

White Paper: Cotton Genome Sequencing

December 11, 2006

(Writing committee members: Jeffrey Chen, Xiaoya Chen, Elizabeth Dennis, Andrew H. Paterson, Brian E. Scheffler, David A. Stelly, Christopher D. Town, and Tianzhen Zhang)

Consulting Writing Members

Australia

Liz Dennis <Liz.Dennis@csiro.au>

Danny Llewellyn <Danny.Llewellyn@csiro.au>

Belgium

Tony Arioli <tony.arioli@bayercropscience.com>

Curt Brubaker <curt.brubaker@bayercropscience.com>

Brazil

Edina Moresco <edina@facual.org.br>

China

Xiaoya Chen <xychen@sibs.ac.cn>

Wangzhen Guo <moelab@njau.edu.cn>

Zhiying Ma <mzhy@hebau.edu.cn>

Gui-Xian Xia <guixianx@yahoo.com>

Shuxun Yu <yu@cricaas.com.cn>

Tianzhen Zhang <cotton@njau.edu.cn>

Xianglong Zhang <xlzhang@mail.hzau.edu.cn>

Jun Zhu <jzhu@zju.edu.cn>

Shuijin Zhu <shjzhu@zju.edu.cn>

Yuxian Zhu <zhuyx@water.pku.edu.cn>

Egypt

Essam Zaki <eazaki@link.net>

France

Jean-Marc Lacape <marc.lacape@cirad.fr>

India

Ishwarappa Katageri <ikatageri@yahoo.com>

B. M. Khadi <cicrngp@rediffmail.com>

Vidyasagar Parchuri <pvsagar@sify.com>

Vijay N. Waghmare <vijayvnw@yahoo.com>

Israel

Yehoshua Saranga <saranga@agri.huji.ac.il>

Pakistan

Mehboob Rahman <mehboob_pbd@yahoo.com>

USA

John. E. Bowers <jebowers@uga.edu>
Roy Cantrell <RCantrell@cottoninc.com>
Peng Chee <pwchee@uga.edu>
Z. Jeffrey Chen <zjchen@mail.utexas.edu>
Leon S. Dure III <ldure@uga.edu>
Deqiu Fang <Deqiu.Fang@deltaandpine.com>
Alan R. Gingle <agingle@uga.edu>
Candace Haigler <candace_haigler@ncsu.edu>
Ramesh Kantety <ramesh.kantety@aamu.edu>
Russell J. Kohel <kohelrj@neo.tamu.edu>
Siva Kumpatla <spkumpatla@dow.com>
Gerald O. Myers <GMyers@agcenter.lsu.edu>
Andrew H. Paterson <paterson@uga.edu>
Richard Percy <RPercy@uswcl.ars.ag.gov>
Daniel G. Peterson <dpeterson@pss.msstate.edu>
Keerti Rathore <rathore@tamu.edu>
Umesh Reddy <ureddy@wvstateu.edu>
Jun-kang Rong <jkrong@yahoo.com>
Sukumar Saha <ssaha@msa-msstate.ars.usda.gov>
Brian Scheffler <bscheffler@msa-stoneville.ars.usda.gov>
David Stelly <stelly@tamu.edu>
J. McD. Stewart <jstewart@uark.edu>
Christopher Town <cdtown@TIGR.org>
Barbara Triplett <bttriplet@srcc.ars.usda.gov>
Mauricio Ulloa <mulloa@pw.ars.usda.gov>
Allen Van Deynze <avandeynze@ucdavis.edu>
Jonathan Wendel <jfw@iastate.edu>
Thea Wilkins <thea.wilkins@ttu.edu>
Robert J. Wright <robert.wright@ttu.edu>
Jinhua Xiao <jinhua.xiao@monsanto.com>
Zhangyou Xu <xuzhanyou@tamu.edu>
Jing Yu <jingyu@algodon.tamu.edu>
John Yu <zyu@qutun.tamu.edu>
Hongbin Zhang <hbz7049@tamu.edu>

Uzbekistan

Abdusattor Abdurakarimov <genomics@uzsci.net>
Ibrokhim Abdurakhmonov <ibr58@hotmail.com>

Introduction

The topic of cotton genome sequencing has been informally and formally discussed in several meetings including the biennial research conferences of International Cotton Genome Initiative (ICGI). At the ICGI meeting on September 18-20, 2006 in Brazil, there was broad and strong support for organizing a community effort on cotton genome sequencing. An *ad hoc* group (white paper writing committee) was formed to coordinate writing a white paper on “cotton genome sequencing”.

Our goal is to build a community effort on the development, public release, and utilization of cotton sequence information. To advance this goal, it was clear the white paper should convey the compelling need to sequence the cotton genomes, and thus to foster participation in current and future sequencing opportunities in all cotton growing countries. An important aspect of the white paper was seen to be the enumeration of options and the pro’s and con’s associated with various strategies of resource development and sequencing strategies, and sequence utilization.

We decided not to endorse specific approaches, but to provide guidance and discussion of plans for cotton genome sequencing. We strongly believe that development of a successful sequencing effort will establish a new era of cotton research and improvement worldwide.

I. Rationale and justification for sequencing cotton genomes

Cotton, a mallow, is the world’s most important natural textile fiber, and it doubles as an oilseed crop. In nature, cotton plants are perennial woody shrubs and trees, yet cotton is mostly cultivated as a domesticated annual crop. Production is practiced on about 2.5% of arable land, and provides the jobs of about 100 million family units. Each cotton fiber is a single and phenomenally elongated cell from epidermal layer of the ovule, with about 25,000 per seed. The seed is an important source of feed and foodstuff. In addition to their widely known uses for apparel and home furnishings, fiber-derived products are found in plastics and many industrial products such as digital screens. World consumption of cotton fiber is ~115-million bales (480 lbs/bale) or ~27-million metric tons per year.

China is both the largest producer and consumer of raw cotton, but more than 80 countries produce cotton, including Australia, Franco-Africa, India, Pakistan, USA, and Uzbekistan. The USA is the second largest producer and grows about \$6 billion/yr. More than 150 countries are involved in import and export of cotton. Economic impact is estimated to be ~\$500 billion/yr worldwide (National Cotton Council, 2005). Moreover, cotton is a major economic driver for some developing countries like Uzbekistan that produces annually ~4 million tons of raw cotton and exports ~\$900 million worth cotton fiber.

Genetic improvements that enhance the economics of production and fiber processing characteristics will ensure competitiveness in the market of this natural-renewable product with petroleum-derived synthetic fibers, and the livelihoods of millions of people worldwide. Moreover, modifications to expand the use of seed derivatives for food and feed could profoundly benefit the diets and livelihoods of millions of people in food-challenged economies.

Biologically, cotton fiber is an excellent model system for the study of plant cell elongation and cell wall and cellulose biosynthesis (Kim and Triplett, 2001). The fiber is composed of nearly pure cellulose, which is also the largest component of plant biomass. Annual world production of cellulose is *ca.* 100 million metric tons, primarily in the cell walls of all higher plants. The basic study of cellulose biosynthesis in fiber cells is highly pertinent to the applied objectives of renewable resource and bioenergy research.

Decoding the cotton genomes will contribute significantly to our understanding of the functional and agronomic significance of polyploidy. *Gossypium hirsutum* and *G. barbadense* are classic natural allopolyploids derived from an interspecific hybridization of an African/Asian A-genome and an American D-genome species about 1-2 mya (Wendel and Cronn, 2003). *G. hirsutum*, Upland or American cotton, represents over 95% of the annual cotton crop worldwide. In the US, over 98% is Upland cotton, and the extra-long staple (ELS) or Pima cotton (*G. barbadense*) accounts for less than 2% (National Cotton Council, 2005).

The genus *Gossypium* occurs naturally throughout tropical and subtropical regions, and includes about 45 species split across two ploidy levels, diploid ($2n = 2x = 26$) and tetraploid ($2n = 4x = 52$). According to meiotic pairing and chromosome size, diploid species ($2n = 26$) fall into genomic groups A, B, C, D, E, F, G, or K. The A-genome clade, also including B, E, and F genome types distinguished from one another based on pairing behavior, chromosome sizes, and relative fertility in interspecific hybrids (Beasley, 1942) occur naturally in Africa and Asia, while the D-genome clade occurs in America. A third diploid clade exists in Australia, including C, G, and K genome types. Subscripts in each group are used to denote related genomes within a genomic group such as *G. herbaceum* L. (A_1) and *G. arboreum* L. (A_2). Both allotetraploids originated in the New World from interspecific hybridization between an A-genome-like ancestral African diploid species and a D-genome-like American diploid species (Skovsted, 1934; Beasley, 1940). The extant species most closely related to these ancestors are *G. herbaceum* L. (A_1), *G. raimondii* (D5) Ulbrich or *G. gossypoides* (Ulbrich) Standley (D6) (Gerstel, 1958; Phillips, 1963). Ancestral hybridization and polyploidization are estimated to have occurred 1–2 mya (Wendel and Cronn, 2003). The resulting disomic polyploid gave rise to the five extant allotetraploid species (Percival et al., 1999). Domesticated A-genome diploid and AD-tetraploid cottons appeared in Indus valley and the New World by 3,500–2,300 B.C, respectively (Hutchinson et al., 1947; Jiang et al., 1998). Cotton fiber production is affected by polyploid formation and crop domestication. The A-genome species produce spinnable fiber, whereas the D-genome species do not (Applequist et al., 2001). Significant impact on fiber traits in the allotetraploids by their D-subgenomes has been indicated by marker-assisted QTL localization (Jiang et al., 1998) and chromosome substitution line performance (Saha et al., 2006).

Efficient strategies for capturing the sequence diversity represented within the *Gossypium* genus will be greatly influenced by large differences in genome size and organization across the genus. The diploid genomes vary about 3-fold in DNA content, but have the same chromosome number and similar gene content. This variation in genome size appears to have accumulated in about 5-10 million years since the diploid clades are thought to have diverged from a common ancestor (Senchina et al., 2003). The haploid genome sizes are estimated to be ~880-Mb for *G. raimondii* Ulbrich, ~1.75-Gb for *G. arboreum* L., and ~2.5 Gb for *G. hirsutum* L. (Hendrix and Stewart,

2005). DNA content of the allopolyploids is approximately the sum of those of the A and D-genome progenitors, and nearly all of >22,000 AFLP fragments surveyed are additive in the allopolyploids (Liu et al., 2001). The variation in DNA content in the diploid species might be a net result of both increases and decreases in copy number of various repeat families.

Sequenced cotton genomes will ultimately stimulate fundamental research on genome evolution, polyploidization and associated re-diploidization, gene expression, cell differentiation and development, cellulose synthesis, cell growth, and molecular determinants of cell wall biogenesis. Practical ramifications will include improvement of biological processes key to agricultural productivity, economic yield, health, and ecologically-safe production practices, e.g., water use efficiency, abiotic and biotic stress tolerance/resistance, fertilizer and pesticide requirements, and new opportunities as a source of foodstuffs, expanded use as a feed and specialization of fiber types. While some are more tangible than others, the economic, health and ecological and thus societal impacts are truly compelling on both national and international scales.

II. Overview of cotton genomics research and resources

Researchers in the cotton community have developed many genomics resources and have coordinated efforts on developing DNA markers and genetic maps.

Germplasm resources

A narrow genetic base is a bottleneck for cotton breeding and cultivar improvement. Developing cotton genomic resources should facilitate better characterization and utilization of diverse cotton germplasm collections in many countries such as China, India, France, Mexico, USA, and Uzbekistan. For example, the Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD, France) maintains a germplasm collection of 3200 *Gossypium* accessions, including ca 1000 cultivars created through collaborative projects of CIRAD in developing countries, and ca 1000 perennial accessions collected worldwide. The Cotton Research Institutes of the Republic of Uzbekistan have collected and maintained ~17,000 cotton germplasm accessions including isogenic and inbred lines, elite allotetraploid varieties, monosomic and translocation lines, and wild, primitive and extant representatives of A to G genome groups. A comprehensive nursery has been established by Instituto Nacional de Investigaciones Forestales y Agropecuarias (INIFAP, Mexico) and the US for preservation and study of *Gossypium* D genome species (Ulloa et al., 2006). India maintains approximately 5,990 *G. hirsutum*, 1,049 *G. barbadense*, 1,867 *G. arboreum*, 568 *G. herbaeceum*, and 173 wild derivatives of crosses from cultivated and wild types. The Commonwealth Scientific and Industrial Research Organisation (CSIRO) Plant Industry in Australia maintains 2,000 *G. hirsutum* accessions for improving disease and fiber traits and additional 500 accessions of 17 indigenous Australian species covering 3 genomic groups. Commercial varieties, marketed locally as Sicot, Sicala and Siokra and internationally by Bayer as FiberMax is an excellent example of germplasm utilization for cotton fiber improvement.

Unraveling the cotton genome will contribute significantly to the characterization and utilization of germplasm which is needed for the future of cotton improvement using molecular marker-assisted introgression (MAI) and marker-assisted selection (MAS) for the genes that control fiber

length and quality, agronomic traits, and pest resistance (Wright et al., 1998; Rong et al., 2005; Waghmare et al., 2005; Abdurakhomonov et al., 2006; Wang et al., 2006). Classical and molecular breeders could use these molecular tools to locate and characterize the sequences of DNA and DNA binding regions that control expression of these traits through MAS.

Mapping populations

High-density genetic maps in tetraploid cotton have been initially developed from interspecific *G. hirsutum* x *G. barbadense* F₂ (Reinisch et al., 1994) and backcross (Lacape et al., 2005) populations. Incorporating diverse types and sources of DNA markers has served as reference maps for many other studies. An interspecific *G. hirsutum* x *G. tomentosum* F₂ population has been published (Waghmare et al., 2005), and a *G. hirsutum* x *G. mustelinum* F₂ is being mapped (P. Chee and A. Paterson, unpublished). Recently, several groups have developed immortal interspecific RIL populations derived from TM-1 x 3-79 (Frelichowski et al., 2006) and Guazuncho 2 x VH8 (Lacape, unpublished) and chromosome-specific recombinant inbred lines developed for some of the existing disomic substitution lines that contain *G. barbadense* germplasm (S. Saha, unpublished), which will be useful for high-resolution (multi-site) QTL mapping.

A total of 191 recombinant inbred lines (RILs) have been developed by single seed decent from F₂ individuals derived from a cross between TM-1 and 3-79 (R. Kohel and J. Yu, unpublished). The RILs display a wide range of phenotypic variability observed throughout the *Gossypium* genus (S. Hoffman and J. Yu, unpublished). These lines provide stable genetic materials for studying important fiber and morphological traits. A genetic map of 1,238 SSR markers, covering 2,870 cM of 26 chromosomes, has been constructed (J. Yu et al., unpublished). The TM-1 x 3-79 RILs are currently used by several cotton research groups to develop a common genetic map with additional portable markers (SSRs, AFLPs and SNPs).

Other mapping resources include panels of hypoaneuploid *G. hirsutum*, their interspecific F₁ hybrids with three other AD species, *G. barbadense* (Stelly, 1993), *G. tomentosum* (Saha et al., 2006), and *G. mustelinum* (D. M. Stelly, unpublished). The hypoaneuploids cover approximately 75% of the genome and identify 23 of the 26 cotton chromosomes. The other three have long been identified by translocations (Brown, 1980). The hypoaneuploids allow chromosomal localization and validation of marker assignments to chromosomes and linkage groups. A number of disomic substitution lines have already been developed for *G. barbadense* chromosomes (Stelly et al., 2005) in *G. hirsutum* background and more are near release. Disomic substitutions are under construction for *G. tomentosum* and *G. mustelinum* (D. M. Stelly and S. Saha, unpublished).

Development of wide-cross whole-genome radiation hybrid panels has enabled the use of radiation hybrid mapping in cotton (Gao et al., 2004; Gao et al., 2006), which complements other forms of genome mapping in terms of resolution, coverage and independence. Independent panels based on 8-Krad segmentation have been constructed for radiation hybrid maps of *G. barbadense* (Gao et al., 2006) and *G. hirsutum* (D. M. Stelly, unpublished) genomes.

Genetic maps

At least a dozen genetic maps of crosses between diverse cotton species and genotypes are available, most made to map specific traits (QTLs). These maps collectively include ~5,000 public DNA markers (~3,300 RFLP, 700 AFLP, 1,000 SSR, and 100 SNP). The published maps include sequence tagged site (STS)-based maps consisting of 2584 loci at 1.72 cM (~600 kb) intervals based on 2007 probes in tetraploids (AD genomes) and 1014 loci at 1.42 cM (~600 kb) intervals detected by 809 probes in diploid (D genome) (Rong et al., 2004; Rong et al., 2005). There is a high degree of colinearity among the respective genome types (Rong et al., 2005). Another EST-SSR based genetic map was made using 1,710 loci at 1.92 cM intervals in tetraploid cotton (Han et al., 2006). Many A and D homoeologous loci were detected by EST-SSR primer pairs, confirming the expression origins of A and D subgenomes.

Jean-Marc Lacape and his colleagues have integrated linkage maps into TropGENE-DB, <http://tropgenedb.cirad.fr/en/cotton.html>, using 'CMap' comparative map viewer. A similar map viewer has been implemented in CottonDB (<http://cottondb.org>) and in cotton microsatellite database (CMD) (<http://www.mainlab.clemson.edu/cmd>) (Blenda et al., 2006). The integrated map include several linkage maps developed by researchers in China (T. Zhang), France (J. Lacape), Pakistan (M. Rahman), and the US (A. Paterson and M. Ulloa) (Lacape and Nguyen, 2005; Lacape et al., 2005). Additional maps in the community can be integrated. A large number (~4,000) of microsatellites have been developed and are described in CMD (<http://www.mainlab.clemson.edu/cmd/AboutUs.shtml>) and CottonDB. These genetic maps and DNA markers will facilitate the development of comprehensive linkage maps that are valuable for sequence assembly.

Bacterial artificial chromosome (BAC) resources and physical maps

BAC libraries exist for several *G. hirsutum* cultivars, at least one *G. barbadense* cultivar, and A, D, and F genome diploid cottons as well as an outgroup, *Gossypioides kirkii* (summarized below). A total of 10 genome-equivalent coverage of *G. raimondii* BACs have been fingerprinted using standard procedures (Marra et al., 1997). To genetically-anchor the fingerprints into an integrated physical map, virtually all genetically mapped probes have been applied to the fingerprinted BACs using the overlapping oligonucleotides (overgo) method (Cai et al., 1998). Manual editing and revision of the physical map is in progress, incorporating genetic marker hybridization data with BAC fingerprint data, and assembly into contigs using finger-printed contigs (FPC). The assembly will be publicly available via a WebFPC site. Research is underway to expand the *G. arboreum* BAC library from 6x to 10x, and to genetically-anchor the entire library by hybridization to genetically-mapped DNA probes. These data will be incorporated into the existing 'BACMan resource' at the Plant Genome Mapping Laboratory web site (www.plantgenome.uga.edu).

BAC libraries (*G. hirsutum* L. cv. TM-1) were used to develop an integrated physical map (R. Khoel, Z. Xu, R. J. Yu, H. Zhang, unpublished) using the latest physical mapping technology (Xu et al., 2004) through a collaborative research project among Kohel (USDA-ARS), Yu (USDA-ARS), and Zhang (Texas A&M) laboratories. A total of 103,979 clones (6x) have been fingerprinted on capillary sequencers, and 6,031 contigs were assembled and manually edited. About 800 markers and 3,490 ESTs were mapped on the contigs by overgo hybridization and sequence comparison. The results indicate that 69% of the contigs are separated by chromosome-specific markers, and 31% of the contigs contain clones with duplicated loci

originating from homoeologous chromosomes. Additional BAC-end sequences and TM-1 BAC libraries with a larger insert size (>150kb) will help resolve genomic complexity in these regions.

Another BAC library was constructed from a male-sterile fertility restorer line 0-613-2R (*G. hirsutum* L.). The library has been successfully used for localizing *Rf₁* gene in a 100-kb region (Yin et al., 2006). The assignment of linkage groups to identified chromosomes has been completed using BAC-fluorescence *in situ* hybridization (FISH) (Wang et al., 2006).

Additional physical resources known to be publicly available include:

G. hirsutum TM-1 (15.6X, *Bam*HI, *Hind*III, and *Eco*RI libraries, average insert size 130-150 kb);
G. hirsutum Acala Maxxa (8.3X, *Hind*III library, average insert size 137 kb);
G. barbadense (Pima S6, 5X, average insert size 100 kb);
G. barbadense (Pima 90, 6.5X, *Bam*HI / *Hind*III libraries, average insert size 130 kb);
G. hirsutum Auburn 623 (2.7X, *Bam*HI library, average insert size ~140 kb);
G. hirsutum Tamcot HQ95 (2.3X, *Hind*III library, average insert size 93 kb);
G. hirsutum 0-613-2R (5.7X, *Hind*III library, average insert size 130 kb);
G. arboreum AKA8401 (6X, *Mbo*I library, average insert size 120 kb);
G. arboreum Jinglinzhongmian (6.2X, *Hind*III library, average insert size 100.2 kb);
G. longicalyx (being validated);
G. kirkii (being validated).

EST resources

As of October 22, 2006, 352,298 *Gossypium* sequences were in Genbank. A total of 67,060 ESTs derived from the *G. raimondii* D-genome, 39,966 ESTs from the A genome, 230,244 from the AD tetraploid (*G. hirsutum*), and with a few from other members of the genus. Several analyses have been published (Arpat et al., 2004; Udall et al., 2006; Yang et al., 2006). A recent study indicated that during early stages of fiber development there is preferential accumulation of ESTs encoding putative transcription factors such as MYB and WRKY and the genes predicted to encode proteins involved in auxin, brassinosteroid (BR), gibberellic acid (GA), abscisic acid (ABA) and ethylene signaling pathways (Yang et al., 2006). The data agree with the known roles of MYB and WRKY transcription factors in *Arabidopsis* leaf trichome development and the well-documented phytohormonal effects on fiber cell development in immature cotton ovules cultured *in vitro* (Beasley and Ting, 1974). A-subgenome ESTs of all functional classifications including cell cycle control and transcription factor activity are dramatically enriched in *G. hirsutum* L. (Yang et al., 2006), a result consistent with the production of long lint fibers in A-genome species. Additional analyses of ESTs and expression patterns using microarrays has led 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).

Synteny/colinearity information

The Malvales (including cotton) are the nearest relative to *Arabidopsis* outside of the Brassicales for which detailed genetic and physical maps have been described. Among the 2337 probes in the inferred cotton map, 2162 (92.5% of) probes representing 2800 (92.8% of) loci could be sequenced. A total of 1738 (62.1%) of the sequenced loci had matches ($E < 10^{-10}$) in *Arabidopsis*. Comparative analyses using this inferred gene order reveal a considerable degree of

synteny/colinearity of the ancestral cotton genome to the *Arabidopsis* genome (Rong et al., 2005). Sequence data will permit further substantial improvement because the diploid *Gossypium* species are paleopolyploids tracing to an event thought to have occurred 30 mya or more (Bowers et al., 2003; Rong et al., 2005).

III. Sequencing strategies and methodologies

It is likely that a strong case can be made for complete sequencing of one or more representatives of each *Gossypium* genome type, including a tetraploid-derived AD ($n = 26$) genome, given the expected continuing progress in improving sequencing throughput and reducing cost (Paterson, 2006). A comprehensive strategy needs to consider present needs along with long-term goals in relation to economics and technology.

Sequencing of representatives from each diploid clade, and preferably each genome, 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, arid regions of the world. Sequences from diploid species, especially certain A and D genome species will aid AD genome sequence assembly, and could prove to be invaluable in revealing differences in gene content and expression patterns across the ploidy levels, and providing piercing insight into polyploid genome evolution. The high degree of conservation of gene order and sequence between diploids and tetraploids suggests that the vast majority of data from diploids will extrapolate directly to tetraploids.

Sequencing of diploid and tetraploid species will be invaluable for practical genomics-based applications such as marker development for high-throughput marker assisted analyses, e.g., trait dissection, QTL mapping, gene identification and marker-assisted selection. Comparisons of AD to A, D and other genomes will be essential for clarifying the genetic basis of adaptations for increased lint fiber yield and lint quality. Sequencing of 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 over 95% of world cotton production. Sequencing of elite *G. arboreum* and perhaps *G. herbaceum* genomes, A_2 and A_1 , will valuable data on fiber genes. Comparisons of four species across two ploidy levels, including A_1 , A_2 , D_5 , and AD tetraploid subgenomes, could provide clues regarding how polyploidy affects domestication. Parallel comparisons of domesticated versus non-domesticated forms of the A- and AD-genome species may help shed light on the effects of artificial selection versus natural selection.

Efficient approaches to capturing the information available from the genus will need to consider several constraints that may require the use of different sequencing strategies for different taxa.

1. The diploid clades diverged approximately 5-10 mya (Cronn et al., 2002), and are known to have preserved a high degree of genomic content and arrangement. At the whole-genome level, a high degree of colinearity and synteny among the A, D, and tetraploid genomes (Reinisch et al., 1994; Brubaker et al., 1999; Rong et al., 2004; Desai et al., 2006) suggests that complete sequencing of a small number of genotypes together with reduced-representation sequencing of representatives of additional nodes might be a cost-effective

interim step. There has been some additional rearrangement of tetraploid chromosomes relative to their diploid progenitors (Brubaker et al., 1999; Rong et al., 2004; Desai et al., 2006), including some evidence of cryptic rearrangements that are supported by cytological observations (D. M. Stelly, unpublished) but may not be obvious based on genetic maps (Waghmare et al., 2005). Nonetheless, the high degree of transferability of information about gene content and order among the respective genome types suggests that whole-genome efforts in diploid taxa will provide strong guidance for future efforts in tetraploid taxa.

2. Among the diploid cottons alone, there is approximately 3-fold variation in genome size. This has obvious ramifications for costs associated with whole-genome sequencing, although it invites reduced representation approaches to jump-start progress in the larger genomes (see below).
3. Much of the genome size variation among the diploid genome types is due to dispersed repetitive DNA (Zhao et al., 1998), especially retrotransposon-like elements (Hawkins et al., 2006). There appears to have been large expansions of repetitive DNA content in the A/B/E/F and C/G/K genome clades in the 5-10 million years since the divergence of the diploid clades, thus many of these element families may include large numbers of relatively recently-derived members that are problematic for whole-genome shotgun approaches. By contrast, the D genome clade appears to have only a minimum of such recently-amplified repetitive DNA, and may be more amenable to whole-genome shotgun approaches. A survey of about 100 of the most abundant families in the tetraploid genome showed only 4 to be abundant in the D genome but rare or absent in the 'A' genome, which diverged from the 'D' genome of *G. raimondii* about 5-10 mya. Thus, most high-copy repetitive DNA families in the D genome are at least 5-10 million years old and likely to be amenable to assembly by a whole-genome shotgun approach. By contrast, the alternative A genome progenitor contains about 50 repetitive element families that are rare or absent from the D genome, suggesting that these families amplified in this same 5-10 million year period. Most of these A-genome repetitive element families contain thousands of members, and have continued to amplify and transpose since polyploid formation about 1-2 mya (Zhao et al., 1998), rendering the A and tetraploid 'AD' genomes less amenable to whole-genome shotgun approaches.
4. The tetraploid clades combine the properties of the A and D genome diploids with modification by intergenomic concerted evolution. Concerted evolution of the repetitive DNA fraction (Wendel et al., 1995, 1995; Cronn et al., 1996; Zhao et al., 1998) has been clearly shown. The possibility of intergenomic exchange of low-copy DNA remains somewhat unclear, with evidence for (Reinisch et al., 1994) and against it (Cronn et al., 1999), but growing data from other taxa strongly suggest that it may be an important dimension of polyploid evolution (Hughes and Hughes, 1993; Moore and Purugganan, 2003; Gao and Innan, 2004; Chapman et al., 2006) (X. Wang, H. Tang, J. E. Bowers, F. A. Feltus, A. H. Paterson, submitted).

Based on these considerations, one can envision an efficient strategy by which to proceed toward molecular dissection of the genomic and morphological diversity of this economically and scientifically important genus, with some initial steps detailed below.

1. Extensive EST data exist for the A, D, and tetraploid genomes but the remaining genomes are virtually unexplored. Current ESTs were mainly derived from *G. arboreum* (A genome), *G. raimondii* (D genome), and *G. hirsutum* (AD genome) species. Other species in the diploid (e.g., C, G, and K genomes) and tetraploid (e.g., *G. barbadense*, AD) clades should be considered for generating additional ESTs. Comparative analysis of ESTs in diploid and tetraploid species may help predict SNPs and genome-specific polymorphism (GSP) and may provide new insights into relative abundance of progenitors' transcripts in cotton tetraploids (Yang et al., 2006).
2. Sequencing using gene-enrichment techniques such as methylation filtration and C₀t-based cloning will generate novel genomic sequences that are absent in EST collections. To jump-start progress toward dissecting genomic and morphological diversity in *Gossypium*, several reduced-representation approaches (Rabinowicz et al., 1999; Peterson et al., 2002) might be applied to representatives of diploid and tetraploid species, in order to take advantage of the relative merits of each (Palmer et al., 2003; Whitelaw et al., 2003; Rabinowicz et al., 2005).

A pilot study in methylation filtration (B. E. Scheffler, S. Saha, and Orion Genomics, unpublished) using *G. raimondii*, *G. arboreum*, *G. hirsutum*, and *G. barbadense* has been conducted. The study illustrates two important facts that might be useful in developing a sequencing strategy for *Gossypium*. First, methylation filtration is more efficient for *G. arboreum*, requiring only 452,480 sequences for 1X coverage compared to 787,457 for the smaller D genome of *G. raimondii*. While these results appear to be counter intuitive because the D genome is almost half the size of the A genome, methylation filtration often works better on those genomes with more junk DNA that is methylated. Second, the combined number of sequences needed for the D genome of *G. raimondii* and the A genome of *G. arboreum* (1,239,937) is relatively equivalent to the number of sequences needed for the AD genomes of *G. hirsutum* and *G. barbadense*. This indicates the total amount of methylation has not been significantly altered in the creation of the AD tetraploids.

Likewise, pilot studies in C₀t-based cloning and sequencing were effective at separating repetitive DNA from low-copy DNA in representatives of the A, D, and C clades (T. Wicker, J. Robertson, A. H. Paterson, unpublished). Methylation filtration and Cot-based cloning and sequencing appear to offer complementary coverage of the low-copy DNA (Whitelaw et al., 2003), integration of the two methods may be especially efficient in the discovery of novel *Gossypium* sequences.

3. The whole-genome shotgun sequence of the smallest *Gossypium* genome would provide fundamental information about gene content and organization. *G. raimondii* has a relatively small genome size (~880Mb) and low amount of repetitive DNA sequences. A partially or fully sequenced *G. raimondii* genome will establish an initial 'template/backbone' toward the long-term goal of characterizing the spectrum of diversity among the eight *Gossypium* genome types and three polyploid clades. Whole-genome shotgun based characterization of the smallest genome is in theory the most cost effective and easiest of the whole genome approaches at the present stage. For these and other reasons the U.S. Department of Energy Joint Genome Institutes has selected *G. raimondii* for a pilot study for shotgun sequencing

0.5x coverage to better define the genome and a workable strategy for its complete sequencing.

4. The economic importance of cotton fibers and scientific interests in polyploidy suggest an ultimate goal of sequencing *G. hirsutum* tetraploid. A BAC-based sequence of a tetraploid will elucidate the types and frequencies of changes that have distinguished polyploid from diploid cottons. The process could be greatly enhanced by using a finished diploid genome as a template and guide. The possibility of intergenomic concerted evolution, much like the presence of recently-amplified repetitive DNA families, would tend to support the need for a BAC-based rather than a whole-genome shotgun approach. A reasonable approach is to establish minimum tiling path of FPC of *G. hirsutum* L. homoeologous chromosomes. This can be achieved by developing integrated homoeologous chromosome maps using anchored DNA markers in genetic maps and BAC-end sequences in physical maps, which can be further validated using BAC fluorescence *in situ* hybridization (FISH) (Hanson et al., 1995; Stelly et al., 1995; Zwick et al., 1998; Kim et al., 2005; Kim et al., 2005; 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 basis of FISH signal strength (D. M. Stelly, unpublished). Large duplicated segments have been reported within individual corresponding homoeologous chromosomes, suggesting ancient or recent genome expansion in cotton genomes (Rong et al., 2005)(K. Wang, W. Z. Guo, and T. Z. Zhang, unpublished).
5. Other approaches. An alternative approach is to develop chromosome-specific BAC libraries using chromosome sorting, which has been widely used in sequencing animal and human genomes and has recently been demonstrated in polyploid plants with large chromosomes such as wheat (Kubalakova et al., 2005). The available cytological facilities and resources in cotton may permit testing the feasibility and utility of this approach.

Note that enrichment techniques may not produce many regulatory sequences important to the control of gene expression. The sequences generated from the enrichment approaches are difficult to assemble in continuous chromosomal segments. Establishing minimum tiling path using FPC and integrated DNA markers and BAC FISH will provide chromosomal views of genomic sequences but they are time-consuming and relatively expensive. At the initial stage, the technical effectiveness should be compared and evaluated by analyzing the sequences being generated in *G. raimondii* and *G. arboreum* diploids and *G. hirsutum* tetraploid using enrichment techniques. Predictably, it is essential to sequence and assemble homoeologous BACs and/or a few pairs of homoeologous chromosomes prior to large-scale sequencing of *G. hirsutum* tetraploid genomes.

IV. Resource Development

Successful development of cotton genome sequencing projects and utilization of cotton genome sequences will require additional resources be developed, integrated, coordinated with one another and with existing resources, and rendered much more accessible to the research community, both *in silico* and physically. Some priorities for future resource development and coordination include the following areas.

1. The community needs to experimentally determine the difficulty of sequence assembly from the *G. hirsutum* genome. The various possible means of rendering the *G. hirsutum* genome sequence data amenable to direct assembly need to be experimentally determined with adequate sampling of the genome to allow for a legitimate basis for extrapolated inference; their efficacy in terms of time, cost and accuracy need to be weighed against alternative options such as shotgun, fractionation by reassociation, methylation, and minimum-tiling, and/or reliance on sequences from related diploid species.
2. Maps of the various *Gossypium* genomes should be more closely integrated with each other and with closely related species. There is a strong need to further develop a *consensus map* of at least the [AD]₁ genome, which is highly populated with public markers. The map should include all data of linkage maps, cytogenomic maps, contig maps, and genes (ESTs).
3. EST collections need to be expanded, and where possible, converted into integrated and comparative genomic markers.
4. One or more minimum-tiling paths need to be constructed using large-insert libraries. If inaccuracies arise at the initial step due to within genome redundancy, additional steps such as mapping of all ESTs onto BACs using hybridization and fingerprinting approaches should help detect within-genome redundancy.
5. A comprehensive cytogenomic map, including genome coverage, orientation, positions of specific loci relative to huge blocks of euchromatin versus pericentromeric heterochromatin, relative sizes of chromosomes and arms, and valuable comparisons among related genomes, will be for a valuable complement that adds a biological dimension to contig and sequence assembly.
6. Means of mapping low-recombination regions would be highly valuable as these will be most problematic in terms of sequence assembly. This suggests that more physical mapping is needed, e.g., by FISH, radiation hybrid mapping and/or chromosome sorting might be appropriate solutions.
7. In silico mechanisms for integration of other genomic and morphological data would be desirable to have in place prior to sequencing, while databases and annotation systems should be established when large-scale sequences start to become available.
8. The cotton community needs to develop additional functional genomics resources such as genome-specific gene chips, genomic tiling arrays, and bioinformatic tools that will precisely predict small RNAs and microRNAs and their targets in cotton.

V. Data management (annotation, curation, and dissemination)

The amount of data generated from various sequencing projects will be extremely large and difficult to comprehend for many end users in the cotton community. Therefore, it is essential to develop data management system that can facilitate access and utilization of the genomic and

sequence data. Several cotton project databases are currently available.

CottonDB (<http://cottondb.org>, prior location <http://cottondb.tamu.edu/>) is a comprehensive database that was established years ago. Through a website interface, it provides genomic, genetic, and taxonomic information, including germplasm, markers, genetic and physical maps, trait studies, sequences, bibliographic citations. The Cotton Portal (<http://gossypium.info>) offers the community a single port of entry to participating Cotton Web resources. The Cotton Diversity Database (<http://cotton.agtec.uga.edu>) (Gingle et al., 2006) provides for integrative queries 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 with the capacity for additional integrative queries. CMD (<http://www.mainlab.clemson.edu/cmd/AboutUs.shtml>) provides centralized access to all publicly available cotton microsatellites. TropGENE-DB (<http://tropgenedb.cirad.fr/en/cotton.html>) integrates a subset of published mapping data. Several project websites such as cotton functional genomics (<http://cottongenomecenter.ucdavis.edu/>), cotton fiber genomics (<http://www.cottongenomics.org/>), and genetic and physical mapping (www.plantgenome.uga.edu) are primarily used for disseminating genomic resources being generated from individual research projects. Some research groups have coordinately distributed genomic resources to the cotton research community. For example, cotton oligo-gene microarrays consisting of ~23,000 70-mer oligos designed from 250,000 ESTs are coordinately developed in two NSF-funded projects (PI: Chen and PI: Wendel). The microarray information and slide distribution can be found in the website (<http://cottonrevolution.info/microarray>). EST information can be found in several project websites including Cotton Gene Indices (CGI) (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=cotton), cotton portal (<http://cottonrevolution.info/pages/estlib/overview.aspx>), and Arizona Genomics Institute (<http://www.genome.arizona.edu/>).

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. Some community database examples include the Arabidopsis Information Resource (TAIR, <http://www.arabidopsis.org/>), Maize Genetics and Genomics Database (MaizeGDB, <http://www.maizegdb.org/>), Soybase (<http://soybase.agron.iastate.edu/>), and GrainGenes (<http://wheat.pw.usda.gov/GG2/index.shtml>). Future cotton sequence database should be able to host and manage 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 sequence flow and community requests, but additional resources should be sought to support some key bioinformatic needs.

At the beginning, at least one community website should be identified to establish a “newsgroup” list-server that will allow researchers to express and discuss their views and ideas about cotton genome sequencing and genomic research including conceptual and technical issues. Some important issues should be collectively discussed in the International Cotton Genome Initiative (ICGI) steering committee with a prompt reply to the entire community.

VI. Goals of cotton genome sequencing projects

Short-term goals

1. Develop a set of community-wide recommendations for sequencing that is supported by cotton researchers from the majority of countries with an interest in cotton genomics.
2. Develop workshops and communication methods for planning, coordinating and executing sequencing and post-sequencing activities. This will involve developing a Web site or utilizing existing resources such as the ICGI website. Coordinate all activities on the same web site or have multiple sites and locations connected.
3. Evaluate and identify suitable strategies for sequencing the genome of cotton, both in terms of the species chosen and the physical process of generating the sequence. Given the high level of repetitive sequences in plants methyl and/or C₀t filtration to target low-copy gene sequences in the genome may be useful as an initial sequencing approach for many of the *Gossypium* species. As cultivated cotton (*G. hirsutum* L.) is a tetraploid species this could present technical challenges to current sequencing methodologies, and sequencing could initially start from the nearest living diploid progenitors *G. raimondii* and *G. arboreum*.

Long-term goals

4. Fully sequence representatives of each of the *Gossypium* clades. Among these, a singularly important goal is to establish the complete genome sequence of allotetraploid cotton (*Gossypium hirsutum*, 2n=52, ~2.5-Gb), using the progenitor diploid species genomes as frameworks.
5. Integrate functional and structural genomic resources at molecular and *in silico* levels.
6. Sequence full-length cDNAs for genome annotation and expression assays.
7. Perform detailed annotation of the cotton genome sequence to support gene discovery and map based cloning in this species.
8. Develop and implement a large-scale platform for identifying DNA sequence diversity (SNPs and GSPs) on a genome-wide scale.
9. Develop a tiling array for the whole cotton genome to support gene expression analysis studies of cotton traits of biological and agronomic interest.
10. Sequence and annotate small RNAs and microRNAs and identify their targets.

References

- Abdurakhomonov IY, Kohel RJ, Saha S, Pepper AE, Yu J, Buriev ZT, Shermatov SE, Abdullaev AA, Kushanov FN, Jenkins JN, Scheffler BE, Abdugarimov A** (2006) Linkage disequilibrium based association mapping of fiber quality traits in cotton using diverse cotton germplasm from Uzbekistan. *In* ICGI, ed, Proceedings of the IV International Cotton Genome Initiative, Brasilia, Brazil
- 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. *Amer J Bot* **61**: 188-194
- Beasley JO** (1940) The origin of American tetraploid *Gossypium* species. *Amer Naturalist* **74**: 285-286
- Beasley JO** (1942) Meiotic chromosome behavior in species hybrids, haploids, and polyploids of *Gossypium*. *Genetics* **27**: 25-54
- Blenda A, Scheffler J, Scheffler B, Palmer M, Lacape JM, Yu JZ, Jesudurai C, Jung S, Muthukumar S, Yellambalase P, Ficklin S, Staton M, Eshelman R, Ulloa M, Saha S, Burr B, Liu S, Zhang T, Fang D, Pepper A, Kumpatla S, Jacobs J, Tomkins J, Cantrell R, Main D** (2006) CMD: a Cotton Microsatellite Database resource for *Gossypium* genomics. *BMC Genomics* **7**: 132
- Bowers JE, Chapman BA, Rong J, Paterson AH** (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**: 433-438
- Brown MS** (1980) Identification of the chromosomes of *Gossypium hirsutum* L. by means of translocations. *J Hered* **71**: 266-274
- 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
- Chapman BA, Bowers JE, Feltus FA, Paterson AH** (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication. *Proc Natl Acad Sci U S A* **103**: 2730-2735
- Cronn RC, Small RL, Haselkorn T, Wendel JF** (2002) Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *American Journal of Botany* **89**: 707-725
- Cronn RC, Small RL, Wendel JF** (1999) Duplicated genes evolve independently after polyploid formation in cotton. *Proc Natl Acad Sci U S A* **96**: 14406-14411
- Cronn RC, Zhao X, Paterson AH, Wendel JF** (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, 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 LZ, Innan H** (2004) Very low gene duplication rate in the yeast genome. *Science* **306**: 1367-1370
- 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
- Gao W, Chen ZJ, Yu JZ, Raska D, Kohel RJ, Womack JE, Stelly DM** (2004) Wide-cross whole-genome radiation hybrid mapping of cotton (*Gossypium hirsutum* L.). *Genetics* **167**: 1317-1329
- Gerstel DU** (1958) Chromosomal translocations in interspecific hybrids of the genus *Gossypium*. *Evolution* **7**: 234-244
- 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
- Han Z, Wang C, Song X, Guo W, Gou J, Li C, Chen X, Zhang T** (2006) Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. *Theor Appl Genet* **112**: 430-439
- 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
- Hughes MK, Hughes AL** (1993) Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Mol Biol Evol* **10**: 1360-1369
- Hutchinson JB, Silow AR, Stephens SG** (1947) The evolution of *Gossypium* and differentiation of the cultivated cottons. Oxford University Press, London
- 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 U S A* **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
- Kim JS, Islam-Faridi MN, Klein PE, Stelly DM, Price HJ, Klein RR, Mullet JE** (2005) Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. *Genetics* **171**: 1963-1976
- Kim JS, Klein PE, Klein RR, Price HJ, Mullet JE, Stelly DM** (2005) Chromosome identification and nomenclature of *Sorghum bicolor*. *Genetics* **169**: 1169-1173
- Kubalaková M, Kovarova P, Suchankova P, Cihalikova J, Bartos J, Lucretti S, Watanabe N, Kianian SF, Dolezel J** (2005) Chromosome sorting in tetraploid wheat and its potential for genome analysis. *Genetics* **170**: 823-829
- Lacape JM, Nguyen TB** (2005) Mapping quantitative trait loci associated with leaf and stem pubescence in cotton. *J Hered* **96**: 441-444

- Lacape JM, Nguyen TB, Courtois B, Belot JL, Giband M, Gurlot JP, Gawryziak G, Roques S, Hau B** (2005) QTL analysis of cotton fiber quality using multiple *Gossypium hirsutum* x *Gossypium barbadense* backcross generations. *Crop Science* **45**: 123-140
- Lee JJ, Hassan OSS, Gao W, Wang J, Wei EN, Russel JK, Chen XY, Payton P, Sze SH, Stelly DM, Chen ZJ** (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
- Moore RC, Purugganan MD** (2003) The early stages of duplicate gene evolution. *Proc Natl Acad Sci U S A* **100**: 15682-15687
- Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR** (2003) Maize genome sequencing by methylation filtration. *Science* **302**: 2115-2117
- Paterson AH** (2006) Leafing through the genomes of our major crop plants: strategies for capturing unique information. *Nat Rev Genet* **7**: 174-184
- Percival AE, Wendel JF, Stewart JM** (1999) Taxonomy and germplasm resources. In CW Smith, JT Cothren, eds, *Cotton: Origin, History, Technology, and Production*. John Wiley & Sons, Inc., New York, pp 33-63
- Peterson DG, Schulze SR, Sciara EB, Lee SA, Bowers JE, Nagel A, Jiang N, Tibbitts DC, Wessler SR, Paterson AH** (2002) Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res* **12**: 795-807
- Phillips LL** (1963) The cytogenetics of *Gossypium* and the origin of New World cottons. *Evolution* **17**: 460-469
- Rabinowicz PD, Citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy AL, Nascimento LU, McCombie WR, Martienssen RA** (2005) Differential methylation of genes and repeats in land plants. *Genome Res* **15**: 1431-1440
- Rabinowicz PD, Schutz K, Dedhia N, Yordan C, Parnell LD, Stein L, McCombie WR, Martienssen RA** (1999) Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat Genet* **23**: 305-308
- Reinisch AJ, Dong JM, Brubaker CL, Stelly DM, Wendel JF, Paterson AH** (1994) A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. *Genetics* **138**: 829-847
- Rong J, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW, Delmonte TA, Ding X, Garza JJ, Marler BS, Park CH, Pierce GJ, Rainey KM, Rastogi VK, Schulze SR, Trolinder NL, Wendel JF, Wilkins TA, Williams-Coplin TD, Wing RA, Wright RJ, Zhao X, Zhu L, Paterson AH** (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (*Gossypium*). *Genetics* **166**: 389-417
- 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

- Rong J, Pierce GJ, Waghmare VN, Rogers CJ, Desai A, Chee PW, May OL, Gannaway JR, Wendel JF, Wilkins TA, Paterson AH** (2005) Genetic mapping and comparative analysis of seven mutants related to seed fiber development in cotton. *Theor Appl Genet* **111**: 1137-1146
- Saha S, Jenkins JN, Wu J, McCarty JC, Gutierrez OA, Percy RG, Cantrell RG, Stelly DM** (2006) Effects of chromosome-specific introgression in upland cotton on fiber and agronomic traits. *Genetics* **172**: 1927-1938
- Saha S, Raska DA, Stelly DM** (2006) Upland cotton (*Gossypium hirsutum* L.) x Hawaiian cotton (*G. tomentosum* Nutt. ex. Seem) F₁ hybrid hypoaneuploid chromosome substitution series. *J. Cotton Science* **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
- 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
- Skovsted A** (1934) Cytological studies in cotton. II. Two interspecific hybrids between Asiatic and New World cottons. *J. Genet.* **28**: 407-424
- Stelly DM** (1993) Interfacing cytogenetics with the cotton genomic mapping. *In* DJ Herber, DA Richter, eds, *Proc. Beltwide Cotton Conf.*, pp 1545-1550
- Stelly DM, Price HJ, McKinght TD** (1995) Molecular-meiotic cytogenetic analysis of cotton. *In* BS Gill, J Raupp, eds, *Classical and Molecular Cytogenetic Analysis of Cereal Genomes*, pp 148-156
- 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 Science* **45**: 2663-2665
- Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J, Yu Y, Wu Y, Dowd C, Arpat AB, Sickler BA, Wilkins TA, Guo JY, Chen XY, Scheffler J, Taliencio E, Turley R, McFadden H, Payton P, Klueva N, Allen R, Zhang D, Haigler C, Wilkerson C, Suo J, Schulze SR, Pierce ML, Essenberg M, Kim H, Llewellyn DJ, Dennis ES, Kudrna D, Wing R, Paterson AH, Soderlund C, Wendel JF** (2006) A global assembly of cotton ESTs. *Genome Res* **16**: 441-450
- Ulloa M, Stewart JM, Garcia-C. EA, Godoy-A. S, Gaytan-M. A, Acosta-N. S** (2006) Cotton genetic resources in the western states of Mexico: In situ conservation status and germplasm collection for ex situ preservation. *Genet Resour Crop Evol* **53**: 653-668
- 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 C, Ulloa M, Roberts PA** (2006) Identification and mapping of microsatellite markers linked to a root-knot nematode resistance gene (*rkn1*) in Acala NemX cotton (*Gossypium hirsutum* L.). *Theor Appl Genet* **112**: 770-777
- 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, Cronn RC** (2003) Polyploidy and the evolutionary history of cotton. *Advances in Agronomy* **78**: 139-186

- Wendel JF, Schnabel A, Seelanan T** (1995) Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). *Proc Natl Acad Sci U S A* **92**: 280-284
- Wendel JF, Schnabel A, Seelanan T** (1995) An unusual ribosomal DNA sequence from *Gossypium gossypoides* reveals ancient, cryptic, intergenomic introgression. *Mol Phylogenet Evol* **4**: 298-313
- Whitelaw CA, Barbazuk WB, Perteza G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, SanMiguel P, Lakey N, Bedell J, Yuan Y, Budiman MA, Resnick A, Van Aken S, Utterback T, Riedmuller S, Williams M, Feldblyum T, Schubert K, Beachy R, Fraser CM, Quackenbush J** (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science* **302**: 2118-2120
- Wright RJ, Thaxton PM, El-Zik KM, Paterson AH** (1998) D-subgenome bias of Xcm resistance genes in tetraploid *Gossypium* (cotton) suggests that polyploid formation has created novel avenues for evolution. *Genetics* **149**: 1987-1996
- 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
- Xu Z, Sun S, Covalada L, Ding K, Zhang A, Wu C, Scheuring C, Zhang HB** (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
- Yang SS, Cheung F, Lee JJ, Ha M, Wei NE, Sze SH, Stelly DM, Thaxton P, Triplett B, Town CD, Chen ZJ** (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
- Zwick MS, Islam-Faridi MN, Czeschin DG, Jr., Wing RA, Hart GE, Stelly DM, Price HJ** (1998) Physical mapping of the liguleless linkage group in *Sorghum bicolor* using rice RFLP-selected sorghum BACs. *Genetics* **148**: 1983-1992