The *Journal of Cotton Science 15:22–32 (2011)*
http://journal.cotton.org, © The Cotton Foundation 2011

BREEDING AND GENETICS

Development and Mapping of Gene-Tagged SNP Markers For Gland Morphogenesis In Cotton

Jaemin Cho, Cairui Lu, Russell J. Kohel, and John Z. Yu*

ABSTRACT

The genus *Gossypium* has small, pigmented glands filled with a terpenoid aldehyde, gossypol, in many parts of the plant including the seed. Gossypol limits food utilization of cotton seed due to its toxicity to human and non-ruminant animals. Because the genesis of gossypol is closely related to gland morphogenesis, we report the results on the identification of functional markers for gland morphogenesis in cotton. A total of eighty-five unique, gland development-related protein (GDRP) mRNA sequences were collected from NCBI GenBank. Intron-flanking regions within GDRP genes were amplified using 156 primers designed from exon/exon junction sites inferred from *Arabidopsis* homologous sequences. Intron size polymorphism was not detected in this study between *G. hirsutum* cv. TM-1 and *G. barbadense* cv. 3-79. Therefore, 30 purposely selected GDRP PCR products were subsequently sequenced to identify single nucleotide polymorphisms (SNPs). A total of forty SNPs including six indels were identified in 16,328 nucleotides of GDRP coding and intron sequences. We genotyped all SNP markers against 188 recombinant inbred lines (RILs) previously developed from a cross between TM-1 and 3-79 and located 36 SNP markers (24 loci) to the cotton genome. The 24 loci mapped in this study were distributed among the 15 chromosomes of allotetraploid cotton. Localization of functional genes in the cotton genome would facilitate the discovery of candidate genes associated with gland/gossypol traits and would increase the utility of previously developed genetic markers for cotton seed improvement.

J. Cho, R.J. Kohel, and J.Z. Yu*, USDA/ARS Southern Plains Agricultural Research Center, 2881 F&B Road, College Station, TX 77845; C. Lu, Cotton Research Institute, CAAS Anyang, Henan 455004, China
*Corresponding author: John.Yu@ars.usda.gov

Although cotton (*Gossypium* spp.) is a well-known natural fiber crop, the seed is the second most important source of plant protein after soybean, and the importance of its oil ranked just below soybean, palm-tree, colza and sunflower (Li et al., 2009). Cotton seed oil comprises 18-20 % of the seed by weight and cotton seed meal, the remnant after oil extraction, contains 22-24 % high nutrient-value protein which constitutes nearly half of the seed's weight, and is a rich source of protein supplement to livestock (Risco and Chase, 1997). However, utility of cotton seed as a resource of food protein and oil is hampered by the presence of a terpenoid aldehyde, gossypol, which is toxic to humans and non-ruminant animals (Stipanovic et al., 2005). While gossypol limits food utilization of cotton seed, it is still beneficial to the agricultural and pharmaceutical industries; gossypol induces natural resistance to insects and pathogens (Howell et al., 2000; Stipanovic et al., 1977), and gossypol also has several biological activities, such as anti-cancer, anti-microbial, anti-HIV, anti-oxidant and non-hormonal male contraceptive (Sun et al., 2009). Due to gossypol's benefits and disadvantages, selectively reducing the level of gossypol in cotton seed has been considered by cotton research scientists (Altman et al., 1987; Sunilkumar et al., 2006; Townsend and Llewellyn, 2007).

Gossypol is concentrated within glands which are small pigment-bearing tissues located in sub-epidermal tissues in many parts of the plant including the seed (Boatner and Hall, 1946; Lee, 1962). The pattern of glandular forms in gossypol content variation has well established the correlation between the number of glands and the amount of gossypol or other terpenoid aldehyde contents (Benbouza et al., 2009). The glandular patterns in cotton species (*Gossypium* spp.) are determined by the combination of at least six independent loci which are gl_1, gl_2, gl_3, gl_4, gl_5, and gl_6 (Pauly, 1979). In allotetraploid Upland cotton (*G. hirsutum*), combinations of two major alleles Gl_2 and Gl_3 establish the formation of fully developed gossypol glands and the Gl_2 allele mainly controls seed glanding (Lee, 1965; McCarty et al.,

1996). The glandless plant is completely devoid of pigment glands and the phenotype is controlled by two recessive mutant alleles, gl_2 and gl_3. The glandless plant produces nontoxic, glandless cotton seed when both alleles are present in a recessive homozygous state (McCarty et al., 1996; McMichael, 1960). Furthermore, an incomplete dominant allele, Gl_2^e, at the Gl_2 locus was discovered in Egyptian cotton, *G. barbadense* (Kohel and Lee, 1984). Many reports have shown that gland morphogenesis is closely related to the genesis of gossypol and other terpenoid aldehyde contents (Benedict et al., 2004; Lee, 1973, 1977, 1978; Lee et al., 1968; Zhu et al., 2005).

Elucidation on the development mechanism of pigment glands at the molecular level has been recently initiated with the construction of a normalized cDNA library during gland formation (Xie et al., 2007). Gland development-related protein (GDRP) sequences have been released to NCBI's GenBank, awaiting further characterization. It is highly desirable to present genomic distribution of GDRP sequences to assist in further structural and functional characterization of the genes in gland morphogenesis. Molecular markers such as restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs) and simple sequence repeats (SSRs) have been used in cotton genome mapping and genetic diversity analysis (Chen, 2007). Their relatively low levels of polymorphism and limited association with candidate genes have hampered mapping studies of important candidate genes. Single nucleotide polymorphism (SNP), a single nucleotide difference including indels (insertion or deletion) at specific nucleotide positions between members of a species, has been shown to be the most abundant class of DNA polymorphism (Van Deynze et al., 2009). SNP marker development from candidate genes could provide valuable information regarding the complete sequence, allelic variation as well as mapping, evolution and effects of complex quantitative trait loci (QTLs).

In this study, GDRP gene markers were developed and located in the cotton genome. Gene markers are more intriguing than predominantly used molecular markers such as SSRs in that gene markers are directly indicative of genes of interest and further linkage analysis is not needed as in other types of markers. If the limit of low polymorphism rates of gene markers can be overcome when developed from less variable coding regions than non-coding regions, gene markers are highly preferable for mapping. Here, we report the results on identification of SNP markers that were developed for twenty GDRP genes and subsequently validated in a mapping population of 188 recombinant inbred lines (RILs) previously developed from an interspecific cross between *G. hirsutum* and *G. barbadense*.

MATERIALS AND METHODS

GDRP sequences used in this study. A total of eighty-eight GDRP mRNA sequences were collected from NCBI GenBank entries. Identical sequences were assembled using CAP3 sequence assembly program at http://pbil.univ-lyon1.fr/cap3.php (Huang and Madan, 1999). GDRP mRNA sequences were developed through normalized cDNA library construction of a cotton mutant (characterized by delayed pigment gland morphogenesis) during the gland development stage (Xie et al., 2007).

Primer design and touchdown PCR amplification. Primers were designed to amplify intron-flanking GDRP genomic regions by following the method of Wei et al. (2005) used to develop EST-PCR markers in *Rhododendron*. The forward and reverse PCR primers were developed from the 5' upstream and 3' downstream sequences of one or several exon/exon junction sites. The primer was therefore used to amplify a PCR product containing expected intron regions. All primers were designed using Primer3 (Rozen and Skaletsky, 2000) (http://frodo.wi.mit.edu/primer3/input.htm) with the default settings; listed in Supplementary Table 1.

PCR amplification of cotton genomic DNA from *G. hirsutum* cv. TM-1 (PI 607172) or *G. barbadense* cv. 3-79 (GB 1585) was carried out by "touchdown PCR" (Don et al., 1991) with slight modification as follows: PCR reaction mixtures were incubated at 94°C for 2 min, followed by 10 cycles of denaturation at 94°C for 30 sec, annealing for 30 sec at selected temperatures (see below) and elongation at 72°C for 1 min and another 30 cycles of 94°C for 30 sec, 61°C for 30 sec and 72°C for 1 min, followed by a final 10 min extension at 72°C. The annealing temperature was decreased by 0.8°C per cycle during the first 10 cycles from 69 to 61°C. PCR products were separated by 1.5 % agarose gel electrophoresis; band sizes were determined against HyperLadder™ I (separation range, 200 ~ 10,000 nucleotides) and V (25 ~ 500 nucleotides) (Bioline USA Inc., Taunton, MA, USA).

Table 1. The list of cotton GDRP mRNA sequences with *A. thaliana* homologs mapped in the cotton genome through the development of SNP markers.

Acc. Number	*At* homolog ID	Definition	E value
EU219610	NM_128977.3	transducin family protein / WD-40 repeat family protein (AT2G34260)	5.00E-113
EU372996	NM_121321.3	ANAC083 (*Arabidopsis* NAC domain containing protein 83); transcription factor (ANAC083)	8.00E-65
EU373010	NM_101247.2	XH/XS domain-containing protein / XS zinc finger domain-containing protein (AT1G13790)	7.00E-29
EU373011	NM_113304.4	HSP60 (Heat shock protein 60); ATP binding / protein binding / unfolded protein binding (HSP60)	0
EU373018	NM_124404.3	RNA-binding protein RNP-T, putative / RNA-binding protein 1/2/3, putative (AT5G50250)	7.00E-105
EU373021	NM_123090.3	unknown protein (AT5G37310)	9.00E-91
EU373025	NM_101551.3	sugar binding / transferase, transferring glycosyl groups (AT1G16900)	0
EU373029	NM_129335.4	aldo/keto reductase family protein (AT2G37790)	1.00E-177
EU373036	NM_124149.3	protein kinase, putative (AT5G47750)	0
EU373042	NM_101609.3	ATDRG1 (*ARABIDOPSIS THALIANA* DEVELOPMENTALLY REGULATED G-PROTEIN 1); GTP binding (ATDRG1)	0
(EU373049, EU373058)	NM_202808.2	YT521-B-like family protein (AT4G11970)	7.00E-100
EU373051	NM_100591.3	unknown protein (AT1G07170)	4.00E-85
EU373052	NM_118602.3	ubiquitin-associated (UBA)/TS-N domain-containing protein (AT4G24690)	4.00E-65
EU373053	NM_179491.2	CAT2 (CATIONIC AMINO ACID TRANSPORTER 2); amino acid transmembrane transporter (CAT2)	5.00E-113
EU373055	NM_104404.3	oxidoreductase, 2OG-Fe(II) oxygenase family protein (AT1G55290)	2.00E-105
EU373056	NM_100188.1	WD-40 repeat family protein / beige-related (AT1G03060)	9.00E-174
EU373059	AY088609.1	clone 8342 mRNA	0.00E+00
EU373060	NM_202841.2	unknown protein (AT4G19003)	7.00E-128
EU373064	NM_102064.4	unknown protein (AT1G22140)	4.00E-21
EU373074	DQ653295.1	clone 0000016742_0000011667 unknown mRNA	1.00E-06

Sequence analysis of GDRP genomic PCR products for SNP detection. SNP detection was performed by the comparison of GDRP genomic DNA sequences amplified from the templates of two different genotypes, TM-1 and 3-79. The selected PCR products were purified using QIAquick spin columns (QIAGEN USA Inc., Valencia, CA, USA) and were cloned into TOPO TA vector (Life Technologies Corporation, Carlsbad, CA, USA). Up to twenty individual clones from each recombinant construct were selected for sequencing on ABI 3130xl Genetic Analyzer (50 cm 16-capillary array, Applied Biosystems Inc, Foster City, CA, USA) to check the insert sequences. The sequence alignment was conducted using ClustalW2 at http://www.ebi.ac.uk/Tools/clustalw2/_(Larkin et al., 2007). SNP identification was performed by manual sequence-alignment comparison of at least ten individual sequences from both TM-1 and 3-79 GDRP DNA. The replications allowed identification of a mixture of slightly different duplicate sequences from a single band of GDRP PCR product.

SNaPshot analysis and mapping. SNaPshot primers were designed from upstream SNP sites in different regions, so that SNP markers could be detected by single base extension. ABI Prism® SNaPshot™ multiplex kit (Applied Biosystems Inc, Foster City, CA, USA) was used for SNP genotyping with the following modifications: PCR products from the genomic DNA of 188 TM-1 X 3-79 RILs were PCR-purified and treated with shrimp alkaline phosphatase (SAP) and *Exo* I (2 unit of SAP and 2 unit of *Exo* I, respectively) at 37°C for 1 hour to ensure a low background. Enzymes were then inactivated at 75°C for 15 min and the treated PCR product (up to 0.40 pmol) was used as template in a SNaPshot

reaction. Single base extension was performed in 10 µl reaction mixtures containing 5 µl of SNaPshot Multiplex Ready Reaction Mix, 0.4 µl of purified PCR product, 1 µl of SNaPshot primer (2 µM), and 3.6 µl of deionized water. The thermal cycling parameters included 25 cycles of 96°C for 10 sec, 50°C for 5 sec, and 60°C for 30 sec. After post-extension treatment with 1 unit of SAP at 37°C for 1 hour and 75°C for 15 min, the SNaPshot reaction product was denatured with HiDi™ Formamide (Applied Biosystems Inc, Foster City, CA, USA) at 95°C for 5 min and run using an ABI 3130xl Genetic Analyzer loaded with 36-cm 16-capillary array. The SNP was analyzed using GeneMapper® software v3.7 (Applied Biosystems Inc, Foster City, CA, USA).

Mapping of GDRP SNPs was performed with a population of 188 TM-1/3-79 RILs by employing regression mapping based on the Kosambi mapping function of JoinMap 4.0 (Van Ooijen, 2006). Strong linkages with previously mapped molecular markers were then investigated.

RESULTS AND DISCUSSION

Test for intron size polymorphism. Eighty-eight GDRP mRNA sequences representing differentially expressed genes during gland formation were collected from NCBI GenBank entries. Three pairs of two identical GDRP sequences (EU373000 + EU373066, EU373009 + EU373027, EU373049 + EU373058) were assembled to give a final total of eighty-five unique GDRP sequences. To test the polymorphism size of GDRP genes between TM-1 and 3-79, intron-flanking PCR primers were designed to target more variable intron regions within the GDRP gene as described in Wei's report (2005). All cotton GDRP sequences were blasted against *Arabidopsis thaliana* protein sequences using NCBI BLASTX with default parameters' value (expect cutoff = 0.01), and the result of the mapped GDRP genes is listed in Table 1. The best hit *A. thaliana* genomic sequence (found in its protein sequence) was aligned to the corresponding cotton GDRP mRNA sequences to infer exon/exon (E/E) junction sites that helped amplify intron regions of GDRP genes. In this way we found 279 expected E/E junction sites from 85 GDRP sequences (112,339 nucleotides in total length), averaging one site per 403 nucleotides. Sixteen GDRP sequences were not hit by any *A. thaliana* homologs that fit the blast criterion. Eleven GDRP sequences were not found with any expected E/E sites and other eleven GDRP sequences were found with only one E/E site. For those thirty-eight GDRP sequences without or with only one E/E site, one primer pair was designed from each GDRP sequences to amplify a product up to 699 nucleotides of mRNA sequences. For the other forty-seven GDRP sequences containing two or more E/E sites, multiple primer pairs were designed; primers that include one E/E site or primers that contain more than two E/E sites within the length of a product. A total of 156 primer pairs were designed to cover 54,255 nucleotides of cotton GDRP mRNA sequences including 279 E/E sites (Table 2).

Among 156 primer pairs, 134 pairs were successfully amplified from both TM-1 and 3-79 templates and 22 (14.1 %) were not amplified. Based on the size analysis from the 1.5 % agarose gel, 86.8 % of the amplified PCR products were found to contain introns in the corresponding sequences. Introns were contained in 95.6 % of PCR products amplified from primers designed on at least one E/E site. Products generated with 49 primer pairs were less than 800 nucleotides in size, permitting detection of size polymorphisms via agarose gel electrophoresis; these products did not show any detectable polymorphisms between TM-1 and 3-79 (Table 2).

The size of intron polymorphism between TM-1 and 3-79 was not ascertained in this study. This was expected as in Wendel's observation (2002) that 76 introns investigated from 5 genomes (3 diploid and 2 polyploid) of cotton showed no case of scoring intron size differences. Sequence analyses of GDRP genomic PCR products clearly revealed that all introns from TM-1 and 3-79 are the same size with only single nucleotide changes (6 indels observed) and the nucleotide sequences at exon/intron junctions are consistent with 'gt-ag' rule. Sizes of 45 introns ranged from 72 to 392 nucleotides with a mean length of 132 and a median length of 101 nucleotides (Supplementary Table 2). The observed lower range of cotton intron size (72 nucleotides) is similar to the reported minimal size (71 nucleotides) for the introns of *AdhC* gene (Wendel et al., 2002). Similarly, it nearly matched the minimum intron size (~70 nucleotides) as Deutsch and Long (1999) suggested for correct splicing of introns. All intron sequences surveyed in this study started with 'gt' nucleotides at the 5' end and finished with 'ag' at the 3' end, matching with the 'gt-ag' intron splice-site found in plant and animal genes (Simpson et al., 1993; Thanaraj, 1999).

Table 2. PCR amplification of genomic products using GDRP mRNA–based primers.

Description	Total	Subtotal
No. of GDRP mRNA sequences collected	88	
No. of GDRP mRNA sequences after assembly	85	
No. of designed primers to amplify GDRP genomic PCR products	156	
based on Exon/Exon (E/E) junction sites		129
based on without E/E sites		27
No. of primers that produced genomic GDRP PCR products	134	
in size of <800bp		49
in size of >800bp		85
no. of primers that did not amplify a product	22	
% introns amplified	86.8	
by primers based on E/E junction sites		95.6
by primers based on without E/E sites		40.9
% polymorphism found between TM-1 and 3-79 based on the gel picture, in size of <800bp	0	

Identification of SNP loci. Since intron size polymorphism was not found in the cotton GDRP PCR products, we compared the sequences of those products to identify SNP loci. Forty-nine GDRP sequences were selected for the SNP survey based on the size (<800 nucleotides) suitable for TOPO cloning and one-time sequencing from both 5' and 3' directions (Table 2). For 30 out of 49 GDRP sequences, at least ten clones from each individual sequence were picked up to find SNPs between TM-1 and 3-79 GDRP PCR products. Table 3 shows the result of SNP screening with these 30 GDRP sequences. A total of forty SNPs including six indels were identified among the 16,328 nucleotides of GDRP coding and intron sequences. The rates of SNPs were 1:612 in coding regions and 1:258 in intron regions. Clones of three GDRP sequences (EU373025, EU373029 and EU373053) were found to have one PCR product containing two different sequences in the same size and to have SNPs found from both sequences. Two groups of sequences could be easily identified through the phylogram derived from alignment of sequences when more than ten individual sequences from one PCR product were aligned (Figure 1a and 1b). Considering the two subsets of partially homoeologous chromosomes in allotetraploid cotton, existence of two slightly different DNA sequences may indicate amplification of homoeologous counterparts in the genome. Although the homoeologous relationship did not appear in all of the mixed GDRP sequences (EU373025 located on chromosomes 3 and 4), it was observed in chromosomal locations of EU373029 and EU373053 sequences (Table 4).

Table 3. SNP discovery from GDRP sequences between TM-1 and 3-79.

Description	Total	Subtotal
No. of unique GDRP screened	30	
No. of GDRP PCR products screened[z]	33	
No. of bases screened	16,328	
no. of bases in coding region		10,399
no. of bases in intron region		5,929
No. of SNPs found	40	
within coding region		17
within intron		23
SNP frequency	1/408.2	
within coding region		1/611.7
within intron		1/257.8

[z] Two different regions within the gene were amplified for three GDRP genes.

Table 4. The list of SNP markers with closely linked SSR markers through linkage analysis (JoinMap 4.0).

Identifier	Primer sequences for SNaPshot analysis	SNP		Chr	Linked SSR markers
		TM-1	3-79		
EU219610	AGGTAAAGCTTTTTGTTTTGTATTTTTT	-	T	22	TMB1958
EU372996	GGGACCAAAAATTAGATTTTTTTTT	-	T	22	MUSB1050c
EU373010-1_2	GATGAAACTATGTTCAAACTGGCT	A	G	18	MUSB1135e
EU373011_1	TATACAGTTTCAAAAAGCTGGTGATTT	-	T	9	MUSS083
EU373018-2_2	CCCATATTAAGCTATCATCTTCTAAATA	T	C	7	NAU1048
EU373021-1_1	GTTCTGTTCCTCAGGTTTGATTTT	A	T	7	NAU1048
EU373021-1_2	GTGTCGATTTCTATTTTATGATTTTTTTTT	-	T	7	NAU1048
EU373025-2_1	AAGTATGCCCCTGACTTGCT	T	C	4	BNL0530
EU373025-2_2	GAACATTCCAATATATGACTTTTAGAAGTTT	G	C	3	BNL3441
EU373025-2_4	GTGGGCATCTTCTACACCAAT	T	A	4	BNL0530
EU373029-1_1	TGGCGCCGCCGTTTC	T	C	4	MUSB1050a
EU373029-1_2	AGTTCGCTGGTCGTATTTATTTT	G	T	22	JESPR230
EU373029-3_3	GCCTGCAAATGGGTCA	A	T	4	MUSB1050a
EU373036-1	CGAAAATGCCACAGTTTCTG	-	T	25	TMB1448
EU373042-3	GGAGAGGCCGCTTCAAA	A	T	19	TMB0189b
EU373049-2_1	CATGTTTTCTTTTTAGGAATTACCTC	T	G	8	BNL3255
EU373051-1	CCGCCCTTGCACGCT	T	C	9	MUSS547
EU373052-2_1	GGGGGAGACAAGTTAACAAAT	G	T	17	MUSB0964a
EU373052-2_2	TTAATTGCAATCTGCTATATCATTTC	A	T	17	MUSB0964a
EU373052-2_3	ACTAAATGATTTTATTGCTGTGCTT	G	A	17	MUSB0964a
EU373052-2_4	TTGTTGTTGATTTTACTGCACCA	A	G	17	MUSB0964a
EU373053-1_1	TAATTACACTTGAGATCCCTTTCA	G	A	11	TMB0628a
EU373053-3	TTTGGATGCTACTGAAATTGGTCT	C	T	21	NAU1014
EU373055-1_1	GGAAAGATTACCTTAGCTTGTTTTATGT	C	A	24	JESPR157a
EU373055-1_2	GTAAATATATATTTTCATTATAAAAGATCCTAACAT	T	A	24	JESPR157b
EU373055-1_3	CCTAACCATATATTTTCTTAAAAAATCTTAA	A	T	24	JESPR157a
EU373055-1_4	ATTATGTCTTTGATTAATTGTCTAATTTTTT	-	T	24	JESPR157a
EU373055-1_5	CTAAATTTTTTATTATATATAATCATATCAGATAC	T	C	24	JESPR157b
EU373055-1_6	TTGATGGGTTCGATGAGG	A	G	24	JESPR157b
EU373056	TTCATGGTGGCAACCAAATTC	A	T	11	TMB2803
EU373059-1_1	GGGTCTGCCCTTCTTATTCTT	G	C	9	JESPR248b
EU373059-1_2	CAAAAACATATTTTTCTCCTTTTTTTTTT	G	A	9	JESPR248b
EU373060-2	ACATTTCATGATCCAAAACTCTTGTAAAT	A	G	3	MUCS547
EU373064-1_2	AATTTCCAGTTTATGCCCCAAA	A	G	6	TMB1530
EU373074-1	CTTACAATGGCGTAACAGGATC	C	G	5	NAU2296
EU373074-2	GTCCTATTAACCGGCAGTGG	T	A	5	NAU2296
(EU373000, EU373066)-2	GAACATCAAATTCCTTTGTGC	A	C	-	-
EU372995_2	TCAATCCAAGGACTTGCTTAA	A	T	-	-
EU373007-2_1	GAATTATCTAGGACAAGACCTCTAGTCC	A	G	-	-
EU373007-2_2	CCTCTACTACAATGCTAAAAATTGTCTTT	A	G	-	-

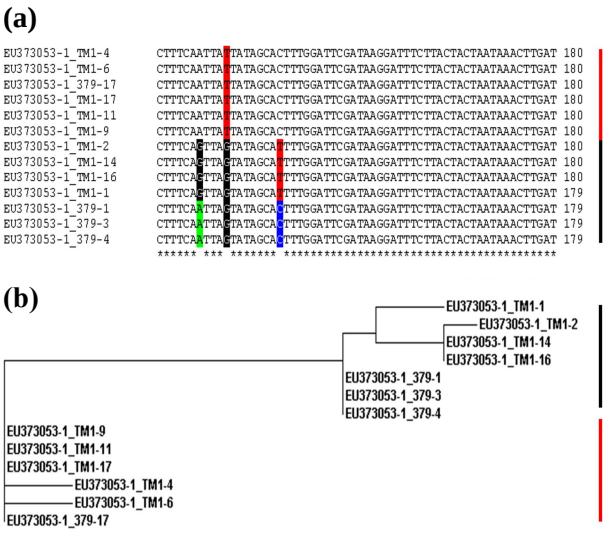

Figure 1. Sequence alignment and phylogram of GDRP SNP markers. SNPs are found in only one group of sequences (indicated by vertical black line bar). (a) Grouping of sequences identified through sequence alignment.. (b) Phylogram grouping of sequences.

Sequencing of GDRP genomic PCR products revealed the mixture of two homoeologous sequences was present as a single PCR band, indicating amplification of those contributed by the A and D genome donors at the time of polyploid formation. Homoeologous pairs could be separated by two distinctive phylogram groupings as we aligned and analyzed the sequences (at least 10 clones compared) of one GDRP PCR product. The separated homoeologs were mapped to pairs of homoeologous chromosomes, suggesting that sequence alignment and phylogram analysis can be used to identify sub-genome specific sequences of single-copy genes in allotetraploid cotton genome.

SNP loci were detected from the partial fragments of GDRP mRNA sequences at a mean frequency of 0.25 % (one SNP per 408 nucleotides) between TM-1 and 3-79. The rate (0.25 %) is slightly lower than 0.35 % of Rong et al (2004) rate of variation per nucleotide of sequence-tagged sites (STSs) between *G. hirsutum* and *G. barbadense* species. Differences in SNP frequencies observed between coding region (0.16 %) and introns (0.39 %) in this study were not as significant as those detected in Hsu's (2008) SNP characterization of GhMyb8 and GhMyb10 genes between coding region (0.55 %) and 3' UTR (5.14 %). However, the lower SNP frequency in the coding region exists with the GDRP sequences.

Mapping of SNP markers. To locate the above 40 SNP markers identified in the cotton genome, SNaPshot primers were developed from the upstream sequences just before SNP site with at least 50°C of melting temperature for the complementary region. A population of 188 individual RILs derived from the TM-1/3-79 cross was used for SNaPshot genotyping and mapping analysis. The chromosome location was determined by adding the genotype scores of 40 SNP markers into a previously developed cotton genetic mapping database (not published). Thirty-six SNP markers were mapped to 24 loci in 15 chromosomes. Four SNPs showed no linkages to any of the previously mapped markers on the 26 cotton chromosomes. Table 4 lists the SNP markers mapped on the corresponding chromosomes and closely linked SSR markers. SNP markers of EU373029 and EU373053 were mapped to the loci of the homoeologous chromosomes (A4 – D4 and A11 – D11), while SNP markers of EU373025 were mapped to the loci of non-homoeologous chromosome pairs (A3 - A4) (Figure 2).

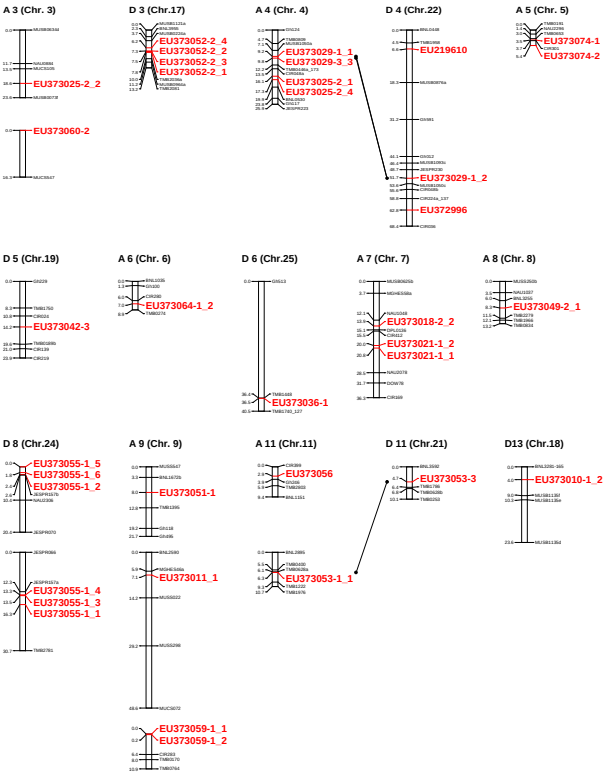

Figure 2. Map locations of GDRP SNP markers. The map locations show that the GDRP SNP markers developed from two mixed sequences in one PCR product are located on two different chromosomes in either a homoeologous (indicated by line) or non-homoeologous relationship.

Supplementary Table 1. Primer sequences developed from exon/exon junction sites of cotton GDRP mRNA sequences.

Identifier	Primer sequences	Length	E/E junction
EU219610_F1	CCTTTGGCATCGACTTTCAT	325	E/E
EU219610_R1	CAGCCTTCATCGTCTCCAGT	325	E/E
EU219610_F2	AATCAACTGTTGCCACTGGA	308	E/E
EU219610_R2	TCCCCACGAATATAACAGCA	308	E/E
EU219610_F3	TTGTGGATCACAAGGTGGAA	378	E/E
EU219610_R3	CGTCCATCCCATCACTATCA	378	E/E
EU372996_F1	TGCCTGCCTCTATCATACCC	178	E/E
EU372996_R1	TGTTTGTCGATTCCAGTTGC	178	E/E
EU373010_F1	AGAGCAGCCTTGGAACAAAA	190	E/E
EU373010_R1	CACCATTCATGTGCTCCATC	190	E/E
EU373010_F2	ATGGGGATGCAGACACAAAG	213	E/E
EU373010_R2	CTCCCATTTTCTTGACACCAA	213	E/E
EU373010_F3	GCTGGCATCCTTTTAAAATCC	213	E/E
EU373010_R3	TTGCCTTCCTTCCTTCATTG	213	E/E
EU373011_F1	TGGAATAGGAACTATGCTGCAA	358	E/E
EU373011_R1	TGTCACCATTCCCATCAAGA	358	E/E
EU373011_F2	TCTTGATGGGAATGGTGACA	386	E/E
EU373011_R2	CAACTCCAGCATTTGAAGCA	386	E/E
EU373011_F3	GGCCAACTTTGACCAGAAGA	339	E/E
EU373011_R3	TAATCCATGCCACCCATACC	339	E/E
EU373018_F1	TTGTTGGGAATTTGCCTTTC	312	E/E
EU373018_R1	TGTCAACATCCCATGGAAGA	312	E/E
EU373018_F2	GTGGTTTCGGTTTCGTGACT	198	E/E
EU373018_R2	TGGGAAAAATGAAAAGGAAAAA	198	E/E
EU373021_F1	GCGTACCTCAGGTCAGATCC	200	E/E
EU373021_R1	CACTAGTCGATCGCCATTCA	200	E/E
EU373025_F1	GCTCTGCTTGGCTAGTGGTT	355	E/E
EU373025_R1	ATGGCTCTCACCACCTCCTA	355	E/E
EU373025_F2	TGCTTTTCCTGGGAATTCTG	328	E/E
EU373025_R2	GGAGCAGCATAGCCATTGAT	328	E/E
EU373025_F3	AATGGCTATGCTGCTCCAAT	390	E/E
EU373025_R3	TGAGAGCTCCCTGTCCAAGT	390	E/E
EU373029_F1	AGCGCTTTCTCTTTGTTCCA	179	E/E
EU373029_R1	ACACTTGAGCGCAGTCGATA	179	E/E
EU373029_F2	GCGCTCAAGTGTACGGAAAT	364	E/E
EU373029_R2	AGCAGATCCCCCAATTTCTT	364	E/E
EU373029_F3	TTGCAAATCCAAGGGAGTTC	301	E/E
EU373029_R3	TGAATTTCCCCTGAGCAATC	301	E/E
EU373036_F1	CGCTTTACACGCATTTTGAA	205	E/E
EU373036_R1	TTTTCGGGTTTAAGGTCACG	205	E/E
EU373042_F1	AAGGTGCTTCTGAAGGCAAA	319	E/E
EU373042_R1	TCATCCACTGTTGCATCCTC	319	E/E
EU373042_F2	TCGAAGGAAACCGCAAATAC	328	E/E
EU373042_R2	GAGCCCCAGACCAGAACATA	328	E/E
EU373042_F3	CCATTCCTTGCAGGATGAAG	167	E/E
EU373042_R3	TTCCTGCCGACATCTTTTGT	167	E/E
(EU373049, EU373058)_F1	TCATCGGTGACTGAATGGAA	341	E/E
(EU373049, EU373058)_R1	TTATCTCGCCTCCATCCAAC	341	E/E
(EU373049, EU373058)_F2	GCTTGAATCACCTGCCTTTC	337	E/E
(EU373049, EU373058)_R2	TGCGGATGTGTTAAATGGAA	337	E/E
EU373051_F1	GTCATGGCTAAGCACCATCC	384	E/E
EU373051_R1	CCTGTGACATAAGCACAGCAA	384	E/E
EU373052_F1	TTGTGGGGTTCTTCCCATTA	350	E/E
EU373052_R1	ACCAGGGCAGAGTACCATTG	350	E/E
EU373052_F2	CAATGGTACTCTGCCCTGGT	348	E/E
EU373052_R2	GCATTCATATCCGCAATGTG	348	E/E
EU373052_F3	TGGGATCCGATACTTGAGGA	163	E/E
EU373052_R3	AAACCAAGAGCCATCATCCA	163	E/E
EU373053_F1	AACCATTGGAGCGGTTACTG	216	E/E
EU373053_R1	ATCCCTGCCAACTGTGAAAC	216	E/E
EU373053_F2	GTGGGAAAAACCCTGAGACA	325	E/E
EU373053_R2	CCGAAATTATGCCTTGCATC	325	E/E
EU373053_F3	TTAACATGTTGCGGAGTTGG	324	E/E
EU373053_R3	GTCACCCGAGGATCGATAAA	324	E/E
EU373055_F1	CTTTGGGTTAGCAGCTGAGG	354	E/E
EU373055_R1	GTGAGTTCGGGGTTAGGACA	354	E/E
EU373056_F	AAAAGGCGGTCTGACAGAAA	668	No E/E
EU373056_R	GATGGCAGCTGGGATGTACT	668	No E/E
EU373059_F1	TATTCAAGTGTGGGGGAAAAA	301	E/E
EU373059_R1	GTCTCCCGTCAGCTGCTAAT	301	E/E
EU373059_F2	ACTCGGTTCATCGTCATTCC	357	E/E
EU373059_R2	TGCAGGTCATTTCTGACAGC	357	E/E
EU373059_F3	GCAGGGAAGGGGTTATTGA	163	E/E
EU373059_R3	CTAGCTGCAACCTTGGCTTC	163	E/E
EU373060_F1	CTGAGAAGGAACGCCGATAG	335	E/E
EU373060_R1	GAATTCGGTGCCAAAGGATA	335	E/E
EU373060_F2	CTTATCCTTTGGCACCGAAT	160	E/E
EU373060_R2	GCACTAGAATCGTCCGGTCT	160	E/E
EU373064_F1	GGGGAGAAAGGAAAAAGGAA	162	E/E
EU373064_R1	ACGCTTTCTAGCTGCCTCAG	162	E/E
EU373074_F	GGTTCGTTTCTTTGGATGGA	636	No E/E
EU373074_R	AACCGAGCTTACCATTGTGG	636	No E/E

Supplementary Table 2. Intron sizes (bp) within GDRP genomic PCR products.

ID	Length of sequenced region	Length w/o intron	No. of Introns	Total intron size	Intron 1	Intron 2	Intron 3	Intron 4
EU219610-1	729	325	4	404	84	147	86	87
EU372995	504	504	0	0	0			
EU372996	267	178	1	89	89			
(EU373000, EU373066) -2	480	312	2	168	96	72		
(EU373000, EU373066)-4	258	169	1	89	89			
EU373001	677	677	0	0	0			
EU373007-2	389	184	2	205	118	87		
(EU373009, EU373027)	476	389	1	87	87			
EU373010-1	274	190	1	84	84			
EU373011-3	677	339	3	338	161	95	82	
EU373013	654	654	0	0	0			
EU373018-2	427	198	1	229	229			
EU373021	364	200	2	164	77	87		
EU373022-2	196	196	0	0	0			
EU373025-2	713	328	2	385	303	82		
EU373029-1	322	179	1	143	143			
EU373029-3	526	301	2	225	88	137		
EU373036	293	205	1	88	88			
EU373040-2	646	324	3	322	118	101	103	
EU373042-2	764	328	2	436	346	90		
EU373042-3	325	167	1	158	158			
EU373047-1	266	168	1	98	98			
(EU373049, EU373058)-2	526	337	2	189	110	79		
EU373051	384	384	0	0	0			
EU373052-2	561	348	2	213	120	93		
EU373053-1	528	216	1	312	312			
EU373053-3	533	324	2	209	101	108		
EU373055	577	354	1	223	223			
EU373056	668	668	0	0	0			
EU373059-1	627	301	3	326	98	107	121	
EU373060-2	571	160	2	411	102	309		
EU373064	496	162	1	334	334			
EU373074	630	630	0	0	0			
Total	16,328	10,399	45	5,929	3,856	1,594	392	87

We have mapped 36 SNP markers that represent 20 GDRP mRNA sequences in the cotton genome. The 24 mapped loci were distributed among the 15 chromosomes in the allotetraploid cotton. There appears to be a gene network in the cotton genome that contributes to gland morphogenesis. While gossypol glanding loci Gl_2 and Gl_3 and the glandless locus Gl_2^e were previously mapped on chromosomes A12 and D12 (Kohel and Lee, 1984; Samora et al., 1994), our study has not located any SNP markers in these two chromosomes. Because of this gap, however, launching GDRP SNP markers on these specific chromosomes would be intriguing to reveal the relationship of these genes with new GDRP mRNA sequences. The genetic maps developed to date have been saturated with mostly SSR markers, which were largely developed from non-coding regions of the genome and generally were not linked to genes of known function or to traits of economic importance. Locating functional genes on the maps may increase the utility of previously developed genetic markers. Localization of functional gene markers of interest is hindered by the lack of size polymorphism as shown in case of the GDRP genes. Considering that SNP frequency is higher than the rate of size polymorphism, SNP markers derived from candidate genes associated with gland morphogenesis would be another strategy for the discovery and eventual cloning of genes for gossypol or gland (or glandless) traits.

ACKNOWLEDGEMENTS

This research was supported by USDA-ARS CRIS Project 6202-21000-030-00D. We thank Darcey Klaahsen for technical assistance.

DISCLAIMER

Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA.

REFERENCES

Altman, D.W., D.M. Stelly, and R.J. Kohel. 1987. Introgression of the glanded-plant and glandless-seed trait from *Gossypium sturtianum* Willis into cultivated upland cotton using ovule culture. Crop Sci. 27:880-884.

Benbouza, H., G. Lognay, J. Scheffler, J. Baudoin, and G. Mergeai. 2009. Expression of the " glanded-plant and glandless-seed " trait of Australian diploid cottons in different genetic backgrounds. Euphytica. 165:211-221.

Benedict, C.R., G.S. Martin, J. Liu, L. Puckhaber, and C.W. Magill. 2004. Terpenoid aldehyde formation and lysigenous gland storage sites in cotton: variant with mature glands but suppressed levels of terpenoid aldehydes. Phytochem. 65:1351-1359.

Boatner, C., and C. Hall. 1946. The pigment glands of cotton seed. I. Behavior of the glands toward organic solvents. Journal of the American Oil Chemists' Society. 23:123-128.

Chen, Z.J. 2007. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev of Plant Biol. 58:377-406.

Deutsch, M., and M. Long. 1999. Intron-exon structures of eukaryotic model organisms. Nucl. Acids Res. 27:3219-3228.

Don, R.H., P.T. Cox, B.J. Wainwright, K. Baker, and J.S. Mattick. 1991. 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucl. Acids Res. 19:4008.

Howell, C.R., L.E. Hanson, R.D. Stipanovic, and L.S. Puckhaber. 2000. Induction of terpenoid synthesis in cotton roots and control of *Rhizoctonia solani* by seed treatment with *Trichoderma virens*. Phytopath 90:248-252.

Hsu, C.-Y., C. An, S. Saha, D.-P. Ma, J. Jenkins, B. Scheffler, and D. Stelly. 2008. Molecular and SNP characterization of two genome specific transcription factor genes *GhMyb8* and *GhMyb10* in cotton species. Euphytica. 159:259-273.

Huang, X., and A. Madan. 1999. CAP3: A DNA Sequence Assembly Program. Genome Res. 9:868-877.

Kohel, R.J., and J.A. Lee. 1984. Genetic analysis of Egyptian glandless cotton. Crop Sci. 24:1119-1121.

Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. ClustalW and ClustalX version 2.0. Bioinformatics. 23:2947-2948.

Lee, J.A. 1962. Genetical studies concerning the distribution of pigment glands in the cotyledons and leaves of upland cotton. Genetics. 47:131-142.

Lee, J.A. 1965. The genomic allocation of the principal foliar-gland loci in *Gossypium hirsutum* and *Gossypium barbadense*. Evol. 19:182-188.

Lee, J.A. 1973. The inheritance of gossypol level in *Gossypium* II: Inheritance of seed gossypol in two strains of cultivated *Gossypium barbadense* L. Genetics. 75:259-264.

Lee, J.A. 1977. Inheritance of gossypol level in *Gossypium*. III. Genetic potentials of two strains of *Gossypium hirsutum* L. differing widely in seed gossypol level. Crop Sci. 17:827-830.

Lee, J.A. 1978. Inheritance of gossypol level in *Gossypium*. IV. Results from the reciprocal exchange of the major gossypol-gland alleles between *G. hirsutum* L. and *G. barbadense* L. Crop Sci. 18:482-484.

Lee, J.A., C.C. Cockerham, and F.H. Smith. 1968. The inheritance of gossypol level in *Gossypium* I. Additive, dominance, epistatic, and maternal effects associated with seed gossypol in two varieties of *Gossypium hirsutum* L. Genetics. 59:285-298.

Li, W., Z. Zhou, Y. Meng, N. Xu, and M. Fok. 2009. Modeling boll maturation period, seed growth, protein, and oil content of cotton (*Gossypium hirsutum* L.) in China. Field Crops Res. 112:131-140.

McCarty, J.C., P.A. Hedin, and R.D. Stipanovic. 1996. Cotton *Gossypium* spp. plant gossypol contents of selected Gl_2 and Gl_3 Alleles. J Agric and Food Chem. 44:613-616.

McMichael, S.C. 1960. Combined effects of glandless genes *gl2* and *gl3* on pigment glands in the cotton plant. Agron J. 52:385-386.

Pauly, G. 1979. Les glandes à pigments du cotonnier: aspects gènètique et sèlection des variètès glandless et high gossypol. Cotton Fibres Trop. 34:379-402.

Risco, C.A., and C.C. Chase, Jr. 1997. Gossypol. p. 87-98. *In* J.P.F. D'Mello (ed.) Handbook of plant and fungal toxicants. CRS Press, Boca Raton, FL.

Rong, J., C. Abbey, J.E. Bowers, C.L. Brubaker, C. Chang, P.W. Chee, T.A. Delmonte, X. Ding, J.J. Garza, B.S. Marler, C.-h. Park, G.J. Pierce, K.M. Rainey, V.K. Rastogi, S.R. Schulze, N.L. Trolinder, J.F. Wendel, T.A. Wilkins, T.D. Williams-Coplin, R.A. Wing, R.J. Wright, X. Zhao, L. Zhu, and A.H. Paterson. 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.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. p. 365-386. *In* S. Krawetz and S. Misener (ed.) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ.

Samora, P.J., D.M. Stelly, and R.J. Kohel. 1994. Localization and mapping of the Le_1 and Gl_2 loci of cotton (*Gossypium hirsutum* L.). Journal of Heredity. 85:152-157.

Simpson, C.G., D.J. Leader, and J.W.S. Brown. 1993. Characteristics of plant pre-mRNA introns and transposable elements. p. 183-252. *In* R.R.D. Croy (ed.) Plant molecular biology labfax. BIOS Scientific Publishers, Oxford, OX1 1SJ, UK.

Stipanovic, R.D., B. AA, and L. MJ. 1977. Natural insecticides from cotton (*Gossypium*). p. 197-214. *In* P.A. Hedin (ed.) Host plant resistance to pests. American Chemical Society, Washington, D. C.

Stipanovic, R.D., L.S. Puckhaber, A.A. Bell, A.E. Percival, and J. Jacobs. 2005. Occurrence of (+)- and (−)-gossypol in wild species of cotton and in *Gossypium hirsutum* var. *marie-galante* (Watt) Hutchinson. J Agric and Food Chem. 53:6266-6271.

Sun, Q., Y. Cai, Y. Xie, J. Mo, Y. Yuan, Y. Shi, S. Li, H. Jiang, Z. Pan, Y. Gao, M. Chen, and X. He. 2009. Gene expression profiling during gland morphogenesis of a mutant and a glandless upland cotton. Molecular Biology Reports. [online]. Available at http://dx.doi.org/10.1007/s11033-009-9918-3 (verified 4 Nov. 2009).

Sunilkumar, G., L.M. Campbell, L. Puckhaber, R.D. Stipanovic, and K.S. Rathore. 2006. Engineering cotton seed for use in human nutrition by tissue-specific reduction of toxic gossypol. Proceedings of the National Academy of Sciences. 103:18054-18059.

Thanaraj, T. 1999. A clean data set of EST-confirmed splice sites from *Homo sapiens* and standards for clean-up procedures. Nucl. Acids Res. 27:2627-2637.

Townsend, B.J., and D.J. Llewellyn. 2007. Reduced terpene levels in cotton seed add food to fiber. Trends in Biotech. 25:239-241.

Van Deynze, A., K. Stoffel, M. Lee, T.A. Wilkins, A. Kozik, R.G. Cantrell, J.Z. Yu, R.J. Kohel, and D.M. Stelly. 2009. Sampling nucleotide diversity in cotton. BMC Plant Biol. 9:125.

Van Ooijen, J.W. 2006. JoinMap ® 4, Software for the calculation of genetic linkage maps in experimental populations. [online]. Available at http://www.kyazma.nl/index.php/mc.JoinMap/ (verified 25 Jul. 2006).

Wei, H., Y. Fu, and R. Arora. 2005. Intron-flanking EST–PCR markers: from genetic marker development to gene structure analysis in *Rhododendron*. Theoretical and Applied Genetics. 111:1347-1356.

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.

Xie, Y.-F., B.-C. Wang, B. Li, Y.-F. Cai, L. Xie, Y.-X. Xia, P.-A. Chang, and H.-Z. Jiang. 2007. Construction of cDNA library of cotton mutant (Xiangmian-18) library during gland forming stage. Colloids and Surfaces B: Biointerfaces. 60:258-263.

Zhu, S.J., N. Reddy, and Y.R. Jiang. 2005. Introgression of a gene for delayed pigment gland morphogenesis from *Gossypium bickii* into upland cotton. Plant Breeding. 124:590-594.