TITLE: **Toward Comprehensive Understanding of the Cotton Genome in Structure, Organization, Function and Evolution**

DISCIPLINE: **Genomics and Biotechnology**

AUTHORS:

H.-B. Zhang (Corresponding author)
Department of Soil and Crop Sciences
2474 TAMU
Texas A&M University
College Station, TX 77843
Phone: 979-862-2244
Fax: 979-845-0456
E-mail: hbz7049@tamu.edu

M.-K. Lee
C. Scheuring
Y. Rong
M. Goebel
Y.-H. Wu
L. Zhang
D. M. Stelly
C. W. Smith
Department of Soil and Crop Sciences
2474 TAMU
Texas A&M University
College Station, TX 77843

Toward Comprehensive Understanding of the Cotton Genome in Structure, Organization, Function and Evolution

ABSTRACT

Cottons are a leading fiber and oilseed crop and a model system for studies of plant genome size evolution, polyploidization, cell extension, and cellulose biosynthesis. To comprehensively understand structure, organization, function and evolution of the cotton genome, we are studying the cotton genome in different aspects. These include development of large-insert BAC and BIBAC resources, development of high-throughput techniques for genome, gene and QTL mapping, development of RIL mapping populations, physical mapping of the Upland cotton genome, isolation and characterization of repeated element repertoire, reconstruction of phylogeny and re-examination of the genome origin of polyploid cottons, analysis of underlying molecular mechanisms of genome size evolution and impact of polyploidization, domestication and breeding on genome evolution, and association analysis between gene expression and fiber trait performance. The results of these studies provide not only infrastructure and tools essential for advanced genomics research, but also novel insights into structure, organization, function and evolution of the cotton genome.

KEY WORDS:

BACs and BIBACs, gene expression, genome evolution, *Gossypium*, DNA marker, phylogeny, and physical mapping

Cottons (*Gossypium* ssp.) are important economically and biologically (for review, see Zhang et al., 2007). Economically, cottons are a world's leading textile fiber and oilseed crop. In the years 2004/2005, for instance, cottons were grown in an area of about 35 million hectares worldwide and the total world's production of cottons reached a record of about 23 million metric tons (http://www.fao.org/). Cotton fibers sustain the world's textile industry whereas cotton seeds are a major source of the world's food oil industry. Furthermore, cotton fiber production also significantly influences the production of artificial "synthetic" fibers that consumes nearly a billion of barrels of fossil oil annually in the world. Hence, enhanced production of cotton fibers will replace or significantly reduce the consumption of fossil oil that is used for synthetic fiber production, thus being saved for energy production. Finally, cottonseed oil also could be potentially used as biofuels. Biologically, cottons have been used as a model system to address many fundamental questions of biology. These include plant genome size evolution (e.g., Hendrix and Stewart, 2005; Grover et al., 2004; Wendel et al., 2002; Hawkins et al., 2006), plant polyploidzation (e.g., Beasley, 1942; Wendel, 1989; Wendel et al., 1995; Jiang et al., 1998; Cronn et al., 1999; Adams et al., 2004; Desai et al., 2006), plant cell expansion and cellulose biosynthesis (for review, see Kim and Triplett, 2001), and the molecular basis of biomass-based bioenergy production because celluloses are a major source of bioenergy production (for review, see Kim and Triplett, 2001).

Research of molecular genetics and genomics has been demonstrated in several crop plants and agricultural animals to provide powerful, essential and unprecedented tools for continued genetic improvement. Because of their importance in economy and biology, significant efforts have been made to develop genomic tools and resources in cotton for genetic improvement and biological research (for detail, see Zhang et al., 2007). These include DNA

markers, genetic maps, mapped genes and QTLs (quantitative trait loci), ESTs (expressed sequence tags), microarrays, gene expression profiles, large-insert bacterial artificial chromosome (BAC) and plant-transformation-competent binary BAC (BIBAC) libraries, and genome physical maps. Nevertheless, the research of cotton molecular genetics and genomics, in particular genomics, is far behind other world's leading crops, such as maize, wheat, rice and soybean. For instance, the total number of ESTs, which is a crucial resource for expression profiling and functional analysis of genes, for *Gossypium* spp. was 281,233 in GenBank as of April 27, 2007, which is only 71.8% of that of soybean, 26.1% of that of wheat, 24.2% of that of maize, and 23.1% of that of rice (Zhang et al., 2007). Whole genome integrated physical and genetic maps, which has been considered to be the central platform and "freeway" of many aspects of advanced genetics and genomics research (Zhang and Wing, 1997; Zhang and Wu, 2001; Wu et al., 2005), have been developed for rice (Tao et al., 2001; Chen et al., 2001; Li et al., 2007), soybean (Wu et al., 2004a), and maize (Nelson et al., 2005), whereas the development of whole-genome integrated physical maps for cotton is still in its infancy stage.

Here, we report the efforts of my laboratory at Texas A&M University in development of genomic resources and tools for and toward comprehensive understating of the cotton genome in structure, organization, function and evolution. Especially, we focus the presentation on the construction and characterization of large-insert BAC and BIBAC resources, development of recombinant inbred line (RIL) mapping populations for cotton gene and QTL refine and fine mapping and positional cloning, development of new tools for high-throughput cotton genome mapping, trait fine mapping and marker-assisted selection (MAS), physical mapping of the Upland cotton genome, reconstruction of phylogeny of the *Gossypium* species, estimation of impacts of polyploidization, domestication and breeding on the nucleotide binding site (NBS)-

leucine-rich repeat (LRR)-encoding gene family in the *Gossypium* genomes, and association analysis between gene expression and fiber trait performance.

Progress of cotton genomics research. The cotton research of my laboratory at Texas A&M University is focused on several aspects of structural, functional and evolutionary genomics. Reported below are the current statuses of several of these research areas.

1. Construction and characterization of large-insert BAC and BIBAC resources.

Large DNA fragments (>100 kb) cloned in BAC or BIBAC vectors and arrayed in microplates, i.e., large-insert, arrayed BAC and BIBAC libraries, have been demonstrated to be essential and desirable resources and tools for many aspects of advanced genomics and genetics research (Zhang et al., 1996; Zhang and Wu, 2001; Wu et al., 2004b; Ren et al., 2005; He et al., 2007). BAC and BIBAC libraries have widely been used in many research areas of genomics and molecular biology, including DNA marker development (Lichtenzveig et al., 2005), DNA marker conversion from one type to another for MAS (Cregan et al., 1999), positional cloning of genes and QTLs (for review, Zhang, 2007), isolation and characterization of structural and regulatory genes (Chen et al., 1997; Patocchi et al., 1999), long-range genome analysis (Chen et al., 1997; Patocchi et al., 1999), organization and evolution of multigene families (Patocchi et al., 1999), cytologically physical mapping (Zwick et al., 1998), clone-based genome physical mapping (Zhang and Wing, 1997; Marra et al., 1999; Mozo et al., 1999; Hoskins et al., 2000; Chang et al., 2001; International Human Genome Mapping Consortium, 2001; Tao et al., 2001; Zhang and Wu, 2001; Ren et al., 2003; Wu et al., 2004a; Xu et al., 2004, 2005; Wallis et al., 2005; Wu et al., 2005; Zhang et al., 2006; Li et al., 2007), and large-scale genome sequencing (Adams et al., 2000; The Arabidosis Genome Initiative, 2000; International Human Genome

Sequencing Consortium, 2001; International Rice Genome Sequencing Project, 2005; Tyler et al., 2006).

To facilitate cotton genomics research, we constructed two BAC and one BIBAC libraries from two genotypes of Upland cotton, which accounts for >90% of the world's cotton production. One BAC library was constructed with *Bam*HI in the BAC vector pBeloBAC11 from cv. Auburn 623, the root-knot nematode resistance source and a parent of a newly developed RIL mapping population (see below). The library consists of 44,160 clones with an average insert size of 140 kb, representing 2.7 x Upland cotton genome equivalents. The other BAC library was constructed with *Eco*RI in the BAC vector pECBAC1 from cv. Texas Marker-1 (TM-1). The library consists of 76,800 clones and has an average insert size of about 175 kb, representing a 6.0 x genome coverage of Upland cotton. The BIBAC library was also constructed from the cultivar TM-1 with *Bam*HI in the plant-transformation-competent BIBAC vector pCLD04541. The library consists of 76,800 clones with an average insert size of 130 kb and covers about 4.1 x of the Upland cotton genome. Therefore, the combined TM-1 BAC and BIBAC libraries contain a total of 153,600 clones arrayed in 400 384-well microplates, together covering > 10 x of the Upland cotton haploid genome. As shown above in other plant and animals species, these BAC and BIBAC libraries provide resources essential for advanced genomics and genetics research of cotton in many aspects. Using these BAC and BIBAC libraries, we are developing SSR (Simple Sequence Repeat) and SNP (Single Nucleotide Polymorphism) markers that have been proven well-suited for marker-assisted germplasm analysis and breeding in cotton, constructing a whole-genome physical map of the Upland cotton genome, cloning and characterization of genes and QTLs important to cottons, and long-rang analysis of the cotton genomes (see below).

2. Development of RIL mapping populations for cotton gene and QTL refine and fine mapping and positional cloning. A large number of genes and QTLs have been genetically mapped in cottons (for review, see Ulloa et al., 2006 and Zhang et al., 2007). Nevertheless, almost all of the genes and QTLs mapped were mapped using mapping populations at F_2 or early generations. Consequently, variations were observed in the mapping results, including those among different populations and different environments within a population. It was also observed that the markers for most of the genes and QTLs mapped were distant from the targeted genes or QTLs, which is not well-suited for MAR of the genes and QTLs in cotton breeding. Therefore, refine mapping of the genes and QTLs, especially for those of quantitative traits, using populations that can be assayed in multiple environments such as RILs is essential to position them to the cotton genome accurately. It is also necessary to develop closer linked DNA markers and fine map the loci using large populations to make use of the developed DNA markers for MAS in cotton breeding and finally, for positional cloning of the corresponding genes and QTLs.

It has been demonstrated that RIL populations are desired for gene and QTL mapping in multiple environments (including years and locations) (for review, see Zhang, 2007). This is because each RIL is homologous while the variation among RILs is maintained. The seeds of same genotype that are required for multiple-environment experiments can be readily reproduced. The multiple-environment mapping of traits has been proven and widely used as an effective approach to mapping genes and QTLs accurately (for review, see Zhang, 2007). For fine mapping of QTLs, it has been shown that nearly-isogenic RIL populations developed from nearly-isogenic lines for the targeted locus are most desired (Alpert and Tanksley, 1996; Fridman et al., 2000; Uauy et al., 2006) because such populations can be readily developed, allow

phenotypic assay experiments to be conducted in multiple environments and, most importantly, provide homogenous genetic backgrounds that differ only in the targeted locus for phenotypic assay of each line. Therefore, we, in collaboration with A. F. Robinson, USDA/ARS, College Station, Texas, developed a RIL population from an interspecific cross between Pima S6 and Auburn 623, and a nearly-isogenic RIL population from a cross between M-240 and Deltapine 61 to refine and fine map the genes and QTLs of importance, especially those for root-knot nematode resistance. The Pima S6 x Auburn 623 population consists of 190 RILs to facilitate mapping operation by PCR (polymerase chain reaction) in two 96-well plates (190 RILs plus 2 parents). This population separates not only in root-knot nematode resistance, but also in many other agronomic traits including fiber traits. The M-240 x Deltapine 61 population consists of over 5,000 nearly-isogenic RILs for root-knot nematode resistance. This population is especially developed for high-resolution mapping and map-based cloning of the genes for root-knot nematode resistance in cotton. Since we are also working on a couple of other RIL populations developed by C. W. Smith, Laboratory for Cotton Genetic Improvement, Texas A&M University, College Station, Texas, particularly for refine mapping of major traits of cotton fiber yield and quality (see below), the Pima S6 x Auburn 623 RIL population, along the RIL populations developed by C. W. Smith, will provide resources for refine mapping of all major traits of agronomic importance in cultivated cottons.

3. Development of new tools for high-throughput genotyping, trait high-resolution mapping and marker-assisted selection. SSRs and SNPs have been demonstrated to be desirable DNA markers for cotton genome mapping, germplasm analysis and marker-assisted breeding because they are PCR-based and thus require only a small amount of DNA for assay, co-dominant and highly polymorphic in the cotton genome. Previous community efforts have

resulted in at least 5,000 SSR markers in cottons (for review, see Zhang et al., 2007). Because polyploid cultivated cottons have a genome size of about 2,400 Mb/haploid (Hendrix and Stewart, 2005) and their genetic linkage maps cover approximately 5,000 cM (for review, see Zhang et al., 2007), it would be, on average, 1 cM or less than 500 kb per marker if 5,000 of the SSR markers could be mapped to the cotton genome. However, only a small percentage of the markers have been mapped genetically to date and use of the markers for germplasm analysis and MAS is also limited (for review, see Zhang et al., 2007). Therefore, it is of significance to develop a high-throughput, universal and economical method to map the SSR and SNP markers to the cotton genome and to use them to refine and high-resolution map the traits of agronomic importance, analyze the cotton germplasm in a large scale, and conduct MAS for enhanced cotton breeding.

We developed a high-throughput, universal and economical method using capillary sequencers that is well-suited for high-throughput mapping of the cotton genomes, germplasm analysis and MAS using SSR and SNP markers (M. Goebel and H.-B. Zhang, unpublished). Using this method, a single PCR reaction amplifies 4 SSR or SNP marker loci and a single labeling kit labels all SSR or SNP markers at a rate of 4 SSR or SNP markers per reaction. The labeled PCR products are fractionated on a capillary sequencer in a multiplex format from a minimum of 4 to 40 samples per channel. Therefore, 5,000 SSR or SNP markers could be mapped to the cotton genome within a few months using a mapping population consisting of 100 RILs if the method is used. Furthermore, because this method only needs nanograms of DNA from each line, has a high resolution (0.20 nucleotide) and is economical, it is well-suited for large-scale genome mapping, gene and QTL fine mapping, germplasm analysis, and MAS in cottons.

4. Physical mapping of the Upland cotton genome. Whole-genome, BAC and/or BIBAC-based, integrated physical and genetic maps have played a central role in genomics research of human, plants, animals and microbes (Zhang and Wing, 1997; Zhang and Wu, 2001; Wu et al., 2005). They provide not only central platforms , but also "freeways" for many aspects, if not all, of modern genomics research, including large-scale transcript or gene mapping, region-targeted marker development for fine mapping and MAS of genes and QTLs, map-based gene and QTL cloning, local and whole genome comparative analysis, genome sequencing, and functional analysis of genomic sequences (Wu et al., 2005). Therefore, whole-genome, BAC/BIBAC-based, integrated physical and genetic maps have been developed for a number of plant and animal species. In plants, these include *Arabidopsis thaliana* (Marra et al., 1999; Chang et al., 2001), indica rice (Tao et al., 2001), japonica rice (Chen et al. 2002; Li et al., 2007), soybean (Wu et al., 2004a), and maize (Nelson et al., 2005).

To advance cotton genomics and genetics research, we, in collaboration with R. J. Kohel, USDA/ARS, College Station, Texas, and D. M. Stelly, Texas A&M University, College Station, Texas, USA, are working toward development of a whole-genome physical map of Upland cotton from the TM-1 BAC and BIBAC libraries described above by fingerprint analysis using capillary sequencers (Xu et al., 2004; Wu et al., 2005). Nearly 120,000 BIBACs and BACs (~7 x) selected from the TM-1 BIBAC and BAC libraries have been fingerprinted and a draft BAC/BIBAC contig map has been constructed (unpublished). Currently, additional clones are being analyzed to reach about 10 x genome coverage that has been demonstrated to be most efficient for high-quality physical map construction (Xu et al., 2004, 2005; Ren et al., 2005). Moreover, Upland cotton is an allotetraploid containing two subgenomes A and D, and each of the subgenomes is likely an ancient polyploid. The nature of the cotton genome makes the

physical map construction much more complicated than the genome physical map construction of diploid species. Therefore, we are further verifying and characterizing each of the map contigs and sorting them according to their origin of subgenomes using several methods to develop the physical map into a reliable and robust one. This integrated physical and genetic map of Upland cotton is expected to provide a platform not only for targeted DNA marker development, cloning and characterization of genes and QTLs, and long-range genome analysis, but also for whole-genome sequencing of Upland cotton.

5. Reconstruction of the phylogeny of *Gossypium* species and inference of the genome origin of the polyploid cottons. Knowledge of phylogenetic relationships among taxa or genomes is not only crucial for plant genetics research and breeding, but also essential for comprehensive studies of the plant genomes, especially genome organization, function and evolution. Due to this reason, the genus *Gossypium* has been studied extensively in phylogeny and a consensus phylogenetic tree of the genus constructed from the data generated using different methods (Wendel and Cronn, 2003). These included chloroplast (cp) DNA restriction site variation, and nucleotide sequence variations of selected chloroplast genes, nuclear ribosomal DNA (5S gene and spacer, 5.8S gene and its flanking internal transcribed spacers) and low-copy nuclear genes (Wendel and Albert, 1992; Cronn et al., 1996, 2002). Nevertheless, several significant questions and/or uncertainties about their phylogeny need to be further investigated. Variation of nuclear DNA repeated sequences has been successfully used to decipher intractable questions in the genome origin and evolution of several plant species complexes (e.g., Dvorak and Zhang, 1990, 1992a, b; Zhang and Dvorak, 1991, 1992; Zhao and Kochert, 1993). This is because they evolve more rapid than gene sequences, often constitute a large portion of the genomes of higher plants, largely disperse throughout the entire genome, and

usually are genome- or species-specific (Dvorak and Zhang, 1990, 1992a, b; Zhang and Dvorak, 1991, 1992). It was shown that the genomes of *Gossypium* are abundant in repeated sequences and well dispersed in the genomes (Zhao et al., 1995, 1998; Hawkins et al., 2006). These results suggested that the variation of nuclear repeated sequences in the cotton genomes could provide a useful tool for phylogenetic analysis of the genus *Gossypium*.

We, in collaboration with A. E. Percival, USDA/ARS, College Station, Texas, re-examined the genome origin and evolution of the genus *Gossypium* and reconstructed the phylogenetic tree of the genus using the variation of 22 nuclear repeated sequence families randomly selected from the cotton genome repeat element repertoire isolated recently (L. Zhang, M.-L. Lee, X. Zhang, H.-B. Zhang, and D. M. Stelly, unpublished). DNA was isolated from 87 accessions of 35 species representing all eight genome groups of the genus (A through G and K) and analyzed. A total of 642 restriction fragments of repeated sequences were used for phylogenetic analysis of the species. A phylogenetic tree of the species was constructed and shown to be consistent with those existing trees in major clades (Wendel and Cronn, 2003); however, significantly different branching among some species and low correspondence in the A and D genomes between diploid and polyploid species were observed.

First, comparative analysis of the repeated sequence fragments that are genome- or species-specific (defined genome or species marker fragments) showed that only the A-genome and D-genome diploid species exclusively shared maker fragments with the polyploid species, suggesting that only the D- and A-genome diploid species possibly contributed to the subgenomes of the polyploid species. However, none of the extant A- and D-genome diploid species could be claimed the donor of the A or D subgenome of the polyploid species due to their low correspondence. Further analysis of the data suggested that the polyploid species of

Gossypium originated before the split between the A-genome diploid species and the split among the D-genome diploid species at the early stage of the D-and A-genome linage evolution.

Second, the phylogenetic tree of polyploid species constructed in this study was significantly different from that of Wendel and Cronn (2003). Wendel and Cronn (2003) classified the five polyploid species into three branches, one consisting of *G. mustelinum* (AD_4), one consisting of *G. tomentosum* (AD_3) and *G. hirsutum* (AD_1), and the third containing *G. barbadense* (AD_2) and *G. darwinii* (AD_5). However, this study showed that *G. barbadense* (AD_2) and *G. tomentosum* (AD_3) formed one branch, *G. mustelinum* (AD_4) and *G. darwinii* (AD_5) formed the second branch, and *G. hirsutum* (AD_1) alone formed the third branch. Our result was supported by the data of flavonoid (Parks et al., 1975) and DNA fingerprinting (Khan et al., 2000).

Third, the positions of the B-, F-, and E-genome species in the diploid tree constructed in this study differed from that of Wendel and Cronn (2003). Wendel and Cronn (2003) showed the B-genome species is sister branches to the A- and F-genome species, and the E-genome is basal to the A-, F-, and B-genome lineage. In our study, the B-genome species with the E-genome species was found to form a sister branch to the F- and A-genome species.

Finally, the phylogenetic relationships among the thirteen diploid species of the D-genome linage were also different between the tree constructed in this study and that of Wendel and Cronn (2003). Although several branches of this lineage between the species agree with the one of Wendel and Cronn (2003), the order of the lineage branches was very different.

6. Estimation of impacts of polyploidization, domestication and breeding on the NBS-LRR-encoding gene family in the *Gossypium* genomes. Domestication, breeding and polyploidization represent three major forces of crop plant genome evolution. However, little

is known about how these activities influence the evolution of crop plant genomes at the genomic level, the fate and evolution of a multigene family in particular. The nucleotide-binding site (NBS)-leucine-rich repeat (LRR)-encoding gene family, probably consisting of hundreds of gene members in the cotton genome (He et al., 2004), represents a large multigene family in plants. Because the family contributes to plants at least 80% of the genes conferring resistance to various pathogens including bacteria, fungi, viruses and nematodes, its fate and evolution are readily subjected to the three evolutionary forces of crop plants, domestication, breeding and polyploidization.

We investigated the variation and fate of the NBS-LRR-encoding gene family in the course of cotton evolution using 108 lines representing 35 species and all eight basic genome groups of the genus *Gossypium*. Preliminary analysis showed that the number of genes of the family is significantly affected by genome size, genome polyploidization, domestication, and breeding. The results provide the first insights into how crop plant domestication, breeding and polyploidization affect the fate and evolution of a multigene family that is subjected to both natural and artificial selections. The knowledge provides a novel basis and raises a novel challenge for plant genetic improvement and breeding, especially breeding for biotic stress resistance.

7. Association analysis between gene expression and fiber trait performance.

Development of a large number of ESTs from developing fibers and ovaries in cottons has provided useful resources for cotton fiber gene microarray fabrication and high-throughput tools to assay the activities or expression of the genes (Arpat et al., 2004; Haigler et al., 2005; Lee et al., 2006; Shi et al., 2006; Udall et al., 2006; Yang et al., 2006). Using the microarrays, the expression of the fiber genes was profiled and comparatively analyzed at several stages of

fiber development, including initiation (Wu et al., 2006), elongation (Arpat et al., 2004; Shi et al., 2006) and second cell wall deposition (Arpat et al., 2004). However, little is known about what the significantly increased or decreased expression activities of the genes at a particular developmental stage mean with regard to the final fiber yield and quality. The goals of our effort into this project are to address this question, thus advancing our knowledge of cotton fiber development and developing a genomics-assisted system to help effectively manipulate the genes involved in fiber development for cotton breeding.

As the initial step of this effort, we, in collaboration with T. Wilkins, Texas Tech University, Lubbock, Texas, and C. W. Smith, Texas A&M University, College Station, Texas, profiled the expression of fiber genes in the 10-dpa (days post-anthesis) fibers collected from a panel of four cultivars of Upland and Sea Island cottons, with two lines for each species, using the cotton fiber gene-specific long-oligo arrays newly developed by T. Wilkins. The four cultivars include NMSI-1311, Sea Island Barbados, Acala 1517-99, and 94 L-25. They are the parents of three RIL populations developed by C. W. Smith for refine mapping, cloning and characterization of the QTLs important to fiber yield and quality. We are currently analyzing the microarray data and expect to derive a list of genes that differentially expressed in 10-dpa fibers between Upland and Sea Island cottons as well as between different cultivars of Upland or Sea Island cotton. These results will provide knowledge and tools to translate the expression activities of the fiber genes into fiber yield and quality, thus providing knowledge to design the genomics-assisted system to effectively manipulate desirable fiber genes from both Upland and Sea Island cottons for high fiber yield and quality cultivar breeding.

CONCLUDING REMARKS

We have developed resources and tools that are essential for comprehensive studies of the cotton genomes in structure, organization, function and evolution. These resources and tools include the repeated sequence repertoire that constitutes a large portion of the cotton genome, large-insert BAC and BIBAC libraries, robust RIL mapping populations, and a high-throughput, universal and economical method that is well-suited for high-resolution mapping of the cotton genome, genes and QTLs, large-scale germplasm analysis, and MAS in cotton breeding. These resources and tools, along with the community-previously developed ESTs, microarrays, DNA markers, and BAC libraries (for more information, see Zhang et al., 2007), will allow us to study the cotton genome comprehensively.

Using a sample of 22 nuclear repeated sequence families randomly selected from the cotton genome repeat element repertoire developed by collaboration between the laboratories of M. D. Stelly and H.-B. Zhang, we have reconstructed the phylogenetic tree of the major *Gossypium* species and re-examined the origin of the polyploid cotton genomes. The differences between the phylogenetic tree of the cotton species and the origin of the polyploid cotton genome inferred in this study and those previously inferred by others (Wendel and Cronn, 2003) have raised a need of further research in these regards. Since the nuclear repeated sequence families used in this study well disperse throughout the cotton genome and represent a large portion of the cotton genome, the results of this study likely have a reasonable representation for the relationships of the *Gossypium* genomes, thus providing useful knowledge for our understanding of the cotton genome structure, organization, function, and evolution.

The whole-genome BAC/BIBAC physical map of Upland cotton under construction in our laboratories represents a most advanced status of cotton genome physical mapping even

though significant efforts are needed to develop it into a reliable and robust physical map. According to our previous studies in genome physical mapping in other species (Chang et al., 2001; Tao et al., 2001; Ren et al., 2003; Wu et al., 2004a, 2005; Xu et al., 2004, 2005; Zhang et al., 2006; Li et al., 2007), additional clones are needed to be analyzed for construction of a quality physical map for the Upland cotton genome. It has been shown in our previous genome physical mapping work of soybean (Wu et al., 2004a) that is considered to be an ancient polyploid that many additional efforts are needed to construct the physical map of the polyploid cotton genome. Nevertheless, given the importance of a whole genome physical map for advanced genomics and genetics research, it is imperative to develop a whole-genome reliable and robust physical map of Upland cotton not only for its genome sequencing and analysis, but also for many other aspects of genomics and genetics research in cottons.

We, for the first time, show that the evolution and fate of the NBS-LRR-encoding gene family is significantly influenced by genome polyploidization, domestication, and breeding using the cotton genus as a model system. Although further investigations are needed to test whether this result remains in other gene and sequence families of the cotton species and in other plant species, this study indicates that cotton domestication and breeding have functioned not only on the NBS-LRR gene allele constitution of the cotton genome, as known in general, but also on the number and combination of the NBS-LRR genes in the cotton genome. This finding reveals a novel basis of cotton breeding, thus providing new knowledge for cotton genetic improvement.

Upland and Sea Island cottons have been traditionally shown to differ significantly in fiber yield and quality. The list of the genes that we identified in this study that differentially expressed in 10-dpa fibers between the two major cultivated cotton species may account, in part, for the difference of fiber yield and quality between the species. However, further investigations

are needed to address how the differentially expressed genes between them contribute to the final cotton fiber yield and quality, for which a new experiment is being conducted in our laboratory using the list of the differentially expressed genes. The accomplishment of this project will provide new knowledge and tools for continued and enhanced cotton genetic improvement.

REFERENCES

Adams, K.L., R. Percifield, and J.F. Wendel. 2004. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics 168: 2217-2226.

Adams, M.D., S.E. Celniker, and E.A. Holt, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287: 2185-2195.

Alpert, K.B., and S.D. Tanksley. 1996. High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw*2.2: A major fruit weight quantitative trait locus in tomato. Proc. Natl. Acad. Sci. 93: 15503-15507.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796-815.

Arpat, A., M. Waugh, J.P. Sullivan, M. Gonzales, D. Frisch, D. Main, T. Wood, A. Leslie, R. Wing, and T. Wilkins. 2004. Functional genomics of cell elongation in developing cotton fibers. Plant Mol. Biol. 54: 911-929.

Beasley, J.O. 1942. Meiotic chromosome behavior in species hybrids, haploids, and induced polyploids of *Gossypium*. Genetics 27: 25-54.

Chang, Y.-L., Q. Tao, C. Scheuring, K. Meksem, and H.-B. Zhang. 2001. An integrated map of *Arabidopsis thaliana* for functional analysis of its genome sequence. Genetics 159: 1231-1242.

Chen, M., G. Presting, and W.B. Barbazuk, et al. 2002. An integrated physical and genetic map of the rice genome. Plant Cell 14: 537-545.

Chen, M., P. SanMiguel, A.C. de Oliveira, S.-S. Woo, H.-B. Zhang, R.A. Wing, and J.L. Bennetzen. 1997. Microcolinearity in *sh2*-homologous regions of the maize, rice and sorghum genomes. Proc. Natl. Acad. Sci. 94: 3431-3435.

Cregan, P.B., J. Mudge, E.W. Fickus, L.F. Marek, and D. Danes, et al. 1999. Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor. Appl. Genet. 98: 919-928.

Cronn, R.C., R.L. Small, T. Haselkorn, and J.F. Wendel. 2002. Rapid diversification of the cotton genus (*Gossypium:* Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. Amer. J. Bot. 89: 707-725.

Cronn, R.C., R.L. Small, and J.F. Wendel. 1999. Duplicated genes evolve independently in allopolyploid cotton. Proc. Natl. Acad. Sci. 96: 14406-14411.

Cronn, R.C., X.P. Zhao, A.H. Paterson, and J.F. Wendel. 1996. Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. J. Mol. Evol. 42: 685-705.

Desai, A., P.W. Chee, J.K. Rong, O.L. May, and A. H. Paterson. 2006. Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*. Genome 49: 336-345.

Dvorak, J., and H.–B. Zhang. 1990. Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proc. Natl. Acad. Sci. 87: 9640-9644.

Dvorak, J., and H.-B. Zhang. 1992a. Application of molecular tools for study of the phylogeny of diploid and polyploid taxa in Triticeae. Hereditas 116: 37-42.

Dvorak, J., and H.-B. Zhang. 1992b. Reconstruction of the phylogeny of *Triticum* from variation in repeated nucleotide sequences. Theor. Appl. Genet. 84: 419-429.

Fridman, E., T. Pleban, and D. Zamir. 2000. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proc. Natl. Acad. Sci. 97: 4718-4723.

Grover, C.E., H. Kim, R.A. Wing, A.H. Paterson, and J.F. Wendel. 2004. Incongruent patterns of local and global genome size evolution in cotton. Genome Res. 14: 1474-1482.

Haigler, C. H., D. Zhang, and C.G. Wilkerson. 2005. Biotechnological improvement of cotton fibre maturity. Physiol. Plant. 124: 285-294.

Hawkins, J.S., H. Kim, J.D. Nason, R.A. Wing, and J.F. Wendel. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res. 16: 1252-1261.

He. L., C. Du, L. Covaleda, A.F. Robinson, J.Z.Yu, R.J. Kohel, and H.-B. Zhang. 2004. Cloning, characterization, and evolution of the NBS-encoding resistance gene analogue family in polyploid cotton (*Gossypium hirsutum* L.). Mol. Plant-Microbe Interact. 17: 1234-1241.

He, L., C. Du, Y. Li, C. Scheuring, and H.-B. Zhang. 2006. Large-insert bacterial clone libraries and their applications. *In* Z. Liu (ed.) Aquaculture Genome Technologies. Blackwell Publishing, Ames, IA (in press).

Hendrix, B., and J. McD. Stewart. 2005. Estimation of the nuclear DNA of *Gossypium* species. Ann. Bot. 95: 789-797.

Hoskins, R.A., C.R. Nelson, and B.P. Berman, et al. 2000. A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. Science 287: 2271-2274.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

International Human Genome Mapping Consortium. 2001. A physical map of the human genome. Nature 409: 934-941.

International Rice Genome Sequencing Project, 2005. The map-based sequence of the rice genome. Nature 436: 793-800.

Jiang, C.X., R.J. Wright, K.M. El-Zik, and A.H. Paterson.1998. Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). Proc. Natl. Acad. Sci. 95: 4419-4424.

Khan, S.A., D. Hussain, E. Askari, J.M. Stewart, K.A. Malik, and Y. Zafar. 2000. Molecular phylogeny of *Gossypium* species by DNA fingerprinting. Theor. Appl. Genet. 101: 931-938.

Kim, H.J., and B.A. Triplett. 2001. Cotton fiber growth in planta and *in vitro*. Models for plant cell elongation and cell wall biogenesis. Plant Physiol. 127: 1361-1366.

Lee, J.J.,O.S.S. Hassan, W. Gao, N.E. Wei, R.J. Kohel, X.-Y. Chen, P. Payton, S.-H. Sze, D.M. Stelly, and Z.J. Chen. 2006. Developmental and gene expression analyses of a cotton naked seed mutant. Planta 223: 418-432.

Li ,Y., T. Uhm, C. Ren, C. Wu, T.S. Santos, M.-K. Lee, B. Yan, F. Santos, A. Zhang, Z. Xu, C. Scheuring, A. Sanchez, A.C. Millena, H.T. Nguyen, H. Kou, D. Liu, and H.-B. Zhang. 2007. A plant-transformation-competent BIBAC/BAC-based map of rice for functional analysis and genetic engineering of its genomic sequence. Genome 50: 278-288.

Lichtenzveig, J., C. Scheuring, J. Dodge, S. Abbo, and H.-B. Zhang. 2005. Construction of BAC and BIBAC libraries and their applications for generation of SSR markers for genome analysis of chickpea, *Cicer arietinum* L. Theor. Appl. Genet. 110: 492-510.

Marra, M., T. Kucaba, and M. Sekhon, et al. 1999. A map for sequence analysis of the *Arabidopsis thaliana* genome. Nat. Genet. 22: 265-270.

Mozo, T., K. Dewar, P. Dunn, J.R. Ecker, S. Fischer, S. Kloska, H. Lehrach, M. Marra, R. Martienssen, S. Meier-Ewert, and T. Altmann. 1999. A complete BAC-based physical map of the *Arabidopsis thaliana* genome. Nat. Genet. 22: 271-275.

Nelson, W.M., A.K. Bharti, E. Butler, F. Wei, G. Fuks, H. Kim, R.A. Wing, J. Messing, and C. Soderlund. 2005. Whole-genome validation of high-information-content fingerprinting. Plant Physiol. 139: 27-38.

Parks, C.R., D.E. William, and D.L. Dreyer. 1975. The application of flavonoid distribution to taxonomic problems in the genus Gossypium. Bull. Torrey Bot. Club 102: 350-361.

Patocchi, A., B.A. Vinatzer, S. Gianfranceschi, H.-B. Zhang, S. Sansavini, and C. Gessler. 1999. Construction of a 550-kb BAC contig spanning the genomic region containing the apple scab resistance gene *Vf*. Mol. Gen. Genet 262: 884-891.

Ren, C., M.-K. Lee, B. Yan, K. Ding, B. Cox, M.N. Romanov, J.A. Price, J.B. Dodgson, and H.-B. Zhang. 2003. A BAC-based physical map of the chicken genome. Genome Res. 13: 2754-2758.

Ren, C., Z. Xu, S. Sun, M.-L. Lee, C. Wu, C. Scheuring, T.S. Santos, and H.-B. Zhang. 2005. Genomic DNA Libraries and Physical Mapping. p. 173-213. *In* K. Meksem and G. Kahl (eds.) The Handbook of Plant Genome Mapping: Genetic and Physical Mapping. Wiley-VCH Verlag GmbH, Weinheim, Germany.

Shi, Y.H., S.W. Zhu, X.Z. Mao, J.X. Feng, Y.M. Qin, L. Zhang, J. Cheng, L.P. Wei, Z.Y. Wang, and Y.X. Zhu. 2006. Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. Plant Cell 18: 651-664.

Tao, Q., Y.-L. Chang, J. Wang, H. Chen, C. Scheuring, M.N. Islam-Faridi, B. Wang, D.M. Stelly, and H.-B. Zhang. 2001. Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis. Genetics 158: 1711-1724.

Tyler, B.M., S. Tripathy, and X. Zhang, et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science 313: 1261-1266.

Uauy, C., A. Distelfeld, T. Fahima, A. Blechl, and J. Dubcovsky. 2006. A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science 314: 1298-1301

Udall, J.A., J.M. Swanson, and K. Haller, et al. 2006. A global assembly of cotton ESTs. Genome Res. 16: 441-450.

Ulloa, M., C. Brubaker, and P. Chee. 2006. Cotton. *In* C. Kole (ed.) Genome Mapping & Molecular Breeding. Vol. 7: Technical Crops. Springer, Heidelberg, Berlin, New York, Tokyo.

Wallis, J.W., J. Aerts, and M.A.M. Groenen, et al. 2004. A physical map of the chicken genome. Nature 432: 761-764.

Wendel, J.F. 1989. New World tetraploid cottons contain Old World cytoplasm. Proc. Natl. Acad. Sci. 86: 4132-4136.

Wendel, J.F., and V.A. Albert. 1992. Phylogenetics of the cotton genus (*Gossypium* L.): Character-state weighted parsimony analysis of chloroplast DNA restriction site data and its systematic and biogeographic implications. Syst. Bot. 17: 115-143.

Wendel, J.F., and R.C. Cronn. 2003. Polyploidy and the evolutionary history of cotton. Adv. Agron. 78: 139-186.

Wendel, J.F., R.C. Cronn, I. Alvarez, B. Liu, R.L. Small, and D.S. Senchina. 2002. Intron size and genome size in plants. Mol. Biol. Evol. 19: 2346-2352.

Wendel, J.F., A. Schnabel, and T. Seelanan.1995. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc. Natl. Acad. Sci. 92: 280-284.

Wu, Y., A.C. Machado, R.G. White, D.J. Llewellyn, and S. Dennis. 2006. Expression profiling identifies gene expressed early during lint fiber initiation in cotton. Plant Cell Physiol. 18: 651-664.

Wu, C., S. Sun, M.-L. Lee, Z. Xu, C. Ren, and H.-B. Zhang. 2005. Whole genome physical mapping: An overview on methods for DNA fingerprinting. p. 257-284. *In* K. Meksem, and G. Kahl (eds.) The Handbook of Plant Genome Mapping: Genetic and Physical Mapping. Wiley-VCH Verlag GmbH, Weinheim, Germany.

Wu, C., S. Sun, P. Nimmakayala, F.A. Santos, R. Springman, K. Meksem, K. Ding, D. Lightfoot, and H.-B. Zhang. 2004a. A BAC and BIBAC-based physical map of the soybean genome. Genome Res. 14: 319-326.

Wu, C., Z. Xu, and H.-B. Zhang. 2004b. DNA Libraries. p. 385-425. *In* R. A. Meyers (ed.) Encyclopedia of Molecular Cell Biology and Molecular Medicine. Vol. 3 (2nd Edition). Wiley-VCH Verlag GmbH, Weinheim, Germany.

Xu, Z., S. Sun, L. Covaleda, K. Ding, A. Zhang, C. Scheuring, and H.-B. Zhang. 2004. Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage and contig map quality. Genomics 84: 941-951.

Xu, Z., M. van den Berg, C. Scheuring, L. Colaveda, H. Lu, F.A. Santos, T. Uhm, M.-L. Lee, C. Wu, S. Liu, and H.-B. Zhang. 2005. Genome-wide physical mapping from large-insert clones by fingerprint analysis with capillary electrophoresis: A robust physical map of *Penicillium chrysogenum*. Nucleic Acids Res. 33: e50.

Yang, S.S., F. Cheung, J.J. Lee, M. Ha, N.E. Wei, S.-H. Sze, D.M. Stelly, P. Thaxton, B. Triplett, C.D. Town, and Z.J. Chen. 2006. Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. Plant J. 47: 761-775.

Zhang, H.-B. 2007. Map-based cloning of genes and QTLs. *In* C. Kole and A. Abbott (eds.) Plant Molecular Mapping and Breeding. Springer (in press)

Zhang, H.-B., and J. Dvorak. 1991. The genome origin of tetraploid species of *Leymus* (Poaceae: Triticeae) inferred from variation in repeated nucleotide sequences. Amer. J. Bot. 78: 871-884.

Zhang, H.-B., and J. Dvorak. 1992. The genome origin and evolution of hexaploid *Triticum crassum* and *Triticum syriacum* determined from variation in repeated nucleotide sequences. Genome 35: 806-814.

Zhang, H.-B., Y. Li, B. Wang, and P. Chee. 2007. Recent advances in cotton genomics. International Journal of Plant Genomics (accepted).

Zhang, X., C. Scheuring, S. Tripathy, Z. Xu, C. Wu, A. Ko, S.K. Tian, F. Arredond, M.-K. Lee, F.A. Santos, H.-B. Zhang, and B.M. Tyler. 2006. An integrated BAC and genome sequence physical map of *Phytophthora sojae*. Mol. Plant-Microbe Interact. 19: 1302-1310.

Zhang, H.-B., and R.A. Wing. 1997. Physical Mapping of the rice genome with BACs. Pl. Mol. Biol. 35: 115-127.

Zhang, H.-B., S.-S. Woo, and R.A. Wing. 1996. BAC, YAC and Cosmid Library Construction. p. 75-99. *In* G. Foster and D. Twell (eds.) Plant Gene Isolation: Principles and Practice. Foster John Wiley & Sons, Ltd., England.

Zhang, H.-B., and C. Wu. 2001. BACs as tools for genome sequencing. Plant Physiol. Biochem. 39: 195-209.

Zhao, X., and G. Kochert. 1993. Clusters of interspersed repeated DNA sequences in the rice genome (*Oryza*). Genome 36: 944-953.

Zhao, X., Y. Si, R.E. Hanson, C.F. Crane, H.J. Price, D.M. Stelly, J.F. Wendel, and A.H. Paterson. 1998. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res. 8: 479-492.

Zhao, X., R.A. Wing, and A.H. Paterson. 1995. Cloning and characterization of the majority of repetitive DNA in cotton (*Gossypium* L.). Genome 38: 1177-1188.

Zwick, M.S., M.N. Islam-Faridi, D.G. Czeschin Jr., R.A. Wing, G.F. Hart, D.M. Stelly, and H.J. Price. 1998. Physical mapping of the *liguleless* linkage group in *Sorghum bicolor* using rice RFLP-selected sorghum BACs. Genetics 148: 1983-1992.