of clusters, and the other, largely due to generally weak signal amplitude.

A systematic analysis and graphic plot of GenTrain scores for Type-II markers reveals that over 96% of them are lower than 0.6 (**Figure 26**), which implies that multiple loci were typically detected by the Infinium-II assays for those markers. That suggests the presence of similar sequences in at least two locations per AD genome, i.e., one homologous chromosome pair (e.g. "12"), plus one homeologous chromosome pair (e.g. "26"). The joint presence of such similar sequences could easily interfere the genome assembly and increase the probability of segment misplacement, especially if contextual information is limited. Given relative strength, sequence divergence and use of the diploid D5-genome assembly as a reference during AD genome assembly, versus shortcomings and non-use of the diploid A2-genome assembly, it seems reasonable to infer that during AD genome assembly, sequences from A- locations would be assimilated into the D-subgenome assembly if the corresponding homeologous D-subgenome sequences were absent or significantly more different than the A-subgenome sequence from the D5 reference genome assembly, at least barring contextual sequence contig information to avoid such an error. Under this scenario, one would expect some A-subgenome sequences to be assimilated into the D-subgenome assembly, whereas there no mechanism or expectation for D-subgenome sequences to be assimilated into the A-subgenome assembly. Thus, assembly errors would expectedly have a subgenome-specific bias, but should be detectable where SNPs are available and linkage mapped. Thus, subgenome bias could be expected in Type-II marker distributions, and where similar A- and D-subgenome sequences exist, the polyploid-dependent duplications would expectedly lead to lower GenTrain scores. A straightforward

approach to verify these ideas was to examine the differences between the top two best BLAST results for each marker and also to determine the association between BLAST differences and GenTrain score.

In BLAST analysis, the BIT score serves as an index to describe the quality of alignment for a target sequence against genome database. The true definition of BIT score is the required size of search space in which the current match could be found merely by chance. For example, if a BIT score for an alignment were 30, it would imply that the required size for having such alignment by chance would 2^{30} = 1 billion base pairs (bp) or more. Higher scores correspond to better sequence alignments. In this study, the BIT scores were used for calculating the differences between to BLAST results: BIT score of 1st BLAST hit minus BIT score of 2nd BLAST hit and then divided by the BIT score of 1st BLAST hit. The new created index, BIT difference ratio, ranged from 0 to 1. Zero represents the two best alignments are equally good and have the same BIT score; BIT difference ratio of 1 represents there is only one BLAST alignment return. For the 1,996 Type-II markers of our interest, the average BIT score difference ratio was 0.12 with standard deviation of 0.183.

For comparison, same investigations on Type-I markers were also conducted. A visual examination suggests three possible peaks in the histogram plot for Type-I marker GenTrain scores, at approximately 0.85, 0.5, and 0.35. Among them, over 36% of the 13,211 markers had score above 0.6 (**Supplemental Figure S7**). The average BIT difference ratio for Type-I markers was 0.24 and the standard deviation was 0.195. According to the one-way ANOVA result, the F ratio of 640.251 and p-value less than

0.0001 strongly suggested that alignment differences (BIT indices) for Type-II markers were significantly smaller than for Type-I markers (**Supplemental Figure S8**). This supports the idea that relative to Type-1 markers, Type-II markers generally involve sequences for which identical or sequences occur elsewhere in the genome, e.g., in a homeolog or paleo-homeolog.

Correlation analysis between BIT score difference ratios and GenTrain scores was conducted using three different data set combinations. When combining all Type-I and Type-II markers, the correlation coefficient was 0.52 ($p < 0.0001$). The second and third correlation analysis took into considerations about the huge difference in numbers of Type-I versus Type-II markers, as well as skewness toward low GenTrain scores among Type-II markers. Therefore, the second correlation analysis only combined Type-II markers with the top 1,996 markers from Type-I markers, based on the order of their GenTrain scores. The correlation coefficient became 0.623, and was highly significant. A third correlation analysis compared the top 1,996 markers from Type-I sorted according to their BIT score difference ratios, and all 1,996 markers from Type-II; the resulting correlation coefficient increased to 0.699 and was highly significant, too (**Supplemental Figure S9**).

Positive correlations were found between the GenTrain scores and the differences between first two best BLAST results from all three analyses. Comparison between second and third correlation analyses demonstrated that correlation coefficient increased when selection of Type-I markers was based on their BIT score difference ratio rather than based on the order of GenTrain scores. Markers that had the greatest

differences between BLAST results had the highest GenTrain scores, but not all markers that had high GenTrain scores had high difference between BLAST results. From the one-way ANOVA for BIT score difference ratio between two types and the correlation analysis above, Type-II markers had more repetitive sequences in the genome than Type-I markers, and the GenTrain score of Type-II markers were lowered by their repetitive sequences. The repetitive sequences influenced the genome assembly accuracy and possibly resulted the segment misplacement into homeologous chromosome instead. Despite of the fact that GenTrain scores mainly reflected the cluster files of markers and the distribution of samples' genotypes, some other factors were evolved in the score. An advanced index or a combination of indexes should be referred to purely describe the marker quality and to establish a more accurate marker pattern classification.

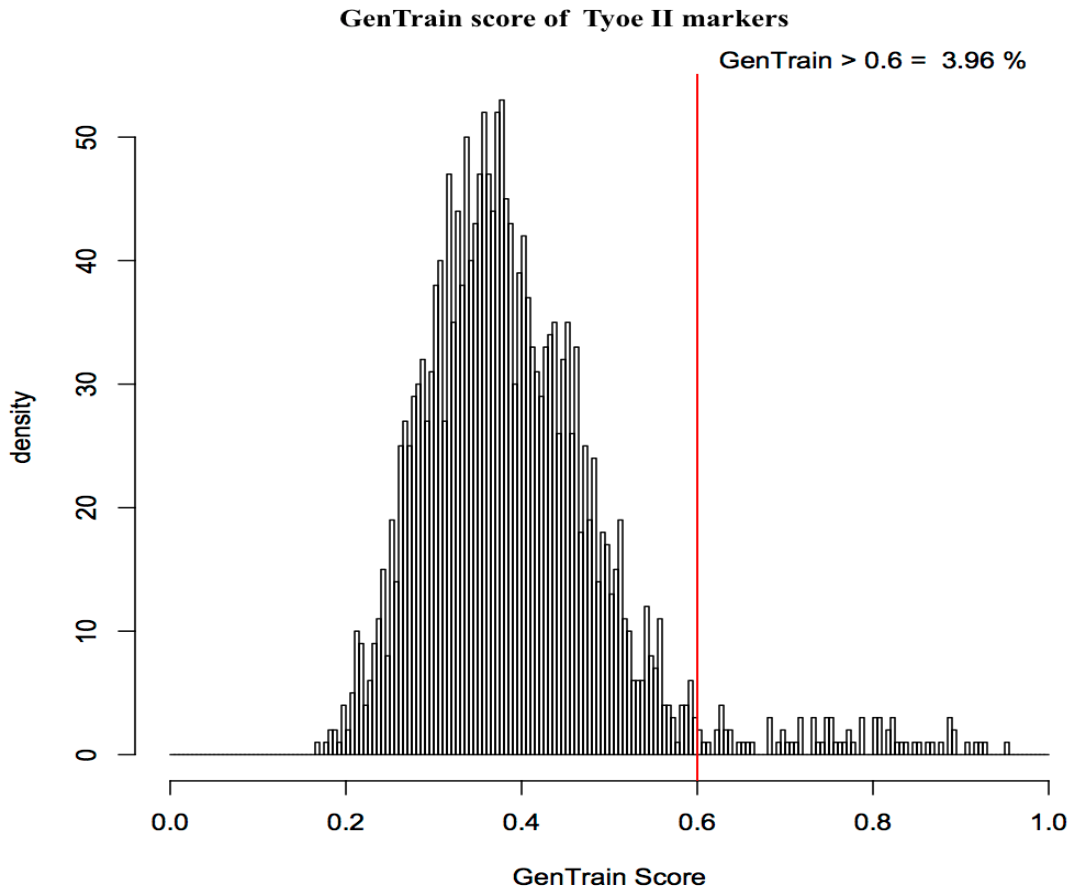


Figure 26. Histogram of GenTrain scores for linkage-mapped markers that were detected on chromosomes homeologous to the linkage-group chromosome (Type-II markers). The mean of the distribution is 0.39 and the standard deviation is 0.111. Over 96% of the 1,996 markers have GenTrain score lower than 0.6.

CHAPTER IV

CONCLUSIONS

Generating a great amount of molecular marker information in a short period of time, high-throughput genotyping technologies undoubtedly enhances the efficiency and accuracy in plant breeding and genetic research. The facility of high-density genetic mapping makes it possible to locate QTLs to very narrow region with strong confidence, which greatly improves the feasibility of marker-assisted selection (MAS) in breeding for superior trait introgression as well as linkage drag prevention. But a common constraint on this process is statistical insensitivity, where only QTLs with major effects can be reliably detected, and experimental "noise" compromises efforts to detect other QTLs. This type of complication is accentuated in 52-chromosome cottons by high levels of genetic redundancies that exist due to the polyploid and paleopolyploid nature of its genome. Because the strengths of such QTLs can vary widely and can be much stronger in other genetic contexts, the detection of QTLs with lesser effects in a given situation tends to be more important than might be anticipated. In fact, we think the positions of major and minor QTLs can be used proactively in genetic research and breeding. Methods that refine and increase sensitivity of QTL definition in cotton are thus needed.

Chromosome substitution is an approach for germplasm introgression that leads to development of isogenic materials, the isogenicity of which enhances in QTL analysis by greatly reducing genetic background noise and most G x E interactions. The degrees

of isogenicity can be affected by various factors, e.g., the degree of backcrossing, chromosome number, marker-based selection. Having a relatively high chromosome number, the effects are quite strong in cotton. While this approach restricts each experiment's target to a given chromosome, it increases the QTL detection power for detecting minor QTLs and the consistency of analysis results across environments. In this study, fiber quality QTLs on chromosome 17 were detected consistently, but have not been reported previously from other research using the same parental lines. Minor FOV 4 resistance QTLs were identified consistently and concordantly in this study, and the position corresponds to the preliminary research using SSR markers. According to their additive effect, many QTLs here might be considered as minor, and in most cases the beneficial allele was from *G. hirsutum* rather than *G. barbadense*. Marker-assisted selection methods make it quite feasible to select concomitantly at such QTLs, as well as at QTLs with major effects, which is extremely difficult to do without the aid of markers.

Chromosome-specific RILs are true-breeding and mostly Upland cotton, so those with desirable segments from *G. barbadense* can be used easily for traits introgression into Upland cotton and for broadening the genetic base of *G. hirsutum*.

The availability of a high-density genetic map of shared markers facilitated examination of newly bred germplasm (CS-B17 RILs and *G. mustelinum* BC1F1 individuals) based on genotype visualization; this enabled a facile accounting of recombination events in each individual; in the CS-B17 study, it made it possible to detect and eliminate the catastrophic presence of a genetically polluted RIL. The high degree of sensitivity afforded by the CS-RIL approach made it especially susceptible to

statistical effects of the polluted RIL, so the detection and removal were paramount to success of the entire study.

Compared to domesticated cotton, *G. barbadense*, many wild cotton species are still unfamiliar territories as genetic resources for Upland cotton genetic improvement in the perspective of abiotic/biotic stress tolerance and disease resistance. The high resolution *G. hirsutum* - *G. mustelinum* genetic map in this study will fundamentally facilitate research in genetic and breeding with this Brazilian AD-genome species, and will have ramifications that extend to analysis and use of the other tetraploid species, too. We are now well positioned to study segregation-distorted markers in detail, e.g., differences in marker co-segregation between reciprocal crosses, *G.h.* x *G.b.* versus *G.h.* x *G.m.*, to uncover the genomic basis of cotton species. High collinearity is observed between linkage map and physical map, but the present of markers detected on homeologous chromosomes suggests the presence of assembly errors, especially the inclusion of A sub-genome segments into the D-subgenome assembly.

Future work should include CS-B17 QTL validation and the comparisons to QTL analysis to a normal RIL population with genome-wide segregation for the same parental lines. These should document the effects of different levels of genetic background noise. Using modern resources for marker assisted selection, another chromosome 17 substitution RIL population with *G. barbadense* genetic background could be developed as the opposite of the CS-B17-RILs in this study for genic interaction and epistasis research for our detected QTLs. This approach could also be extended to other

chromosomes to detect more minor QTLs and fine map the major QTLs, e.g., chromosome 22.

The *G.h.* x *G.m.* linkage map will be extremely useful as framework for developing chromosome segment substitution line (CSSL) from an advanced backcross inbred line (BIL) population. With the understanding of alien segment locations in the BILs, it will be possible to genotype each line for its line-specific area(s) of interest only, which can reduce the cost for genotyping and enable increased population sizes. Additionally, the *G.h.* x *G.m.* genetic map can be compared with another three linkage maps, *G.h.* x *G.b., G.h.* x *G. tomentosum,* and *G.h.* x A_2D_1, synthetic tetraploid cotton, for chromosomal structure difference in cotton evolutionary research.

REFERENCES

Altschul, S. F. et al. (1990) 'Basic local alignment search tool.', Journal of Molecular Biology, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Alves, M. F. et al. (2013) 'Diversity and genetic structure among subpopulations of *Gossypium mustelinum* (Malvaceae)', Genetics and Molecular Research, 12(1), pp. 597–609. doi: 10.4238/2013.February.27.9.

Baker, K. (1957) 'U.C. system for producing healthy container-grown plants', Calif Agric Exp Stn Man. University of California, Division of Agricultural Sciences, Agricultural Experiment Station, Extension Service, pp. 1–332.

Barone, A. and Frusciante, L. (2007) 'Molecular marker-assisted selection for resistance to pathogens in tomato', in Marker-assisted selection: current status and future perspectives in crops, livestock, forestry and fish, pp. 153–164.

Beasley, J. O. (1941) 'Hybridization, cytology, and polyploidy of *Gossypium*.', Chron. Bot., 6, pp. 394–395.

Bell, A. A. et al. (2017) 'Genetic diversity , virulence , and meloidogyne incognita interactions of *fusarium oxysporum* isolates causing cotton wilt in Georgia', Plant Disease, 101(6), pp. 948–956.

Bianco, L. et al. (2014) 'Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus* × *domestica* Borkh)', PLoS ONE, 9(10). doi: 10.1371/journal.pone.0110377.

Blenda, A. et al. (2012) 'A high density consensus genetic map of tetraploid cotton that

integrates multiple component maps through molecular marker redundancy check', PLoS ONE, 7(9). doi: 10.1371/journal.pone.0045739.

Brubaker, C. L., Bourland, F. M. and Wendel, J. F. (1999) 'The origin and domestication of cotton', in Cotton: Origin, history, technology, and production., pp. 3–31.

Brubaker, C. L. and Wendel, J. F. (1993) 'On the specific status of *Gossypium lanceolatum* Todaro', Genetic Resources and Crop Evolution, 40(3), pp. 165–170. doi: 10.1007/BF00051121.

Camacho, C. et al. (2009) 'BLAST+: architecture and applications', BMC Bioinformatics., 10(1), p. 421. doi: 10.1186/1471-2105-10-421.

Campbell, B. T. et al. (2010) 'Status of the global cotton germplasm resources', Crop Science, 50(4), pp. 1161–1179. doi: 10.2135/cropsci2009.09.0551.

Carvalho, M. R. et al. (2011) 'Paleocene Malvaceae from northern South America and their biogeographical implications', American Journal of Botany, 98(8), pp. 1337–1355. doi: 10.3732/ajb.1000539.

Charcosset, A. et al. (2004) 'Use of molecular markers for the development of new cultivars and the evaluation of genetic diversity', Euphytica, 137(1), pp. 81–94. doi: 10.1023/B:EUPH.0000040505.65040.75.

Chen, G. K. et al. (2006) 'Accommodating chromosome inversions in linkage analysis', American Journal of Human Genetics, 79(2), pp. 238–251. doi: 10.1086/505540.

Chen, H. et al. (2014) 'A high-density snp genotyping array for rice biology and molecular breeding', Molecular Plant, 7(3), pp. 541–553. doi: 10.1093/mp/sst135.

Chen, H. et al. (2015) 'A high-density SSR genetic map constructed from a F_2 population of *Gossypium hirsutum* and *Gossypium darwinii*', Gene. Elsevier B.V., 574(2), pp. 273–286. doi: 10.1016/j.gene.2015.08.022.

Cherry, J. P. and Leffler, H. R. (1984) 'Seed', in Cotton, Agronomy Monograph. 24th edn. ASA-CSSA-SSSA, Madison, WI., pp. 511–558.

Churchill, G. A. and Harbor, B. (2001) 'A statistical framework for quantitative trait mapping', Genetics, 159, pp. 371–387.

Cianchetta, A. N. et al. (2015) 'Survey of *Fusarium oxysporum* f. sp. *vasinfectum* in the United States', The Journal of Cotton Science, 19, pp. 328–336.

CottonGen (2010) QTL Nomenclature. Available at: https://www.cottongen.org/help/nomenclature_qtl (Accessed: 13 October 2017).

Cowen, N. M. and Frey, K. J. (1987) 'Relationships between three measures of genetic distance and breeding behavior in oats (*Avena sativa* L.)', Euphytica, 36(1968), pp. 413–424.

Cox, T. S., Murphy, J. P. and Rodgers, D. M. (1986) 'Changes in genetic diversity in the red winter wheat regions of the United States', Proceedings of the National Academy of Sciences, 83, pp. 5583–5586. doi: 10.1073/pnas.83.15.5583.

Cronn, R. C. et al. (2002) 'Rapid diversification of the cotton genus (*Gossypium* : *Malvaceae*) revealed by analysis of sixteen nuclear and chloroplast genes', American Journal of Botany, 89(4), pp. 707–725.

Dalton-Morgan, J. et al. (2014) 'A high-throughput SNP array in the amphidiploid species *Brassica napus* shows diversity in resistance genes', Functional and

Integrative Genomics, 14(4), pp. 643–655. doi: 10.1007/s10142-014-0391-2.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) 'Maximum Likelihood from incomplete data via the EM algorithm', Journal of the Royal Statistical Society. Series, 39(1), pp. 1–38.

Desai, A. et al. (2006) 'Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*.', Genome, 49(4), pp. 336–345. doi: 10.1139/G05-116.

Diamond, J. M. (1998) Guns, germs and steel: a short history of everybody for the last 13,000 years. Random House.

Doerge, R. W. and Weir, B. S. (1994) 'Constructing genetic maps by rapid chain delineation', Journal of Quantitative Trait Loci, 2, pp. 1–14.

Dowd, C., Wilson, I. W. and McFadden, H. (2004) 'Gene expression profile changes in cotton root and hypocotyl tissues in response to infection with Fusarium oxysporum f. sp. vasinfectum.', Molecular Plant-Microbe Interactions, 17(6), pp. 654–667. doi: 10.1094/MPMI.2004.17.6.654.

Duvick, D. N. (1984) 'Genetic diversity in major farm crops on the farm and in reserve', Economic Botany, 38(2), pp. 161–178. Available at: http://www.jstor.org/stable/4254451.

Eathington, S. R. et al. (2007) 'Molecular markers in a commercial breeding program', Crop Science, 47(SUPPL. DEC.). doi: 10.2135/cropsci2007.04.0015IPBS.

Endrizzi, J. . E., Turcotte, E. L. and Kohel, R. J. (1985) 'Genetics, cytology, and evolution of *Gossypium*', Advances in genetics (USA).

Esbroeck, G. Van and Bowman, D. T. (1998) 'Cotton germplasm diversity and its

importance to cultivar development', Journal of Cotton Science, 2, pp. 121–129.

Flagel, L. E., Wendel, J. F. and Udall, J. A. (2012) 'Duplicate gene evolution, homoeologous recombination, and transcriptome characterization in allopolyploid cotton', BMC Genomics, 13(1), p. 302. doi: 10.1186/1471-2164-13-302.

Francia, E. et al. (2005) 'Marker assisted selection in crop plants', Plant Cell, Tissue and Organ Culture, 82(3), pp. 317–342. doi: 10.1007/s11240-005-2387-z.

Frelichowski, J. E. et al. (2006) 'Cotton genome mapping with new microsatellites from Acala "Maxxa" BAC-ends', Molecular Genetics and Genomics, 275(5), pp. 479–491. doi: 10.1007/s00438-006-0106-z.

Frost, J. (2014) How to interpret a regression model with low R-squared and low p values, The Minitab Blog. Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values (Accessed: 5 July 2017).

Fryxell, P. A. ., Craven, L. A. . and Stewart, J. M. (1992) 'A revision of *Gossypium* sect . *Grandicalyx* (Malvaceae), including the description of six new species', Systematic Botany, 17(1), pp. 91–114.

Gallagher, J. P. et al. (2017) 'A new species of cotton from Wake atoll, *Gossypium stephensii* (Malvaceae)', Systematic Botany, 42(1), pp. 115–123. doi: 10.1600/036364417X694593.

Ganal, M. W. et al. (2011) 'A large maize (*Zea mays* L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with

the B73 reference genome', PLoS ONE, 6(12). doi: 10.1371/journal.pone.0028334.

Garber, R. H. et al. (1979) 'Interaction of population levels of *Fusarium oxysporum* f. sp. *vasinfectum* and meloidogyne incognita on cotton', Journal of Nematology, 11(2), pp. 133–137.

Gardunia, B. W. (2006) Introgression from *Gossypium mustelinum* and *G. tomentosum* into Upland cotton, *G. hirsutum*.

Grover, C. E. et al. (2012) 'Assessing the monophyly of polyploid *Gossypium* species', plant systematics and evolution, 298(6), pp. 1177–1183. doi: 10.1007/s00606-012-0615-7.

Grover, C. E. et al. (2014) 'Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack', Genetic Resources and Crop Evolution, 62(1), pp. 103–114. doi: 10.1007/s10722-014-0138-x.

Grover, C. E. et al. (2015) 'Re-evaluating the phylogeny of allopolyploid *Gossypium* L.', Molecular Phylogenetics and Evolution. Elsevier Inc., 92(June), pp. 45–52. doi: 10.1016/j.ympev.2015.05.023.

Guo, W. et al. (2007) 'A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in *Gossypium'*, Genetics, 176(1), pp. 527–541. doi: 10.1534/genetics.107.070375.

Haley, C. S. and Knott, S. A. (1992) 'A simple regression method for mapping quantitative trait loci in line crosses using flanking markers', The Genetical

Society of Great Britain, 69, pp. 315–324.

Hall, C., Heath, R. and Guest, D. (2013) 'The infection process of *Fusarium oxysporum* f.sp. *vasinfectum* in Australian cotton', Australasian Plant Pathology, 42(1), pp. 1–8. doi: 10.1007/s13313-012-0169-8.

Hamilton, J. P. et al. (2011) 'Single nucleotide polymorphism discovery in elite north american potato germplasm', BMC Genomics, 12(1), p. 302. doi: 10.1186/1471-2164-12-302.

Han, Z. et al. (2006) 'Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton', Theoretical and Applied Genetics, 112(3), pp. 430–439. doi: 10.1007/s00122-005-0142-9.

Hans Van Os, Piet Stam, R. G. F. V. H. J. V. E. (2005) 'RECORD : a novel method for ordering loci on a genetic linkage map', Theoretical and Applied Genetics, 112, pp. 30–40. doi: 10.1007/s00122-005-0097-x.

Hendrix, B. and Stewart, J. M. (2005) 'Estimation of the nuclear DNA content of *Gossypium* species', Annals of Botany, 95(5), pp. 789–797. doi: 10.1093/aob/mci078.

Hoffman, K. L., Padberg, M. and Rinaldi, G. (2013) 'Traveling salesman problem', Encyclopedia of Operations Research and Management Science, 3(1), pp. 1573--1578. doi: 10.1201/b11506-7.

Holland, J. B. (2001) 'Epistasis and plant breeding', in Plant Breeding Review, pp. 27–92.

Hou, M. et al. (2013) 'Construction of microsatellite-based linkage map and mapping of

nectarilessness and hairiness genes in *Gossypium tomentosum*', Journal of Genetics, 92(3), pp. 445–459. doi: 10.1007/s12041-013-0286-3.

Hulse-Kemp, A. M. et al. (2015) 'BAC-end sequence-based SNP mining in allotetraploid cotton (*Gossypium*) utilizing resequencing data, phylogenetic inferences, and perspectives for genetic mapping.', G3: Genes|Genomes|Genetics, 5(6), pp. 1095–105. doi: 10.1534/g3.115.017749.

Hulse-Kemp, A. M. et al. (2015) 'Development of a 63K SNP array for cotton and high-density mapping of intra- and inter-specific populations of *Gossypium* spp.', G3: Genes|Genomes|Genetics, 5(June), p. g3.115.018416. doi: 10.1534/g3.115.018416.

Hutchinson, J. B., Silow, R. A. and Stephens, S. G. (1947) 'The Evolution of *Gossypium* and the differentiation of the cultivated cottons', The Quarterly Review of Biology, 24(2), pp. 143–144.

Ibitoye, D. O. and Akin-Idowu, P. E. (2010) 'Marker-assisted-selection (MAS): A fast track to increase genetic gain in horticultural crop breeding', African Journal of Biotechnology, 9(52), pp. 8889–8895. doi: 10.5897/AJB10.302.

Jenkins, J. N. et al. (2007) 'Genetic effects of thirteen *Gossypium barbadense* L. chromosome substitution lines in topcrosses with Upland cotton cultivars: II. Fiber quality traits', Crop Science, 47(2), pp. 561–572. doi: 10.2135/cropsci2006.06.0396.

Jiang, C. and Zeng, Z. B. (1997) 'Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines.', Genetica, 101(1), pp.

47–58. doi: 10.1023/A:1018394410659.

John, M. E. and Stewart, J. M. (1992) ‘Genes for jeans: biotechnological advances in cotton’, Trends in Biotechnology, 10(C), pp. 165–170. doi: 10.1016/0167-7799(92)90205-A.

Jorgenson, E. C. et al. (1978) ‘Influence of soil fumigation on fusarium-root-knot nematode disease complex of cotton in California’, Journal of Nematology, (10), pp. 228–231.

Kaur, S., Francki, M. G. and Forster, J. W. (2012) ‘Identification , characterization and interpretation of single-nucleotide sequence variation in allopolyploid crop species’, Plant Biotechnology Journal, 10, pp. 125–138. doi: 10.1111/j.1467-7652.2011.00644.x.

Khan, M. A., Stewart, J. M. D. and Murphy, J. B. (1999) ‘Evaluation of the *Gossypium* gene pool for foliar terpenoid aldehydes’, Crop Science, 39(1), pp. 253–258.

Kim, Y., Hutmacher, R. B. and Davis, R. M. (2005) ‘Characterization of California isolates of *Fusarium oxysporum* f. sp *vasinfectum*’, Plant Disease, 89(4), pp. 366–372. doi: Doi 10.1094/Pd-89-0366.

Kirkpatrick, M. (2010) ‘How and why chromosome inversions evolve’, PLoS Biology, 8(9). doi: 10.1371/journal.pbio.1000501.

Knott, D. R. (1987) ‘Transferring alien genes to wheat’, Wheat and Wheat Improvement., 2a(7), pp. 462–471.

Kohel, R. J. et al. (2001) ‘Molecular mapping and characterization of traits controlling fiber quality in cotton’, Euphytica, 121(2), pp. 163–172. doi:

10.1023/A:1012263413418.

Kohel, R. J., Endrizzi, J. E. and White, T. G. (1977) 'Evaluation of *Gossypium barbadense* L. chromosomes 6 and 17 in the *G. hirsutum* L. genome', Crop Science, 17(3), pp. 404–406.

Kosambi, D. D. (1944) 'The estimation of map distance from recombination values', Annuaire of Eugenetics, 12, pp. 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x.

Krapovickas, A. and Seijo, G. (2008) '*Gossypium ekmanianum* (Malvaceae)', Bonplandia, 17(1), pp. 55–63.

Lacape, J. et al. (2005) 'QTL analysis of cotton fiber quality using multiple × backcross generations', Crop Science, 45(1), pp. 123–140. doi: 10.2135/cropsci2005.0123a.

Lacape, J.-M. et al. (2003) 'A combined RFLP-SSR-AFLP map of tetraploid cotton based on a *Gossypium hirsutum* x *Gossypium barbadense* backcross population.', Genome, 46(4), pp. 612–626. doi: 10.1139/g03-050.

Lee, J. A. (1984) 'Cotton as a world crop', in Kohel RJ, L. C. (ed.) Cotton, Agronomy Monograph. American Society of Agronomy, Madison, pp. 1–25. doi: 48/1233/167 [pii]\r10.1126/science.48.1233.167.

Li, C. et al. (2016) 'Genome-Wide SNP linkage mapping and QTL analysis for fiber quality and yield traits in the Upland cotton recombinant inbred lines population', Frontiers in Plant Science, 7, pp. 1–16. doi: 10.3389/fpls.2016.01356.

Li, F. et al. (2014) 'Genome sequence of the cultivated cotton *Gossypium arboreum*', Nature Genetics, 46(6), pp. 567–572. doi: 10.1038/ng.2987.

Li, Z. K. et al. (2003) 'QTL x environment interactions in rice. I. Heading date and plant height', Theoretical and Applied Genetics, 108(1), pp. 141–153. doi: 10.1007/s00122-003-1401-2.

Liao, C. Y. et al. (2001) 'Effects of genetic background and environment on QTLs and epistasis for rice (*Oryza sativa* L.) panicle number', Theoretical and Applied Genetics, 103, p. 104111.

Lopez-Lavalle, L. A. B. et al. (2012) 'Molecular mapping of a new source of Fusarium wilt resistance in tetraploid cotton (*Gossypium hirsutum* L.)', Molecular Breeding, 30(2), pp. 1181–1191. doi: 10.1007/s11032-012-9705-z.

Luan, M. et al. (2009) 'QTL mapping for agronomic and fibre traits using two interspecific chromosome substitution lines of Upland cotton', Plant Breeding, 128(6), pp. 671–679. doi: 10.1111/j.1439-0523.2009.01650.x.

Luitel, K. P., Hudson, D. and Ethridge, D. (2015) 'Evaluating Cotton Utilisation in Nonwoven Textiles', The Journal of Cotton Science, 19, pp. 298–306.

Manjarrez-Sandoval, P. et al. (1997) 'RFLP genetic similarity estimates and coefficient of parentage as genetic variance predictors for soybean yield', Crop Science, 37(3), pp. 698–703. doi: 10.2135/cropsci1997.0011183X003700030002.

Mano, Y., Omori, F. and Takeda, K. (2012) 'Construction of intraspecific linkage maps, detection of a chromosome inversion, and mapping of QTL for constitutive root aerenchyma formation in the teosinte *Zea nicaraguensis*', Molecular Breeding, 29(1), pp. 137–146. doi: 10.1007/s11032-010-9532-z.

Margarido, G. R. A., Souza, A. P. and Garcia, A. A. F. (2007) 'OneMap : software for

genetic mapping in outcrossing species', Hereditas, 144, pp. 78–79. doi: 10.1111/j.2007.0018-0661.02000.x.

May, O. L., Bowman, D. T. and Calhoun, D. S. (1995) 'Genetic diversity of U.S. Upland cotton cultivars released between 1980 and 1990', Crop Science, 35(6), pp. 1570–1574.

McKenzie, W. . H. (1970) 'Fertility Relationships Among Interspecific Hybrid Progenies of *Gossypium*', Crop Science, 10, pp. 571–574.

Menezes, I. P. P. de et al. (2014) 'Genetic diversity and structure of natural populations of *Gossypium mustelinum*, a wild relative of cotton, in the basin of the De Contas River in Bahia, Brazil', Genetica, 142(1), pp. 99–108. doi: 10.1007/s10709-014-9757-6.

Meredith, Jr., W. R. (2000) 'Cotton yield progress – why has it reached a plateau?', Better Crops, 84(4), pp. 2–5.

Moricca, S. et al. (1998) 'Detection of *Fusarium oxysporum* f.sp. *vasinfectum* in cotton tissue by polymerase chain reaction', Plant Pathology, 47(4), pp. 486–494. doi: 10.1046/j.1365-3059.1998.00262.x.

Nagata, K. et al. (2015) 'Advanced backcross QTL analysis reveals complicated genetic control of rice grain shape in a *japonica* × *indica* cross', Breeding Science, 65(4), pp. 308–318. doi: 10.1270/jsbbs.65.308.

Park, Y. H. et al. (2005) 'Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population', Molecular Genetics and Genomics, 274(4), pp. 428–441. doi: 10.1007/s00438-

005-0037-0.

Patel, J. D. et al. (2014) 'Alleles conferring improved fiber quality from EMS mutagenesis of elite cotton genotypes', Theoretical and Applied Genetics, 127(4), pp. 821–830. doi: 10.1007/s00122-013-2259-6.

Paterson, A. H. et al. (2004) 'Reducing the genetic vulnerability of cotton', Crop Science, 44, pp. 1900–1901.

Paterson, A. H. et al. (2012) 'Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres.', Nature, 492(7429), pp. 423–7. doi: 10.1038/nature11798.

Percy, R. G. et al. (2014) 'The U. S. national cotton germplasm collection – its contents, preservation, characterization, and evaluation', in Abdurakhmonov, I. Y. (ed.) World Cotton Germplasm Resources. InTech, pp. 167–201.

Percy, R. G. and Kohel, R. J. (1999) 'Qualitative genetics', in cotton: origin, history, technology, and production. New York: John Wiley and Sons, Inc., pp. 319–360.

Perlak, F. J. et al. (2001) 'Development and commercial use of Bollgard® cotton in the USA - early promises versus today's reality', Plant Journal, 27(6), pp. 489–501. doi: 10.1046/j.1365-313X.2001.01120.x.

Pundir, N. S. (1972) 'Experimental embryology of *Gossypium arboreum* L. and *G. hirsutum* L. and their reciprocal crosses', Botanical Gazette, 133(1), pp. 7–26.

Rasmusson, D. C. and Phillips, R. L. (1997) 'Plant breeding progress and genetic diversity from de novo variation and elevated epistasis', Crop science (USA), pp. 303–310.

Reinisch, A. J. et al. (1994) 'A detailed RFLP map of cotton, *Gossypium hirsutum* X *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome', Genetics, 138(3), pp. 829–847.

Rong, J. et al. (2007) 'Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development', Genetics, 176(4), pp. 2577–2588. doi: 10.1534/genetics.107.074518.

Rooney, W. L. and Stelly, D. M. (1991) 'Preferential transmission and somatic elimination of a *Gossypium sturtianum* chromosome in *G. hirsutum*', Journal of Heredity, 82(2), pp. 151–155.

Rousset, M. et al. (2001) 'Use of recombinant substitution lines for gene mapping and qtl analysis of bread making quality in wheat', in Wheat in a global environment, pp. 195–202.

Saha, S. et al. (2004) 'Effect of chromosome substitutions from *Gossypium barbadense* L. 3-79 into *G. hirsutum* L. TM-1 on agronomic and fiber traits', The Journal of Cotton Science, 69, pp. 2–9.

Saha, S. et al. (2006) 'Effects of chromosome-specific introgression in Upland cotton on fiber and agronomic traits', Genetics, 172(3), pp. 1927–1938. doi: 10.1534/genetics.105.053371.

Saha, S. et al. (2010a) 'Genetic dissection of chromosome substitution lines of cotton to discover novel *Gossypium barbadense* L. alleles for improvement of agronomic traits', Theoretical and Applied Genetics, 120(6), pp. 1193–1205. doi:

10.1007/s00122-009-1247-3.

Saha, S. et al. (2010b) 'Genetic dissection of chromosome substitution lines of cotton to discover novel *Gossypium barbadense* L . alleles for improvement of agronomic traits', Theoretical and Applied Genetics, 120(6), pp. 1193–1205. doi: 10.1007/s00122-009-1247-3.

Saha, S. et al. (2011) 'Delineation of interspecific epistasis on fiber quality traits in *Gossypium hirsutum* by ADAA analysis of intermated *G. barbadense* chromosome substitution lines', Theoretical and Applied Genetics, 122(7), pp. 1351–1361.

Saha, S. et al. (2017) 'Registration of Two CS-B17-derived Upland cotton recombinant inbred lines with improved fiber micronaire', Journal of Plant Registrations, 0(0), p. 0. doi: 10.3198/jpr2015.09.0061crg.

Saha, S., Stelly, D. and Raska, D. (2011) 'Chromosome substitution lines: concept, development and utilization in the genetic improvement of Upland cotton', Plant breeding, pp. 107–128. doi: 10.1038/054138a0.

Said, J. I. et al. (2013) 'A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits , drought tolerance , and disease resistance in tetraploid cotton', BMC Genomics, 14, p. 776.

Said, J. I. et al. (2015) 'Cotton QTLdb: a cotton QTL database for QTL analysis, visualization, and comparison between *Gossypium hirsutum* and *G. hirsutum* × *G. barbadense* populations', Molecular Genetics and Genomics, 290(4), pp. 1615–1625. doi: 10.1007/s00438-015-1021-y.

Saunders, J. H. (1961) 'The wild species of *Gossypium* and their evolutionary history', Oxford University Press, p. 66.

Senchina, D. S. et al. (2003) 'Rate variation among nuclear genes and the age of polyploidy in *Gossypium*', Society for Molecular Biology and Evolution, 20(4), pp. 633–643. doi: 10.1093/molbev/msg065.

Shang, L. et al. (2016) 'Genetic analysis and QTL detection on fiber traits using two recombinant inbred lines and their backcross populations in Upland cotton', G3: Genes|Genomes|Genetics, 6(9), pp. 2717–2724. doi: 10.1534/g3.116.031302.

Shi, Y. et al. (2015) 'Constructing a high-density linkage map for *Gossypium hirsutum* x *Gossypium barbadense* and identifying QTLs for lint percentage', Journal of Integrative Plant Biology, 57(5), pp. 450–467. doi: 10.1111/jipb.12288.

Smith, C. W. and Cothren, J. T. (1999) Cotton : origin, history, technology, and production. 1st edn, Wiley series in crop science. 1st edn. Edited by C. W. Smith and J. T. Cothren. New York : Wiley.

Smith, S. et al. (1981) 'Fusarium wilt of cotton', in Cook, R. J., Toussoun, T. A., and Nelson, P. E. (eds) Fusarium, diseases, biology, and taxonomy. University Park, PA: Pennsylvania State University Press, pp. 29–38. Available at: file://catalog.hathitrust.org/Record/000110774.

Snider, J. L. et al. (2013) 'Quantifying genotypic and environmental contributions to yield and fiber quality in Georgia: data from seven commercial cultivars and 33 yield environments', Journal of Cotton Science, 292, pp. 285–292.

Song, Q. et al. (2013) 'Development and Evaluation of SoySNP50K, a high-density

genotyping array for soybean', PLoS ONE, 8(1), pp. 1–12. doi: 10.1371/journal.pone.0054985.

Souza, E. and Sorrells, M. E. (1991) 'Prediction of progeny variation in oat from parental genetic relationships', Theoretical and Applied Genetics, 82(2), pp. 233–241. doi: 10.1007/BF00226219.

Steele, K. A. et al. (2006) 'Marker-assisted selection to introgress rice QTLs controlling root traits into an Indian Upland rice variety', Theoretical and Applied Genetics, 112(2), pp. 208–221. doi: 10.1007/s00122-005-0110-4.

Steemers, F. J. et al. (2006) 'Whole-genome genotyping with the single-base extension assay', Nature Methods, 3(1), pp. 2005–2007. doi: 10.1038/NMETH842.

Stelly, D. M. et al. (2005) 'Registration of 17 Upland (*Gossypium hirsutum*) cotton germplasm lines disomic for different chromosome or arm substitutions', Crop Science, 45(6), p. 2663. doi: 10.2135/cropsci2004.0642.

Stephens, S. G. (1949) 'The cytogenetics of speciation in *Gossypium*; selective elimination of the donor parent genotype in interspecific backcrosses', Genetics, 34(5), pp. 627–637.

Stewart, J. M. (1995) 'Potential for crop improvement with exotic germplasm and genetic engineering', in Proceeding of the world cotton research conference-I. Brishbane, Australia, pp. 313–327.

Sun, F.-D. et al. (2012) 'QTL mapping for fiber quality traits across multiple generations and environments in Upland cotton.', Molecular Breeding. Springer Science & Business Media B.V., 30(1), pp. 569–582.

Tan, Y.-D. and Fu, Y.-X. (2006) 'A novel method for estimating linkage maps', Genetic Society of America, 173(4), pp. 2383–2390. doi: 10.1534/genetics.106.057638.

Tanksley, S. D. and McCouch, S. R. (1997) 'Seed banks and molecular maps: unlocking genetic potential from the wild', Science, 277(5329), pp. 1063–1066. doi: 10.1126/science.277.5329.1063.

Tanksley, S. D. and Nelson, J. C. (1996) 'Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines', Theoretical and Applied Genetics, 92(2), pp. 191–203. doi: 10.1007/s001220050114.

Tinker, N. A. et al. (2014) 'A SNP genotyping array for hexaploid oat', The Plant Genome, 7(3), p. 0. doi: 10.3835/plantgenome2014.03.0010.

Truco, M. J. et al. (2013) 'An ultra-high-density, transcript-based, genetic map of lettuce', G3: Genes|Genomes|Genetics, 3(4), pp. 617–631. doi: 10.1534/g3.112.004929.

Ulloa, M. et al. (2006) 'Breeding for Fusarium wilt race 4 resistance in cotton under field and greenhouse conditions', The Journal of Cotton Science, 10, pp. 114–127.

Ulloa, M. et al. (2009) 'Registration of four Pima cotton germplasm lines having good levels of Fusarium wilt race 4 resistance with moderate yields and good fibers', Journal of Plant Registrations, 3(2), pp. 198–202. doi: 10.3198/jpr2008.09.0530crg.

Ulloa, M. et al. (2011) 'Mapping Fusarium wilt race 1 resistance genes in cotton by

inheritance, QTL and sequencing composition', Molecular Genetics and Genomics, 286(1), pp. 21–36. doi: 10.1007/s00438-011-0616-1.

Ulloa, M. et al. (2013) 'Inheritance and QTL mapping of Fusarium wilt race 4 resistance in cotton', Theoretical and Applied Genetics, 126(5), pp. 1405–1418. doi: 10.1007/s00122-013-2061-5.

Ulloa, M. et al. (2016) 'Analysis of root-knot nematode and Fusarium wilt disease resistance in cotton (*Gossypium* spp.) using chromosome substitution lines from two alien species', Genetica. Springer International Publishing, 144(2), pp. 167–179. doi: 10.1007/s10709-016-9887-0.

Ulloa, M. et al. (2016) 'Registration of five Pima cotton germplasm lines (Pima SJ-FR05– Pima SJ-FR09) with improved resistance to Fusarium wilt race 4 and good lint yield and fiber quality', Journal of Plant Registrations, (10), pp. 154–158. doi: 10.3198/jpr2015.07.0043crg.

USDA-FAS (no date) Reports and Data - Cotton. Available at: https://apps.fas.usda.gov/psdonline/app/index.html#/app/downloads (Accessed: 20 July 2017).

Voorrips, R. E. (2002) 'MapChart : Software for the Graphical Presentation of Linkage Maps and QTLs', The Journal of Heredity, 93(1), pp. 77–78.

Waghmare, V. N. et al. (2005) 'Genetic mapping of a cross between *Gossypium hirsutum* (cotton) and the Hawaiian endemic, *Gossypium tomentosum*', Theoretical and Applied Genetics, 111(4), pp. 665–676. doi: 10.1007/s00122-005-2032-6.

Wang, B. et al. (2012) 'Molecular diversity, genomic constitution, and QTL mapping of fiber quality by mapped SSRs in introgression lines derived from *Gossypium hirsutum* x *G. darwinii* Watt', Theoretical and Applied Genetics, 125(6), pp. 1263–1274. doi: 10.1007/s00122-012-1911-x.

Wang, B. et al. (2016) 'A genetic map between *Gossypium hirsutum* and the Brazilian endemic *G. mustelinum* and its application to QTL mapping', G3: Genes|Genomes|Genetics, 6(June), pp. 1673–1685. doi: 10.1534/g3.116.029116.

Wang, B. et al. (2017) 'QTL analysis of cotton fiber length in advanced backcross populations derived from a cross between *Gossypium hirsutum* and *G. mustelinum*', Theoretical and Applied Genetics. Springer Berlin Heidelberg, 130(6), pp. 1297–1308. doi: 10.1007/s00122-017-2889-1.

Wang, C. and Roberts, P. A. (2006) 'A Fusarium wilt resistance gene in *Gossypium barbadense* and its effect on root-knot nematode-wilt disease complex.', Phytopathology, 96(7), pp. 727–734. doi: 10.1094/PHYTO-96-0727.

Wang, H. et al. (2015) 'QTL mapping for fiber and yield traits in Upland cotton under multiple environments', PLoS ONE, 10(6), pp. 1–14. doi: 10.1371/journal.pone.0130742.

Wang, H.-M. et al. (2008) 'Mapping and quantitative trait loci analysis of verticillium wilt resistance genes in cotton', Journal of Integrative Plant Biology, 50(2), pp. 174–182. doi: 10.1111/j.1744-7909.2007.00612.x.

Wang, K. et al. (2012) 'The draft genome of a diploid cotton *Gossypium raimondii*', Nature Genetics, 44(10), pp. 1098–1103. doi: 10.1038/ng.2371.

Wang, P. et al. (2012) 'Inheritance of long staple fiber quality traits of *Gossypium barbadense* in *G. hirsutum* background using CSILs', Theoretical and Applied Genetics, 124, pp. 1415–1428. doi: 10.1007/s00122-012-1797-7.

Wang, S. et al. (2014) 'Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array', Plant Biotechnology Journal, 12(6), pp. 787–796. doi: 10.1111/pbi.12183.

Watt, G. (1907) 'The wild and cultivated cotton plants of the world', Longmans, green and co.

Wendel, J. F. (1989) 'New world tetraploid cottons contain old world cytoplasm', Evolution, 86, pp. 4132–4136. doi: 10.1073/pnas.86.11.4132.

Wendel, J. F. et al. (2009) 'Evolution and natural history of the cotton genus', in Genetics and genomics of cotton. Springer US, pp. 3–22. doi: 10.1007/978-0-387-70810-2.

Wendel, J. F. and Cronn, R. C. (2003) 'Polyploidy and the evolutionary history of cotton.', in Advances, pp. 139–186. doi: 10.1016/S0065-2113(02)78004-8.

Wendel, J. F. and Grover, C. E. (2015) 'Taxonomy and evolution of the cotton genus, *Gossypium*', Cotton, 2nd edition. Madison, Wisconsin: Soil Science Society of America, Inc., (2015). Web.

Wendel, J. F., Rowley, R. and Stewart, J. M. (1994) 'Genetic diversity in and phylogenetic relationships of the Brazilian endemic cotton, *Gossypium mustelinum* (*Malvaceae*)', Plant Systematics and Evolution, 192(1–2), pp. 49–59. doi: 10.1007/BF00985907.

Xu, J. (2014) Transmission rates of *Gossypium mustelinum* and *G. tomentosum* SNP markers in early-generation backcrosses to cotton (*G. hirsutum* L.).

Yu, J. et al. (2013) 'Mapping quantitative trait loci for lint yield and fiber quality across environments in a *Gossypium hirsutum* x *Gossypium barbadense* backcross inbred line population', Theoretical and Applied Genetics, (126), pp. 275–287. doi: 10.1007/s00122-012-1980-x.

Yu, J. Z. et al. (2014) 'Mapping genomic loci for cotton plant architecture, yield components, and fiber properties in an interspecific (*Gossypium hirsutum* L. x *G. barbadense* L.) RIL population', Molecular Genetics and Genomics, pp. 1347–1367. doi: 10.1007/s00438-014-0930-5.

Zeng, Z.-B. (1993) 'Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci', Genetics, 90, pp. 10972–10976.

Zhai, H. et al. (2016) 'Identification of chromosome segment substitution lines of *Gossypium barbadense* introgressed in *G . hirsutum* and quantitative trait locus mapping for fiber quality and yield traits', PLoS ONE, 11(9), pp. 1–14. doi: 10.1371/journal.pone.0159101.

Zhang, J., Percy, R. G. and McCarty, J. C. (2014) 'Introgression genetics and breeding between Upland and Pima cotton: A review', Euphytica, 198(1), pp. 1–12. doi: 10.1007/s10681-014-1094-4.

Zhang, K. et al. (2012) 'Genetic mapping and quantitative trait locus analysis of fiber quality traits using a three-parent composite population in Upland cotton (*Gossypium hirsutum* L .)', Molecular Breeding, 29, pp. 335–348. doi:

10.1007/s11032-011-9549-y.

Zhang, Z. et al. (2011) 'QTL alleles for improved fiber quality from a wild Hawaiian cotton, *Gossypium tomentosum*', Theoretical and Applied Genetics, 123(7), pp. 1075–1088. doi: 10.1007/s00122-011-1649-x.

Zhang, Z. et al. (2016) 'Construction of a high-density genetic map by specific locus amplified fragment sequencing (SLAF-seq) and its application to quantitative trait loci (QTL) analysis for boll weight in Upland cotton (*Gossypium hirsutum*.)', BMC Plant Biology. 16(79). doi: 10.1186/s12870-016-0741-4.

Zhang, Z. et al. (2017) 'Construction of a high-density genetic map and its application to QTL identification for fiber strength in Upland cotton', Crop Science, 57, pp. 774–778. doi: 10.2135/cropsci2016.06.0544.

Zhiyuan, N. et al. (2014) 'Molecular tagging of QTLs for fiber quality and yield in the Upland cotton cultivar Acala-Prema', Euphytica, 195(1), pp. 143–156. doi: 10.1007/s10681-013-0990-3.

Zhu, Q.-H. et al. (2014) 'Transcriptome and complexity-reduced, DNA-based identification of intraspecies single-nucleotide polymorphisms in the polyploid *Gossypium hirsutum* L.', G3: Genes|Genomes|Genetics, 4(10), pp. 1893–1905. doi: 10.1534/g3.114.012542.

Zhu, T. et al. (2017) 'CottonFGD : an integrated functional genomics database for cotton', BMC. BMC Plant Biology, 17(1), pp. 1–9. doi: 10.1186/s12870-017-1039-x.