The *Journal of Cotton Science 10:29–38 (2006)*
http://journal.cotton.org, © The Cotton Foundation 2006

BREEDING AND GENETICS

Predicting Intron Sites by Aligning Cotton ESTs with *Arabidopsis* Genomic DNA

Pawan Kumar, Andrew H. Paterson, and Peng W. Chee*

ABSTRACT

Among the transcribed regions of a gene, introns are the most polymorphic; therefore, they are ideally suited for developing molecular markers. The identification of intron regions, however, is not a straight-forward process and involves the alignment of expressed sequence tags (EST) or cDNA with their genomic counterpart. In cotton, this process is exacerbated by the scarcity of cotton genomic DNA sequences in Genbank. In this study, the possibility of utilizing genomic sequences from *Arabidopsis*, whose genome has been completely sequenced, to locate intron regions in cotton was evaluated. Cotton ESTs were searched in BLAST against the *Arabidopsis* database to identify orthologous genes. Cotton introns were identified with a 92% success rate, based on the alignment of cotton ESTs with *Arabidopsis* genomic DNA, which demonstrated that this approach is both feasible and practical for predicting the locations of introns in cotton ESTs. A majority of cotton introns had the canonical GT-AG splice site junctions, facilitating their identification in the sequence alignment process. Comparison of sequences between *G. arboreum* L. and *G. raimondii* Ulbr. indicated that introns had an almost four-fold greater variation in nucleotides than exons. A majority of the differences were due to a repeating thymine (T) or to the number of simple sequence repeat motifs.

Eukaryotic genes consist of protein coding sequences called exons and non-protein coding intervening sequences called introns. In the initial step of DNA transcription, the genetic information of a gene is copied from DNA into heterogeneous nuclear RNA, which contains both exon and intron sequences. Introns are then removed by a precise cleavage-ligation reaction, called splicing, to produce a functional messenger RNA (mRNA) that carries only the genetic code that specifies a sequence of amino acids.

Major changes in intron structure, such as large insertions or deletions, have been shown to affect gene expression (Fridman et al., 2000). Like other DNA sequences that do not code for protein products, introns are more prone to accumulate nucleotide variation because the fitness consequences of mutations are small (Holland et al., 2001). Because intron sequences evolve more rapidly than exon sequences in both plants (Small and Wendel, 2000) and animals (Hughes and Yeager, 1997), they are a valuable tool in population genetics studies, such as in phylogenetic reconstruction (Liu et al., 2001) and in developing “molecular clocks” to estimate the time of divergence between species (He and Haymer, 1997; Johnson and Soltis, 1994). The prospect of utilizing intron sequences as genetic markers, however, has not been adequately studied in most species. Despite their abundance in eukaryotic genomes, most introns have been found during cloning and sequencing of specific genes; therefore, their use as genetic markers has been restricted to a few genes.

Introns were first discovered in 1977 based on the observation that mRNA was always shorter than the DNA template from which it had been transcribed (Williamson, 1977). Comparing transcribed mRNA sequences with their template genomic sequences is the primary means of identify introns. Regions in the genomic sequence that correspond with the transcribed sequence delimit exons, while the sequences present in the genomic region but absent in the transcribed counterpart are indicative of introns (Loraine and Helt, 2002). In the past, transcribed sequences could be obtained from the sequences of cDNA libraries developed for restriction fragment length polymorphism (RFLP) analysis. More recently, a large number of ESTs generated from the sequencing of randomly selected cDNA clones have been developed in many species, including cotton, and are available in the Genbank database. As of 7

P. Kumar and P.W. Chee, Department of Crop and Soil Sciences, Coastal Plains Experiment Station, University of Georgia, Tifton, GA 31794; A. H. Paterson, Plant Genome Mapping Laboratory, University of Georgia, Athens GA 30602
*Corresponding author: pwchee@uga.edu

Jan. 2005, the Genbank database contained more than 126,000 EST sequences from the *Gossypium* genera, including 39,007 from *G. arboreum* L., 134 from *G. barbadense* L., and 23,899 from *G. hirsutum* L.

At present, the limiting factor in identifying and studying cotton introns is the lack of genomic sequences. Although a few cotton genomic DNA sequences exist in the Genbank database, most were generated from either BAC-end sequences (Tomkins et al., 2001) or from microsatellite marker discovery (Reddy et al., 2001), which rarely contain coding gene sequences. Chee et al. (2004) reported that 23% of the polymerase chain reaction (PCR) primers developed from random EST sequences amplify products that were more than 150 nucleotides longer than expected from the EST sequence. DNA sequencing confirmed that many of those PCR products contain intron sequences. The unexpectedly larger amplicon size could offer a means of predicting the presence of an intron without the need to perform the costly DNA sequencing on each amplicon, but the low frequency of introns harbored in cotton ESTs renders this strategy too costly for developing intron spanning PCR markers.

Since the draft sequences of the *Arabidopsis* and *Oryza* (rice) genomes have been recently completed, a large number of genomic DNA sequences are now available from these species. Since *Arabidopsis* and *Gossypium* diverged from a common ancestor in less than 100 million years ago (Bowers et al., 2003), they are close relatives and conservation in both gene structure and sequence is expected to be high between these species. *Arabidopsis* genomic DNA sequences may offer a template from which to locate the positions of exons and introns of cotton genes, but this has not been tested. The objective of this study was to study the structure of cotton introns and to explore the possibility of identifying cotton intron positions with the help of *Arabidopsis* genomic sequences.

MATERIALS AND METHODS

Plant material and DNA extraction. Seeds of *G. arboreum* accession A2-47 (PI 213373) and *G. raimondii* accession D5-4 (PI 530901) were obtained from the National Cotton Germplasm Collection, USDA/ARS, College Station, Texas. Young leaves from 3 to 5 seedlings grown in a greenhouse were collected in bulk for genomic DNA following the protocol described by Paterson et al. (1993).

EST sequences and PCR primers. A total of 170 EST sequences from *G. arboreum* and G *hirsurum* were randomly selected from GenBank (National Center for Biotechnological Information; http://www.ncbi.nlm.nih.gov/GeneBank/index/html). Only sequences that were more than 100 nucleotides in length were selected, because short sequences may detect partial homology for genes that have similar motifs but different functions. Also, a short PCR product may have less chance of amplifying intron sequences. PCR primers were designed for 64 ESTs that showed more than 80% nucleotide identity to previously described genes. PCR primers were designed using the Primer Express software (Perkin-Elmer Applied Biosystems; Foster City, CA). The average length of the primers was 20 nucleotides with an annealing temperature of 55 °C. PCR amplification was performed using genomic DNA from *G. arboreum* and *G. raimondii*. Conditions for PCR amplification have been described elsewhere (Chee et al., 2004).

Chee et al. (2004) showed that amplicons for PCR primers derived from ESTs that were at least 150 nucleotides larger than predicted from EST sequences contained introns. They also observed that individual intron size in cotton could vary from 77 to 611 nucleotides in length, which indicates that the 150 nucleotides threshold for identifying amplicons that might be harboring introns may need to be reduced. In this study, amplicons with 50 nucleotides greater than predicted from ESTs were assumed to harbor introns. In total, 43 (67%) primers produced PCR fragments that were 50 nucleotides longer than expected. These amplicons were selected for DNA sequencing.

Sequencing of PCR products. The genomic sequence for each of the 43 ESTs was obtained by sequencing the PCR fragments in both the 3' and 5' directions using BigDye Terminator cycle sequencing kit (Perkin-Elmer Applied Biosystems). Electrophoretic separation of the sequencing products, after filtering through Sephadex filters, was performed on an ABI Prism 3700 DNA analyzer 96-capillary automated sequencer (Perkin-Elmer Applied Biosystems).

BLAST search and EST to genomic DNA alignment. All of the 43 selected EST sequences were aligned with their genomic counterpart by ClustalW (European Bioinformatics Institute; http://www.ebi.ac.uk/clustalw/index.html) multiple alignment accessory, BioEdit software ver.

6.0.7. (Hall et al., 1999). EST to genomic DNA alignment of 13 of the 43 sequences yielded complete introns (Table 1). Because the remaining sequences were too short to give the complete sequence of the intron, they were excluded from this study. In addition, this study also included sequences of 7 PCR products (EST-118, -132, -141, -144, -152, -167 and –186) previously described by Chee et al. (2004) and 12 cotton genes obtained from the Genbank that were previously reported by Wendel et al. (2002). The sequences from Chee et al. (2004) were derived by direct sequencing of PCR products amplified from genomic DNA using arbitrary EST-PCR primers, and those from Wendel et al. (2002) were derived from the sequencing of cloned PCR products amplified using either genomic DNA or BAC clones. In total, 32 partial or full-length cDNA-genomic DNA sequences were aligned, allowing the structure of 56 complete introns to be analyzed.

Putative functions of ESTs were assigned by searching in BLAST (Altschul et al., 1990) against the database of all organisms. Cotton EST sequences were also searched in BLAST against a non-redundant database of *Arabidopsis*. The conditions for BLAST were moderately stringent with the expected (E) value set at 10, which is the statistical significance threshold for reporting matches against database sequences. EST sequences were aligned with cotton and *Arabidopsis* genomic DNA sequences.

Table 1. Primer sequences, putative gene orthologs, and Genbank accessions of cotton EST primers

Primer	Genbank accession no.	Putative gene ortholog	Primer sequences
EST206	AA659985	Putative nucleotide-sugar	F-ATAAAGACGAATGTGATCGG
		Dehydrates	R-CATCAAAGTTTCAGCCACTC
EST207	AI726709	Putative alpha-L-	F-TGGGAAGAAGAATTATCGAG
		Arabinofuranosidase	R-TGTCGTGTTGTACGATCAAA
EST208	AF305065	PR protein class 10	F-CATGGGTGTTGTCACTTATAAC
			R-ATACCAAAAGCACACCATTC
EST210	AI728703	P-glycoprotein-2	F-AAAACGTAGGCTTGGTCATT
			R-CCATAATCTCGAACACCG
EST226	AI728009	Annexin-like protein	F-TGCACTAGGTCTTCACATGA
			R-ATTTCTCAGGGACAGTCAAG
EST237	AI727755	Alpha subunit of F-actin	F-CCCCTGATGATAGTGCAA
		capping protein	R-CTGTACCCAAAATTTCCACA
EST245	AI055464	Feebly-like protein	F-ATGGATTCCTATTTCAGGTG
			R-CGTCTTTTCAGTCATTCTGTC
EST268	AI725935	Nodulin-like protein	F-ATGCCCCCTTCTTATTGT
			R-CAAGAGAAAGGTAAAGACTGGA
EST280	AI728393	Putative NADH-	F-TCATGAAACCTGCTGTAATG
		ubiquinone oxireductase	R-GGGTAACCCTTCAACTTATG
EST319	AI730677	Plastid protein	F-CACCTTACTCTCTCCTTTATCC
			R-ACAATTCAGCACCATAGTCC
EST322	AI726457	Putative pollen-specific	F-GAGTTCACAACGTGGGAATA
		Protein	R-CTGGTGGTTTGTTTCTCAA
EST387	AI728296	Nonclathrin coat protein	F-GGAGTTTGCTGAAGTAGCTT
		gamma-like protein	R-CTTGACATTGCCGATGTATA
EST394	AI728954	Germin-like protein	F-TGTTAAATCTCCATGGCTG
			R-GAACACGAAAATGTCTCCTT

RESULTS AND DISCUSSION

Cotton EST and genomic DNA alignment. The alignment between orthologous sequences of the ESTs and their PCR-amplified counterparts from *G. arboreum* and *G. raimondii* was straightforward due to the low level of nucleotide polymorphism. When present, intron structure can easily be detected because of its demarcation as gaps on ESTs when compared with genomic DNA alignment (Fig. 1). In total, 57 complete introns were recognized from the 32 EST genomic DNA alignments that were studied (Table 2 and 3). They included 19 introns newly detected in this study, 11 introns previously

Table 2. Introns size variation between cotton and *Arabidopsis* from EST-derived sequences

Cotton primer[y]	Orthologous *Arabidopsis* genbank accession[z]	Introns number	Intron size (bp)			Splice site
			G. arboreum	*G. raimondii*	*Arabidopsis*	
EST118	U78721	1	106	106	70	GT-AG
		2	156	156	316	GT-AG
		3	102	102	120	GT-AG
EST132[*]	AC021640	1	96	96	106	GT-AG
EST141[*]	AB005241	1	92	92	107	GT-AG
EST144[*]	AC011810	1	90	87	90	GT-AG
		2	80	80	85	GT-AG
EST152[*]	AC002335	1	608	609	305	GT-AG
EST167[*]	AL161562	1	80	80	420	GT-AG
EST186[*]	AC007211	1	73	73	103	GT-AG
		2	219	218	96	GT-AG
EST206	AL133298	1	124	121	82	GT-AG
		2	80	89	86	GT-AG
EST207	AF149413	1	139	140	86	AA-TA
EST208	AP000736	1	77	77	78	GT-AG
EST210	AL049483	1	82	85	120	GT-GC
EST226	AL356332	1	94	93	88	GT-AG
		2	292	301	85	GT-AG
EST237	AC009606	1	77	77	81	GT-AG
EST245	AL161550	1	265	267	81	GT-AG
		2	138	140	96	GT-AG
EST268	AC006569	1	268	268	559	GT-AG
EST280	AL161585	1	101	102	94	GT-AG
		2	72	74	76	GT-AG
		3	319	315	107	GA-GG
		4	101	101	89	GT-AG
EST319	AC010870	1	307	309	144	GT-AG
EST322	AC002332	1	146	144	147	GT-AG
EST387	AL161563	1	206	206	196	GT-AG
EST394	Z97336	1	93	137	96	GT-AG

[y] EST sequences marked with an asterisk (*) are described in Chee et al. (2004).

[z] Orthology determined by BLAST Match.

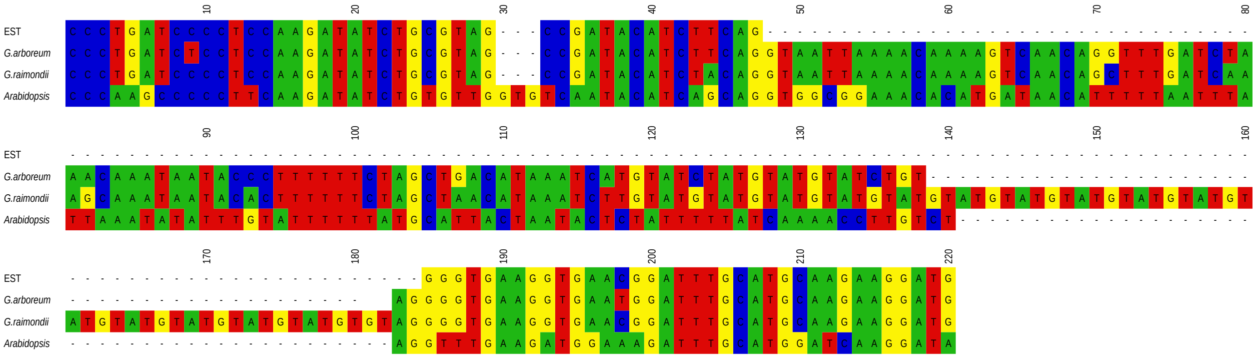
10 20 30 40 50 60 70 80
EST CCCTGATCCCCTCCAAGATATCTGCGTAG---CCGATACATCTTCAG---------------------------------
G.arboreum CCCTGATCTCCTCCAAGATATCTGCGTAG---CCGATACATCTTCAGGTAATTAAAACAAAAGTCAACAGGTTTGATCTA
G.raimondii CCCTGATCCCCTCCAAGATATCTGCGTAG---CCGATACATCTACAGGTAATTAAAACAAAAGTCAACAGCTTTGATCAA
Arabidopsis CCCAAGCCCCCTTCAAGATATCTGTGTTGGTGTCAATACATCAGCAGGTGGCGGAAACACATGATAACATTTTTAATTTA
90 100 110 120 130 140 150 160
EST --
G.arboreum AACAAATAATACCCTTTTTTCTAGCTGACATAAATCATGTATCTATGTATGTATCTGT----------------------
G.raimondii AGCAAATAATACACTTTTTTCTAGCTAACATAAATCTTGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGT
Arabidopsis TTAAATATATTTGTATTTTTTATGCATTACTAATACTCTATTTTTATCAAAACCTTGTCT--------------------
170 180 190 200 210 220
EST ------------------------GGGTGAAGGTGAACGGATTTGCATGCAAGAAGGATG
G.arboreum ----------------------AGGGGTGAAGGTGAATGGATTTGCATGCAAGAAGGATG
G.raimondii ATGTATGTATGTATGTATGTGTAGGGGTGAAGGTGAACGGATTTGCATGCAAGAAGGATG
Arabidopsis ----------------------AGGTTTGAAGATGGAAAGATTTGCATGGATCAAGGATA

Figure 1. Multiple alignments of EST394 (*Germin-like protein*) with genomic DNA from *G. arboreum*, *G. raimondii*, and *Arabidopsis*. Gaps in EST demarcate intron regions.

Table 3. Intron size variation in cotton and *Arabidopsis* gene orthlogous

Cotton gene[y]	Orthologous *Arabidopsis* GB accession[z]	Intron	Intron size (bp)					Splice site
			G. arboreum	*G. hirsutum* A-subgenome	*G. raimondii*	*G. hirsutum* D-subgenome	*Arabidopsis*	
Adh A	AL161556	1	74	74	72	72	87	GT-AG
		2	84	84	84	84	0	GT-AG
		3	89	97	90	97	86	GT-AG
Adh B	AC002291	1	83	83	84	84	103	GT-AG
		2	90	90	91	91	87	GT-AG
		3	118	115	115	108	0	GT-AG
		4	100	100	100	100	0	GT-AG
		5	109	109	107	107	85	GT-AG
		6	143	143	143	143	0	GT-AG
B5	AB016892	1	106	104	98	98	54	GT-AG
B8	AL161491	1	222	222	221	222	84	GT-AG
CesA2	AB006703	1	133	133	133	133	79	GT-AG
		2	85	84	84	84	71	GT-AG
		3	133	133	133	133	543	GC/GT-AG
D1	AF002109	1	115	115	115	115	84	GT-AG
D5	AC073506	1	82	82	82	82	72	GT-AG
GhMYB6	AL161515	1	94	94	93	97	0	GT-AG
		2	153	153	155	155	309	GT-AG
GhCLK1	AB016893	1	124	124	125	125	95	GT-AG
		2	312	312	312	312	151	GT-AG
		3	83	83	83	83	89	GT-AG
		4	140	140	140	140	113	GT-AG
		5	221	221	221	221	135	GT-AG
C7		1	470	470	477	477	NA	GT-AG
D2		1	79	79	80	80	NA	GT-AG
		2	73	73	75	74	NA	GT-AG
D7		1	443	452	420	441	NA	GT-AG

[y] Gene identity determined by Wendel et al. (2002).

[z] Orthology determined by BLAST Match.

detected by Chee et al. (2004), and 27 detected by Wendel et al. (2002). Of the 30 introns found in this study and those reported by Chee et al. (2004), the average length of introns from *G. arboreum* was 156 base pairs (bp) and from *G. raimondii* was 158 bp. These results compare closely with the 149.43 and 149.57 bp length for *G. arboreum* and *G. rainomdii*, respectively, reported by Wendel et al. (2002) based on 76 introns from 28 genes. The average intron size of both species does not differ significantly, despite the *G. arboreum* genome being nearly twice the physical size of the *G. raimondii* genome (Endrizzi et al., 1985). A very similar intron size was observed in the At and Dt subgenome of allotetraploid cottons (Wendel et al., 2002). These data confirm previous findings that indicate genome size and intron size are not correlated in the A and D genome *Gossypium* species (Wendel et al., 2002).

Although average intron size is remarkably consistent across the genome of *G. arboreum* and *G. raimondii*, the size of individual introns varies from 73 bp to 609 bp (Table 2). Within a single gene, the number and size of introns can vary considerably. For example, the *CesA1* gene, which encodes for cellulose synthase, was reported to have 11 introns, and the GhCLK1 gene, which encodes for a protein kinase, was reported to have 5 introns (Wendel et al., 2002) with the shortest being 83 bp and the longest being 312 bp (Table 3). While the presence of introns can be beneficial because of their association with regulatory elements, such as enhancers (Duret and Bucher, 1997), and with alternative splicing, which allows a single stretch of DNA to code for more than one functional protein (Caceres and Kornblihtt, 2002), the large disparity in intron size is less clear. Carvalho and Clark (1999) suggest that mutation tends to increase intron size, while natural selection favors smaller introns because of the burden associated with unnecessary DNA replication and gene transcription.

All introns contain conserved sequences adjoining the exons at the splice sites that direct their correct removal from the initial transcripts when processed to mature RNAs (Lewin, 1997). Splicing involves a precise looping process that is controlled by GT at the start, or 3' end, and AG at the distal (5') 'acceptor' site (Downie et al., 1991). If these sites are changed or removed, the gene product may be altered, resulting in a nonfunctional protein (Brown et al., 1996). The majority of cotton introns (95%) contained the canonical GT-AG splice site junctions, so they follow the standard splicing pathway (U2-type spliceosome) (Hebsgaard et al., 1996). Only 5% of introns were found with non-canonical splice site (Table 2 and 3). Non-canonical splice sites are reported to occur at a frequency of 1.7% in *Arabidopsis* (Zhu et al., 2003) and 1% in mammals (Burset et al., 2000). This study may have overestimated the frequency of non-canonical splice sites in cotton introns because of the smaller sample size.

Multiple alignment with Arabidopsis genomic DNA. It was hypothesized that multiple alignments of cotton EST sequences with *Arabidopsis* genomic sequences could facilitate scanning of cotton introns. Cotton ESTs were searched using BLAST against the *Arabidopsis* database in Genbank to identify orthologous genomic DNA sequences. The *Arabidopsis* orthologs were found for 29 of the 32 (91%) cotton genes used in this study (Table 2 and 3). A high rate of matching was expected because most structural genes have conserved function across widely divergent taxa (Li et al., 2004). Further, since the *Malvaceae* and *Brassicaceae* have diverged from a common ancestor less than 100 million years ago (Bowers et al., 2003), a high level of DNA sequence conservation is expected within protein coding regions (Fulton et al., 2002).

In comparing sequence alignments between cotton ESTs and the genomic sequences of *G. arboreum* and *G. raimondii* (discussed above), the alignment of the genomic sequence of *Arabidopsis* with cotton ESTs was not straightforward (see Fig. 1 and 2). Extensive levels of variation in sequence and fragment size were often observed between cotton and *Arabidopsis* introns, which complicated the search for intron-exon boundaries. Despite the fact that the two species showed a low level of divergence in the exonic regions, gaps were often present in either *Arabidopsis* or cotton genomic sequences and were manually created to establish good alignment of exons. Once aligned, the intron size of cotton and *Arabidopsis* can be unambiguously determined (Fig. 1 and 2). Among the 29 partial or full-length genes analyzed, the average length of the *Arabidopsis* intron was 140 bp or only slightly smaller than the 146 bp previously reported (Hebsgaard et al., 1996). The smaller size was likely due to sampling error in this study. While the average intron size between cotton and *Arabidopsis* was similar (about 15 bp difference), the individual intron size of orthologous genes displayed a much higher level of disparity. For example, the nodulin-like protein gene in both cotton species amplified by primer pairs EST268 showed only a 268 bp intron, but this same intron was 559 bp in *Arabidopsis* (Table 3). A similar disparity but in reverse order was observed for the copalyl diphosphate synthase 1 gene (EST152), which had 608 bp in the cotton intron but only 305 bp in *Arabidopsis* (Table 3).

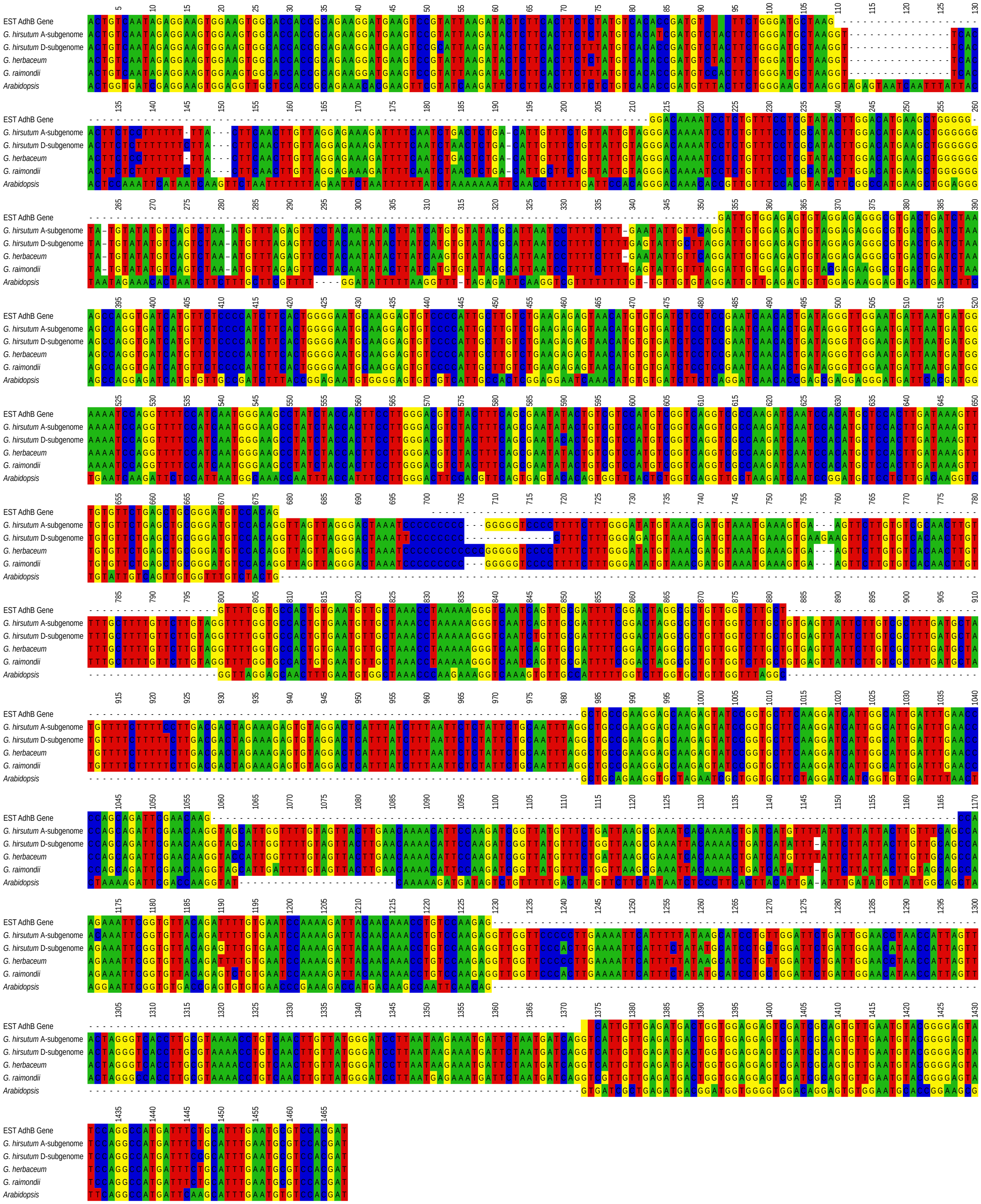

Figure 2. Multiple alignments of *Alcohol dehydrogenase B* gene sequence from *G. hirsutum* A- and D-subgenome, *G. arboreum*, *G. raimondii*, and *Arabidopsis*. Gaps in EST demarcate intron regions. Introns number 3, 4, and 6 are missing in the *Arabidopsis* sequence.

The high success rate (91%) in utilizing genomic DNA from *Arabidopsis* to predict the presence of introns in cotton ESTs, suggest that it is feasible to determine the splice site of a cotton intron by aligning cotton ESTs to the fully sequenced genome of *Arabidopsis*. The ability to predict the presence of introns in cotton ESTs would bypass the necessity of testing each EST by PCR primer development and DNA sequencing to identify the subset that may harbor introns. The lack of association between the intron size of *Arabidopsis* and cotton suggests that DNA sequencing will still be necessary to reveal the intron size in cotton. Interestingly, of the four introns that were not detected in *Arabidopsis*, one was from the alcohol dehydrogenase A gene and 3 were from alcohol dehydrogenase B (Fig. 2). Absence of these introns in *Arabidopsis* is not surprising as intron loss or gain is frequently observed in eukaryotes (Roy and Gilbert, 2005).

Intron spanning PCR markers. The major hindrances in applying DNA marker technology to study and improve Upland cotton (*G. hirsutum* L.) include a large genome size, polyploidy, and lack of DNA variation in the domesticated germplasm. While the first two are inherent problems that are unavoidable in cotton genome research, genetic markers can be developed that can provide a more effective means for detecting DNA polymorphisms. RFLP remains a popular type of DNA marker for use in cotton genomic research (Rong et al., 2004), because it is efficient for genetic map construction and gene discovery and essential for comparative genomics. RFLPs, however, are cumbersome for use in mainstream cotton improvement, because they require large amounts of DNA, tedious blot hybridization, autoradiographic methods, and are only sensitive enough to detect a small fraction of the DNA polymorphisms that occur between genotypes.

Several PCR-based DNA marker systems that are arguably more sensitive in detecting polymorphisms, such as random amplified polymorphic DNA (RAPD) analysis, simple single repeats (SSR), and amplied fragment length polymorphisms (AFLP), have been developed for cotton in recent years. Except for a specific class of SSRs that were derived from ESTs (Han et al., 2004; Qureshi et al., 2004), a majority of these PCR-based markers amplify DNA sequences from the non-expressed portion of the cotton genome. Since cotton has a large genome size, composed largely of non-informative “junk DNA”, the development of markers that specifically target gene sequences would lead to better understanding of the organization and location of gene rich regions of the genome. The development of markers that target protein-coding gene sequences will allow cotton researchers to make the transition from genetic linkage mapping to candidate gene mapping, which has been adopted by many model organisms to dissect complex traits (McCombie et al., 1992; Newman et al., 1994).

Exonic regions, which have large and direct effects on phenotype, are likely to maintain the least amount of nucleotide variation due to selection compared with intronic regions (Holland et al., 2001). Previous studies involving direct comparison of large segments of DNA sequences between two *Arabidopsis* genotypes revealed 3 times greater frequency of polymorphisms in introns than in exons (*Arabidopsis* Genome Initiative, 2001). In *Gossypium* species, Chee et al. (2004) showed that introns harbored 3.7 times more nucleotide substitution rates than exons among the 6 genes that were compared between *G. arboreum* and *G. raimondii*. If the goal were to develop PCR markers based on functional genes, targeting introns regions would increase the probability of detecting polymorphism. Of the 32 partial or full-length gene sequences analyzed for *G. arboreum* and *G. raimondii,* 11% nucleotide variation was detected in the intronic regions and 4% in flanking exonic regions. A close observation on the nucleotide compositions suggest that monomeric T repeats are the major source of nucleotide polymorphism in cotton introns. This is perhaps due to the fact that introns are more AT-rich than exons, but our data shows that the average AT content of introns was 66%, which was slightly lower that the 73% reported for dicotyledonous plants (Hebsgaard et al., 1996). In contrast, flanking exons in cotton have an average AT content of 56%, which was very similar to the 55% for most dicotyledonous plants (Hebsgaard et al., 1996).

An unexpected observation was that a number of insertions/deletions (indels) in cotton introns appear to be in the form of SSRs. For example, the gene B5 (germin E protein precursor) harbored a SSR with a $(CT)_n$ motif, which was repeated 11 times in *G. raimondii* but only 7 times in *G. arboreum*, and resulted in intron sizes of 106 and 98 bp, respectively (Table 3). Likewise, primer pair EST394 amplified a partial sequence for the putative Gemini-like protein gene that contained a intron with a $(GTAT)_n$ motif, which was repeated 16 times in *G. raimondii* but only 5 times in *G. arboreum*, and created an intron length variation of 44 bp between the two species (Fig. 1). These hypervariable SSRs are the marker of choice for genome mapping and dissection of complex traits in many crop species, including cotton. Recently, a large number of cotton SSR markers have been developed from mining cotton EST databases (Han et al., 2004; Qureshi et al., 2004), but a study in *Arabidopsis* showed that of all the SSRs found in transcribed sequences, 63% were found in introns and only 46% in exons (Cardle et al., 2000). Therefore, it is suspected that many additional cotton SSRs targeting transcribed genes could be developed once their intron sequences are identified.

ACKNOWLEDGEMENTS

We thank S. Schulze and D. Williams for technical assistance and the UGA-Research Foundation for financial support.

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Bowers, J. E., B. A. Chapman, J. Rong, and A. H. Paterson. 2003. Unravelling angiosperm chromosome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422: 433-438.

Brown, J. W. S., P. Smith, and C. G. Simpson. 1996. *Arabidopsis* consensus intron sequences. Plant Mol. Biol. 32: 531–535.

Burset, M., I. A. Seledtsov, and V. V. Solovyev. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 28: 4364–4375.

Caceres, J. F., and A. R. Kornblihtt. 2002. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends Genet. 18: 186–193.

Cardle, L., L. Ramsay, D. Milbourne, S. M. Macaulay, D. Marshall and R. Waugh. 2002. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156: 847–854.

Carvalho, A. B., and A. G. Clark. 1999. Intron size and natural selection. Nature 401: 344

Chee, P. W., J. Rong, D. W. Coplin, S. R. Schulze, and A. H. Paterson. 2004. EST derived PCR-based markers for functional gene homologues in cotton. Genome 47: 449-462.

Downie, S. R., R. G. Olmstead, G. Zurawski, D. E. Soltis, P. S. Soltis, J. C. Watson, and J. D. Palmer. 1991. Six independent losses of the chloroplast DNA *rpl*2 intron in dicotyledons: molecular and phylogenetic implications. Evolution 45: 1245-1259.

Duret, L., and P. Bucher. 1997. Searching for regulatory elements in human noncoding sequences. Curr. Opin. Struct. Biol. 7: 399-406.

Endrizzi, J. E., E. L. Turcotte, and R. J. Kohel. 1985. Genetics, cytology, and evolution of *Gossypium*. Adv. Genet. 23: 271-375.

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

Fulton, T. M., R. Van der Hoeven, N. T. Eanetta, and S. D. Tanksley. 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14: 1457-1467.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41: 95–98.

Han, Z. G., W. Z. Guo, X. L. Song, and T. Z. Zhang. 2004. Genetic mapping of EST-derived microsatellites from the diploid *Gossypium arboreum* in allotetraploid cotton. Mol. Genet. Genomics. 272: 308–27.

He, M., and D. S. Haymer. 1997. Polymorphic intron sequences detected within and between populations of the oriental fruit fly (Diptera: Tephritidae). Ann. Entomol. Soc. Am. 90: 825–831.

Hebsgaard, S. M., P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouzé, and S. Brunak. 1996. Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Res. 24:3439–3452.

Holland, J. B., S. J. Helland, N. Sharopova, and D. C. Rhyne. 2001. Polymorphism of PCR-based markers targeting exons, introns, promoter regions, and SSRs in maize and introns and repeat sequences in oat1. Genome 44: 1065–1076.

Hughes, A. L., and M. Yeager. 1997. Comparative evolutionary rates of introns and exons in murine rodents. J. Mol. Evol. 45: 125–130.

Johnson, L. A., and D. E. Soltis. 1994. MatK DNA sequences and phylogenetic reconstruction in *Saxifragaceae*. Syst. Bot. 19: 143–156.

Lewin, B. 1997. Genes VI. Oxford University Press, Oxford.

Li, Y., A. B. Koral, T. Fahima, and E. Nevo. 2004. Microsatellites within genes: structure, function and evolution. Mol. Biol. Evol. 21: 991-1007.

Liu, Q., C.L. Brubaker, A.G. Green, D.R. Marshall, P.J. Sharp, S.P Singh. 2001. Evolution of the FAD2-1 fatty acid desaturase 5 UTR intron and the molecular systematics of *Gossypium* (Malvaceae). Am. J. Bot. 88:92–102

Loraine, A. E., and G. A. Helt. 2002. Visualizing the genome: techniques for presenting human genome data and annotations. BMC, Bioinformatics 3: 19.

McCombie, W. R., M. D. Adams, J. M. Kelley, M. G. FitzGerald, T. R. Utterback, M. Khan, M. Dubnick, A. R. Kerlavage, J. C. Venter and C. Fields. 1992. *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. Nature Genet. 1: 124-131.

Newman, T., F. J. de Bruijn, P. Green, K. Keegstra, H. Kende, L. McIntosh, J. Ohlrogge, N. Raikhel, S. Somerville, and M. Thomashow. 1994. Genes galore: a summary of the methods for accessing the results of large scale partial sequencing of anonymous *Arabidopsis thaliana* cDNA clones. Plant Physiol. 106: 1241-1255.

Paterson, A. H., C. Brubaker and J. F. Wendel. 1993. A rapid method of cotton (*Gossypium* spp) genomic DNA suitable for RFLP and PCR analysis. Plant Mol. Biol. Rep. 11: 122-127.

Qureshi, S. N., S. Saha, R. V. Kantety, and J. N. Jenkins. 2004. EST-SSR: A new class of genetic markers in cotton [Online]. J. Cotton Sci. 8: 112–123. Available at http://www.cotton.org/journal/2004-08/2/112.cfm

Reddy, O. K., A. E. Pepper, I. Abdurakhmonov, S. Saha, J. N. Jenkins, T. Brooks, Y. Bolek, and K. M. El-Zik. 2001. New dinucleotide and trinucleotide microsatellite marker resources for cotton genome research [Online]. J. Cotton Sci. 5: 103-113. Available at http://www.cotton.org/journal/2001-05/2/103.cfm

Rong, J., C. Abbey, J.E. Bowers, C.L. Brubaker, C. Chang, P.W. Chee, T.A. Delmonte, X. Ding, J.J. Garza and B.S. Marler, 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.

Roy, S. W., and W. Gilbert. 2005. Rates of intron loss and gain: Implications for early eukaryotic evolution. Proc. Natl. Acad. Sci.102: 5773-5778.

Small, R. L., and J. F. Wendel. 2000. Copy number liability and evolutionary dynamics of the *Adh* gene family in diploid and tetraploid cotton (*Gossypium*). Genetics 155: 1913–1926.

The *Arabidopsis* Genome Initiative. 2001. Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. Nature 408: 196–815.

Tomkins, J. P., D. G. Peterson, T. J. Yang, D. Main, T. A. Wilkins, A. H. Paterson, and R. A. Wing. 2001. Development of genomic resources for cotton (*Gossypium hirsutum* L.): BAC library construction, preliminary STC analysis, and identification of clones associated with fiber development. Mol. Breeding 8: 255-261.

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

Williamson, B. 1977. DNA insertions and gene structure. Nature 270: 295-297.

Zhu, W., S. D. Schlueter, and V. Brendel. 2003. Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. Plant Physiol. 132:469-484.