

Toward Sequencing Cotton (*Gossypium*) Genomes

Despite rapidly decreasing costs and innovative technologies, sequencing of angiosperm genomes is not yet undertaken lightly. Generating larger amounts of sequence data more quickly does not address the difficulties of sequencing and assembling complex genomes de novo. The cotton (*Gossypium* spp.) genomes represent a challenging case. To this end, a coalition of cotton genome scientists has developed a strategy for sequencing the cotton genomes, which will vastly expand opportunities for cotton research and improvement worldwide.

WHY SEQUENCE COTTON GENOMES?

Cotton is the world's most important natural textile fiber (Fig. 1A) and a significant oilseed crop. The seed is an important source of feed, foodstuff, and oil. World consumption of cotton fiber is approximately 115 million bales or approximately 27 million metric tons per year (National Cotton Council, <http://www.cotton.org/>, 2006). Genetic improvement of fiber production and processing will ensure that this natural renewable product will be competitive with petroleum-derived synthetic fibers. Moreover, modifying cottonseed for food and feed could profoundly enhance the nutrition and livelihoods of millions of people in food-challenged economies.

Cotton production provides income for approximately 100 million families, and approximately 150 countries are involved in cotton import and export. Its economic impact is estimated to be approximately \$500 billion/year worldwide. China is the largest producer and consumer of raw cotton, but more than 80 countries, including Australia, some African countries, India, Pakistan, the United States, Mexico, and Uzbekistan, also produce cotton. The United States is the second largest producer, and grows cotton worth approximately \$6 billion/year for fiber and approximately \$1 billion/year for cottonseed oil and meal. Cotton is a major economic driver for some developing countries, like Uzbekistan, which annually produces approximately 4 million tons of raw cotton and exports fiber worth approximately \$900 million.

Cotton fiber is an outstanding model for the study of plant cell elongation and cell wall and cellulose biosynthesis (Kim and Triplett, 2001). Each seed has approximately 25,000 cotton fibers, each of which is a single and greatly elongated cell from the epidermal layer of the ovule (Fig. 1B). The fiber is composed of nearly pure cellulose, the largest component of plant biomass. Compared to lignin, cellulose is easily convertible to biofuels. Translational genomics of cotton

fiber and cellulose may lead to the improvement of diverse biomass crops.

The genus *Gossypium* includes approximately 45 diploid ($2n = 2x = 26$) and five tetraploid ($2n = 4x = 52$) species, all exhibiting disomic patterns of inheritance. Diploid species ($2n = 26$) fall into eight genomic groups (A–G, and K). The African clade, comprising the A, B, E, and F genomes (Wendel and Cronn, 2003), occurs naturally in Africa and Asia, while the D genome clade is indigenous to the Americas. A third diploid clade, including C, G, and K, is found in Australia. All 52 chromosome species, including *Gossypium hirsutum* and *Gossypium barbadense*, are classic natural allotetraploids that arose in the New World from interspecific hybridization between an A genome-like ancestral African species and a D genome-like American species. The closest extant relatives of the original tetraploid progenitors are the A genome species *Gossypium herbaceum* (A1) and *Gossypium arboreum* (A2) and the D genome species *Gossypium raimondii* (D5) 'Ulbrich' (Brubaker et al., 1999). Polyploidization is estimated to have occurred 1 to 2 million years ago (Wendel and Cronn, 2003), giving rise to five extant allotetraploid species. Interestingly, the A genome species produce spinnable fiber and are cultivated on a limited scale, whereas the D genome species do not (Applequist et al., 2001). More than 95% of the annual cotton crop worldwide is *G. hirsutum*, Upland or American cotton, and the extra-long staple or Pima cotton (*G. barbadense*) accounts for less than 2% (National Cotton Council, http://www.cotton.org, 2006). Understanding the contribution of the A and D subgenomes to gene expression in the allotetraploids may facilitate improving fiber traits (Jiang et al., 1998; Saha et al., 2006; Yang et al., 2006).

Decoding cotton genomes will be a foundation for improving understanding of the functional and agronomic significance of polyploidy and genome size variation within the *Gossypium* genus. The haploid genome sizes are estimated to be approximately 880 Mb for *G. raimondii* 'Ulbrich', approximately 1.75 Gb for *G. arboreum*, and approximately 2.5 Gb for *G. hirsutum* (Hendrix and Stewart, 2005). Variation in DNA content in the diploid species reflects increases and decreases in copy numbers of various repeat families (Zhao et al., 1998), especially retrotransposon-like elements (Hawkins et al., 2006). DNA content of the allopolyploids is approximately the sum of the A and D genome progenitors, and nearly all of the approximately 22,000 amplified fragment length polymorphism fragments surveyed are additive in the allopolyploids (Liu et al., 2001). This suggests a role of genetic and epigenetic mechanisms for gene expression in phenotypic variation and selection of allotetraploid species (Jiang et al.,

Figure 1. Cotton bolls at maturity (A) and cotton fibers under electron microscope (B). Photos courtesy of Mike Dougherty from the National Cotton Council (A) and Barbara Triplett (B).



1998; Wendel, 2000; Adams et al., 2003; Yang et al., 2006; Chen, 2007).

WHAT RESOURCES ARE AVAILABLE?

Genomic resources such as bacterial artificial chromosomes (BACs), ESTs, linkage maps, and integrated genetic and physical maps provide landmarks for sequence analysis and assembly.

Linkage maps in tetraploid cotton have been most densely populated by analysis of interspecific *G. hirsutum* × *G. barbadense* F₂ families (Reinisch et al., 1994; Rong et al., 2004) and backcross lines (Lacape et al., 2005; Guo et al., 2007) due to low levels of DNA polymorphism within cotton species. Mapping populations have also been developed for *G. hirsutum* × *Gossypium tomentosum* F₂ (Waghmare et al., 2005) and *Gossypium mustelinum* × *G. hirsutum* (P. Chee and A. Paterson, unpublished data). Molecular marker linkage groups were localized and orientated with various interspecific hypoaneuploid F₁ hybrids available for most chromosomes, and elsewhere by in situ hybridization (Hanson et al., 1995; Saha et al., 2006; Wang et al., 2006; Ji et al., 2007). Synteny and locus order were also determined by wide-cross whole-genome radiation hybrid mapping, a method complementary to other forms of cotton genome mapping (Gao et al., 2006).

At least a dozen genetic maps of crosses between diverse cotton species and genotypes are available, most made to map specific traits and quantitative trait loci (QTLs). Some of these maps collectively include approximately 5,000 DNA markers (approximately 3,300 restriction fragment length polymorphisms, approximately 700 amplified fragment length polymorphisms, approximately 1,000 simple sequence repeats, and approximately 100 single nucleotide polymorphisms). In addition, sequence-tagged site-based maps consisting of 2,584 loci at 1.72-cM (approximately 600 kb) intervals in tetraploids (AD genomes), 1,014 loci at 1.42-cM (approximately 600 kb) intervals in diploids (D genome; Rong et al., 2004, 2005), and an EST-simple sequence repeat-based genetic map of 1,710 loci at 1.92-cM intervals in tetraploids (AD genomes; Guo et al., 2007) are available. There is a high degree of co-

linearity among the respective genome types (Rong et al., 2005).

Of particular long-term value are permanent recombinant inbred lines (RILs) and chromosome substitution lines. RILs have already begun to contribute to QTL definition, e.g. for a *G. hirsutum* × *G. barbadense* cross (Frelichowski et al., 2006) and intraspecific crosses within *G. hirsutum* (Ulloa et al., 2005; Abdurakhmonov et al., 2007; Shen et al., 2007). Near-isogenic disomic substitution lines of *G. hirsutum* enable the localization of net phenotypic effects. Moreover, chromosome-specific RILs enable high-resolution QTL definition and mapping (Stelly et al., 2005).

Reference maps have incorporated diverse types and sources of DNA markers. Jean-Marc Lacape and his colleagues have integrated linkage maps developed by researchers in China (T. Zhang), France (J.M. Lacape), and the United States (A. Paterson and M. Ulloa) into TropGENE-DB (<http://tropgenedb.cirad.fr/en/cotton.html>) using a CMap comparative map viewer (Nguyen et al., 2004). A similar map viewer has been implemented in the CottonDB (<http://cottondb.org>) and the Cotton Microsatellite Database (<http://www.cottonmarker.org>) that contains approximately 8,000 microsatellites (Blenda et al., 2006). The further development of comprehensive linkage maps will be used to anchor and assemble genomic sequences.

BAC libraries have been developed for several *G. hirsutum* cultivars ('0-613-2R', 'Acala Maxxa', 'Auburn 623', 'Tancot HQ95', and 'TM-1'), *G. barbadense* ('Pima S6'), two *G. arboreum* strains (AKA8401 and Jinglinzhongmian), *G. raimondii*, *Gossypium longicalyx*, and an outgroup (*Gossypioides kirkii*). A total of 10 genome equivalents of *G. raimondii* BACs has been fingerprinted using standard procedures (Marra et al., 1997). All genetically mapped probes have been incorporated into the fingerprint assembly using the overlapping oligonucleotides hybridization method (Cai et al., 1998). The assembly will be publicly available via a WebFPC site and incorporated into the existing BACMan resource at the Plant Genome Mapping Laboratory (www.plantgenome.uga.edu). A *G. hirsutum* L. 'TM-1' library has been used to develop integrated genetic and physical maps (R. Kohel, J. Yu, and

T. Zhang, unpublished data). A *G. hirsutum* L. '0-613-2R' library has been successfully used to locate the restorer of fertility gene in a 100-kb region (Yin et al., 2006) and to assign linkage groups to identified chromosomes using BAC-fluorescence in situ hybridization (FISH; Wang et al., 2006).

As of July 18, 2007, 356,889 *Gossypium* sequences were in GenBank, including 40,069 ESTs from *G. arboreum* (A), 67,098 from *G. raimondii* (D), 232,006 from *G. hirsutum* (AD tetraploid), and a few from other *Gossypium* members (Arpat et al., 2004; Udall et al., 2006; Yang et al., 2006; Taliercio and Boykin, 2007). Among these ESTs, many are from developing fiber and are enriched in putative MYB and WRKY transcription factors and phytohormone regulators (Yang et al., 2006). Transcription factors in these families are known to be important in the development of *Arabidopsis* (*Arabidopsis thaliana*) leaf trichomes, and phytohormonal effects on fiber cell development in immature cotton ovules cultured in vitro are well documented (Beasley and Ting, 1974). Moreover, A subgenome ESTs of all functional classifications are dramatically enriched in *G. hirsutum* fiber (Yang et al., 2006), a result consistent with the production of long lint fibers in A genome species. Some ESTs have been used to develop sequence-specific markers in breeding and to construct microarrays, leading to the identification of many candidate genes involved in fiber cell initiation and elongation (Arpat et al., 2004; Lee et al., 2006; Shi et al., 2006; Wu et al., 2006; Udall et al., 2007).

The Malvales (including cotton) are the nearest relative to *Arabidopsis* outside of the Brassicales for which detailed genetic and physical maps have been described (Bowers et al., 2003). Comparative analyses reveal a considerable degree of synteny/colinearity between the ancestral cotton and *Arabidopsis* genomes. A total of 1,738 (62%) sequenced loci in cotton had matches in *Arabidopsis* (Rong et al., 2005). Gaining access to the unique features that distinguish cotton from other plants both as an economic crop and a botanical model might benefit from translational genomics, leveraging of structural and functional information from *Arabidopsis*.

WHICH SEQUENCING STRATEGIES ARE BEST?

A comprehensive strategy needs to consider present needs along with long-term goals in relation to economics, technology, and priorities. A strong case can be made for complete sequencing of one or more representatives of each *Gossypium* genome group, A, B, C, D, E, F, G, K, and a tetraploid-derived AD ($n = 26$) genome (Paterson, 2006). Continuing progress in sequencing throughput and cost reduction will render this goal increasingly feasible and desirable.

Sequencing representatives from each diploid clade will be important for molecular dissection of evolutionary patterns and biological phenomena, including the genomic and morphological diversity that has permitted species within the genus to adapt to a wide

range of ecosystems in warmer and arid regions of the world. Sequences from A and D genome diploid species will aid tetraploid AD genome sequence assembly and could prove to be invaluable for revealing differences in gene content and expression patterns across the ploidy levels and for providing insight into polyploid genome evolution. Although there is an approximately 3-fold variation in genome size among the diploids, the high degree of conservation of gene order at the macro level between diploids and tetraploids (Brubaker et al., 1999; Rong et al., 2004; Desai et al., 2006) suggests that the vast majority of sequence data from diploids will extrapolate directly to tetraploids. Sequencing an elite *G. hirsutum* genome, AD, will provide the ultimate reference and resource for application-oriented structural, functional, and bioinformatic needs for the species that accounts for >95% of world cotton production. Sequencing an elite *G. arboreum* or *G. herbaceum* genome will provide valuable data on fiber genes. Comparisons of four species across two ploidy levels, including A1, A2, D5, and AD tetraploid subgenomes, will provide clues as to how polyploidy and domestication "interact." Parallel comparisons between domesticated and nondomesticated forms of the A and AD genome species will shed light on the effects of artificial versus natural selection.

Based on these considerations, one can envision multiple and parallel approaches to reveal genome diversity and complete genome information of *Gossypium* genomes. Additional ESTs should be sequenced from other diploid (e.g. C, G, and K genomes) and tetraploid (e.g. *G. barbadense*, AD) clades and in late fiber development stages such as secondary wall biosynthesis (Haigler et al., 2005). Sequencing using gene enrichment techniques such as methylation filtration and C_0t -based cloning that appear to offer complementary coverage of the low-copy DNA will generate novel genomic sequences that are absent in EST collections. A pilot study in methylation filtration comparing *G. raimondii*, *G. arboreum*, *G. hirsutum*, and *G. barbadense* is under way (B.E. Scheffler, S. Saha, and Orion Genomics, unpublished data).

The whole-genome shotgun sequence of the smallest *Gossypium* genome, *G. raimondii* (approximately 880 Mb), will provide fundamental information about gene content and organization. The U.S. Department of Energy Joint Genome Institutes (<http://www.jgi.doe.gov/>) has selected *G. raimondii* for a pilot study for shotgun sequencing at 0.5 \times coverage to better define the genome and establish a workable strategy for its complete sequencing.

A partially or fully sequenced *G. raimondii* genome will establish the critical initial template for characterizing the spectrum of diversity among the eight *Gossypium* genome types and three polyploid clades (Wendel and Cronn, 2003). A survey of approximately 100 of the most abundant repetitive families in the tetraploid genome showed only four to be abundant in the D genome but rare or absent in the A genome (Zhao et al., 1998), which diverged from the D genome

of *G. raimondii* about 5 to 10 million years ago (Senchina et al., 2003). Thus, most high-copy repetitive DNA families in the D genome are at least 5 to 10 million years old and likely to be amenable to assembly by a whole-genome shotgun approach.

A BAC-based AD genome sequence may offer superior opportunities to elucidate the types and frequencies of changes that distinguish polyploid from diploid cottons. The process could be greatly enhanced by using the finished genome sequence of a diploid species as a template and guide. Intergenomic concerted evolution and the presence of recently amplified repetitive DNA families would be problematic for a whole-genome shotgun approach. A reasonable approach is to establish minimum tiling path of fingerprinted contigs of *G. hirsutum* homoeologous chromosomes. This goal can be achieved by developing integrated homoeologous chromosome maps that include anchored DNA markers in linkage maps and BAC-end sequences in physical maps that can be further validated by radiation hybrid mapping and/or BAC-FISH (Hanson et al., 1995; Wang et al., 2006). FISH of landed BACs indicated that homoeologous segments were readily detectable by BAC-FISH for low-copy probes and that they seemed amenable to differentiation on the basis of FISH signal strength (Wang et al., 2007). Large duplicated segments have been reported within individual corresponding homoeologous chromosomes, suggesting ancient or recent genome expansion in cotton genomes (Rong et al., 2005; Wang et al., 2007). It will be prudent to sequence and assemble representative homoeologous BACs and/or a few pairs of homoeologous chromosomes prior to large-scale sequencing of *G. hirsutum* tetraploid genomes.

WHAT ARE THE GOALS?

The cotton community and industry are cooperatively developing workshops and communication methods for planning, coordinating, and executing sequencing and post-sequencing activities. The key questions under consideration are: (1) which species should we sequence; and (2) which techniques should be used for each genome? In the long term, a singularly important goal will be to establish the complete genome sequence of the most widely cultivated cotton, i.e. *G. hirsutum*. Given its genomic redundancies, large size (approximately 2.5 Gb), polyploid nature, and other complexities, we anticipate a need to experimentally assess potential approaches that range from autonomous to heavily reliant on sequence from related genomes, e.g. *G. raimondii* and perhaps *G. herbaceum* or *G. arboreum*.

Toward this long-term goal, we envision the following specific actions.

1. Whole-genome shotgun sequencing of *G. raimondii*, a probable ancestor of cultivated cottons and among the smallest *Gossypium* genomes, to provide fundamental information about gene content and organization.

2. Comparative sequencing of corresponding segments of tetraploid *G. hirsutum* to reveal the technical obstacles likely to be encountered during complete sequencing.
3. Develop and implement a strategy to deliver high-quality sequence of *G. hirsutum*. This may very well require establishment of a minimum tiling path of finger-printed contigs of *G. hirsutum* homoeologous chromosomes.
4. Develop bioinformatic and database tools to assemble, analyze, and make the information useable to the cotton community.

Future characterization and utilization of sequence information should integrate functional and structural genomic resources at the molecular and in silico levels, sequence full-length cDNAs for genome annotation and expression assays, perform detailed annotation of the cotton genome sequence to support gene discovery and map-based cloning in this species, implement a large-scale platform for identifying DNA sequence diversity (single nucleotide polymorphisms and genome-specific polymorphisms), facilitate high-resolution whole-genome association studies, develop genomic tiling arrays to support gene expression and epigenomic analysis of biological and agronomic traits, and sequence and annotate small RNAs and microRNAs and identify their targets.

WHAT ARE THE CHALLENGES IN COTTON GENOME SEQUENCING AND GENOMICS?

To build and take full advantage of comprehensive cotton genomic resources, the most important factors to consider are fund raising, effort coordination, data dissemination and management, and data analysis and utilization. To coordinate genomic research in cotton, the International Cotton Genome Initiative (<http://icgi.tamu.edu/>) was established in 2000 with a mission to increase knowledge of the structure and function of the cotton genome for the benefit of the global community. A single-community Web site will be identified to establish a newsgroup list-server that will allow researchers to express and discuss their ideas about cotton genome sequencing and genomic research.

The amount of data generated from various sequencing projects will be extremely large and difficult to comprehend for many prospective end users, so it is essential to develop a data management system that can facilitate access and utilization of genomic and sequence data. In addition to the CMap and Cotton Microsatellite databases (see above), CottonDB (<http://cottondb.org>) provides genomic, genetic, and taxonomic information, including germplasm, markers, genetic and physical maps, trait studies, sequences, and bibliographic citations. The Cotton Portal (<http://gossypium.info>) offers the community a single port of entry to participating Cotton Web resources. One participating resource, the Cotton Diversity Database

(<http://cotton.agtec.uga.edu>; Gingle et al., 2006), provides for an interface relating to performance trial, phylogenetic, genetic, and comparative data, and is closely integrated with comparative physical, EST, and genomic (BAC) sequence data, expression profiling resources, and the capacity for additional integrative queries. Cotton oligo-gene microarrays consisting of approximately 23,000 70-mer oligos designed from 250,000 ESTs can be found at the Web site (<http://cottonrevolution.info/microarray>).

There is a great need to expand bioinformatic infrastructure for managing, curating, and annotating the cotton genomic sequences that will be generated in the near future. A model community database example is The Arabidopsis Information Resource (<http://www.arabidopsis.org/>). The cotton sequence database of the future should be able to host and manage cotton information resources in cotton using community-accepted genome annotation, nomenclature, and gene ontology. Some existing databases may be upgraded to effectively handle a large amount of data flow and community requests, but additional resources will be sought to support key bioinformatic needs.

A universal challenge for sequencing polyploid genomes is the discrimination among paralogous, orthologous, and homoeologous sequences in diploid and allotetraploid species. *Gossypium* species are paleopolyploids (Bowers et al., 2003; Rong et al., 2005). Moreover, allopolyploids contain two or more sets of homoeologous chromosomes, leading to genetic and epigenetic changes in subgenomes and their functions (Chen, 2007). Developing new bioinformatic tools and software for assembly and annotation of allopolyploid genomes is a prerequisite for sequencing cotton and other polyploid genomes such as wheat, oat, and sugarcane. A completely sequenced cotton genome will provide a reference for re-sequencing many genomes in *Gossypium* species using traditional and new sequencing technologies (e.g. 454, Solexa, and SOLiD; Bentley, 2006). The best combination of technologies will be employed to establish a high-quality reference sequence anchored to physical and genetic maps. This sequence will be used to query homologous and orthologous genomes and to investigate the gene and allele basis of phenotypic and evolutionary diversity for cotton improvement.

CONCLUDING REMARKS

Sequenced cotton genomes will ultimately stimulate fundamental research on genome evolution, polyploidization and associated diploidization, gene expression, cell differentiation and development, cellulose synthesis, cell growth, molecular determinants of cell wall biogenesis, and epigenomics. Practical ramifications will include improvement of biological processes key to safe and sustainable production of high-yielding and high-quality fiber, seed, and biomass crops as well as expanded use of cotton germplasm and products. These advances will be underpinned by practical

improvement in elements key to all of agriculture, e.g. improvement of yield, water-use efficiency, abiotic and biotic stress tolerance/resistance, and reduction of fertilizer and pesticide requirements. While some objectives are more tangible than others, the economic, health, and ecological (and, thus, societal) impacts are truly compelling on both national and international scales. The international community is committed, organized, and convinced of the immediate need and value of sequencing cotton genomes.

ACKNOWLEDGMENTS

We thank Joe Ecker (Salk Institute) for moderating a cotton genome sequencing white paper discussion forum and for insightful and constructive comments received from the members of the International Cotton Genome Initiative. We thank members of the cotton genomics and breeding community for their input and apologize for not citing many enlightening papers owing to space limitations. Support for cotton research is provided by grants from the National Science Foundation, U.S. Department of Agriculture, Cotton Inc., National Science Foundation of China, and additional state support groups and funding agencies in Australia, Belgium, China, India, Pakistan, the United States, Uzbekistan, and represented countries. The Cotton Genome Sequencing White Paper can be found at http://algodon.tamu.edu/sequencing/docs/2WhitePaper12_11_2006.pdf.

Received August 20, 2007; accepted September 10, 2007; published December 6, 2007.

LITERATURE CITED

- Abdurakhmonov IY, Buriev ZT, Saha S, Pepper AE, Musaev JA, Almatov A, Shermatov SE, Kushanov FN, Mavlonov GT, Reddy UK, et al (2007) Microsatellite markers associated with lint percentage trait in cotton, *Gossypium hirsutum*. *Euphytica* **156**: 141–156
- Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc Natl Acad Sci USA* **100**: 4649–4654
- Applequist WL, Cronn R, Wendel JF (2001) Comparative development of fiber in wild and cultivated cotton. *Evol Dev* **3**: 3–17
- Arpat AB, Waugh M, Sullivan JP, Gonzales M, Frisch D, Main D, Wood T, Leslie A, Wing RA, Wilkins TA (2004) Functional genomics of cell elongation in developing cotton fibers. *Plant Mol Biol* **54**: 911–929
- Beasley CA, Ting IP (1974) The effects of plant growth substances on *in vitro* fiber development from unfertilized cotton ovules. *Am J Bot* **61**: 188–194
- Bentley DR (2006) Whole-genome re-sequencing. *Curr Opin Genet Dev* **16**: 545–552
- Blenda A, Scheffler J, Scheffler B, Palmer M, Lacape JM, Yu JZ, Jesudurai C, Jung S, Muthukumar S, Yellambalase P, et al (2006) CMD: a Cotton Microsatellite Database resource for *Gossypium* progenitors. *BMC Genomics* **7**: 132
- Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433–438
- Brubaker CL, Bourland FM, Wendel JF (1999) The origin and domestication of cotton. In CW Smith, JT Cothren, eds, *Cotton: Origin, History, Technology, and Production*. John Wiley & Sons, New York, pp 3–32
- Brubaker CL, Paterson AH, Wendel JF (1999) Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome* **42**: 184–203
- Cai WW, Reneker J, Chow CW, Vaishnav M, Bradley A (1998) An anchored framework BAC map of mouse chromosome 11 assembled using multiplex oligonucleotide hybridization. *Genomics* **54**: 387–397
- Chen ZJ (2007) Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu Rev Plant Biol* **58**: 377–406
- Desai A, Chee PW, Rong J, May OL, Paterson AH (2006) Chromosome structural changes in diploid and tetraploid A genomes of *Gossypium*. *Genome* **49**: 336–345
- Frelichowski JE Jr, Palmer MB, Main D, Tomkins JP, Cantrell RG, Stelly DM, Yu J, Kohel RJ, Ulloa M (2006) Cotton genome mapping with new microsatellites from *Acala* 'Maxxa' BAC-ends. *Mol Genet Genomics* **275**: 479–491

- Gao W, Chen ZJ, Yu JZ, Kohel RJ, Womack JE, Stelly DM (2006) Wide-cross whole-genome radiation hybrid mapping of the cotton (*Gossypium barbadense* L.) genome. *Mol Genet Genomics* **275**: 105–113
- Gingle AR, Yang H, Chee PW, May OL, Rong J, Bowman DT, Lubbers EL, Day JL, Paterson AH (2006) An integrated Web resource for cotton. *Crop Sci* **46**: 1998–2007
- Guo W, Cai C, Wang C, Han Z, Song X, Wang K, Niu X, Wang C, Lu K, Shi B, et al (2007) A microsatellite-based, gene-rich linkage map reveals genome structure, function and evolution in *Gossypium*. *Genetics* **176**: 527–541
- Haigler CH, Zhang DH, Wilkerson CG (2005) Biotechnological improvement of cotton fibre maturity. *Physiol Plant* **124**: 285–294
- Hanson RE, Zwick MS, Choi S, Islam-Faridi MN, McKnight TD, Wing RA, Price HJ, Stelly DM (1995) Fluorescent in situ hybridization of a bacterial artificial chromosome. *Genome* **38**: 646–651
- Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. *Genome Res* **16**: 1252–1261
- Hendrix B, Stewart JM (2005) Estimation of the nuclear DNA content of *Gossypium* species. *Ann Bot (Lond)* **95**: 789–797
- Ji Y, Zhao X, Paterson AH, Price HJ, Stelly DM (2007) Integrative mapping of *Gossypium hirsutum* L. by meiotic fluorescent in situ hybridization of a tandemly repetitive sequence (B77). *Genetics* **176**: 115–123
- Jiang C, Wright RJ, El-Zik KM, Paterson AH (1998) Polyploid formation created unique avenues for response to selection in *Gossypium*. *Proc Natl Acad Sci USA* **95**: 4419–4424
- Kim HJ, Triplett BA (2001) Cotton fiber growth in planta and in vitro: models for plant cell elongation and cell wall biogenesis. *Plant Physiol* **127**: 1361–1366
- Lacape JM, Nguyen TB, Courtois B, Belot JL, Giband M, Gourlot JP, Gawryziak G, Roques S, Hau B (2005) QTL analysis of cotton fiber quality using multiple *Gossypium hirsutum* × *Gossypium barbadense* backcross generations. *Crop Sci* **45**: 123–140
- Lee JJ, Hassan OSS, Gao W, Wang J, Wei EN, Russel JK, Chen XY, Payton P, Sze SH, Stelly DM, et al (2006) Developmental and gene expression analyses of a cotton naked seed mutant. *Planta* **223**: 418–432
- Liu B, Brubaker G, Cronn RC, Wendel JF (2001) Polyploid formation in cotton is not accompanied by rapid genomic changes. *Genome* **44**: 321–330
- Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. *Genome Res* **7**: 1072–1084
- Nguyen TB, Giband M, Brottier P, Risterucci AM, Lacape JM (2004) Wide coverage of the tetraploid cotton genome using newly developed microsatellite markers. *Theor Appl Genet* **109**: 167–175
- Paterson AH (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat Rev Genet* **7**: 174–184
- Reinisch AJ, Dong JM, Brubaker CL, Stelly DM, Wendel JF, Paterson AH (1994) A detailed RFLP map of cotton, *Gossypium hirsutum* × *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. *Genetics* **138**: 829–847
- Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, et al (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). *Genetics* **166**: 389–417
- Rong J, Bowers JE, Schulze SR, Waghmare VN, Rogers CJ, Pierce GJ, Zhang H, Estill JC, Paterson AH (2005) Comparative genomics of *Gossypium* and *Arabidopsis*: unraveling the consequences of both ancient and recent polyploidy. *Genome Res* **15**: 1198–1210
- Saha S, Raska DA, Stelly DM (2006) Upland cotton (*Gossypium hirsutum* L.) × Hawaiian cotton (*G. tomentosum* Nutt. ex. Seem) F₁ hybrid hypoaneuploid chromosome substitution series. *J Cotton Sci* **10**: 146–154
- Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol Biol Evol* **20**: 633–643
- Shen XL, Guo WZ, Lu QX, Zhu XF, Yuan YL, Zhang TZ (2007) Genetic mapping of quantitative trait loci for fiber quality and yield trait by RIL approach in Upland cotton. *Euphytica* **155**: 371–380
- Shi YH, Zhu SW, Mao XZ, Feng JX, Qin YM, Zhang L, Cheng J, Wei LP, Wang ZY, Zhu YX (2006) Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. *Plant Cell* **18**: 651–664
- Stelly DM, Saha S, Raska DA, Jenkins JN, McCarty JC, Gutierrez OA (2005) Registration of 17 upland (*Gossypium hirsutum*) cotton germplasm lines disomic for different *G. barbadense* chromosome or arm substitutions. *Crop Sci* **45**: 2663–2665
- Taliercio EW, Boykin D (2007) Analysis of gene expression in cotton fiber initials. *BMC Plant Biol* **7**: 22
- Udall JA, Flagel LE, Cheung F, Woodward AW, Hovav R, Rapp RA, Swanson JM, Lee JJ, Gingle AR, Nettleton D, et al (2007) Spotted cotton oligo-nucleotide microarrays for gene expression analysis. *BMC Genomics* **8**: 81
- Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J, Yu Y, Wu Y, Dowd C, Arpat AB, et al (2006) A global assembly of cotton ESTs. *Genome Res* **16**: 441–450
- Ulloa M, Saha S, Jenkins JN, Meredith WR Jr, McCarty JC Jr, Stelly DM (2005) Chromosomal assignment of RFLP linkage groups harboring important QTLs on an intraspecific cotton (*Gossypium hirsutum* L.) Joinmap. *J Hered* **96**: 132–144
- Waghmare VN, Rong J, Rogers CJ, Pierce GJ, Wendel JF, Paterson AH (2005) Genetic mapping of a cross between *Gossypium hirsutum* (cotton) and the Hawaiian endemic, *Gossypium tomentosum*. *Theor Appl Genet* **111**: 665–676
- Wang K, Guo W, Zhang T (2007) Detection and mapping of homologous and homeologous segments in homeologous groups of allotetraploid cotton by BAC-FISH. *BMC Genomics* **8**: 178
- Wang K, Song X, Han Z, Guo W, Yu JZ, Sun J, Pan J, Kohel RJ, Zhang T (2006) Complete assignment of the chromosomes of *Gossypium hirsutum* L. by translocation and fluorescence in situ hybridization mapping. *Theor Appl Genet* **113**: 73–80
- Wendel JF (2000) Genome evolution in polyploids. *Plant Mol Biol* **42**: 225–249
- Wendel JF, Cronn RC (2003) Polyploidy and the evolutionary history of cotton. *Adv Agron* **78**: 139–186
- Wu Y, Machado AC, White RG, Llewellyn DJ, Dennis ES (2006) Expression profiling identifies genes expressed early during lint fibre initiation in cotton. *Plant Cell Physiol* **47**: 107–127
- Yang SS, Cheung F, Lee JJ, Ha M, Wei NE, Sze SH, Stelly DM, Thaxton P, Triplett B, Town CD, et al (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
- Yin J, Guo W, Yang L, Liu L, Zhang T (2006) Physical mapping of the Rf1 fertility-restoring gene to a 100 kb region in cotton. *Theor Appl Genet* **112**: 1318–1325
- Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson AH (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* **8**: 479–492

Z. Jeffrey Chen

Section of Molecular Cell and Developmental Biology and Institute for Cellular and Molecular Biology,
University of Texas, Austin, TX 78712

Brian E. Scheffler

U.S. Department of Agriculture Agricultural Research Service, Catfish Genetics Research Unit, Mid South
Area Genomics Facility, Stoneville, MS 38776–0038

Elizabeth Dennis

Commonwealth Scientific and Industrial Research Organization, Plant Industry, Black Mountain, Canberra,
Australian Capital Territory 2601, Australia

Barbara A. Triplett

**U.S. Department of Agriculture Agricultural Research Service, Southern Regional Research Center,
New Orleans, LA 70124**

Tianzhen Zhang and Wangzhen Guo

**National Key Laboratory of Crop Genetics and Germplasm Enhancement and Cotton Research Institute,
Nanjing Agricultural University, Nanjing 210095, China**

Xiaoya Chen

Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai 200032, China

David M. Stelly

Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843

Pablo D. Rabinowicz

**Institute for Genome Sciences and Department of Biochemistry and Molecular Biology, School of Medicine,
University of Maryland, Baltimore, MD 21201**

Christopher D. Town

J. Craig Venter Institute, Rockville, MD 20850

Tony Arioli and Curt Brubaker

Bayer Bioscience NV, B-9052 Gent, Belgium

Roy G. Cantrell

Monsanto Company, St. Louis, MO 63167

Jean-Marc Lacape

CIRAD, UMR-DAP, TA A-96/03, 34398, Montpellier cedex 5, France

Mauricio Ulloa

**U.S. Department of Agriculture Agricultural Research Service, Western Integrated Cropping Systems
Research Unit, Shafter, CA 93263**

Peng Chee

Department of Crop and Soil Sciences, University of Georgia, Tifton, GA 31793

Alan R. Gingle

Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602

Candace H. Haigler

**Department of Crop Science and Department of Plant Biology, North Carolina State University, Raleigh, NC
27695-7620**

Richard Percy

**U.S. Department of Agriculture Agricultural Research Service, Maricopa Agricultural Center, Maricopa, AZ
85239-3101**

Sukumar Saha

**U.S. Department of Agriculture Agricultural Research Service, Genetics and Precision Agriculture Unit,
Mississippi State, MS 39762**

Thea Wilkins and Robert J. Wright

Department of Plant and Soil Science, Texas Tech University, Lubbock, TX 79409-3121

Allen Van Deynze

Seed Biotechnology Center, University of California, Davis, CA 95616

Yuxian Zhu

College of Life Sciences, Peking University, Beijing 100871, China

Shuxun Yu

Cotton Research Institute, Chinese Academy of Agricultural Sciences, Anyang 455112, China

Ibrokhim Abdurakhmonov
Institute of Genetics and Plant Experimental Biology, Academy of Sciences of Uzbekistan, Tashkent District,
02151, Uzbekistan

Ishwarappa Katageri
University of Agricultural Sciences, Dharwad 580005, India

P. Ananda Kumar
National Research Center for Plant Biotechnology, Pusa Campus, New Delhi 110012, India

Mehboob-ur-Rahman and Yusuf Zafar
Plant Genomics and Molecular Breeding Labs, National Institute for Biotechnology
and Genetic Engineering, Faisalabad,
Pakistan

John Z. Yu and Russell J. Kohel
U.S. Department of Agriculture Agricultural Research Service, Crop Germplasm Research Unit,
College Station, TX 77845

Jonathan F. Wendel
Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011

Andrew H. Paterson
Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602