2005 Beltwide Cotton Conferences, New Orleans, Louisiana - January 4 - 7, 2005

PCR MARKERS BASED ON GENE INTRONS

Pawan Kumar and Peng W. Chee
University of Georgia
Tifton, GA

Abstract

Introns are the most variable part of a gene as these accumulate more nucleotide variations. Introns contain more nucleotide polymorphism than exons, therefore are better candidate for marker development. Cotton EST's in GenBank offer the opportunity to identify a large number of genes which contain intron sequences.

We studied 30 cotton introns amplified by 20 EST derived primers. We found that cotton introns are three times more polymorphic than exons and in addition to single nucleotide polymorphisms (SNP's), introns are also a good source of short repeating sequences or SSR.

Introduction

Eukaryotic genes consist of protein coding sequences called exons and non-protein coding sequences called introns. Introns, as like other DNA sequences that do not code for protein products, are potentially more prone to accumulate nucleotide variation due to mutations because the fitness consequences of such mutation are expected to be smaller. Because intron sequences evolve more rapidly than exon sequences in both plants (Small and Wendel 2000) and mammals (Hughes and Yeager 1997), they have been successfully used in population genetics surveys, including "genetic time" estimates in phylogenetic studies (He and Haymer 1997; Johnson and Soltis 1994).

Introns can be identified by comparing transcribed sequences with the genomic sequences. Regions in the genomic sequence that match with the transcribed sequence delimit exons while the sequences present in genomic region but absent in the transcribed counterpart that exceed some re-defined length threshold are delimited as introns (Loraine and Helt, 2002). In the past, cDNA clones used in RFLP analysis has been an good source of transcribed sequences. However, only a small number of cDNA clones have thus far been sequenced in cotton. More recently, a large number of transcribed sequences in the form of Expressed Sequence Tags (EST's) are available in GenBank for most crop species, including cotton. The cotton EST's contains more than 117,000 sequences, thus offering the opportunity to identify a large number of genes which contain intron sequences.

Materials And Methods

G. arboreum and *G. raimondii* represents A & D genomes of cotton respectively and collectively they represent the genome of tetraploid cotton. DNA was extracted from *G. arboreum* (accession No. A2-47) and *G. raimondii* (accession No. D5-4). A total of 170 cotton EST's were downloaded from GenBank. PCR primers were designed for EST's that showed high homology with previously described genes. Genomic sequence for each EST is obtained by sequencing the PCR product amplified by the EST derived primers. PCR fragments were sequenced in both directions. Sequencing results from 20 primers showed the presence of complete intron (Table 1).

Table 1. Details of EST primers.

Primer	GB acc	Ortholog gene	Primer sequences
EST118	BG447438	Copper-binding protein CUTA	F- TTTCCCAGTCTCTGCTCAATT R- GCAGGTCATATACTACATCCGA
EST132	BG447042	Adenosylmethionine decarboxylase	F- CCCTCCACCCTAGATTCTCATA R- GGACTGCAAACTTTATCCGA
EST141	BG447356	Omega-3 fatty acid desaturase	F- TGGATACGAGCAGAAGCTT R- ACAGGAACTTCAACCTT

EST144	BG447092	Glyceraldehyde 3-phosphate dehydrogenase	F- CTCTCCTACTTCCCTCGTTATC R- CGTGCTTGTAGTCTTTCTCGT
EST152	BG447314	Copalyl diphosphate synthase 1	F- GCCTAAGCAGATTCATGAGA R- GTTGCAGAATGACCAGAAA
EST167	BG447274	Peptide methionine sulfoxide reductase	F- CGAAGCCTTGCGCTTACT R- CGGGTGAAATTCCTCTGC
EST186	BG447211	ZAP1	F- AGCCACTGAACCTTCCACTA R- GCAAACAGGACCGTTAAAAT
EST206	AA659985	Putative nucleotide-sugar dehydratas	F- ATAAAGACGAATGTGATCGG R- CATCAAAGTTTCAGCCACTC
EST207	AI726709	Putative alpha-L-arabinofuranosidase	F- TGGGAAGAAGAATTATCGAG 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 capping protein	F- CCCCTGATGATAGTGCAA R- CTGTACCCAAAATTTCCACA
EST245	AI055464	Feebly-like protein	F- ATGGATTCCTATTTCAGGTG R- CGTCTTTTCAGTCATTCTGTC
EST268	AI725935	Nodulin-like protein	F- ATGCCCCCTTCTTATTGT R- CAAGAGAAAGGTAAAGACTGGA
EST280	AI728393	Putative NADH-ubiquinone oxireductase	F- TCATGAAACCTGCTGTAATG R- ACATGCTTTCAAGAACCG
EST319	AI730677	Plastid protein	F- CACCTTACTCTCTCCTTTATCC R- ACAATTCAGCACCATAGTCC
EST322	AI726457	Putative pollen-specific protei	F- GAGTTCACAACGTGGGAATA R- CTGGTGGTTTGTTTCTCAA
EST387	AI728296	Nonclathrin coat protein gamma-like protein	F- GGAGTTTGCTGAAGTAGCTT R- CTTGACATTGCCGATGTATA
EST394	AI728954	Germin-like protein	F- TGTTAAATCTCCATGGCTG R- GAACACGAAAATGTCTCCTT

Putative functions to these EST's were assigned by Basic Local Alignment Search Tool (BLAST) search (Altschul et al 1997) against all organisms database. All the parameters were as in default setting. The conditions for BLAST were moderate stringent. The expect (E) value which is the statistical significance threshold for reporting matches against database sequences, was 10. EST sequences were aligned with cotton genomic DNA sequences by ClaustalW multiple alignment accessory application of software BioEdit v 6.0.7. (Hall et.al. 1999).

RESULTS AND DISCUSSION

Cotton genome alignment

In total 20 EST-genomic DNA alignments were studied which recognized 30 introns in cotton. We deduced that the average length of A genome intron is 151 nucleotides and that of D genome 152 nucleotides. Exonic sequences are expected to harbor lower levels polymorphism than the Intronic regions as the sequences that have large and direct effects on phenotypes are likely to maintain the least amount of variation under selection. Cotton introns had 11% nucleotide variation as compared to only 4% in flanking exonic regions. Intron length variation between A and D

genomes of cotton was not significant, 20 % of the times A genome introns were larger than D genome intron and 37% of the times D genome introns were longer than A genome. But majority of times (43%) the intron length from both genomes was equal.

Intron structure in Cotton

Monomeric Thymine repeats are the major source of polymorphism in cotton introns. Cotton introns contained less Cytosine and Guanine than exons and much more Thymine, while the Adenine content did not change significantly. The average GC content of cotton introns and flanking exons was 34% and 44%, respectively. Introns contain splice sites that direct their correct removal from the initial transcripts when processed into mature RNA's. Most eukaryotes, the splice sites are recognized by a highly conserved GT motif at the 3' end and AG motif at the 5' end, of the transcripts. Our results show that a vast majority of cotton introns contained the canonical GT-AG splice site junctions, thus following same splicing pathway (U2-type spliceosome), although other varieties also existed (Table 2).

Table 2. Details of introns amplified by EST primers

Primer	Introns	A2	D5	Splice Site
EST118	1	106	106	GT-AG
	2	156	156	GT-AG
	3	102	102	GT-AG
EST132	1	96	96	GT-AG
EST141	1	92	92	GT-AG
EST144	1	90	87	GT-AG
	2	80	80	GT-AG
EST152	1	608	609	GT-AG
EST167	1	80	80	GT-AG
EST186	1	73	73	GT-AG
	2	219	218	GT-AG
EST206	1	124	121	GT-AG
	2	80	89	GT-AG
EST207	1	139	140	AA-TA
EST208	1	77	77	GT-AG
EST210	1	82	85	GT-GC
EST226	1	94	93	GT-AG
	2	292	301	GT-AG
EST237	1	77	77	GT-AG
EST245	1	265	267	GT-AG
	2	138	140	GT-AG
EST268	1	268	268	GT-AG

EST280	1	101	102	GT-AG
	2	72	74	GT-AG
	3	319	315	GA-GG
	4	101	101	GT-AG
EST319	1	307	309	GT-AG
EST322	1	146	144	GT-AG
EST387	1	206	206	GT-AG
EST394	1	93	137	GT-AG

Intron based marker

PCR based markers such as RAPD's, SSR and AFLP's, target mostly non-genic regions of the cotton genome. Developing specific markers that target gene sequences will help us between understand the location and organization of gene rich regions of the cotton genome, and aid in candidate gene approach for dissecting complex traits (Chee et al. 2004). Cotton introns contain more nucleotide polymorphism than exons (Fig 1), therefore are better candidate for markers development.

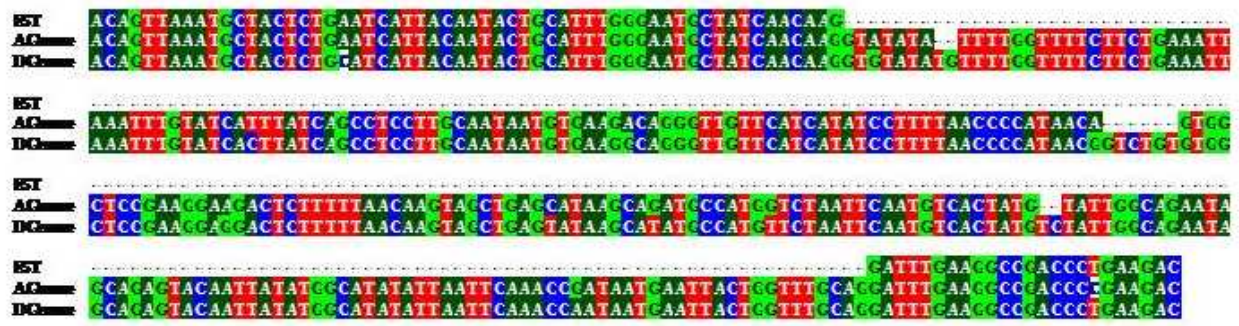

Fig. 1. Intron (demarcated as gaps in EST sequence) amplified by primer EST226 showing higher levels of single nucleotide polymorphism.

In addition to single nucleotide polymorphisms (SNP's), introns are also a good source of short repeating sequences or SSR. These hypervariable SSR's have been extensively used for genome mapping in many crop species, including cotton. For example, SSR markers have been developed for cotton from genomic sequences (Saha et al 2003.) and also more recently, from EST's (Qureshi et al 2004). However, recent studies have shown that introns harbor twice the amount of SSR's than in exons (Cardle et al 2000). Further, SSR's from introns are potentially more variable because unlike exons, intron sequences are selectively neutral. An example of a SSR from intron sequences is shown in Figure 2.

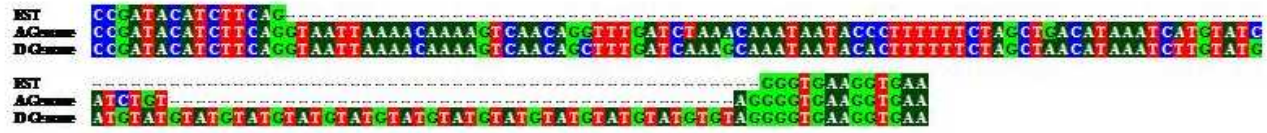

Fig. 2. Genome specific SSR in intron amplified by primer EST394.

The Primer EST394, which amplified an intron from a gene that codes for a germin-like protein, contained a SSR with a (GTAT)n core motif. This motif repeated 16 times in D genome and 5 times in A genome, creating an intron length variation of 44 nucleotide between the two genome.

REFERENCES

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

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D and Waugh R. (2000) Computational and Experimental Characterization of Physically Clustered Simple Sequence Repeats in Plants. *Genetics* 156: 847–854

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

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

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

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

Johnson L A and Soltis D E (1994) matK DNA sequences and phylogenetic reconstruction in Saxifragaceae s.str. *Syst. Bot*. 19: 143–156.

Loraine A E and Gregg A H (2002) Visualizing the genome: techniques for presenting human genome data and annotations. *BMC Bioinformatics* 2002, **3**:19

Qureshi S N, Saha S, Kantety R V and Jenkins J N (2004) EST-SSR: A New Class of Genetic Markers in Cotton. *Journal of Cotton Science 8:112–123*

Saha S, Karaca M, Jenkins J N, Zipf A E, Reddy U K and R V Kantety (2003) Simple sequence repeats as useful resources to study transcribed genes of cotton. *Euphytica **130:** 355–364, 2003.*

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