Theor Appl Genet (2010) 120:587–606
DOI 10.1007/s00122-009-1178-z

ORIGINAL PAPER

Clustering, haplotype diversity and locations of *MIC-3*: a unique root-specific defense-related gene family in Upland cotton (*Gossypium hirsutum* L.)

Zabardast T. Buriev · Sukumar Saha · Ibrokhim Y. Abdurakhmonov · Johnie N. Jenkins · Abdusattor Abdukarimov · Brian E. Scheffler · David M. Stelly

Received: 20 June 2009 / Accepted: 30 September 2009 / Published online: 28 October 2009

Abstract *MIC-3* is a recently identified gene family shown to exhibit increased root-specific expression following nematode infection of cotton plants that are resistant to root-knot nematode. Here, we cloned and sequenced *MIC-3* genes from selected diploid and tetraploid cotton species to reveal sequence differences at the molecular level and identify chromosomal locations of *MIC-3* genes in *Gossypium* species. Detailed sequence analysis and phylogenetic clustering of *MIC-3* genes indicated the presence of multiple *MIC-3* gene members in *Gossypium* species. Haplotypes of a *MIC-3* gene family member were discovered by comparative analysis among consensus sequences across genotypes within an individual clade in the phylogram to overcome the problem of duplicated loci in the tetraploid cotton. Deficiency tests of the SNPs delimited six A_t-genome members of the *MIC-3* family clustered to chromosome arm 4sh, and one D_t-genome member to chromosome 19. Clustering was confirmed by long-PCR amplification of the intergenic regions using A_t-genome-specific *MIC-3* primer pairs. The clustered distribution may have been favored by selection for responsiveness to evolving disease and/or pest pressures, because large variants of the *MIC-3* gene family may have been recovered from small physical areas by recombination. This could give a buffer against selection pressure from a broad range of pest and pathogens in the future. To our knowledge, these are the first results on the evolution of clustering and genome-specific haplotype members of a unique cotton gene family associated with resistant response against a major pathogen.

Communicated by D. Lightfoot.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-009-1178-z) contains supplementary material, which is available to authorized users.

Z. T. Buriev, S. Saha and I. Y. Abdurakhmonov contributed equally to the work.

Z. T. Buriev · I. Y. Abdurakhmonov · A. Abdukarimov
Center of Genomic Technologies, Institute of Genetics and Plant Experimental Biology, Academy of Sciences of Uzbekistan, Yuqori Yuz, Qibray Region, 111226 Tashkent, Uzbekistan
e-mail: zabar75@yahoo.com

I. Y. Abdurakhmonov
e-mail: genomics@uzsci.net

A. Abdukarimov
e-mail: genetics@uzsci.net

S. Saha (✉) · J. N. Jenkins
Crop Science Research Laboratory, United States Department of Agriculture-Agricultural Research Service, Mississippi State, MS 39762, USA
e-mail: Sukumar.Saha@ars.usda.gov

J. N. Jenkins
e-mail: Johnie.Jenkins@ARS.USDA.Gov

B. E. Scheffler
Genomics and Bioinformatics Research Unit, Genomics Laboratory, United States Department of Agriculture-Agricultural Research Service, Stoneville, MS 38776, USA
e-mail: brian.scheffler@ars.usda.gov

D. M. Stelly
Department of Soil and Crop Sciences, Texas A&M University, College Station, TX 77843, USA
e-mail: stelly@tamu.edu

Introduction

A cotton gene, *MIC-3*, was found to exhibit root-specific expression in root-knot nematode (RKN)-resistant plants during nematode infection (Callahan et al. 1997; Zhang et al. 2002). Callahan et al. (2004) demonstrated that *MIC-3* protein accumulation was positively correlated with RKN resistance in selected cotton lines. Wubben et al. (2008) showed that *MIC-3* is a multigene family and associated with a root-specific defense-response mechanism that is independent of other well-known pathways such as gossypol biosynthesis, lipid peroxidation and PR10 expression in cotton. Both of the studies indicated that *MIC-3* expression occurs only in cotton roots and lacks similarity at the sequence level with any other genes of other plant species.

If *MIC-3* can be shown to have a functional role in RKN resistance, then characterization of *MIC-3* genetic variants should help in revealing the genetic mechanisms underlying the RKN defense-response mechanism in cotton. This will facilitate development of 'candidate' gene marker(s) for *MIC-3* genes and marker-assisted selection. Given that nucleotide changes in *MIC-3* coding or regulatory regions could have functional significance, characterization of *MIC-3* genetic variants in different diploid and tetraploid cotton species should reveal underlying genetic mechanisms, origin and evolutionary history of the *MIC-3* family.

Given the uniqueness of the *MIC-3* gene family and apparent exclusive occurrence in *Gossypium* (Callahan et al. 1997; Zhang et al. 2002; Wubben et al. 2008), it seems especially important to determine *MIC-3* gene position(s) and structure(s) as well as their complexities in *G. hirsutum* ($2n = 52$). A common AD 26-chromosome haploid genome structure is shared by the five extant *Gossypium* species, *G. hirsutum*, *G. barbadense* (also cultivated), *G. tomentosum* (wild species endemic to Hawaii), *G. mustelinum* (wild species of Brazil) and *G. darwinii* (wild species endemic to Galapagos Islands). All are partially diploidized evolutionary derivatives of a nascent new world tetraploid ($2n = 4x = 52$) that arose about 1–2 million years ago (Beasley 1940, 1942; Wendel and Cronn 2003). Their origins entailed formation of a tetraploid hybrid between an old world taxon similar to the extant 'A-genome' species *G. herbaceum* and *G. arboreum* ($2n = 2x = 26$), with a taxon of the 'D-genome' group related to the new world species *G. raimondii* Ulbrich ($2n = 2x = 26$). The five extant $2n = 52$ species are thus disomic and have AD haploid genomes dubbed $[AD]_1$, $[AD]_2$, etc., whereas the 45 extant diploid species are $2n = 26$, with haploid genomes falling into eight groups A to G and K that relate to meiotic behavior, dissimilarity and geographic distribution.

Single-nucleotide polymorphism (SNP) at specific nucleotide positions is considered to be an efficient diagnostic marker for candidate genes and reported to be the most abundant and highly polymorphic marker. The identification of SNPs is generally done via sequencing of genes in multiple related organisms and can be complicated by similarities between orthologous loci and, in disomic polyploids, paralogous loci. Disomic tetraploids such as cotton normally contain two divergent paralogous copies of each gene (one per subgenome). In cotton, such sequence differences that trace phylogenetically to the ancestral A- or D-genome can be dubbed genome-specific polymorphisms or GSPs (Yang et al. 2006). Frequently, SNPs from a gene can be grouped into haplotype blocks, so that each haplotype can be tagged by a selected subset of unique SNPs. We have reported and used specific strategies to discover and map several SNP markers associated with different fiber genes in tetraploid cotton (An et al. 2007, 2008; Hsu et al. 2008). We showed that the delineation of SNP–SNP associations and haplotypes which increased the resolution and discovery of multiple alleles at a specific genomic location.

Here, we report SNP discovery to delineate haplotypes of the *MIC-3* gene family, describe the family members in selected diploid and tetraploid cotton species, chromosomally localize the *MIC-3* loci and document clustering of different members of the *MIC-3* gene family.

Materials and methods

Plant materials and DNA extraction

DNA samples were extracted from young leaves of four disomic tetraploid ($2n = 52$) and three diploid ($2n = 26$) species using the DNeasy Plant Mini Kit (Qiagen, Santa Clarita, CA), following the manufacturer's protocol. The materials included individual plants of '3-79' (doubled-haploid) line of *G. barbadense* ($[AD]_2$ genome), *G. tomentosum* ($[AD]_3$ genome) and *G. mustelinum* ($[AD]_4$ genome), and nine lines of *G. hirsutum* ($[AD]_1$ genome), including moderately to highly RKN-resistant lines: Clevewilt, M240 and M315, and *G. hirsutum* lines Golib, ST 215, M8, TM-1, susceptible isoline (Sisoline) and one red mutant line (REDM).

We used deficiency testing to delimit chromosomal locations of *MIC-3* family members, by screening polymorphisms against a panel of monosomic and monotelodisomic interspecific *G. hirsutum* × *G. barbadense* F_1 hybrids following the overall methods of An et al. (2008). Each hypoaneuploid hybrid lacked a specific maternal (*G. hirsutum*) chromosome or chromosome arm and rendered hemizygous the respective chromosome or arm of *G. barbadense*. Each monotelodisomic hybrid was deficient for all or part of the *G. hirsutum* chromosome arm 1Lo (long arm), 2Lo, 2sh (short arm), 3Lo, 3sh, 4sh, 4Lo, 5Lo, 6Lo, 7Lo, 7sh, 9Lo, 11Lo, 14Lo, 15Lo, 15sh, 16sh, 16Lo, 18Lo, 18sh, 20Lo,

22sh, 25Lo or 26sh. Each monosomic hybrid was lacking in *G. hirsutum* chromosome 1, 2, 3, 4, 6, 7, 9, 10, 12, 17, 18, 20, 23 or 25. We also used analogous primary monosomic F_1 hybrids *G. hirsutum* × *G. mustelinum* that were lacking in *G. hirsutum* chromosome 12, 16, 17, 18 or 25; monotelodisomic hybrids deficient in most or all of chromosome arms 14Lo, 20sh, 20Lo, 22sh, 22Lo or 26sh; and similar tertiary monosomic hybrids NTN4-15, NTN16-15, NTN10-19 and NTN12-19 hybrids, which provided partial coverage of the indicated chromosomes and chromosome arms. For example, the NTN4-15 hybrid is deficient for segments of chromosomes 4 and 15 of TM-1 line, and thus hemizygous for the corresponding segments of *G. mustelinum*, but heterozygous for all other genomic regions. The tertiary monosomics involve chromosome translocations previously described by Brown (1980) and Menzel et al. (1985).

Chromosomal locations were also delimited by screening for the polymorphisms across a set of backcross-derived (BC_5) euploid disomic chromosome substitution lines, CS-B lines, which are quasi-isogenic to TM-1, except for the pair of substituted chromosomes or chromosome arms of *G. barbadense* (Stelly et al. 2005).

Two of the diploids included in this study were *G. herbaceum* and *G. raimondii*, which are most closely related to the A_t- and D_t-genomes (respectively) of the extant AD tetraploid species. The African wild cotton, *G. longicalyx*, was included in this study because it is a source of extreme resistance to another root pathogen, the reniform nematode (Yik and Birchfield 1984; Robinson et al. 2007).

PCR amplification

We used two primers, FC1F (5′-AAAAATGGCTTCT CCTCCA-3′) and FC2R (5′-AAGGAGAAACAACGC ACA-3′), from sequences of *MIC-3* (Zhang et al. 2002) to amplify the genomic DNAs of individual plants. The PCR reaction mixture consisted of 1 µl of genomic DNA (25 ng), 1 µl each of FC1F and FC2R (5 µM) primers, 1 µl of dNTPs (2.5 mM each), 5 µl of 10× PCR reaction buffer, 0.25 µl of Taq DNA polymerase (5.0 unit/µl) (Qiagen, Santa Clarita, CA) and 42.5 µl of distilled water. PCR reactions were conducted in a Gene Amp PCR System 9700 with an initial denaturation of 94°C for 3 min, followed by 35 cycles of 94°C for 1 min, 55°C for 1 min and 72°C for 2 min, and final extension at 72°C for 5 min. Based on the sequence information of *MIC-3* genes from this experiment, another pair of primers (5′-GCGCAA ATGGAGTGGGTAATGAAT-3′ and 5′-CAACTTGCTCT TATCATGTGGGGGTG-3′) was used to amplify intergenic regions between *MIC-3* genes using a MasterAmp™ High Fidelity Long PCR kit following the manufacturer's method (Epicentre, Madison, WI, USA). Amplification products of these intergenic regions were visualized by 1.5% agarose (Sigma, USA) gel electrophoresis in 0.5× TBE buffer and staining with ethidium bromide following the standard procedures.

Cloning and sequencing

PCR products were excised from the agarose gel and then purified using the QIAGEN MinElute™ Gel Extraction Kit (Valencia, CA, USA). All PCR products were cloned into the TOPO TA cloning vector as per manufacturer's instructions (Invitrogen, Carlsbad, CA, USA). Bi-directional sequencing was conducted at the Genomics and Bioinformatics Research Unit with standard M13 primers using an ABI Genetic Analyzer 3730XL (Applied Biosystems, USA) following standard protocols using BigDye 3.1.

Sequence analysis

Sequences were analyzed with SEQUENCHER 4.2 software (Gene Codes, USA) and sequences were searched against NCBI GenBank databases using BLASTN (Altschul et al. 1997). Multiple alignments were performed using CLUSTALW (Thompson et al. 1994). The cloned sequences were grouped into contigs on the basis of at least three shared SNPs and, where present, distinct insertion/deletion polymorphisms. A minority of single sequence reads differed from other clones in the contig at one nucleotide position; these were treated as the products of Taq polymerase substitution error.

Phylogenetic analyses and identification of subfamilies

Consensus sequences for each contig (each representing a distinct haplotype) were used for phylogenetic analyses. Phylogenetic analysis of genomic DNA sequences was performed using neighbor-joining (NJ) algorithms with the Juke and Cantor distance option (Saitou and Nei 1987) available in the software package MEGA 4.1 (Tamura et al. 2007). The robustness of the phylogenetic trees was evaluated by bootstrapping with 1,000 repetitions. To group sequences into subfamilies, we first separately constructed NJ-trees for *MIC-3* gene members from all nine *G. hirsutum* genotypes only and classified major subfamilies. Then, an NJ phylogenetic tree, including all 169 *MIC-3* members identified in all cotton genotypes, was constructed and *MIC-3* subfamilies of all *Gossypium* species were classified on the basis of subfamilies determined in *G. hirsutum* genotypes.

SNP marker detection and chromosomal localization

Using a separate phylogenetic and *MIC-3* sequence analyses of two cotton species, TM-1 (a genetic standard line for *G. hirsutum*) and 3-79 (a double haploid line for

G. barbadense), we designed several GSP markers to distinguish the interspecific genomic polymorphisms. Eight A_t-genome- and two D_t-genome-specific GSP marker primer pairs were designed to detect sequence polymorphisms between TM-1 and 3-79 or *G. mustelinum* because interspecific cytogenetic stocks in TM-1 genetic background were available from these two alien species. These candidate gene markers were then used for deletion analysis by screening the cytogenetic stocks including monosomic, monotelodisomic, CS-B (Stelly et al. 2005), and NTN hybrid stocks, following the overall strategy of our previous studies (An et al. 2007, 2008).

The ABI Prism SNaPshot™ multiplex kit (Applied Biosystems, USA) was used to detect candidate *MIC-3* gene marker polymorphisms following the manufacturer's protocol. The excess nucleotides and primers from the amplified *MIC-3* PCR products were first removed by incubating 15 µl PCR product with 5 units of shrimp alkaline phosphatase (SAP) enzyme and two units of *Exo* I enzyme (Amesrsham, Cleveland, Ohio) at 37°C for 1 h, followed by 75°C for 15 min. As per SNaPshot™ multiplex kit protocol, 3 µl of purified PCR product was mixed with 5 µl of SnaPshot multiplex ready reaction mix, 1 µl of SNP primer (2 µM), and 1 µl of distilled water. The thermal cycle reaction was carried out with 25 cycles at 96°C for 10 s, 50°C for 5 s, and 60°C for 30 s for SNP detection. The PCR product was diluted after treatment with SAP enzyme as indicated above, in 1:4 volume with water. As much as 0.5 µl of the diluted SnaPshot product was mixed with 0.5 µl of size standard LIZ 120 (Applied Biosystems USA) and 9 µl of Hi-Di formamide (Applied Biosystems USA), denatured at 95°C for 5 min and then run into a 3100xl Genetic Analyzer (Applied Biosystems USA) to detect SNP markers.

Results

Characterization of the *MIC-3* gene family in *Gossypium* species

Approximately, 150 individual PCR-derived clonal inserts for *MIC-3* gene products were sequenced from each of 15 cotton genotypes. The amplicons from *MIC-3* genes of *Gossypium* species ranged from 653 to 713 bp, including 653–706-bp amplicon length in all allotetraploid genotypes, 703–713-bp amplicons in *G. herbaceum*, 653–691-bp amplicons in *G. raimondii*, and 706-bp amplicons in *G. longicalyx* (Table 1). Multiple sequence alignment created a 721-bp *MIC-3* full-length consensus genomic sequence taking all the species into account, and identities of all the sequences in our results were confirmed with the respective original gene (Zhang et al. 2002) by BLASTN for each of the selected *Gossypium* species. Intronic and exonic parts of the *MIC-3* genes of *Gossypium* species were annotated according to the *MIC-3* nucleotide and protein sequence reported by Zhang et al. 2002. The overall results from multiple sequence alignment of the genomic sequences from all *Gossypium* species demonstrated the sequence identity with an 11-bp fragment from the non-coding 5′UTR region, two coding regions of 221–230 bp in exon 1 and one 193-bp fragment in exon 2, a single noncoding intron region of 65–77 bp and a noncoding 3′UTR region of 151–195 bp (Zhang et al. 2002).

Large insertion and deletion (indel) changes were observed among members of the *MIC-3* gene families of all genotypes studied. For example, all cotton genotypes studied, except *G. herbaceum* (A-genome diploid), had the 9-bp D-genome-specific indel in the first exon. A 9-bp indel was found in A_t-genome grouped *MIC-3* sequences of *G. mustelinum* (Table 1). All allotetraploid lineages except *G. hirsutum* cv. M8 had two other major indels of 15 and 29 bp in the 3′UTR region. The nucleotide length and position of the 29-bp indel was the same in *G. raimondii*, the most closely related extant D-genome diploid to AD tetraploids. The 15-bp indel was unique to the allotetraploid genotypes. Results showed that the tetraploid species had different patterns of indels in the 3′UTR region including a complex pattern combining the 15- and 29-bp indels such as (Table 1): three types in *G. mustelinum*, two types in *G. barbadense* 3-79 and two types in *G. hirsutum* cv. Golib. *G. barbadense* 3-79 and *G. hirsutum* cv. Golib *MIC-3* indel patterns were similar and included an amplicon with unique 15-bp indel suggesting the possibility of introgression between *G. barbadense* into *G. hirsutum*. The 15 and 29-bp indels were absent from the 3′UTR of all 12 *MIC-3* members in *G. hirsutum* cv. M8. Wubben et al. (2008) reported 15 distinct members from the same M8 genotype, suggesting that some *MIC-3* members were missing in our study. The pattern of indels was intriguingly different between AD- and D-genome cottons versus *G. herbaceum*, the extant A-genome species most closely related to AD lineages. *G. herbaceum* completely lacks the 9-bp indel found in the first exon of *MIC-3* genes of AD- and D-genome cottons and has a unique 6-bp indel in the 3′UTR region. Moreover, we found 6–12 bp indels in an intronic region of *G. herbaceum MIC-3* amplicon that were not found in other cotton genotypes studied. These *G. herbaceum*-specific indels in the intronic region of *MIC-3* had a signature of a microsatellite repeat pattern of $(AATT)_{2\text{-}3}$. The largest 31-bp indel found in *G. raimondii* was also unique. None of the above-mentioned indel mutations were found in *G. longicalyx*, a diploid F-genome representative of African origin (Table 1).

Considering all sequences, the *MIC-3* amplicon indel patterns could preliminarily be grouped into F-, A- and D-genomic and A_t- and D_t-genomic groups (Table 1). Further detailed sequence analysis suggested that indel

Table 1 Characteristics of *MIC-3* gene family of *Gossypium* species

Genotype	No. of members	Amplicon length variation	Indel changes			Total SNP
			Exon 1	Intron	3′UTR	
Gh_TM1	10	653 (D_t)	9	–	15, 29	16 (A_t), 21 (D_t)
		706 (A_t, D_t)	–	–	–	
Gh_Sisoline	11	653 (D_t)	9	–	15, 29	16 (A_t), 21 (D_t)
		706 (A_t, D_t)	–	–	–	
Gh_ST213	11	653 (D_t)	9	–	15, 29	16 (A_t), 22 (D_t)
		706 (A_t, D_t)	–	–	–	
Gh_M8	12	697 (D_t)	9	–	–	17 (A_t), 20 (D_t)
		706 (A_t, D_t)	–	–	–	
Gh_REDM	13	653 (D_t)	9	–	15, 29	17 (A_t), 21 (D_t)
		706 (A_t, D_t)	–	–	–	
Gh_Golib	16	653 (D_t)	9	–	15, 29	21 (A_t), 22 (D_t)
		682 (D_t)	9	–	15	
		706 (A_t, D_t)	–	–	–	
Gh_Clevewilt	13	653 (D_t)	9	–	15, 29	17 (A_t), 21 (D_t)
		706 (A_t, D_t)	–	–	–	
Gh_M240	9	653 (D_t)	9	–	15, 29	16 (A_t), 21 (D_t)
		706 (A_t, D_t)	–	–	–	
Gh_M315	15	653 (D_t)	9	–	15, 29	18 (A_t), 27 (D_t)
		697 (D_t)	9	–	–	
		706 (A_t, D_t)	–	–	–	
Gb_3-79	11	653 (D_t)	9	–	15, 2 9	11 (A_t), 29 (D_t)
		682 (D_t)	9	–	15	
		706 (A_t, D_t)	–	–		
G. tomentosum	9	653 (D_t)	9	–	15, 29	10 (A_t), 24 (D_t)
		706 (A_t, D_t)	–	–	–	
G. mustelinum	15	653 (D_t)	9	–	15, 29	12 (A_t), 28 (D_t)
		668 (A_t)	9[a]	–	29	
		682 (D_t)	9	–	15	
		706 (At, D_t)	–	–	–	
G. herbaceum	9	703	–	12	6	19 (A)
		707	–	8	6	
		709	–	6	6	
		713	–	8	–	
G. raimondii	4	651	9	–	31	33 (D)
		653	9	–	29	
		691	–	–	–	
G. longicalyx	11	706	–	–	–	14 (F)

Gh, *G. hirsutum*; Gb, *G. barbadense*; A_t, A subgenome of the tetraploid cotton species; D_t, D subgenome of the tetraploid cotton species; A, A-genome of *G. herbaceum*; D, D-genome of *G. raimndii*; F, F-genome of *G. longicalyx*

[a] Note that 9-bp indel in the first exon of A_t-genome-specific *MIC-3* gene amplicon of *G. mustelinum* has a triplet upstream shift compared to other D_t-genome-specific indels observed in other genotypes

information alone might not be sufficient to determine genomic origins of some amplicons. We had one unique case with two of *G. mustelinum MIC-3* amplicons (designated as *MIC-3*_14 and *MIC-3*_15) that had both D-genome-specific 9 and 29-bp indels, suggesting its origin from a putative D-genome ancestor. Detailed results of sequence analyses have been summarized in Table 1. The tetraploid genotypes had 9–15 distinct *MIC-3* members in each of 9 *G. hirsutum* genotypes, 11 members in *G. barbadense* 3-79, 9 in *G. tomentosum* and 15 members in *G. mustelinum* (Table 2). In diploid genomes, there were 9 distinct *MIC-3* members in *G. herbaceum*, 11 in *G. longicalyx* and 4 in *G. raimondii*.

Table 2 *MIC-3* gene subfamily composition in cotton genotypes

Subfamily	Gh_TM-1	Gh_Sisoline	Gh_ST213	Gh_M8	Gh_REDM	Gh_Golib	Gh_Clevewilt	Gh_M240	Gh_M315	Gb_3-79	*G. tomentosum*	*G. mustelinum*	*G. herbaceum*	*G. raimondii*	*G. longicalyx*
1	2	2	2	2	5	3	2	2	2	–	–	–	–	–	–
2	–	1	1	–	–	–	1	–	–	–	–	–	–	–	–
3	2	1	1	2	2	1	2	1	1	–	–	–	–	–	–
4	1	1	1	1	1	2	1	1	2	1	1	2	–	–	–
5	–	1	–	–	–	1	–	–	–	1	1	–	–	–	–
6	1	2	2	2	1	2	2	1	1	1	2	4	–	–	–
7	–	–	–	–	1	–	–	1	1	–	–	–	–	–	–
8	1	–	–	–	–	–	–	–	2	–	–	–	–	–	–
9	1	1	1	2	1	3	2	1	2	3	2	2	–	–	–
10	1	1	2	2	1	2	2	1	2	3	2	4	–	–	–
11	–	–	–	1	–	–	–	–	1	–	–	–	–	–	–
12	1	1	1	0	1	2	1	1	1	2	1	2	–	–	–
13	–	–	–	–	–	–	–	–	–	–	–	1	3	–	–
14	–	–	–	–	–	–	–	–	–	–	–	–	6	–	–
15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	11
16	–	–	–	–	–	–	–	–	–	–	–	–	–	2	–
17	–	–	–	–	–	–	–	–	–	–	–	–	–	2	–
Total members	10	11	11	12	13	16	13	9	15	11	9	15	9	4	11
Total subfamilies	8	9	8	7	8	8	8	8	10	7	6	6	2	2	1

Gh, *G. hirsutum*; Gb, *G. barbadense*

BLAST results of 150 individual clones from each of 15 cotton genotypes including diploid and tetraploid species confirmed previous reports (Wubben et al. 2008; Callahan et al. 2004; Zhang et al. 2002) of no significant homology of *MIC-3* genes with any gene or sequences other than *Gossypium* species. The results suggest that the multi-gene *MIC-3* gene family diversified by gene duplication following independent modes of evolution within *Gossypium* species (Tables 1, 2). More duplication events have occurred in the A_t-genome than the D_t-genome of the tetraploid species. 169 *MIC-3* members could be grouped into 88 distinctive types from 15 *Gossypium* genotypes based on sequence comparison (Fig. 1).

Classification of *MIC-3* genes into subfamilies in *Gossypium* species

A total of 169 *MIC-3* gene members were grouped into 17 different subfamilies in the phylogram (Fig. 1); 110 *MIC-3* members of *G. hirsut*um genotypes were grouped into12 subfamilies, of which subfamilies 1–9 were specific to the A_t-genome and 10–12 were specific to the D_t-genome (Table 2; Fig. 1). Subfamilies 13 and 14 were specific to *G. herbaceum*, subfamily 15 was specific to *G. longicalyx*, and subfamilies 16 and 17 were specific to *G. raimondii MIC-3* gene family (Fig. S1). All allotetraploid *MIC-3* gene members, except *G. mustelinum*-derived member *MIC-3*_14, were clustered within 12 subfamilies of *G. hirsutum* (Figs. 1, 2, S1). Both *G. herbaceum* and *G. raimondii*, the extant diploids most closely related to AD cottons, had two distinct subfamilies. We also observed some *MIC-3* subfamilies to be species-specific. For example, subfamily 1 was found only in *G. hirsutum*, including susceptible isoline (Sisoline), ST213 and Clevewilt; subfamily 8 included sequences from M315 only, and subfamily 11 contained sequences from M8 and M315 Upland cultivars only (Table 2). This species-specific pattern could also be due to the small number of representative members in the overall sample in an individual species.

Haplotypes of *MIC-3* sequences from *G. hirsutum*

To simplify the identification of putative haplotypes, we used only the sequences of *G. hirsutum* because it represented a diverse group of lines in our study. We broadly divided *G. hirsutum* (AD) haplotypes into: A_t-genome and D_t-genome (Tables 3, 4). The sequences of *G. hirsutum* closely associated in the phylogram with the sequences of *G. herbaceum* were considered putative A_t-genome haplotypes and the haplotypes grouped closely with the sequences of *G. raimondii* were considered putative D_t-genome haplotypes (Fig. 2). Because of their phylogenetic proximity to AD species, we used the consensus sequence of G. *herbaceum* and *G. raimondii* as reference sequences in the tables of haplotype comparisons (Tables 3, 4). Each broad group was subdivided into several subgroups or clades, suggesting multiple duplication events of *MIC-3* genes during their evolution within each of the subgenomes of tetraploid cotton. Evidence of duplications was augmented by subsequent discovery of clustering of *MIC-3* genes on the same chromosome, as discussed later. Results showed that collectively the A_t-genome had nine putative loci (haplotype group) with an average of 3.5 alleles or haplotypes/locus, and the D_t-genome consisted of three putative loci (haplotype groups) with an average of 3.3 putative alleles or haplotypes/locus in *G. hirsutum*. The number of haplotypes or putative alleles per putative locus ranged from two to six in the A_t-genome and four to six in the D_t-genome (Tables 3, 4). The tetraploid *MIC-3* sequences included 25 SNPs in the A_t-genome and 26 SNPs in the D_t-genome, all biallelic, except in positions 113 (A, D), 162 (D) and 568 (D). The average length of each SNP in the A_t-genome was 182.6 bp, and 108.8 bp in the D_t-genome.

Chromosomal location

Based on a separate phylogram (Fig. 3) and *MIC-3* sequence analyses of TM-1 and 3-79, several GSP markers were designed to distinguish the interspecific polymorphisms at the subgenome level between these two lines (Table 5). These candidate gene markers were used in deletion analysis with the cytogenetic stocks (Stelly et al. 2005; Gutierrez et al. 2009). Both parental polymorphisms were present in most of the F_1 hybrids; however, all eight TM-1 GSP alleles, representing seven out of eight different A_t-genome-derived *MIC-3* members of TM-1, were absent in the hypoaneuploid F_1 plants that lacked *G. hirsutum* chromosome-4 (H04 monosomic) or the short arm of chromosome 4 (Te04Lo monotelodisomic) (Table 5). The euploid disomic chromosome substitution line CS-B04 also lacked the same *G. hirsutum MIC-3* alleles for all of these markers. In contrast, all eight GSP loci were heterozygous in all other substitution lines and in the monotelodisomic F_1 substitution plant (Te04sh) deficient for the long arm of chromosome-4. The results collectively indicated that seven out of eight A_t-genome *MIC-3* members of TM-1 are on the short arm of chromosome 4, which may suggest a clustered configuration of the *MIC-3* genes in the cotton genome (Table 5; Fig. 3).

Although we found several D_t-genome *MIC-3* haplotype-specific GSPs between TM-1 and 3-79, they failed to be polymorphic in deletion analysis and chromosomal localization experiments. This was due to the existence of overlapping sites within 3-79 *MIC-3* members that limited our differentiation of GSP polymorphisms or the location of the marker to chromosomes not associated with any of the deletion lines of the cytogenetic stocks. Detailed sequence analysis of *G. mustelinum*-derived *MIC-3*

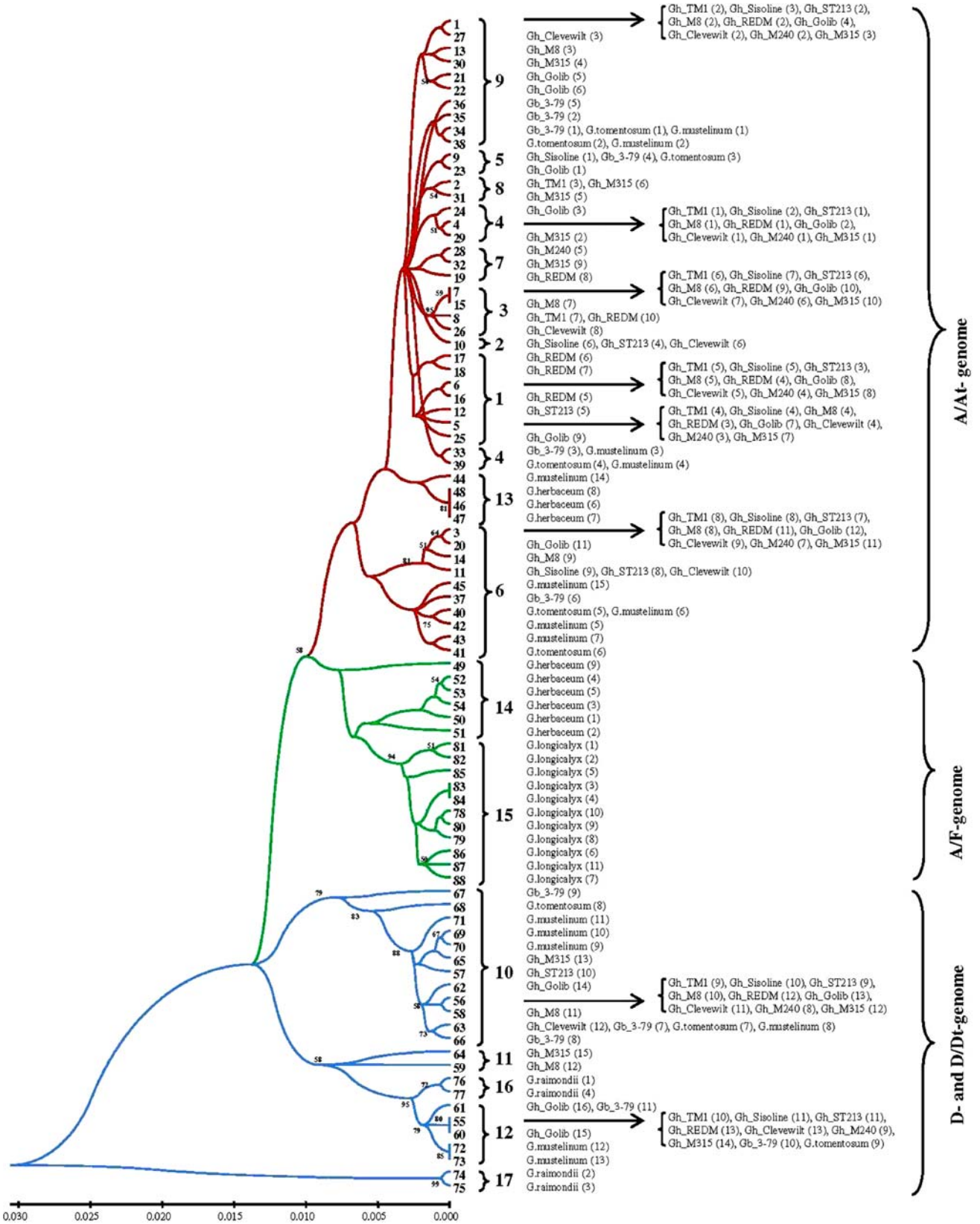

Fig. 1 Curved neighbor-joining phylogenetic tree derived from the sequences of 88 distinct *MIC-3* genes from *Gossypium* species. Branch length and bootstrap values (>50%) are shown

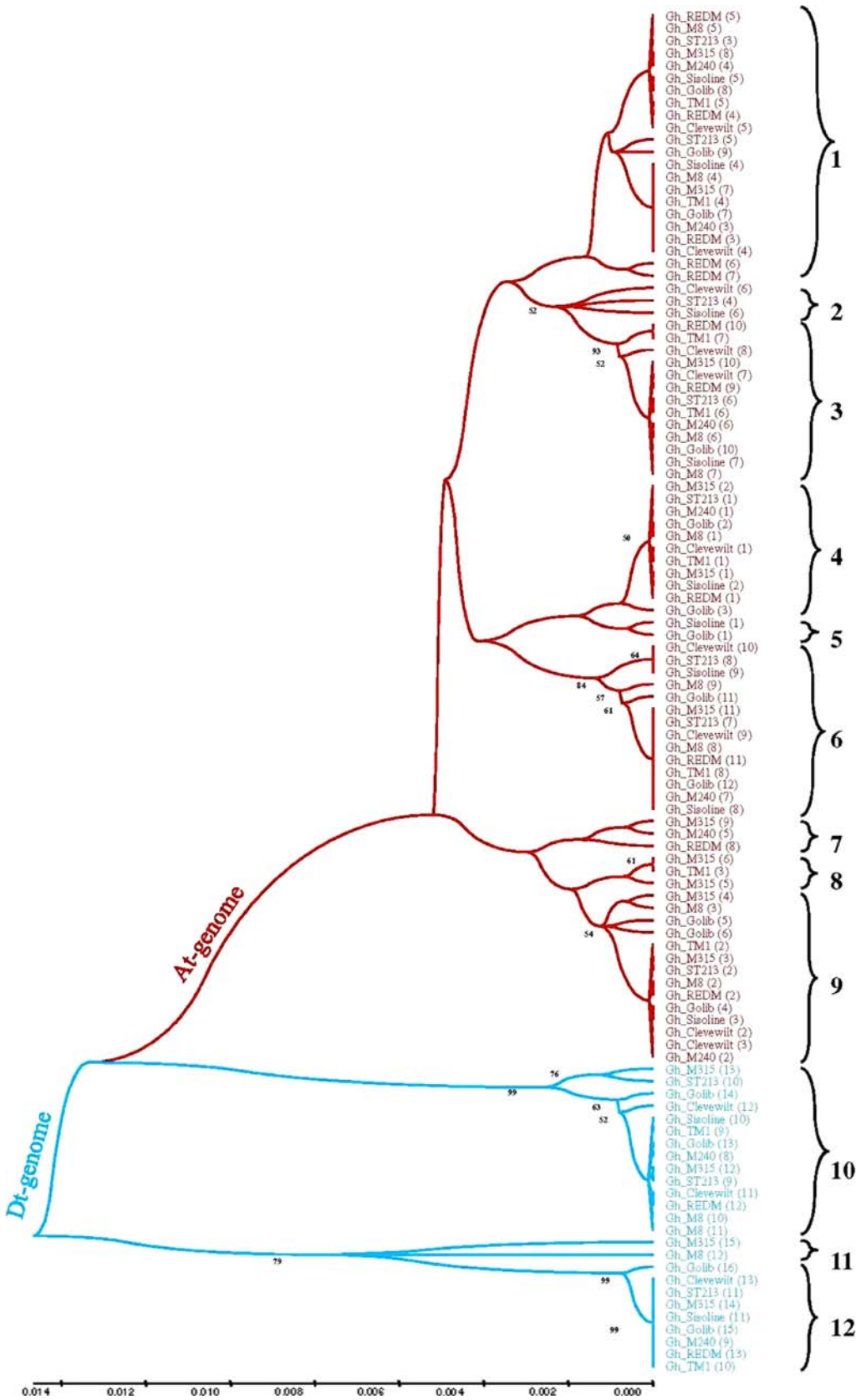

Fig. 2 Curved neighbor-joining phylogenetic tree derived from the sequences of 110 *MIC-3* gene members found in nine genotypes of *G. hirsutum*. Branch length, bootstrap values (>50%) and classified subfamilies (1-12) are shown

Table 3 A haplotype table for A_t-genome *MIC-3* genes of *G. hirsutum* (AD)

SG	Name	22 T[a]	31 T	97 T	105 G	106 T	112 G	128 G	161 S	162 T	163 A	200 G	240 A	286 C	341 C	393 T	428 C	513 A	523 C	524 T	527 G	567 R	631 G	651 T	664 C	674 A	Haplotypes	Genotypes
1	Hap_1	*	C	*	*	*	A	*	*	*	*	*	*	*	*	C	*	*	T	*	*	C	*	*	*	*	BABBBAABBA AAAAAAAABB BABBAA	Gh_TM1 (4), Gh_Sisoline (4), Gh_M8 (4), Gh_REDM (3), Gh_Golib (7), Gh_Clevewilt (4), Gh_M240 (3), Gh_M315 (7)
	Hap_2	*	*	*	*	*	A	*	*	*	*	*	*	*	*	C	*	*	T	*	*	C	*	*		*	BBBBBAABBA AAAAAAAABB BABBAA	Gh_ST213 (5)
	Hap_3	*	C	*	*	*	A	*	*	*	*	*	*	*	*	C	*	*	T	*	*	*	*	*	*	*	BABBBAABBA AAAAAAAABB BBBBAA	Gh_TM1 (5), Gh_Sisoline (5), Gh_ST213 (3), Gh_M8 (5), Gh_REDM (4), Gh_Golib (8), Gh_Clevewilt (5), Gh_M240 (4), Gh_M315 (8)
	Hap_4	*	C	*	*	*	A	*	*	A	*	*	*	*	*	C	*	*	T	*	*	*	*	*	*	*	BABBBAABAA AAAAAAAABB BBBBAA	Gh_REDM (5)
	Hap_5	*	C	*	*	*	A	*	*	*	*	*	*	*	*	C	*	*	T	*	*	C	*	C	G	T	BABBBAABBA AAAAAAAABB BABABB	Gh_Golib (9)
	Hap_6	*	C	*	*	*	C	*	*	*	*	*	*	*	*	C	*	*	T	*	*	*	*	*	*	*	BABBBBABBA AAAAAAAABB BBBBAA	Gh_REDM (7)
	Hap_7	*	*	*	*	*	C	*	*	*	*	*	*	*	*	C	*	*	T	*	*	*	*	*	*	*	BBBBBBABBA AAAAAAAABB BBBBAA	Gh_REDM (6)
2	Hap_8	*	C	*	*	*	*	*	*	*	*	*	*	*	*	C	*	*	T	*	*	*	*	*	*	*	BABBBCABBA AAAAAAAABB BBBBAA	Gh_Sisoline (6), Gh_ST213 (4), Gh_Clevewilt (6)

Table 3 continued

SG	Name	22 T[a]	31 T	97 T	105 G	106 T	112 G	128 G	161 S	162 T	163 A	200 G	240 A	286 C	341 C	393 T	428 C	513 A	523 C	524 T	527 G	567 R	631 G	651 T	664 C	674 A	Haplotypes	Genotypes
3	Hap_9	*	C	A	*	*	*	T	*	*	*	T	*	T	*	C	*	*	T	*	*	*	*	*	*	*	BAABBCBBBA BABAAAAABB BBBBAA	Gh_TM1 (6), Gh_Sisoline (7), Gh_ST213 (6), Gh_M8 (6), Gh_REDM (9), Gh_Golib (10), Gh_Clevewilt (7), Gh_M240 (6), Gh_M315 (10)
	Hap_10	*	C	A	*	*	*	T	*	*	*	T	*	*	*	C	*	*	T	*	*	*	*	*	*	*	BAABBCBBBA BAAAAAAABB BBBBAA	Gh_TM1 (7), Gh_REDM (10)
	Hap_11	*	C	A	*	*	*	T	*	*	*	T	*	T	*	C	*	*	T	C	*	*	*	*	*	*	BAABBCBBBA BABAAAAABA BBBBAA	Gh_M8 (7)
	Hap_12	*	C	*	*	*	*	T	*	*	*	T	*	T	*	C	*	*	T	*	*	*	*	*	*	*	BABBBCBBBA BABAAAAABB BBBBAA	Gh_Clevewilt (8)
4	Hap_13	*	C	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	A	*	*	*	*	*	BABBBCABBA AAAABAAAAB ABBBAA	Gh_TM1 (1), Gh_Sisoline (2), Gh_ST213 (1), Gh_M8 (1), Gh_REDM (1), Gh_Golib (2), Gh_Clevewilt (1), Gh_M240 (1), Gh_M315 (1)
	Hap_14	*	C	*	*	*	A	*	*	*	*	*	*	*	*	*	*	*	*	*	A	*	*	*	*	*	BABBBAABBA AAAABAAAAB ABBBAA	Gh_Golib (3)
	Hap_15	*	C	*	A	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	A	*	*	*	*	*	BABABCABBA AAAABAAAAB ABBBAA	Gh_M315 (2)
5	Hap_16	*	C	*	*	*	*	*	*	*	*	*	*	*	*	*	*	G	*	*	*	*	*	*	*	*	BABBBCABBA AAAABABAAB BBBBAA	Gh_Sisoline (1)
	Hap_17	*	C	*	*	*	A	*	*	*	*	*	*	*	*	*	*	G	*	*	*	*	*	*	*	*	BABBBAABBA AAAABABAAB BBBBAA	Gh_Golib (1)

Table 3 continued

SG	Name	22 T[a]	31 T	97 T	105 G	106 T	112 G	128 G	161 S	162 T	163 A	200 G	240 A	286 C	341 C	393 T	428 C	513 A	523 C	524 T	527 G	567 R	631 G	651 T	664 C	674 A	Haplotypes	Genotypes
6	Hap_18	*	*	*	*	*	C	*	*	*	*	*	*	*	T	*	T	*	*	*	*	*	*	*	*	*	BBBBBBABBA AAABBBABAB BBBBAA	Gh_TM1 (8), Gh_Sisoline (8), Gh_ST213 (7), Gh_M8 (8), Gh_REDM (11), Gh_Golib (12), Gh_Clevewilt (9), Gh_M240 (7), Gh_M315 (11)
	Hap_19	*	*	*	*	*	C	*	*	*	*	*	*	*	T	*	T	*	*	*	*	*	C	*	*	*	BBBBBBABBA AAABBBABAB BBABAA	Gh_Golib (11)
	Hap_20	*	C	*	*	*	C	*	*	*	*	*	*	*	T	*	T	*	*	*	*	*	*	*	*	*	BABBBBABBA AAABBBABAB BBBBAA	Gh_M8 (9)
	Hap_21	*	C	*	*	*	A	*	*	*	*	*	*	*	T	*	T	*	*	*	*	*	*	*	*	*	BABBBAABBA AAABBBABAB BBBBAA	Gh_Sisoline (9), Gh_ST213 (8), Gh_Clevewilt (10)
7	Hap_22	*	*	*	*	*	*	*	A	*	G	*	*	*	*	*	*	*	T	*	*	C	*	*	*	*	BBBBBCAABB AAAABAAABB BABBAA	Gh_REDM (8)
	Hap_23	*	C	*	*	*	*	*	A	*	G	*	*	*	*	C	*	*	T	*	*	*	*	*	*	*	BABBBCAABB AAAAAAAABB BBBBAA	Gh_M240 (5)
	Hap_24	*	C	*	*	*	*	*	A	*	G	*	*	*	*	C	*	*	T	*	*	C	*	*	*	*	BABBBCAABB AAAAAAAABB BABBAA	Gh_M315 (9)
8	Hap_25	*	*	*	*	*	*	*	A	*	G	*	*	*	*	*	*	*	*	*	A	*	*	*	*	*	BBBBBCAABB AAAABAAAAB ABBBAA	Gh_TM1 (3), Gh_M315 (6)
	Hap_26	*	C	*	*	*	*	*	A	*	G	*	*	*	*	*	*	*	*	*	A	*	*	*	*	*	BABBBCAABB AAAABAAAAB ABBBAA	Gh_M315 (5)

Table 3 continued

SG	Name	22 T[a]	31 T	97 T	105 G	106 T	112 G	128 G	161 S	162 T	163 A	200 G	240 A	286 C	341 C	393 T	428 C	513 A	523 C	524 T	527 G	567 R	631 G	651 T	664 C	674 A	Haplotypes	Genotypes
9	Hap_27	*	*	*	*	*	*	*	A	*	G	*	*	*	*	*	*	G	*	*	*	*	*	*	*	*	BBBBBCAABB AAAABABAAB BBBBAA	Gh_TM1 (2), Gh_Sisoline (3), Gh_ST213 (2), Gh_M8 (2), Gh_REDM (2), Gh_Golib (4), Gh_Clevewilt (2), Gh_M240 (2), Gh_M315 (3)
	Hap_28	*	*	*	*	*	*	*	A	*	G	*	T	*	*	*	*	G	*	*	*	*	*	*	*	*	BBBBBCAABB ABAABABAAB BBBBAA	Gh_Clevewilt (3)
	Hap_29	*	C	*	*	*	*	*	A	*	G	*	*	*	*	*	*	G	*	*	*	*	*	*	*	*	BABBBCAABB AAAABABAAB BBBBAA	Gh_M8 (3)
	Hap_30	C	C	*	*	*	*	*	A	*	G	*	*	*	*	*	*	G	*	*	*	*	*	*	*	*	AABBBCAABB AAAABABAAB BBBBAA	Gh_M315 (4)
	Hap_31	*	*	*	*	*	*	*	A	*	G	*	*	*	*	*	*	G	*	*	*	*	C	*	*	*	BBBBBCAABB AAAABABAAB BBABAA	Gh_Golib (5)
	Hap_32	*	C	*	*	G	*	*	A	*	G	*	*	*	*	*	*	G	*	*	*	*	C	*	*	*	BABBACAABB AAAABABAAB BBABAA	Gh_Golib (6)

SG subgroup

(*) Represents identity with the consensus sequence at that nucleotide position

[a] Consensus sequence from A-genome *G. herbaceum*

Table 4 A haplotype table for D_t-genome *MIC-3* genes of *G. hirsutum* (AD)

SG	Name	32 T[a]	88 M	89 T	113 G	162 G	164 A	392 G	394 K	405 A	429 C	437 G	439 A	482 A	509 A	510 G	520 :	525 :	560 G	565 A	580 T	615 T	646 T/:	652 T/:	665 Y/:	674 T	681 C	Haplotypes	Contig IDs[b]
10	Hap_33	C	T	G	*	A	G	A	*	T	*	T	C	*	*	T	C	C	*	*	*	C	*	C	G	*	A	AABACABBAA ACBBABBAAA BBABAABA	Gh_TM1 (9), Gh_Sisoline (10), Gh_ST213 (9), Gh_M8 (10), Gh_REDM (12), Gh_Golib (13), Gh_Clevewilt (11), Gh_M240 (8), Gh_M315 (12)
	Hap_34	C	T	G	*	A	G	A	*	T	*	T	C	*	*	T	C	C	*	*	*	C	C	C	G	*	A	AABACABBAA ACBBABBAAA BBABAABA	Gh_M8 (11)
	Hap_35	C	T	G	*	A	G	A	*	T	*	T	C	*	*	T	C	C	*	*	*	C	*	C	G	*	*	AABACABBAA ACBBABBAAA BBABAABA	Gh_Golib (14)
	Hap_36	*	T	G	*	A	G	A	*	T	*	T	C	*	*	T	C	C	*	*	*	C	*	C	G	*	A	BABACABBAA ACBBABBAAA BBABAABA	Gh_Clevewilt (12)
	Hap_37	*	T	G	C	C	*	A	*	T	*	T	C	*	*	T	C	C	*	*	*	C	*	C	G	*	A	BABABBABAA ACBBABBAAA BBABAABA	Gh_ST213 (10)
	Hap_38	C	T	G	A	C	*	A	*	T	*	T	C	*	*	T	C	C	*	*	*	C	*	C	G	*	A	AABAABABAA ACBBABBAAA BBABAABA	Gh_M315 (13)
11	Hap_39	*	*	*	*	*	*	*	*	*	T	*	C	*	*	*	C	C	A	G	X	*	*	C	G	*	A	BBABCCAABB CCABBABBAA AABAABBA	Gh_M8 (12)
	Hap_40	*	*	*	*	*	*	*	C	*	*	*	C	G	*	*	A	T	A	G	C	*	*	*	*	A	*	BBABCCAABB CAABAABBBA AABAABAB	Gh_M315 (15)
12	Hap_41	*	*	*	*	*	*	C	*	*	*	*	*	*	G	*	:	:	*	*	*	*	:	:	:	*	*	BBABCCAABB BBAAAAABAB ABABBBBB	Gh_TM1 (10), Gh_Sisoline (11), Gh_ST213 (11), Gh_REDM (13), Gh_Golib (15), Gh_Clevewilt (13), Gh_M240 (9), Gh_M315 (14)
	Hap_42	*	*	*	*	*	*	C	*	*	*	*	*	*	G	*	:	:	*	*	*	*	*	*	*	A	*	BBABCCAABB BBAAAAABAB ABABBBAB	Gh_Golib (16)

SG subgroups

(*) Represents identity with the consensus sequence at that nucleotide position

(:) Missing bases at regions of insertions or deletion events

[a] Consensus sequence from D- genome *G. raimondii*

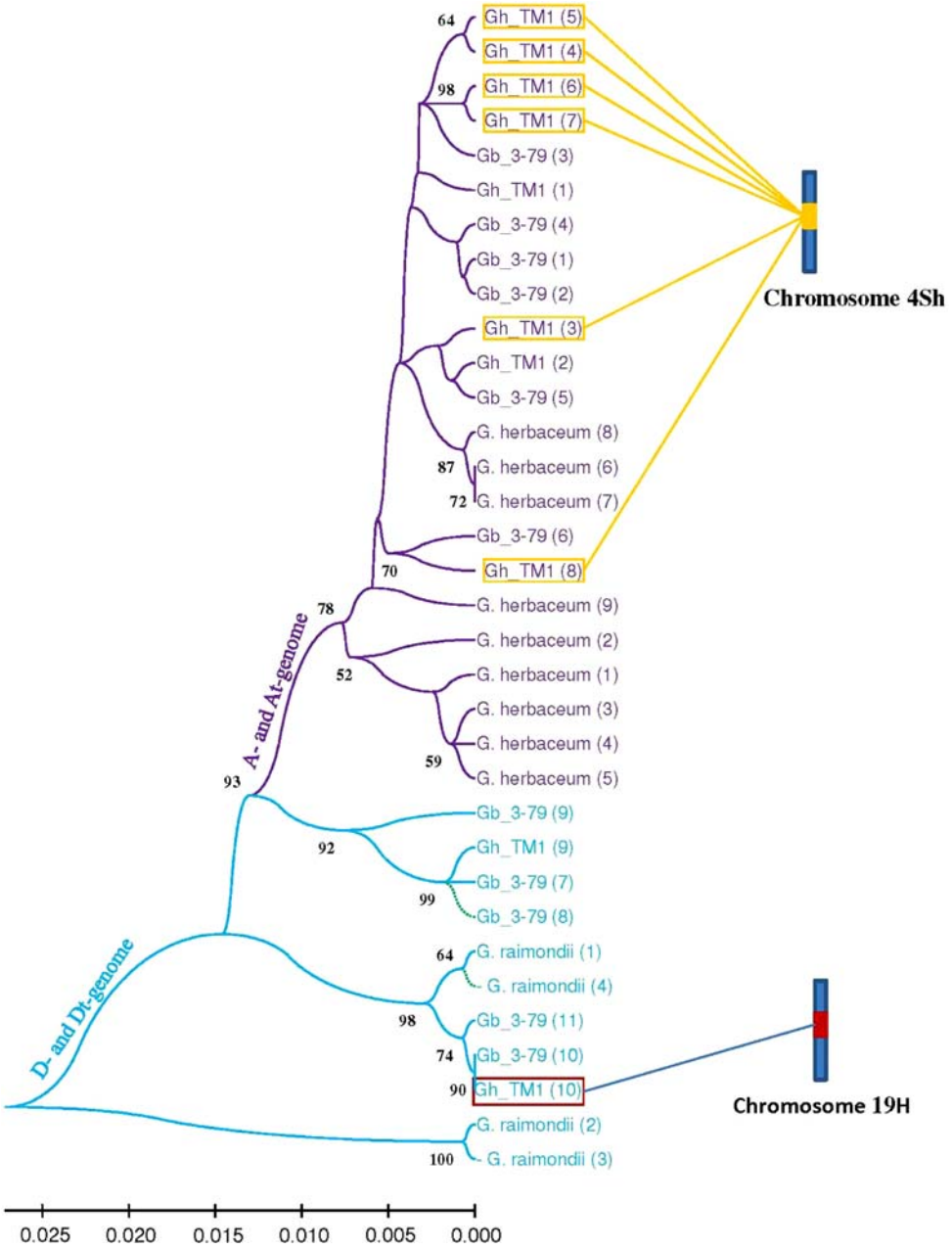

Fig. 3 Curved neighbor-joining phylogenetic tree derived from the sequences of *MIC-3* genes of *G. hirsutum (TM-1), G. barbadense (Pima 3-79)* and two putative diploid ancestors of AD cottons. Branch length, expressed as bootstrap support values (>50%), chromosomal positions and genome origin are shown and color coded for simplicity

members, however, helped us to design two different GSP primers (not polymorphic between TM-1 and 3-79), that were polymorphic between TM-1 and *G. mustelinum MIC-3* gene members, showing the value of utilizing *G. mustelinum*-derived cytogenetic stocks in deletion analysis (Table 5). Tertiary monosomic F_1 interspecific hybrids NTN12-19 and NTN10-19 differentially lacked the TM-1 allele specific to this SNP marker, but not the *G. mustelinum* allele. The NTN12-19 plant is segmentally deficient in parts of TM-1 chromosomes 12 and 19, whereas NTN10-19 is segmentally deficient in parts of the TM-1 chromosomes 10 and 19. Together, they delimit one of the D_t-genome *MIC-3* member (*MIC-3*-10) of TM-1 to chromosome 19. In contrast, primary monosomic interspecific hybrids, deficient in *G. hirsutum* chromosomes 10 and 12, respectively, contained the TM-1 and *G. mustelinum* GSP *MIC-3* markers. The result indicated that this D_t-genome-derived *MIC-3* member is not located in either of these chromosomes and further supports localization to chromosome-19. The other hypoaneuploid F_1 plants that lacked other chromosomes contained both TM-1 and 3-79 alleles for this specific SNP marker. Although a primary monosomic for chromosome-19 is not yet available for *G. hirsutum*, the collective results, nevertheless, clearly indicate that this D_t-genome *MIC-3* member is located on chromosome 19 of *G. hirsutum* (Table 5).

Table 5 Candidate gene (polymorphic between TM-1 parent with 3-79 and *G. mustelinum* parent, respectively) primers for *MIC-3* genes of TM-1 showing their chromosomal locations

No.	Primer name	GSP primer (5′–3′)	TM1/3-79/ *G. mustelinum*	Alleles	Chromosomal location	Genome
1	MIC3-87[a]	CCTTGCGAGCTTGATGTATTCGAA	A/T/T	TM1_6, TM1_7	4 short	A_t
2	MIC3-118	GCGCAAATGGAGTGGGTAATGAAT	T/G/G	TM1_6, TM1_7	4 short	A_t
3	MIC3-190	CGGAGGCCTGGTGTATGATGTT	T/G/G	TM1_6, TM1_7	4 short	A_t
4	MIC3-102	TTCGAATTACCAAGGTGCTCC	A/C/G/G	TM1_4, TM1_5, TM1_8	4 short	A_t
5	MIC3-331	GGTGGCTACTAAGATCATTCCGA	T/C/C	TM1_8	4 short	A_t
6	MIC3-383	GAGCATCCTTCAGCCCTACGA	C/T/T	TM1_4, TM1_5, TM1_6, TM1_7	4 short	A_t
7	MIC3-418	CAACTTGCCCTTATCATGTGGGGGTG	T/C/C	TM1_8	4 short	A_t
8	MIC3-517	GGATGGAGCAATATATGATG	A/G/G	TM1_1, TM1_3	4 short	A_t
9[b]	MIC3-222	GTTAGGGTGAAGTGGATTAT	T/T/C	TM_10	19	D_t
10[b]	MIC3-551	CACAGCATAAACCGATAGAGAT	T/T/G	TM_10	19	D_t

[a] Number in primer names refers to genome-specific polymorphism (GSP) positions in the multiple sequence alignment of *MIC-3* genes of *Gossypium* species

[b] These two primers were designed from the same haplotype or putative locus

Confirmation of A_t-genome *MIC-3* clustering

To confirm the clustering of *MIC-3* genes on the short arm of chromosome four, we utilized forward and reverse primer pairs from the conserved open reading frame of the A_t-genome to amplify the intergenic regions of the clustered genes using genomic DNA as a template (Fig. 4). A PCR product was expected if *MIC-3* genes were located closely side by side; otherwise the specific designed primer pair should not generate an amplicon of any type. Interestingly, TM-1 yielded a very strong polymorphic band of 3 kb and a weak monomorphic band of 5 kb. The 3-kb band size is similar to that of *G. herbaceum*, suggesting that this intergenic region of the tetraploid cotton descends from an A-genome diploid ancestral species. The 5-kb bands of TM-1 (weak) and 3-79 (strong) were of similar size as the *G. raimondii* band, suggesting that they originated from an ancestral D-genome, similar to that of *G. raimondii*. *G. barbadense* L. 3-79 has one strong band of 5 Kb in size, similar to the band of *G. raimondii*, a relative of the putative D-genome diploid ancestral species. Results also suggested that intergenic regions were almost of the same sizes.

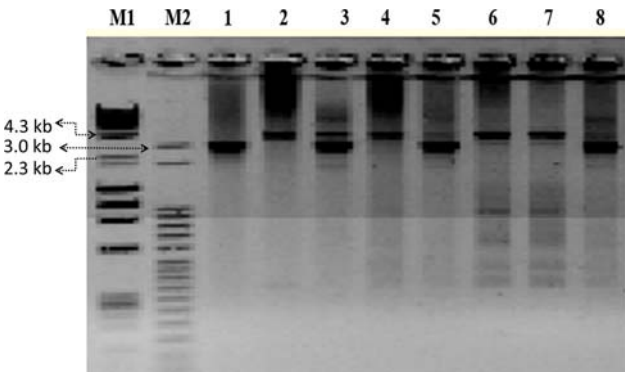

Fig. 4 Agarose gel showing the amplified fragments used to test chromosomal localization of intergenic regions of *MIC3* gene cluster. Lanes in the gel from *left* to *right*: *M1* marker λ Hind III/φX- Hae III, *M2* 1 kb marker, *1 G. herbaceum* A-genome, *2 G. raimondii* D_5-genome, *3 G. hirsutum* L var. TM1, *4 G. barbadense* L var. 3-79, *5* F_1 (TM1x3-79), *6* F1 (TM1x3-79) monosomic for chromosome-4 (lacking TM-1 chromosome-4), *7* F1 (TM1x3-79) monotelodisomic Te04Lo that lacks all or most of the short arm of chromosome-4 (arm 4sh), *8* F_1 (TM1x3-79) monotelodisomic Te04sh that lacks all or most of the long arm of chromosome-4 (arm 4Lo)

In the deletion analysis, results showed that the polymorphic 3-kb TM-1 band was absent from two interspecific F_1 hypoaneuploid plants, the monosomic (H04) missing TM-1 chromosome-4 and also the monotelodisomic (Te04Lo) F_1 plant deficient in TM-1 chromosome arm 4sh. In contrast, the TM-1 3-kb band was differentially present in the monotelodisomic (Te04sh) hybrid deficient in TM-1 chromosome arm 4Lo, as well as in all other monosomic and monotelodisomic plants, which showed typical F_1 heterozygous banding patterns including both 3 and 5-kb fragments. The results indicate that all of the *MIC-3* gene family members specific to A_t-genome are located on the short arm of chromosome four (4sh), and the putative origin of this amplicon from the A-genome diploid ancestral species of *G. herbaceum*. The association of the 3-kb amplicon of intergene region with chromosome arm 4sh is congruent with the localization of nine A_t-genome *MIC-3* SNP markers specific to the same arm.

It was not possible to localize the 5-kb size band of D_t-genome *MIC-3* intergenic region using the aneuploid cytogenetic set of 3-79 because of the monomorphic phenotype of this band between TM-1 and 3-79. However, one can assume that it would only be possible to amplify the D_t-subgenome *MIC-3* intergenic region with the specially

designed primer pair, if *MIC-3* D_t-subgenome members of approximately 721-bp sizes are clustered on a chromosome. Furthermore, the similarity with the 5-kb size band of *G. raimondii* in gel analysis suggested the origin of this band from the putative diploid D_t-subgenome ancestral species of *G. raimondii*, thereby indirectly supporting the clustering of these gene members.

Discussion

Although clustering of genes is well known and has been reported in other crops, this is to our knowledge the first evidence of such clustering of a major pest (RKN) resistance gene family in the cotton genome. SNP-based characterization of the RKN-resistance *MIC-3* gene family revealed the existence of islands containing clustered *MIC-3* genes. These findings are significant because the *MIC-3* genes have several unique and desirable features. Based on sequence analysis of diploid and tetraploid cotton species, *MIC-3* is unique and expressed only in cotton root (Wubben et al. 2008; Zhang et al. 2002). The function(s) of *MIC-3* is unknown (Wubben et al. 2008; Callahan et al. 2004; Zhang et al. 2002). As a small gene (721 bp) with only two exons, *MIC-3* should be amenable to genetic manipulation at the molecular level. The clustering feature of *MIC-3* genes may have broader evolutionary significance with the possibility of allowing *MIC-3* genes to recombine, creating new variants in response to ever-evolving RKN, provided that *MIC-3* genes have a role against RKN infection. Positive correlation of *MIC-3* protein accumulation with RKN resistance in selected cotton lines is a recent discovery (Wubben et al. 2008; Callahan et al. 2004; Zhang et al. 2002).

The intralocus and interlocus duplication mechanism in *MIC-3* genes offers a possible model for studying the origin, evolution and genetics of gene duplication. Gene duplications are one of the major driving forces in the evolution of genomes and genetic systems. In higher plants, a very high percentage of the genome is made of duplicated segments including 80% in the Arabidopsis genome and 77% in the rice genome (Hu and Wise 2008). Our results suggest that multiple duplication events occurred before polyploidization in *MIC-3* in several diploid cotton species. Previous investigations have revealed gene duplication events affecting many low-copy genes in the cotton genome (Pfeil et al. 2004; Udall et al. 2006; Adams and Wendel 2005; An et al. 2008). Hovav et al. (2008) reported that duplicated genes provided temporal partitioning of gene expression in fiber development in tetraploid cotton. Pfeil et al. (2004) reported that early polyploidy events were not the only cause of gene duplication in the cotton genome. Our results also indicate the occurrence of independent duplications in *MIC-3* genes in diploid and tetraploid cotton species and in A_t- and D_t-genomes of tetraploid cotton. A previous study reported that the tandem duplications that have occurred in the Brassica lineage did so before the divergence of *B. rapa* and *B. oleracea*, but after the separation of Brassica and Arabidopsis from a common ancestor (Mayerhofer et al. 2005).

SNPs are highly abundant and efficient diagnostic markers for candidate gene manipulation. Efficient SNP marker discovery must include ways to distinguish differences between paralogous and orthologous loci across genotypes. It is difficult to identify haplotypes that can distinguish allelic differences at a single locus in polyploid species especially considering multiple duplication events both in inter- and intrachromosomal regions, such as in *MIC-3* genes. We used a specific strategy based on the overall methods of our previous studies with other low copy genes to identify haplotypes in tetraploid cotton (An et al. 2007, 2008). We analyzed the sequences from *G. hirsutum* using the NJ cluster analysis method. Sequences of *G. hirsutum* from an individual clade of the phylogenetic tree were aligned and compared to detect putative SNPs. The unique combinations of SNPs in a sequence within a clade of the phylogram were considered as haplotypes.

The discovery of putative haplotypes of *MIC-3* gene family helped to identify the variants for detecting individual *MIC-3* gene family members, and thus to determine if individual *MIC-3* gene members have specific functional roles in the response of cotton against root diseases. The main evolutionary forces contributing to formation of haplotype blocks include selection, recombination, mutation and population structure. Significant differences exist in the haplotype blocks and number between A_t- and D_t-genomes of *G. hirsutum* (Tables 3, 4), suggesting different patterns of evolution at the subgenomes following polyploidization in the tetraploid cotton. Previous studies reported a rapid evolution of R (resistance) genes by the high degree of haplotype diversity within a species, as shown in sequence analyses of *RPS2* and *Rpm1* alleles from different ecotypes of *Arabidopsis thaliana* (Caicedo et al. 1999; Stahl et al. 1999).

Clustering

There are many resistance (R) genes in plant hosts and some confer a unique specificity to various pathogen isolates (Deyoung and Innes 2006). Many of these R genes are physically clustered as complex loci that exhibit gene-family specificities (Baumgarten et al. 2003; Wei et al. 1999; Parniske et al. 1999). Results revealed that putative A_t-genome *MIC-3* members cluster to chromosome arm

4sh, and D_t-genome member to chromosome 19 in tetraploid cotton. PCR amplification of intergenic regions of multiple sizes from the *MIC-3* gene indicates that many of these genes are very closely linked, both physically and recombinationally. The pattern of relationship among *MIC-3* sequences, coupled with the clustering evidence, suggested that the intrachromosomal duplication was at a small scale, e.g., segmental or single gene, rather than genome-wide.

Other disease resistance genes also cluster to linked genomic locations in other plant species (Baumgarten et al. 2003; Parniske et al. 1999; Wei et al. 1999). For example, Baumgarten et al. (2003) found that 81.1% of nucleotide-binding site, leucine-rich repeat (NBS-LRR) duplication events resulted in gene copies on the same chromosome in *Arabidopsis thaliana*. Parniske et al. (1999) reported that the majority of members of a resistance gene family *Cf-9* of the pathogen *Cladosporium fulvum* were located on the short arm of chromosome 1 in tomato. Milligan et al. (1998) reported that *Prf* and *Mi-1.2* genes in tomato share several structural motifs, including an NBS-LRR region, which are characteristic of a plant protein family required for resistance against viruses, bacteria, fungi and nematodes. The *MIC-3* gene family members are among the clustered gene families that are not directly associated with any specific motif or NBS-LRR-like domain of other plant species. Plant proteins belonging to the NBS-LRR family are normally used for pathogen detection and some NBS-LRR proteins directly bind with pathogen proteins (Deyoung and Innes 2006). Although there is no report that the *MIC-3* gene family has direct connection with any other RKN resistance genes, previous studies have revealed that *MIC-3* genes are intimately involved in the resistance response against RKN in cotton (Wubben et al. 2008; Callahan et al. 2004; Zhang et al. 2002).

Williams and Bowles (2004) suggested that genes involved in a specific metabolic pathway that needs coordinated regulation are found to be clustered in some higher eukaryotes. Wisser et al. (2005) reported that genes implicated in defense often show altered expression in response to pathogen challenge in rice and found several defense-associated genes belonging to common biosynthetic pathway clustered in a segment of a chromosome. Lee and Sonnhammer (2003) reported that the orientation of gene pairs in clustered genes played a significant role in the coexpression of neighboring genes in *Arabidopsis*. The balance between diversification and conservation within the *HCR9* gene family is accommodated by the physical distribution of clustered sequences and exploitation of the standard recombination machinery (Song et al. 1997). Identification of *MIC-3* variants provides a tool for the study of whether any *MIC-3* members have biological roles like other clustered resistant genes in plant species.

The clustering of *MIC-3* genes associated with resistance has several other generic biological implications. The duplication events of *MIC-3* genes provided a mechanism against the loss of this gene in evolution and created some unique avenues by clustering, so that large variants of the *MIC-3* gene family can be recovered from small physical areas by recombination. This could provide a buffer against selection pressure from a broad range of pest and pathogens in the future and may have played a role in the evolutionary dynamics of *MIC-3* genes after their origin in cotton. Clustering, duplicated events and changes in *MIC-3* genes may provide the potential needed for the evolution of *MIC-3* members to function like a PR gene family, as suggested by Wubben et al. (2008), with the potential against a broad range of pests or pathogens.

The results of this study showed that the locations of the *MIC-3* genes were not on the homologous chromosomes of the A_t and D_t genomes in tetraploid cotton. Chromosome 19 (D_t-genome) is homologous to chromosome 5 (A_t-genome), and chromosome 4 (A_t subgenome) is homologous to chromosome 22 (D_t-subgenome) in tetraploid cotton (A_tD_t). Recently, Guo et al. (2007) reported the evidence of significant similarities among these four chromosomes in molecular maps, suggesting that significant resemblance exists between chromosome 4 and 19 where the *MIC-3* genes are located. Partial homology of these chromosomes is suggested by cytogenetic and linkage mapping evidence of an ancestral genome translocation that rearranged chromosomes 4 and 5 of the tetraploids relative to the ancestral A- and D-genomes (Menzel and Brown 1954). Positions of redundant molecular markers have more recently revealed that chromosomes 4 and 5 are collectively, but not wholly, homologous to chromosomes 19 and 22 (Gutierrez et al. 2009; Guo et al. 2007).

Recently, Ynturi et al. (2006) reported the chromosomal locations of two RKN resistance loci, one on the long arm of chromosome 11 and one on the short arm of chromosome 14 in cotton. Our results showed that *MIC-3* is not directly associated with either of these two chromosomes. Accordingly, it can be hypothesized that other genes are also involved in RKN resistance; that *MIC-3* is a regulatory gene family with indirect effect on RKN resistance genes or that *MIC-3* is related with some unknown function of broad spectrum resistance genes that are highly expressed in infected resistant lines. We have cloned and sequenced the intergenic regions and are currently studying the upstream regulatory regions of *MIC-3* genes to understand the molecular mechanism on the regulation of *MIC-3* genes.

Wubben et al. (2008) suggested that cotton resistance to RKN involves novel defense signaling pathways and that *MIC-3* genes represent a group of root-specific defense-related genes in cotton. They observed in a study of cDNA

abundance that *MIC* expression was not induced by wounding or by reniform nematode infection during a compatible interaction. *MIC-3* expression was also undetectable in leaves undergoing a hypersensitive response to *Xanthomonas campestris* infection in cotton. Their experiments with expression levels of other known defense genes (*PR10*, *ERF5*, *CDNS*, *LOX1*, *POD4*, *POD8*) in resistant and susceptible cotton roots demonstrated that RKN infection specifically increased the induction of *MIC-3* in resistant roots and not other common defense signaling pathways. It has been reported that activity levels of the enzymes, phenylalanine ammonia-lyase and anionic peroxidase, which are induced early in the resistance response to many pathogens, also increase in a resistant tomato line after nematode infection (Brueske 1980; Zacheo et al. 1993).

Cotton is a major crop in diverse ecological conditions including parts of the African tropics, Australia, China, Egypt, India, Mexico, Pakistan, the Sudan, former Soviet Union countries (including Uzbekistan), the USA and warmer regions of Central and South America. Root diseases and pests including RKN have always been major problems in cotton growing areas. Given the devastating potential of RKN disease in cotton, the exclusive occurrence of *MIC-3* genes among the diploid and tetraploid cotton species and their involvement in the resistance response with a group of root-specific genes (Wubben et al. 2008), it can be postulated that the ancestral *MIC-3* gene evolved due to a serious root pest or disease similar to RKN in an early period of cotton evolution and these genes were maintained in the genome as the guardian of *Gossypium* species with unique features of gene duplication and clustering to buffer against evolving cotton root pathogens and pests.

Success in the breeding program for RKN will depend on the identification of useful resistance sources and understanding the inheritance of resistance. BLAST results of *MIC-3* sequences from the diploid and tetraploid cotton species in our research confirmed the uniqueness of the *MIC3* gene that has only been found to be expressed in cotton roots (Wubben et al. 2008). We detected: (1) SNP markers and haplotypes associated with the *MIC-3* gene family; (2) the first molecular evidence for the clustering of genes associated with a major pest resistance response in cotton; and (3) the chromosomal locations of the *MIC-3* gene family. *MIC-3*-derived candidate gene markers should be very useful in future molecular mapping, functional gene analysis and marker-assisted selection: an important endeavor considering the economic impact of RKN diseases on cotton.

Acknowledgments The authors gratefully acknowledge the reviews and helpful suggestions made by David Fang, USDA/ARS, Stoneville, MS, and Martin Wubben, Frank Callahan, Mike McLaughlin and Marilyn Warburton USDA/ARS, Mississippi State, MS. We are grateful to the ARS-FSU Scientific Cooperation Program under the office of International Research Programs, USDA-ARS for the financial support of this research. This paper was approved for publication as Journal Article No. J-11611 of the Mississippi Agricultural and Forestry Experiment Station, Mississippi State University.

References

Adams KL, Wendel JF (2005) Allele-specific, bidirectional silencing of an alcohol dehydrogenase gene in different organs of interspecific diploid cotton hybrids. Genetics 171:2130–2142

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

An C, Saha S, Jenkins JN, Scheffler BE, Wilkins TA, Stelly D (2007) Transcriptome profiling, sequence characterization, and SNP-based chromosomal assignment of the EXPANSIN genes in cotton. Mol Genet Genomics 278(5):539–553

An C, Saha S, Jenkins JN, Ma DP, Scheffler BE, Kohel RJ, Yu JZ, Stelly DM (2008) Cotton R2R3-MYB (*Gossypium* spp.) transcription factors SNP identification, phylogenomic characterization, chromosome localization, and linkage mapping. Theor Appl Genet 116(7):1015–1026

Baumgarten A, Cannon S, Spangler R, May G (2003) Genome-level evolution of resistance genes in *Arabidopsis thaliana*. Genetics 165:309–319

Beasley JO (1940) The origin of American tetraploid *Gossypium* species. Am Nat 74:285–286

Beasley JO (1942) Meiotic chromosome behavior in species, species hybrids, haploids and induced polyploidy of *Gossypium*. Genetics 27:25–54

Brown MS (1980) The identification of chromosomes of *Gossypium hirsutum* L. by means of translocations. J Hered 71:266–274

Brueske CH (1980) Phenylalamine ammonia-lyase activity in tomato roots infected and resistant to the root-knot nematode, *Meloidogyne incognita*. Physiol Plant Pathol 16:409–414

Caicedo AL, Schaal BA, Kunkel BN (1999) Diversity and molecular evolution of the *RPS2* resistance gene in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 96:302–306

Callahan FE, Jenkins JN, Creech RG, Lawrence GW (1997) Changes in cotton root proteins correlated with resistance to root-knot nematode development. J Cotton Sci 1:38–47

Callahan FE, Zhang X-D, Ma D-P, Jenkins JN, Hayes RW, Tucker ML (2004) Comparison of MIC-3 protein accumulation in response to root-knot nematode infection in cotton lines displaying a range of resistance levels. J Cotton Sci 8:186–190

Deyoung JB, Innes RW (2006) Plant NBS-LRR proteins in pathogen sensing and host defense. Nat Immunol 7:1243–1249

Guo W, Cai C, Wang C, Han Z, Song X, Wang K, Niu X, Wang C, Lu K, Shi B, Zhang T (2007) A microsatellite-based, gene-rich linkage map reveals genome structure, function, and evolution in *Gossypium*. Genetics 176:527–541

Gutierrez OA, Stelly DM, Saha S, Jenkins JN, McCarty JC, Raska DA, Scheffler BE (2009) Integrative placement and orientation of non-redundant SSR loci in cotton linkage groups by deficiency analysis. Mol Breed 23:693–707

Hovav R, Chaudhary B, Udall JA, Flagel L, Wendel JF (2008) Parallel domestication, convergent evolution and duplicated gene recruitment in allopolyploid cotton. Genetics 179:1725–1733

Hsu CY, An C, Saha S, Ma DP, Jenkins JN, Scheffler B, Stelly DM (2008) Molecular and SNP characterization of two genome-specific transcription factor genes GhMyb8 and GhMyb10 in cotton species. Euphytica 159:259–273

Hu P, Wise RP (2008) Diversification of Lrk/Tak kinase gene cluster is associated with subfunctionalization and cultivar-specific transcript accumulation in barley. Funct Integr Genomics 8:199–209

Lee JM, Sonnhammer EL (2003) Genomic gene clustering analysis of pathways in eukaryotes. Genome Res 13(5):875–882

Mayerhofer R, Wilde K, Mayerhofer M, Lydiate D, Bansal VK, Good AG, Parkin IA (2005) Complexities of chromosome landing in a highly duplicated genome: toward map-based cloning of a gene controlling blackleg resistance in *Brassica napus*. Genetics 171:1977–1988

Menzel MY, Brown MS (1954) The significance of multivalent formation in three-species *Gossypium* hybrids. Genetics 39:546–557

Menzel MY, Richmond KL, Dougherty BJ (1985) A chromosome translocation breakpoint map of the *Gossypium hirsutum* genome. J Hered 76:406–414

Milligan SB, Bodeau J, Yaghoobi J, Kaloshian I, Zabel P, Williamson VM (1998) The root-knot nematode resistance gene Mi from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. Plant Cell 10(8):1307–1319

Parniske M, Wulff BH, Bonnema G, Thomas CM, Jones DA, Jones JDG (1999) Homologous of the Cf-9 disease resistance gene (Hcr9s) are present at multiple loci on the short arm of tomato chromosome 1. Mol Plant Microbe Interact 12(2):93–102

Pfeil BE, Brubaker CL, Craven LA, Crisp MD (2004) Paralogy and orthology in the Malvaceae *rpb2* gene family: investigation of gene duplication in *Hibiscus*. Mol Biol Evol 21:1428–1437

Robinson AF, Bell AA, Dighe ND, Menz MA, Nichols RL, Stelly DM (2007) Introgression of resistance to nematode *Rotylenchulus reniformis* into upland cotton (*Gossypium hirsutum*) from *Gossypium longicalyx*. Crop Sci 47:1865–1877

Saitou M, Nei N (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Song WY, Pi LY, Wang GL, Gardner J, Holsten T, Ronald PC (1997) Evolution of the rice Xa21 disease resistance gene family. Plant Cell 9(8):1279–1287

Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J (1999) Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. Nature 400:667–671

Stelly DM, Saha S, Raska DA, Jenkins JN, McCarty JC, Gutierrez OA (2005) Registration of 17 Upland (*Gossypium hirsutum*) germplasm lines disomic for different *G. barbadense* chromosome or arm substitutions. Crop Sci 45:2663–2665

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 40. Mol Biol Evol 24:1596–1599

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Udall JA, Swanson JM, Nettleton D, Percifield RJ, Wendel JF (2006) A novel approach for characterizing expression levels of genes duplicated by polyploidy. Genetics 173(3):1823–1827

Wei F, Gobelman-Werner K, Morrol SM, Kurth J, Mao L, Wing R, Leister D, Schulze-Lefert P, Wise RP (1999) The *Mla* (powdery mildew) resistance cluster is associated with three NBS-LRR gene families and suppressed recombination within a 240-kb DNA interval on chromosome 5S (1HS) of barley. Genetics 153:1929–1949

Wendel JF, Cronn C (2003) Polyploidy and the evolutionary history of cotton. Adv Agron 78:139–186

Williums EJB, Bowles DJ (2004) Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. Genome Res 14:1060–1067

Wisser JR, Sun Q, Hulbert SH, Kresovich S, Nelson RJ (2005) Identification and characterization of regions of the rice genome associated with broad-spectrum, quantitative disease resistance. Genetics 169:2277–2293

Wubben MJ, Callahan FE, Hayes RW, Jenkins JN (2008) Molecular characterization and temporal expression analyses indicate that the MIC (*Meloidogyne*-induced cotton) gene family represents a novel group of root-specific defense-related genes in upland cotton (*Gossypium hirsutum* L.). Planta 228:111–123

Yang SS, Cheung F, Lee JJ, Ha M, Wei NE, Sze SH, Stelly DM, Thaxton P, Triplett B, Town CD, Chen JZ (2006) Accumulation of genome-specific transcripts, transcription factors and phytohormonal regulators during early stages of fiber cell development in allotetraploid cotton. Plant J 47:761–775

Yik CP, Birchfield W (1984) Resistant germplasm in *Gossypium* species and related plants to *Rotylenchulus reniformis*. J Nematol 16:146–153

Ynturi P, Jenkins JN, McCarty JC, Gutierrez OA, Saha S (2006) Association of root-knot nematode resistance genes with simple sequence repeat markers on two chromosomes in cotton. Crop Sci 46:2670–2674

Zacheo G, Orlando C, Bleve-Zacheo T (1993) Characterization of anionic peroxidases in tomato isolines infected by *Meloidogyne incognita*. J Nematol 25:249–256

Zhang XD, Callahan FE, Jenkins JN, Ma D-P, Karaca M, Saha S, Creech RG (2002) A novel root-specific gene, *MIC-3* with increased expression in nematode-resistant cotton (*Gossypium hirsutum* L.) after root-knot nematode infection. Biochim Biophys Acta 1576:214–218