Resource

A global assembly of cotton ESTs

Joshua A. Udall,[1] Jordan M. Swanson,[1] Karl Haller,[2] Ryan A. Rapp,[1] Michael E. Sparks,[1] Jamie Hatfield,[2] Yeisoo Yu,[3] Yingru Wu,[4] Caitriona Dowd,[4] Aladdin B. Arpat,[5] Brad A. Sickler,[5] Thea A. Wilkins,[5] Jin Ying Guo,[6] Xiao Ya Chen,[6] Jodi Scheffler,[7] Earl Taliercio,[7] Ricky Turley,[7] Helen McFadden,[4] Paxton Payton,[8] Natalya Klueva,[9] Randell Allen,[9] Deshui Zhang,[10] Candace Haigler,[10] Curtis Wilkerson,[11] Jinfeng Suo,[12] Stefan R. Schulze,[13] Margaret L. Pierce,[14] Margaret Essenberg,[14] HyeRan Kim,[3] Danny J. Llewellyn,[4] Elizabeth S. Dennis,[4] David Kudrna,[3] Rod Wing,[3] Andrew H. Paterson,[13] Cari Soderlund,[2] and Jonathan F. Wendel[1,15]

[1]Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011, USA; [2]Arizona Genomics Computational Laboratory, BIO5 Institute, [3]Arizona Genomics Institute, Department of Plant Sciences, University of Arizona, Tucson, Arizona 85721, USA; [4]CSIRO Plant Industry, Canberra City ACT 2601, Australia; [5]Department of Plant Sciences, University of California–Davis, Davis, California 95616, USA; [6]Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Shanghai, 200032, China; [7]United States Department of Agriculture–Agricultural Research Service, Stoneville, Mississippi 38776, USA; [8]United States Department of Agriculture–Agricultural Research Service, Lubbock, Texas 79415, USA; [9]Department of Biology, Texas Tech University, Lubbock, Texas 79409, USA; [10]Department of Crop Science and Department of Botany, North Carolina State University, Raleigh, North Carolina 27695, USA; [11]Bioinformatics Core Facility, Michigan State University, East Lansing, Michigan 48824, USA; [12]Institute of Genetics and Developmental Biology, Beijing, 100101, China; [13]Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602, USA; [14]Oklahoma Agricultural Experiment Station, Oklahoma State University, Stillwater, Oklahoma 74078, USA

Approximately 185,000 *Gossypium* EST sequences comprising >94,800,000 nucleotides were amassed from 30 cDNA libraries constructed from a variety of tissues and organs under a range of conditions, including drought stress and pathogen challenges. These libraries were derived from allopolyploid cotton (*Gossypium hirsutum*; A_T and D_T genomes) as well as its two diploid progenitors, *Gossypium arboreum* (A genome) and *Gossypium raimondii* (D genome). ESTs were assembled using the Program for Assembling and Viewing ESTs (PAVE), resulting in 22,030 contigs and 29,077 singletons (51,107 unigenes). Further comparisons among the singletons and contigs led to recognition of 33,665 exemplar sequences that represent a nonredundant set of putative *Gossypium* genes containing partial or full-length coding regions and usually one or two UTRs. The assembly, along with their UniProt BLASTX hits, GO annotation, and Pfam analysis results, are freely accessible as a public resource for cotton genomics. Because ESTs from diploid and allotetraploid *Gossypium* were combined in a single assembly, we were in many cases able to bioinformatically distinguish duplicated genes in allotetraploid cotton and assign them to either the A or D genome. The assembly and associated information provide a framework for future investigation of cotton functional and evolutionary genomics.

[Supplemental material is available online at www.genome.org. The ESTs from GR_Ea and GR_Eb were deposited in GenBank under accession nos. CO069431–CO100583 and CO100584–CO132899.]

Cotton is the world's most important fiber plant, being grown in more than 80 countries with a record forecast of 119.8 million 480-pound bales in world production during the 2004–2005 growing season (United States Department of Agriculture–Foreign Agricultural Service [USDA–FAS] 2005). Genetic improvement of cotton fiber and agricultural productivity will be enhanced by the availability of rapidly developing genetic resources and tools, including a high-density genetic map for *Gossypium hirsutum* (Rong et al. 2004; Lacape et al. 2005). Several studies have reported genes that are highly or exclusively expressed in cotton fibers (Orford and Timmis 1998; Orford et al. 1999; Zhao and Liu 2001; Kim et al. 2002; Li et al. 2002; Ji et al. 2003; Suo et al. 2003; Zhang et al. 2004). To stimulate further progress in cotton genetics and for other purposes including expression profiling, we initiated a project designed to identify a significant portion of the *Gossypium* transcriptome.

Most modern cotton varieties are forms of *G. hirsutum*, or Upland cotton, although three other species are also utilized to a lesser extent, *Gossypium barbadense*, *Gossypium arboreum*, and *Gossypium herbaceum*. *G. barbadense* and *G. hirsutum* are allotetraploids, each containing both an A_T and a D_T genome (Skovsted 1934; Wendel and Cronn 2003), where the T subscript indicates "tetraploid." *G. arboreum* and *G. herbaceum* are diploid, and their constituent genomes (A_2 and A_1, respectively) are phylogenetically equidistant to the A_T genome of allopolyploid cotton (Cronn et al. 2002; Wendel and Cronn 2003). *Gossypium raimondii* is the D-genome species most closely related to the modern-day allopolyploid D_T genome (Endrizzi et al. 1985; Wendel 1995;

[15]Corresponding author.
E-mail jfw@iastate.edu; fax (515) 294-1337.
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4602906.

Wendel and Cronn 2003). A single hybridization event between the A and D genome diploid cottons likely gave rise to modern allotetraploid cotton. Genetic divergence between these diploid groups and divergence between their genomes and the allopolyploid have been estimated (Senchina et al. 2003; Wendel and Cronn 2003), and phylogenetic relationships among the genome groups and species have been determined (Cronn et al. 2002). These relationships make *Gossypium* an attractive model for studying polyploid gene and genome evolution.

EST sequencing projects have been completed or are under way for many plant species. These projects have provided useful tools for intragenomic comparisons (Schlueter et al. 2004) and intergenomic comparisons (Fulton et al. 2002), gene discovery (Ewing et al. 1999; Ronning et al. 2003; Hughes and Friedman 2004), molecular marker identification (Michalek et al. 2002), and microarray development (Wisman and Ohlrogge 2000; Kawasaki et al. 2001; Alba et al. 2004; Arpat et al. 2004; Close et al. 2004). An initial survey of ~42,000 fiber ESTs based on a single fiber library from diploid *G. arboreum* (A genome) proved extremely useful for identifying genes, and led to the development of a 70-mer oligonucleotide cotton fiber microarray. A more thorough description of the *Gossypium* transcriptome, involving a wide array of tissues and organs, would facilitate additional gene discovery for diverse applications.

Here we report the sequencing, clustering, and analysis of 30 EST libraries generated by an international consortium of research groups. While many of these libraries are relatively small and from specialized tissues or growth conditions, we included two larger cDNA libraries (floral and seedling) from *G. raimondii* (D genome) and the previously mentioned A-genome cDNA fiber library. Our strategy was to simultaneously include EST sequences from allopolyploid (AD genome) cotton and species representing its two progenitor genomes (A, D genomes), thereby facilitating the identification of duplicated A_T and D_T (i.e., homoeologous) transcripts for numerous genes. The resulting assembly enables an examination of sequence divergence within a well-defined system of diploid and polyploid plant species on an unprecedented scale, provides insight into gene expression in numerous different tissues and environmental conditions, and sets the stage for the development of a cotton oligonucleotide microarray with deep genomic coverage.

Results

EST assembly

A total of 185,198 EST sequences from 30 cDNA libraries were collected from 14 different research groups across the globe (Table 1). These libraries were constructed from a variety of tissues and organs under a range of conditions, including drought stress and pathogen challenges, and include representation of allopolyploid cotton as well as its two diploid progenitors. Most cDNA libraries were derived from *G. hirsutum* and were relatively small (from 576 to 8643 ESTs). Collectively, these *G. hirsutum* EST collections comprised 38% of the total used in the assembly. The remaining ESTs were derived from three, more deeply sampled cDNA libraries generated from the two diploids (one library from 7–10 dpa fiber of *G. arboreum* and two libraries of *G. raimondii*), comprising 24% and 38% of the total number of ESTs, respectively.

Of that initial set of ESTs, 153,969 were selected as input for the global EST assembly based on length, complexity, and sequence quality (see Methods). Nearly all of the cDNA clones of diploid libraries were sequenced from both the 5′-end and 3′-end of the transcript as were portions of other *G. hirsutum* libraries. After the EST selection process, a total of 87,697 clones were included as input into the assembly pipeline, where 41% of the 153,959 selected ESTs had a mate-pair (a cDNA clone was sequenced in both directions).

Individual ESTs were assembled using the Program for Assembling and Viewing ESTs (PAVE). A conservative philosophy was used to align the ESTs and form a consensus sequence, that is, aligned portions of ESTs must share 95% sequence identity with <20% of overhanging sequence. Hence, alleles, homoeologs, orthologs, and paralogs were only combined into the same contig if they have a low level of divergence. Most alleles and homoeologs generally were expected to coalesce into the same contig, except the relatively rare cases of alternatively spliced transcripts. When the assembly was based on less stringent sequence similarity, it resulted in massive contigs that were joined because of similar domains (data not shown).

The PAVE assembly process yielded 22,030 contigs and 29,077 singletons (51,107 unigenes) in 40.4 Mb of transcribed sequence with an average length of 791 bp (SD = 374). The number of ESTs in a contig ranged from two to 714, with a median of three sequences per contig (Fig. 1); 10,624 contigs contained forward and reverse sequence pairs from at least one cDNA clone. As expected, contigs with four or more EST members exhibited a higher percentage of mate-pairs (51%) than contigs with two (37%) or three (37%) EST members.

The assembly of the ESTs into contigs used multiple libraries from three different *Gossypium* species (Fig. 2), of which 60% of the contigs (13,268) had EST members from more than one library and 40% of the contigs had EST members from more than one species. The values of these two numbers suggested that interspecific nucleotide variation did not have much of an effect on the global assembly process. However, other factors, such as RNA quality, library construction, indels, paralogy, differential gene expression, and systematic sequencing errors may have played a role in the EST assembly, resulting in library biases among the EST members of a contig (Supplemental Table 1). The extent that library bias reflected technical issues and not differential gene expression was unknown.

Several aspects of the assembly were evaluated to assess its quality: (1) the frequency of chimeric contigs; (2) the frequency of mate-pairs in the same contig; (3) phylogenetic analysis using known genes and their relationships; and (4) the amount of redundancy among the assembly's contigs and singletons. In an ideal assembly, only ESTs transcribed from a single gene are conjoined into a single contig. However, spurious EST-contig associations can be generated through the complexity of multigene families (along with the attendant issue of paralogy) and technical errors such as EST misnaming, resulting in chimeric contigs. A straightforward means of visualizing spurious associations is to inspect contigs containing the largest number of EST sequences (Supplemental Table 2). On average, these 20 well-sampled contigs (from 136 to 714 members) contained forward and reverse sequence pairs spanning 91% of the respective contig length, suggesting that nearly the entire length of these contigs could be attributed to a single cDNA clone (i.e., single gene).

Perhaps a better indication of spurious associations could be found within contigs having a poorly sampled interior region. In well-sampled contigs, most of the consensus sequence was represented by four or more individual ESTs. A possible spurious

Table 1. Summary of *Gossypium* EST libraries

Species	Authors	*Gossypium* accession	Library name	Library description	# ESTs	Selected ESTs	SEQ	QUAL
G. arboreum (A_2)	Wing et al., Arpat et al.	8401	GA_Ea	7–10 dpa developing fibers (normalized)	46,603	31,242	f+r	y
	Subtotal no. ESTs				46,603	31,242		
G. raimondii (D_5)	Udall et al.	GN34	GR_Ea	Whole seedlings with first true leaves	33,671	29,177	f+r	y
	Udall et al.	GN34	GR_Eb	−3 dpa buds to +3 dpa bolls	35,061	31,036	f+r	y
	Subtotal no. ESTs				68,732	60,213		
G. hirsutum (AD_1)	Allen	Coker 312	GH_MDI	8–10 dpa boll (irrigated)	1144	868	f	y
	Allen	Coker 312	GH_MDDS	8–10 dpa boll (drought stressed)	1238	773	f	y
	Allen & Payton	Coker 312	GH_LDI	15–20 dpa boll (irrigated)	1799	1324	f	y
	Allen & Payton	Coker 312	GH_LDDS	15–20 dpa boll (drought stressed)	1409	753	f	y
	Blewitt & Burr	Acala Maxxa	GH_BNL	Fiber 5 d post-anthesis (normalized)	8022	7590	x	n
	Chapman	Stv 7A gl	GH_ECT	18 h etiolated seedlings	2880	2685	x	y
	Dowd & McFadden	Delta Emerald	GH_CRH	Root and hypocotyls	1464	1309	f	y
	Dowd & McFadden	Delta Emerald	GH_CFUS	RH tissues infected with *Fusarium oxysporum*	820	641	f+x	y
	Faivre-Nitschke & Dennis	Sicot	GH_LSL	Sicot S9i leaves, late season	1810	1707	f	y
	Gou & Chen	Xu-142	GH_FOX	Ovule (0–5 dpa) and fiber (1–22 dpa)	7997	6277	f	y
	Haigler & Wilkerson	Delta Pine 90	GH_SCW	Secondary vs. primary fibers (suppr. subtr. hyb.)	7385	7372	x	y
	Klueva et al.	Coker 312	GH_SDL	Seedling (control)	1918	1502	f	y
	Klueva & Nguyen	Coker 312	GH_SDLD	Seedling (drought stressed)	1142	475	f	y
	Klueva & Nguyen	Coker 312	GH_SDCH	Seedling (chilling stressed)	576	138	f	y
	Liu & Dennis	Delta Pine 16	GH_IME	Immature embryo	1536	856	x	y
	Patil, Essenberg & Pierce	Im216	GH_IMX	Leaf 8, 14, 20, 30, 45, 60 hpi *Xanthomonas*	1134	685	x	y
	Phillips, Essenberg & Pierce	AcB4BInb7	GH_ACXE	Leaf 8+14 hpi *Xanthomonas*	647	439	x	y
	Phillips, Essenberg & Pierce	AcB4BInb7	GH_ACXM	Leaf 20+30 hpi *Xanthomonas*	1328	863	x	y
	Phillips, Essenberg & Pierce	AcB4BInb7	GH_ACXL	Leaf 45+60 hpi *Xanthomonas*	862	682	x	y
	Suo & Xue	Zhongmian12	GH_SUO	0-dpa ovule	1240	1217	x	n
	Trolinder	T25	GH_pAR	Leaves	1230	904	x+y	y
	Taliercio	DES119	GH_STEM	Mature stem	8643	6187	x+y	y
	Ni & Trelease	DP62	GH_ECOT	Etiolated cotyledon	2772	2338	x	y
	Wan & Wing	91-D-92	GH_CBAZ	Cotton boll abscission zone cDNA Library	1306	1306	f	y
	Wu & Dennis	Delta Pine 16	GH_CHX	Ovules −3 to 0 dpa cycloheximide	7631	7472	x+y	y
	Wu & Dennis	Delta Pine 16	GH_OCF	Ovules 0 dpa	867	820	f	y
	Wu & Dennis	Delta Pine 16	GH_ON	Ovules 0 dpa normalized	5903	5321	f+r,f	y
	Subtotal no. ESTs				69,853	62,504		
	Total no. ESTs				185,198	153,969		

ESTs were included in the assembly after removing short (<300 bp) and low-complexity sequences. Sequencing (SEQ) was in one or both directions (f = 5′; r = 3′; x,y = unspecified). Quality values (QUAL) are PHRED scores (Ewing et al. 1998) calculated from the shape and area of fluorescence intensity from each base pair when sequenced on a fluorescent automated sequencer.

association may be where three or fewer ESTs tie together two flanking sequence segments containing four or more member ESTs. Such occurrences resulted in a "barbell" shape of the contig's EST alignment. The present assembly contained 1397 such contigs (6% of contigs), and a few of these may represent large genes that simply had poor sampling of the internal sequences. However, a subset of these contigs ($n = 100$) had a pair of ESTs belonging to a single cDNA clone (sequenced in both directions probably representing the 5′ and 3′ boundaries of a gene) and also had at least one EST member whose 5′-end did not overlap this clonal pair—rather, it (and usually other sequences) was erroneously tied to the contig by a few other ESTs bridging the two regions. Both types of these contigs were flagged as "suspicious" contigs in PAVE, and based on this annotation it is possible for researchers to exclude these sequences while using PAVE.

The overall distribution of the forward-reverse EST pairs also provided general insight regarding the assembly quality. From the clones sequenced in both directions, 65% of the sequence pairs had both directional reads in the same contig, 11% of the sequence pairs had both reads in different contigs, 17% of sequence pairs had one in a contig and another as a singleton, and 7% of sequence pairs had both reads as singletons, although this final percentage may partially reflect insert size and transcript frequency rather than the assembly process. The fact that only two-thirds of the forward-reverse EST pairs had both directional reads in the same contig may be explained by a combination of a conservative percent-identity parameter during the assembly process, short EST reads (or a long gene), low frequency of rare transcripts, and misnaming.

A phylogenic approach was also used to assess EST assembly quality (Close et al. 2004). During earlier iterations of the EST

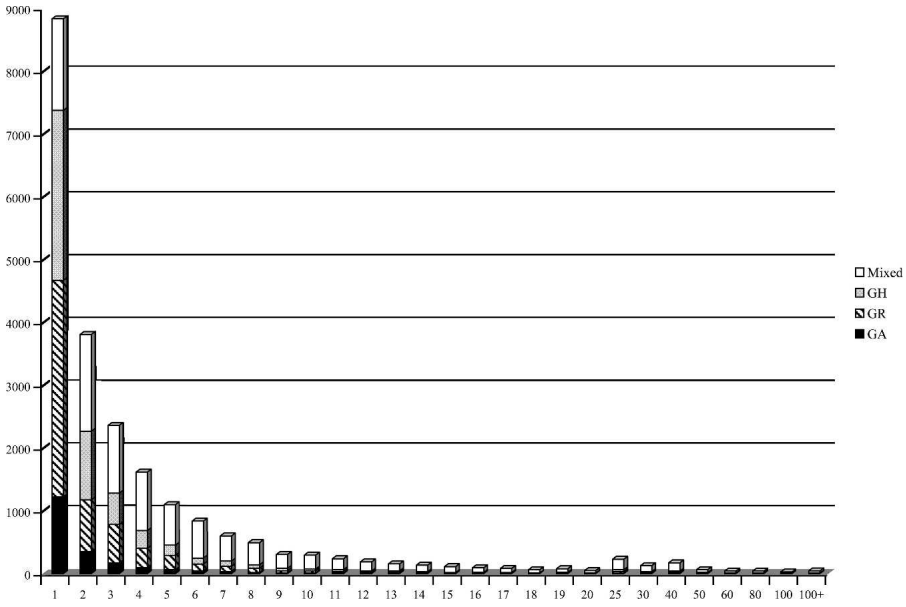

Figure 1. Histogram of number of EST members in a contig. Different patterns and shading of the bars indicate contigs composed of ESTs from a single species and those derived from ESTs from more than one species. Contigs with more than 100 EST members are not illustrated.

assembly process, we evaluated different assembly methods and parameters (data not shown), and at the same time assessed the effects of these permutations on phylogenetic topologies of previously characterized gene sequences whose relationships were already known (Small and Wendel 2000a,b). Alcohol dehydrogenase sequences from other plant species were used to find (BLASTN) homologous sequences ($<1e^{-100}$ and longer than 150 amino acids) in our assemblies. The homologous *Gossypium* contigs and singletons were aligned using MUSCLE (Edgar 2004). Once aligned, PHYLIP was used to create protein distances and neighbor joining trees representing the unigenes (Felsenstein 2004). The final assembly we are reporting here was the one that best recovered the *Adh* topology generated by more direct approaches to gene isolation and comparative sequence analysis (Small and Wendel 2000a,b).

While there appeared to be a low percentage of chimeric contigs in the assembly, it was more difficult to assess whether the number of contigs and singletons could be accurately reduced by further refinement. Ideally, each unigene in the assembly will correspond to a single version of an expressed gene; however, some level of gene redundancy often remains in EST assemblies when conservative parameters are used (Whitfield et al. 2002; Vettore et al. 2003). As a consequence of our efforts to minimize the number of cryptically chimeric contigs, some portion of the assembly appeared to have been "over-split." To reduce the transcript redundancy of the *Gossypium* unigene set, we used BLASTN on the assembly against itself, and pooled all contigs and singletons that had an 80% identity over 75% of "both" sequence lengths. Using these liberal parameters, unigene pools were created that had related sequence and potentially related biological function. From each unigene pool, the longest sequence was chosen as an exemplar (i.e., representative), resulting in 33,665 exemplar sequences. This difference between the number of unigenes (51,107) and the number of exemplars reflected the conservative approach used during the assembly process and possible causes of the library bias enumerated above.

These 33,665 exemplar sequences represent a nonredundant set of putative *Gossypium* genes containing partial or full-length coding regions and, usually, one or two identifiable UTRs. The coding and UTR regions were identified using ESTScan (Iseli et al. 1999; Lottaz et al. 2003). ESTScan uses a hidden Markov model (HMM) to correct sequencing errors common in ESTs and to identify the translated regions, including profiles for start and stop sites flanked by untranslated regions (UTRs). The ESTScan HMM was based on 28,953 *Arabidopsis* RefSeq annotations obtained from GenBank (4/2005). An open-reading frame (ORF) was found for 31,006 of the exemplar sequences with an average length of 613 bp (min = 51, max = 3846; SD = 382). *Arabidopsis thaliana* and *Oryza sativa* have an average ORF length of 1254 and 1374 bp, respectively, suggesting that these *Gossypium* exemplar sequences constitute, on average, approximately half of each of

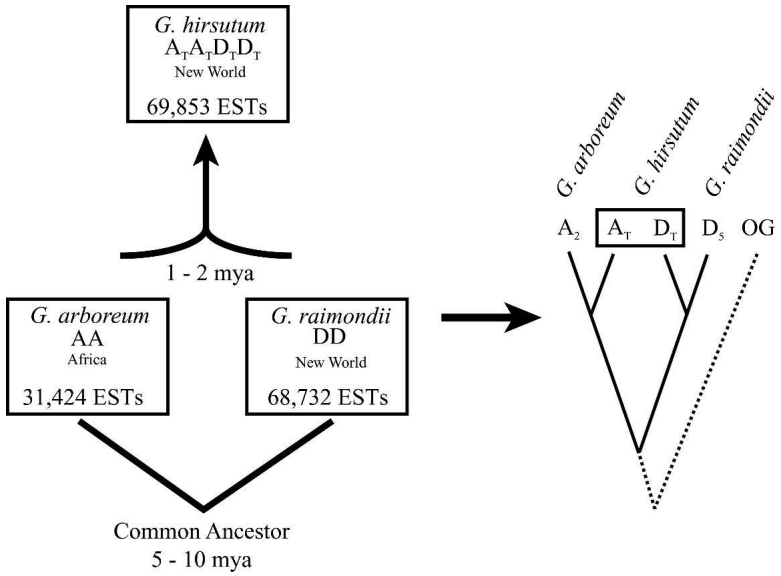

Figure 2. A framework to investigate the genomes of domesticated cotton species. The progenitor genomes of allopolyploid cotton (including *G. hirsutum*, AD genome) are represented by diploid A-genome (*G. arboreum*) and D-genome (*G. raimondii*) lineages, which united ~1–2 million years ago. Nucleotide sequence divergence between diploid A and D genomes (or their corresponding descendants in the allopolyploid) is ~4% (Senchina et al. 2003; OG = Outgroup). Shown also are the number of ESTs derived from each of the three species used in the assembly.

their respective full-length gene-coding products (ftp://ftp.tigr.org). Of these sequences, 66% had a BLASTX hit ($<1\ e^{-20}$) to the UniProt database (Apweiler et al. 2004). Those exemplar sequences that had good ESTScan models but no significant BLASTX hit (34%) represent genes that may have undergone sequence evolution specific to the *Gossypium* clade or they may be ESTScan false positives resulting from an incorrect modeling of a *Gossypium* gene with the *Arabidopsis* data set. The majority of these exemplar sequences are probably in the former category, although a less stringent BLASTX threshold would obviously identify more evolutionarily distant homologs. There was a small set of genes (61) that had a very good BLASTX hit ($<1e^{-100}$), but ESTScan did not identify a potential ORF. These sequences identified in the databases were likely false negatives of ESTScan in which the *Gossypium* coding frame was not correctly identified with the ESTScan model derived from the *Arabidopsis* gene set.

Gene annotation and Pfam

The set of 33,665 exemplar sequences was annotated using both the Gene Ontology (GO) (Ashburner et al. 2000) and Protein families (Pfam) indices (Bateman et al. 2004). As mentioned previously, all sequences were used to search for similar protein sequences in the UniProt database (BLASTX). Using the best hits found by BLASTX ($<1e^{-20}$), an inferred putative GO annotation was found for 64% of the cotton exemplar sequences, and these putative gene functions were categorized into high-level functional categories (Supplemental Fig. 1). The most abundant high-level categories within the biological processes, cellular components, and molecular function groups were cellular process, membrane, and catalytic activity, respectively. Many cotton exemplars appeared to be involved with transcription, including the high-level categories of transcription factor activity (407), RNA binding (638), DNA binding (1508), and nucleotide binding (3244).

The exemplar sequences were also analyzed for their protein domains to assess assignment to characterized protein families, of which 1815 protein domains with a Pfam cutoff threshold of $<1e^{-10}$ were identified in 6797 (20%) exemplar sequences (Fig. 3). Here, the Pfam cutoff threshold of $1e^{-10}$ was used because the conserved, characterized Pfam domains are "average domains" from many divergent species (Bateman et al. 2004). Perhaps the number of identifiable domains was limited because of incomplete gene sequence in the exemplars and an abundance of protein models that were not based on plants. In rice and *A. thaliana*, 4557 (http://rice.tigr.org/) and 2780 (Wortman et al. 2003) protein domains were identified by Pfam analysis, respectively. Many of the exemplar sequences in which a protein domain was recognized (3816) also had a GO annotation; however, the greatest value of Pfam analyses was the characterization of many exemplar sequences that did not have a significant BLASTX hit ($<1e^{-20}$) or, consequently, a GO annotation. In the set of otherwise uncharacterized exemplar sequences, 2981 were found to putatively contain 1421 different functional domains (Fig. 3). Transcription factor domains were included in the set of domains used within each Pfam analysis. In 398 exemplar sequences, 78 "transcription factor" or "DNA-binding" Pfam domains (http://pfam.wustl.edu) were identified, and 72 of these domains were identified in 237 otherwise unannotated exemplar sequences. The most abundant types of Pfam transcriptional domains found in the collection of exemplar sequences were MYB DNA-binding (Myb_DNA-binding), APETALA2 (AP2), auxin in-

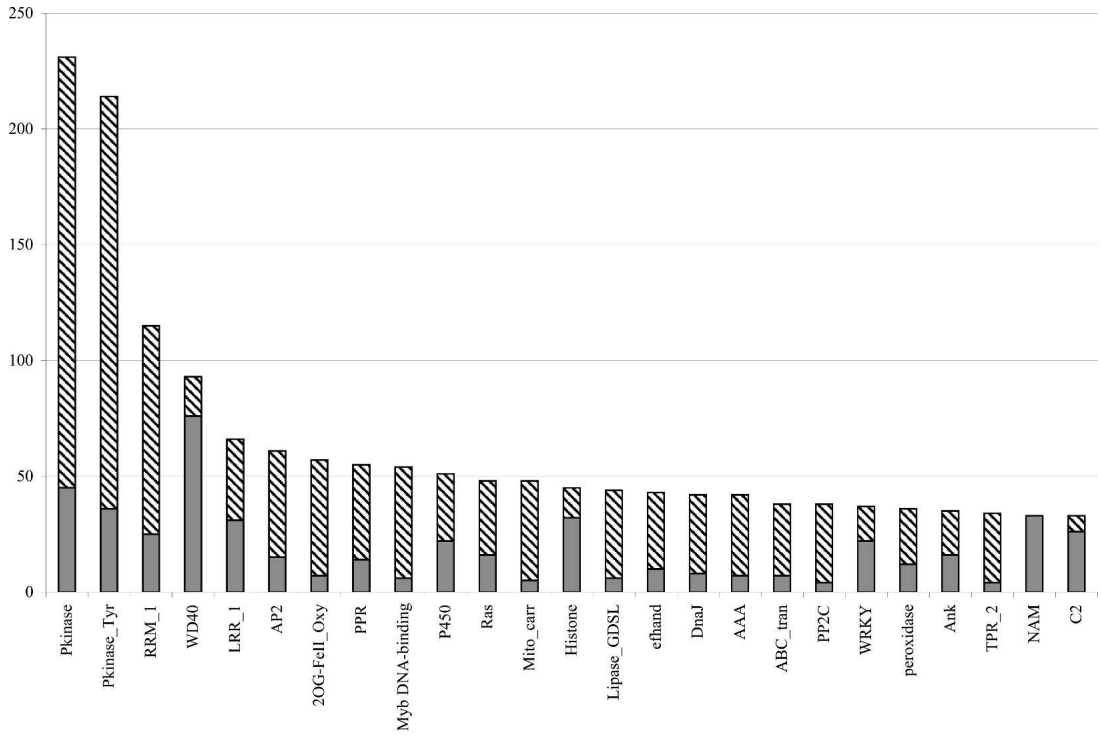

Figure 3. The top 25 categories of protein domains as identified by Pfam analysis of the exemplar sequences. The total bar height indicates the number of exemplar sequences containing each domain. The height of the solid area indicates the number of exemplar sequences that had a Pfam annotation but no significant BLASTX hit or gene ontology information. Categories with <33 members are not shown.

duced (AUX_IAA), WRKY DNA-binding (WRKY), and RING zinc finger domains (zf_C3HC4).

Identification of putative homoeologs

Because *G. hirsutum* is an allopolyploid formed from progenitor A- and D-genome diploids, its genome is expected to contain duplicated or homoeologous genes (A_T and D_T) for most genes (Fig. 2). This expectation has been confirmed for all genes studied to date (Cronn et al. 2002; Small and Wendel 2002; Cedroni et al. 2003; Senchina et al. 2003) with as yet no case of duplicate gene loss having been detected. When ESTs are sequenced from a polyploid, the genomic origin of each sequence is initially unknown. Because our *Gossypium* assembly included ESTs from allotetraploid as well as both diploid genome groups, for many unigenes it was possible to identify homoeologs and assign them to their proper genome (A_T or D_T) through comparisons with their orthologous counterparts from the progenitor diploid genomes (Fig. 2). Comparisons are most feasible for genes that are broadly expressed and at levels that they are readily captured in cDNA libraries. For this subset of housekeeping genes, up to four sequence variants (orthologous A and D copies and homoeologous A_T and D_T copies) are expected within the global *Gossypium* EST collection. These frequently are assembled into a single contig because of the low amount of divergence between sequences from the respective genomes (~4% overall between A and D or A_T and D_T) (Senchina et al. 2003). This sequence divergence, though, provides the signal necessary to confidently assign ESTs from *G. hirsutum* to their appropriate homoeolog. In the simplest cases, differences between the A- and D-genome ESTs were detected, and these were retained, evolutionarily unmodified, until the present in allopolyploid cotton, in which case the ancestry of each homoeolog in *G. hirsutum* was readily inferred. However, not all contigs included full representation of all four sequence types (A, D, A_T, and D_T), and in addition, a modest level of sequence evolution has arisen in both diploid and allopolyploid cotton since polyploid formation, perhaps 1–2 million years ago (Wendel and Cronn 2003).

The number of contigs that had one, two, three, or all four of the relevant sequence types is shown in Table 2. For 309 contigs, A, D, A_T, and D_T sequences were each identified. For 1870 contigs, ESTs from either one or both genomes of allopolyploid cotton were not identified. For the remaining 1966 ortholog-containing contigs, ESTs were found from only one of the two diploid species and its orthologous counterpart in *G. hirsutum* (i.e., A and A_T, or D and D_T). Because of the deep sampling of cDNA libraries from *G. arboreum* and *G. raimondii*, gene discovery was particularly rich in these species, leading to the detection of 3928 and 8626 contigs, respectively, for which only sequences from that species were recovered.

Orthologous and homoeologous sequences were occasionally split among two or more contigs or singletons because of imperfect contig assembly. To identify these cases, the pools of unigenes (above) were further examined to identify more cases of either orthology or homoeology. Within each pool of unigenes, all possible pairwise alignments were made for all contigs (or singletons) containing A and D ESTs. Alignments having 95% or greater sequence similarity (total or coding) and fewer than five gaps were designated as putative ortholog pairs. Using these criteria, 1464 additional pairs of putatively orthologous sequences were identified in the assembly, along with the position and composition of genome-diagnostic single nucleotide polymorphisms (SNPs) and small insertions or deletions (indels) that distinguish the A- and D-genome ESTs (Table 2). These ESTs were joined into a single contig, along with their counterparts from *G. hirsutum*, increasing the number of putatively orthologous pairs from 2179 to 3643. Within this orthologous set, polymorphisms between the A (and A_T where possible) and D (and D_T where possible) sequences were recorded, resulting in 2342 orthologous gene pairs distinguished by ~10,000 SNPs and indels. The numerical difference between the total number of orthologous loci and those with distinguishing polymorphisms was mostly due to cases in which the A and/or D EST transcripts had little to no overlap within the contig, or lack of polymorphism in the region of overlap.

Discussion

A global collection of cotton EST sequences and unigene collection

EST assemblies have previously been published for cotton (*G. hirsutum* and *G. arboreum*), but these have either been limited to one library of cotton fiber (Arpat et al. 2004), or to two pairs of relatively small libraries developed for comparisons between different experimental tissue treatments (Dowd et al. 2004; Zuo et al. 2005). Here we combined these previously reported ESTs with

Table 2. Number of homoeologous gene pairs identified in the cotton EST data set

A	D	A_T	D_T	Contig category totals within unmodified contigs (AGCol)	Number of changes to contig categories by merging putatively orthologous contigs/singletons into a single contig	Total number of classified contigs (ISU)
x				3928	−979	2949
	x			8626	−1140	7486
x		x		1033	−314	719
	x		x	933	−149	784
x	x	x	x	**309**	**+244**	**553**
x	x	x		**412**	**+367**	**779**
x	x		x	**479**	**+261**	**740**
x	x			**979**	**+592**	**1571**
		x		5271	0	5271
Total no. of contigs				22,030		20,852

AGCoL and ISU refer to their respective EST assemblies (http://agcol.arizona.edu/pave/cotton/). The ISU assembly was derived from the AGCoL Assembly as described in the text. Contigs containing only *G. hirsutum* ESTs could not be assigned to either A_T or D_T, thus the last row of the table has these two columns combined. Numbers in bold represent contigs with both A- and D-genome ESTs.

other *G. hirsutum* and *G. raimondii* ESTs to create a large global collection of *Gossypium* ESTs, effectively tripling the number of previously available *Gossypium* EST sequences in GenBank and leading to more robust bioinformatics inferences of gene content.

The assembly process resulted in a collection of 51,107 unigenes, which were further reduced by sequence similarity using BLASTN to 33,665 *Gossypium* exemplar sequences (http://agcol.arizona.edu/pave/cotton/). This set of exemplar sequences represents a nonredundant collection of cotton genes, and the total number of genes was close to the number of expected genes in diploid *Gossypium*. Wortman et al. (2003) and the International Rice Genome Sequencing Project (2005) recently estimated the number of genes in *Arabidopsis* and rice to be 28,952 and 37,544, respectively. In light of these gene number estimates and the fact that gene number inflation may be a common artifact within EST assemblies (Close et al. 2004; Lazo et al. 2004), one might expect that the 33,665 exemplar sequences from *Gossypium* identified here may be an overestimate of the number of putative genes in the assembly. By their very nature, ESTs undersample the genic content of genomes; at the end of the year 2000, for example, 105,000 *Arabidopsis* ESTs identified only ~60% of the genes (The *Arabidopsis* Genome Initiative 2000). We note, however, that the number of genes in the diploid *Gossypium* genomes has recently been independently estimated at 53,550 (Rabinowicz et al. 2005), and 35,283 ± 5423 (J. Hawkins, J. Nason, and J.F. Wendel, pers. comm.), both using a strategy of sequencing random genomic clones. Future additions of ESTs to the *Gossypium* assembly will undoubtedly improve the convergence between exemplar sequence total and gene number.

Because multiple libraries were used in the assembly, the EST collection provides a starting point for comparisons of expression differences between specific tissue treatments, environmental conditions, stress challenges, or plant organs (Supplemental Table 1). Statistical methods have been developed to correlate transcript frequency among libraries with differential gene expression (Claverie 1999; Greller and Tobin 1999; Stekel et al. 2000), and these methods have been experimentally verified (Hughes and Friedman 2004; Pavy et al. 2005). The software PAVE has these statistical tools incorporated into its Web-based functionality (http://agcol.arizona.edu/pave/cotton), facilitating comparison between libraries. For example, contig 00001_225 (a putative chalcone synthase) was identified by selecting only the *G. arboreum* fiber library and two *G. raimondii* seedling and flower libraries. In this contig, the frequency of ESTs from each library were significantly different ($R = 0.0000$), containing three transcripts from the *G. arboreum* fiber library compared to 63 from the *G. raimondii* flower and seedling libraries. This approach holds promise for elucidating various aspects of gene expression variation, particularly when two or more libraries were simultaneously developed to address a particular experimental treatment or hypothesis.

Cotton ESTs as a foundation for expression profiling

The present assembly of cotton ESTs provides a foundation for cotton genomics and functional genomics tools. As in other experimental and crop species (Meyers et al. 2004), the exemplar sequences derived from the assembly can be used as a template for microarray design for cotton functional genomics. In the present assembly, ESTs were generated from cDNA libraries representing many tissues of cotton under varied treatment conditions, thus providing a more broadly applicable data set for functional genomic applications using microarrays than is represented by earlier efforts in this regard (Arpat et al. 2004). To make the resource broadly useful, we have selected and synthesized 12,006 oligonucleotides (60–70-mers) using the software Picky (Chou et al. 2004), to include on a new cotton "longmer" oligonucleotide microarray that is available to the community on a cost-recovery basis (http://cottonevolution.info).

Because ESTs from diploid and allotetraploid *Gossypium* were combined in a single assembly and because the genomic origin of the diploid ESTs is known, we were often able to bioinformatically determine the genomic origin of ESTs from allotetraploid cotton. Intra- and intercontig polymorphisms were identified between and within putative genes, resulting in 3644 orthologous loci, whereas only 2052 loci had ESTs represented from one or both of the homoeologs. This expanded set of orthologous genes may provide novel resources for quantifying homoeologous transcript levels in allotetraploid cotton. Using single-strand conformational polymorphisms (SSCPs) to separate similarly sized sequences containing SNPs, Adams et al. (2003) showed that homoeologs are not equally expressed in all parts of the plant, and that reciprocal silencing of alternative homoeologs may characterize even the separate whorls of single flowers (see also Adams et al. 2004). Mochida et al. (2003) demonstrated similar principles in wheat using pyrosequencing, to show that 80% of the genes showed biased expression from certain genomes and that the preferred homoeolog can vary from tissue to tissue. The genome-specific polymorphisms identified in this study (both SNPs and indels) could be used to quantify A_T and D_T homoeologous transcript biases on a larger scale (e.g., using custom microarrays) than what was possible by SSCP and perhaps address the "recruitment" of D-genome loci in disease resistance and fiber development (Jiang et al. 1998; Wright et al. 1998).

The *Gossypium* EST assembly presented here provides an unprecedented look at the cotton transcriptome and contributes tools for cotton genetics and genomics efforts. The unigene set (contigs and singletons) has been reduced to a set of ~33,000 exemplar sequences for use in numerous applications, which will be aided in many cases by assignments of putative gene function based on GO annotation and protein domains. This set of processed EST sequences provides a framework for future investigation of cotton functional genomics using both long and short oligonucleotide microarrays.

Methods

Plant material, RNA extraction, and cDNA libraries

Various methods were used to create cDNA libraries and perform subtractive hybridization to produce the ESTs reported in this study. Details for 28 of these libraries, including cloning vectors, sequencing methods, and library normalization (if applicable) and GenBank accession numbers have either been published (Wu et al. 2002; Arpat et al. 2004; Haigler et al. 2005) or may be obtained upon request. Construction and sequencing of the two *G. raimondii* cDNA libraries are presented here because they are not reported elsewhere and because they constitute a substantial percentage of the total *Gossypium* EST collection. Whole seedlings and flowering plants of *G. raimondii* were grown from seed in the Pohl Conservatory at Iowa State University. Seeds were planted in a peat:vermiculite (50% Canadian sphagnum peat; 40% coarse Perlite; 10% Iowa dirt) and grown under supplemental light (16-h days) until 2 wk after their first true leaves had

emerged. Entire seedlings were collected and stored at −80°C until RNA extraction once the excess soil had been removed. Buds [−3 days post-anthesis (dpa) to −1 dpa], flowers, and developing bolls (+1 dpa to +3 dpa) were also collected from three full-sized *G. raimondii* plants. The collected tissue was also wrapped in aluminum foil and stored at −80°C until RNA extraction. RNA from both the seedlings and the floral tissues from each time point (−3, −2, −1, 0, +1, +2, +3 dpa) was extracted using a modified hot-borate method (Wilkins and Smart 1996) and checked for integrity on formaldehyde gels. Equimolar amounts of RNA (A_{260}) from each extraction/time point of floral tissue were combined into a single sample for cDNA library construction. The two cDNA libraries (whole seedlings, GR_Ea; flower, GR_Eb) were created using a pCMV.SPORT-6.1 vector and transformed into DH10B-Ton A *Escherichia coli* (Invitrogen).

EST sequencing and processing

Clone picking, arraying, and sequencing of two *G. raimondii* libraries were performed at the Arizona Genomics Institute (AGI). The ESTs were sequenced using T3 and T7 primers and Big Dye Terminator (V3.1) sequencing chemistry. The ESTs from GR_Ea and GR_Eb were deposited in GenBank (CO069431–CO100583 and CO100584–CO132899, respectively), and their corresponding trace files used to calculate quality values are available upon request. We also obtained quality values for most EST sequences and if we could not, we assigned a "neutral" quality score of 20. Less than 7% of the total bases were assigned this neutral quality value. Vector and low-quality bases (20-bp window with an average quality value < PHRED score 16) were trimmed from the libraries by LUCY (Chou and Holmes 2001).

EST assembly and accessibility

Assembly of *Gossypium* EST sequences was accomplished by Program for Assembling and Viewing ESTs (PAVE), which used a unique combination of PAcE (Kalyanaraman et al. 2003) and CAP3 software (Huang and Mandan 1999). PAVE is an EST pipeline created by the Arizona Genomics Computational Laboratory. Forward and reverse pairs of EST clones (i.e., those sequenced in both directions) were first checked for an overlap of at least 100 bases. If such an overlap was found, the "mate-pairs" were regarded as a single EST throughout the assembly process. The resulting set of contigs and singletons is considered as a set of putative unique genes (unigenes) and can be viewed and searched at http://agcol.arizona.edu/pave/cotton/.

The orientation of unigenes was determined by maximizing a weighted score derived from ESTScan (Lottaz et al. 2003) output, BLASTX alignments, and clone directionality. All unigenes were analyzed by ESTScan, which had been trained on coding and noncoding sequences of *Arabidopsis thaliana*. A score was calculated for each possible coding frame, where the ESTScan results were weighted twice as heavily as the BLASTX hits and EST clone direction. If the clone direction and ESTScan agreed (93% of all cases), then that frame and orientation were generally used. The final frame chosen was the one with the highest score, and the protein sequence was written to a separate file for Pfam analysis (see below).

Putative Gene Ontology and Pfam domain analysis

Each cotton unigene was assigned a putative Gene Ontology (GO) and a high-level functional category based on the UniProt Gene Ontology (Camon et al. 2004). The best UniProt BLASTX hit (E-value < $1e^{-20}$) and its corresponding GO annotation were determined for each cotton unigene. A list of gene associations between the UNIPROT database entries and their gene annotations are maintained by the Gene Ontology Consortium (http://www.geneontology.org/GO.current.annotations.shtml). These annotations were mapped to higher level function categories using custom PERL scripts. A Pfam search (Bateman et al. 2004) was done to examine the protein domains present in the unigene sequences. The Pfam database contains alignments and hidden Markov models for 7868 protein families (v. 17), based on the UniProt protein domains. The trusted cutoffs (--cut_tc), noise score cutoffs (--cut _nc), and the parallel virtual machine (--pvm) parameters with all the Pfam models for finding global or complete matches to the domain or family (pfam_ls) were used in the Pfam analyses (http://pfam.wustl.edu/) on 36 nodes of the Mountain cluster at ISU.

Identification of *G. hirsutum* homoeologs

Facilitated by BIOPERL (Stajich et al. 2002), a sequence scanning approach was used to identify homoeologous sequences from the A_T and D_T genomes of allopolyploid cotton by distinguishing polymorphisms within the contigs of the EST assembly. SNPs and indels were assessed at every base in each alignment through comparisons with their orthologous counterparts from the A-genome and D-genome diploids. If multiple diploid ESTs were present, each one was checked for nucleotide consistency. If there was a discrepancy among the A ESTs or among the D ESTs at any given position, the SNP was not counted unless the discrepancy was present within fewer than 25% of the A- or D-genome ESTs (i.e., 1 out of 4 ESTs).

Acknowledgments

We thank J. Mesterhazy and S. Aluru for use of the Mountain cluster at ISU for Pfam analysis. We also gratefully acknowledge the support of the National Science Foundation Plant Genome Program.

References

Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci.* **100:** 4649–4654.

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

Alba, R., Fei, Z., Payton, P., Liu, Y., Moore, S.L., Debbie, P., Cohn, J., D'Ascenzo, M., Gordon, J.S., Rose, J.K.C., et al. 2004. ESTs, cDNA microarrays, and gene expression profiling: Tools for dissecting plant physiology and development. *Plant J.* **39:** 697–714.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. 2004. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **32:** D115–D119.

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

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

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25–29.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L.L., et al. 2004. The Pfam protein families database. *Nucleic Acids Res.* **32:** D138–D141.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. 2004. The Gene Ontology Annotation (GOA) database: Sharing knowledge in

Uniprot with Gene Ontology. *Nucleic Acids Res.* **32:** D262–D266.
Cedroni, M.L., Cronn, R.D., Adams, K.L., Wilkins, T.A., and Wendel, J.F. 2003. Evolution and expression of *MYB* genes in diploid and polyploid cotton. *Plant Mol. Biol.* **51:** 313–325.
Chou, H.H. and Holmes, M.H. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics* **17:** 1093–1104.
Chou, H.-H., Hsia, A.-P., Mooney, D.L., and Schnable, P.S. 2004. Picky: Oligo microarray design for large genomes. *Bioinformatics* **20:** 2893–2902.
Claverie, J.-M. 1999. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **8:** 1821–1832.
Close, T.J., Wanamaker, S.I., Caldo, R.A., Turner, S.M., Ashlock, D.A., Dickerson, J.A., Wing, R.A., Muehlbauer, G.J., Kleinhofs, A., and Wise, R.P. 2004. A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol.* **134:** 960–968.
Cronn, R.C., Small, R.L., Haselkorn, T., and Wendel, J.F. 2002. Rapid diversification of the cotton genus (*Gossypium*: *Malvaceae*) revealed by analysis of sixteen nuclear and chloroplast genes. *Am. J. Bot.* **89:** 707–725.
Dowd, C., Wilson, I.W., and McFadden, H. 2004. Gene expression profile changes in cotton root and hypocotyl tissues in response to infection with *Fusarium oxysporum* f. sp. *vasinfectum*. *Mol. Plant Microbe Inter.* **17:** 654–667.
Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.
Endrizzi, J.E., Turcotte, E.L., and Kohel, R.J. 1985. Genetics, cytology, and evolution of *Gossypium*. *Adv. Genet.* **23:** 271–375.
Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.
Ewing, R.M., Kahla, A.B., Poirot, O., Lopez, F., Audic, S., and Claverie, J.-M. 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* **9:** 950–959.
Felsenstein, J. 2004. *PHYLIP (Phylogeny Inference Package) version 3.6*. Department of Genome Sciences, University of Washington, Seattle, http://evolution.genetics.washington.edu/phylip.html.
Fulton, T.M., Van der Hoeven, R., Eannetta, N.T., and Tanksley, S.D. 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* **14:** 1457–1467.
Greller, L.D. and Tobin, F.L. 1999. Detecting selective expression of genes and proteins. *Genome Res.* **9:** 282–296.
Haigler, C.H., Zhang, D., and Wilkerson, C.G. 2005. Biotechnological improvement of cotton fibre maturity. *Physiol. Plant.* **124:** 285–294.
Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* **9:** 868–877.
Hughes, A. and Friedman, R. 2004. Expression patterns of duplicate genes in the developing root in *Arabidopsis thaliana*. *J. Mol. Evol.* **60:** 247–256.
International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. **436:** 793–800.
Iseli, C., Jongeneel, C.V., and Bucher, P. 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequence. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **138:** 48.
Ji, S.-J., Lu, Y.-C., Feng, J.-X., Wei, G., Li, J., Shi, Y.-H., Fu, Q., Liu, D., Luo, J.-C., and Zhu, Y.-X. 2003. Isolation and analyses of genes preferentially expressed during early cotton fiber development by subtractive PCR and cDNA array. *Nucleic Acids. Res.* **31:** 2534–2543.
Jiang, C., Wright, R.J., El-Zik, K.M., and Paterson, A.H. 1998. Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). *Proc. Natl. Acad. Sci.* **95:** 4419–4424.
Kalyanaraman, A., Aluru, S., Kothari, S., and Brendel, V. 2003. Efficient clustering of large EST data sets on parallel computers. *Nucleic Acids Res.* **31:** 2963–2974.
Kawasaki, S., Borchert, C., Deyholos, M., Wang, H., Brazille, S., Kawai, K., Galbraith, D., and Bohnert, H.J. 2001. Gene expression profiles during the initial phase of salt stress in rice. *Plant Cell* **13:** 889–906.
Kim, H.J., William, M.Y., and Triplett, B.A. 2002. A novel expression assay system for fiber-specific promoters in developing cotton fibers. *Plant Mol. Biol. Rep.* **20:** 7–18.
Lacape, J.-M., Nguyen, T.-B., Courtois, B., Belot, J.-L., Giband, M., Gourlot, J.-P., Gawryziak, G., Roques, S., and Hau, B. 2005. QTL analysis of cotton fiber quality using multiple *Gossypium hirsutum* x *Gossypium barbadense* backcross generations. *Crop Sci.* **45:** 123–140.
Lazo, G.R., Chao, S., Hummel, D.D., Edwards, H., Crossman, C.C., Lui, N., Matthews, D.E., Carollo, V.L., Hane, D.L., You, F.M., et al. 2004. Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. *Genetics* **168:** 585–593.
Li, X.-B., Cai, L., Cheng, N.-H., and Liu, J.-W. 2002. Molecular characterization of the cotton *GhTUB1* gene that is preferentially expressed in fiber. *Plant Physiol.* **130:** 666–674.
Lottaz, C., Iseli, C., Jongeneel, C.V., and Bucher, P. 2003. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* **19:** ii103–ii112.
Meyers, B.C., Galbraith, D.W., Nelson, T., and Agrawal, V. 2004. Methods for transcript profiling in plants. Be fruitful and replicate. *Plant Physiol.* **135:** 637–652.
Michalek, W., Weschke, W., Pleissner, K.-P., and Graner, A. 2002. EST analysis in barley defines a unigene set comprising 4,000 genes. *Theor. Appl. Genet.* **104:** 97–103.
Mochida, K., Yamazaki, Y., and Ogihara, Y. 2003. Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol. Gen. Genomics* **270:** 371–377.
Orford, S.J. and Timmis, J.N. 1998. Specific expression of an expansin gene during elongation of cotton fibers. *Biochem. Biophys. Acta* **1398:** 342–346.
Orford, S.J., Carney, T.J., Olenicky, N.S., and Timmis, J.N. 1999. Characterization of a cotton gene expressed late in fibre cell elongation. *Theor. Appl. Genet.* **98:** 757–764.
Pavy, N., Laroche, J., Bousquet, J., and Mackay, J. 2005. Large-scale statistical analysis of secondary xylem ESTs in pine. *Plant Mol. Biol.* **57:** 203–224.
Rabinowicz, P.D., Citek, R., Budiman, M.A., Numberg, A., Bedell, J.A., Lakey, N., O'Shaughnessy, A.L., Nacimiento, L.U., McCombie, W.R., and Martienssen, R.A. 2005. Differential methylation of genes and repeats in land plants. *Genome Res.* **15:** 1431–1440.
Rong, J., Abbey, C., Bowers, J.E., Brubaker, C.L., Chang, C., Chee, P.W., Delmonte, T.A., Ding, X., Garza, J.J., Marler, B.S., 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.
Ronning, C.M., Stegalkina, S.S., Ascenzi, R.A., Bougri, O., Hart, A.L., Utterbach, T.R., Vanaken, S.E., Riedmuller, S.B., White, J.A., Cho, J., et al. 2003. Comparative analyses of potato expressed sequence tag libraries. *Plant Physiol.* **131:** 419–429.
Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J., and Shoemaker, R. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47:** 868–876.
Senchina, D.S., Alvarez, I., Cronn, R.C., Liu, B., Rong, J., Noyes, R.D., Paterson, A.H., Wing, R.A., Wilkins, T.A., and Wendel, J.F. 2003. Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Mol. Biol. Evol.* **20:** 633–643.
Skovsted, A. 1934. Cytological studies in cotton. II. Two interspecific hybrids between Asiatic and New World cottons. *J. Genet.* **28:** 407–424.
Small, R.L. and Wendel, J.F. 2000a. Phylogeny, duplication, and intraspecific variation of Adh sequences in new world diploid cottons (*Gossypium* L., Malvaceae). *Mol. Phylo. Evol.* **16:** 73–84.
———. 2000b. Copy number lability and evolutionary dynamics of the Adh gene family in diploid and tetraploid cotton (*Gossypium*). *Genetics* **155:** 1913–1926.
———. 2002. Differential evolutionary dynamics of duplicated paralogous Adh loci in allotetraploid cotton (*Gossypium*). *Mol. Biol. Evol.* **19:** 597–607.
Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12:** 1611–1618.
Stekel, D.J., Git, Y., and Falciani, F. 2000. The comparison of gene expression from multiple cDNA libraries. *Genome Res.* **10:** 2055–2061.
Suo, J., Liang, X., Pu, L., Zhang, Y., and Xue, Y. 2003. Identification of *GhMYB109* encoding a R2R3 MYB transcription factor that expressed specifically in fiber initials and elongating fibers of cotton (*Gosspyium hirsutum* L.). *Biochem. Biophys. Acta* **1630:** 25–34.
USDA–FAS. 2005. Cotton: World markets and trade. United States Department of Agriculture Foreign Agricultural Service. FC-07-05, http://www.fas.usda.gov/cotton/circular/2005/07/CottonWMT.pdf
Vettore, A.L., da Silva, F.R., Kemper, E.L., Souza, G.M., da Silva, A.M., Ferro, M.I.T., Henrique-Silva, F., Giglioti, E.A., Lemos, M.V.F., Coutinho, L.L., et al. 2003. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Res.* **13:** 2725–2735.
Wendel, J.F. 1995. Cotton. In *Evolution of crop plants* (eds. N. Simmonds and J. Smartt), pp. 358–366. Longman, London.
Wendel, J.F. and Cronn, R.C. 2003. Polyploidy and the evolutionary history of cotton. *Adv. Agronomy* **78:** 139–186.
Whitfield, C.W., Band, M.R., Bonaldo, M.F., Kumar, C.G., Liu, L.,

Pardinas, J.R., Robertson, H.M., Soares, M.B., and Robinson, G.E. 2002. Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. *Genome Res.* **12:** 555–566.

Wilkins, T.A. and Smart, L.B. 1996. Isolation of RNA from plant tissue. In *A laboratory guide to RNA: Isolation, analysis, and synthesis* (ed. P.A. Krieg), pp. 21–41. Wiley-Liss, New York.

Wisman, E. and Ohlrogge, J. 2000. *Arabidopsis* microarray service facilities. *Plant Physiol.* **124:** 1468–1471.

Wortman, J.R., Haas, B.J., Hannick, L.I., Smith Jr., R.K., Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A., et al. 2003. Annotation of the *Arabidopsis* genome. *Plant Physiol.* **132:** 461–468.

Wright, R.J., Thaxton, P.M., El-Zik, K.M., and Paterson, A.H. 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., Llewellyn, D.J., and Dennis, E.S. 2002. A quick and easy method for isolating good-quality RNA from cotton (*Gossypium hirsutum* L.) tissues. *Plant Mol. Biol. Rep.* **20:** 213–218.

Zhang, D., Hrmova, M., Wan, C.-H., Wu, C., Balzen, J., Cai, W., Wang, J., Densmore, L.D., Fincher, G.B., Zhang, H., et al. 2004. Members of a new group of chitinase-like genes are expressed preferentially in cotton cells with secondary walls. *Plant Mol. Biol.* **54:** 353–372.

Zhao, G. and Liu, J. 2001. Isolation of a cotton RGP gene: A homolog of reversibly glycosylated polypeptide highly expressed during fiber development. *Biochem. Biophys. Acta* **1574:** 370–374.

Zuo, K., Wang, J., Wu, W., Chai, Y., Sun, X., and Tang, K. 2005. Identification and characterization of differentially expressed ESTs of *Gossypium barbadense* infected by *Verticillium dahliae* with suppression of subtractive hybridization. *Mol. Biol.* **39:** 191–199.

Received August 25, 2005; accepted in revised form November 14, 2005.

A global assembly of cotton ESTs

Joshua A. Udall, Jordan M. Swanson, Karl Haller, et al.

Genome Res. 2006 16: 441-450
Access the most recent version at doi:10.1101/gr.4602906