

Transcription Factors in Cotton

Jie Chen¹, Mingzhou Song², and Jinfa Zhang¹



(1)Plant and Environmental Sciences, New Mexico State University, Las Cruces, NM, (2) Computer Science, New Mexico State University, Las Cruces, NM

Introduction

Transcription factors (TFs) regulate gene expressions via interacting with regulatory sequences located in the promoter regions, and they are classified into ~60 families based on their DNA-binding domains. In this study, several recently sequenced genomes of cotton species were analyzed to identify TF-coding genes for a phylogenetic and comparative structural genomic analysis. Single nucleotide polymorphisms (SNPs) are being further identified, using one of the homologous TF-coding gene families as an example. A genome-wide gene expression study will be performed to reveal their regulations associated with abiotic and biotic stresses, fiber development and male fertility in cotton. Table 1 A comparison among the three versions of PlantTFDB database(Jin JP et al., 2013)

PlantTFDB	Version 1.0	Version 2.0	Version 3.0
Species	22	49	83
Species with genome sequences	5	28	67
Species without genome sequences	17	21	16
TF family	64	58	58
TF number	26,402	53,574	129,288
TF prediction server	No	No	Yes

	1 1p	20	3,0	4 <mark>0</mark>	5 <mark>0</mark>	60	7ρ	
Cotton A 0	ATGTCAAACAATC	CAAAATTCGAA	TCTTCCCTTC	JATAGATATA	AGGAGGAGGTT	FCATCCAGAA	CAAGTTCGCCG	STA
Gblcdslsca	ATGTCAAACAATC	CAAAATTCGAA	TCTTCCCTT	GATAGATATA	GGAGGAGGT	FCATCCAGAA	CAAGTTCGCCC	GTA'
COBAD AA	ATGTCAAACAATC	~AAAATTCGAA	TCTTCCCTTC	TATAGATATA	GGAGGAGGT	TCATCCAGAA	CAAGTTCGCCC	3ΤΔ
OL ANTOAT	ATGTCAAACAATC	- A A A A TTCCAA	TCTTCCCTTC			TCATCCAGAA	CAAGTTCGCCC	
Gn_A05G17	AIGICAAACAAIC	AAAATICGAA	TCHCCCT	JATAGATATA	GGAGGAGGI	ICAICCAGAA	CAAGIICGCCC	AIC
	80 90	0 100) 11			130		150
Cotton A 0	TTTTCCACCTGGT	TGTCGATTTT			2.		GGGAAAAACAA	ATTO
	TTTTCCACCTGGT	TCTCCATTT	ACCCTTCCC	ACCACCACCT	TCTAAACCAT	TATCTCACT	CCCAAAAACAA	
Gb cds sca	TTTTCCACCTGGT	TOTCOATTT	ACCUTICUS	AGGAGGAGCI	TCTARACCA	TATCTCACT	GGGAAAAACAA	
GOBAR_AA	TTTTCCACCTGGT	IGICGAIIII	ACCCTTCCGA	AGGAGGAGCI	ICIAAACCA	TATCICACT	GGGAAAAACAA	8110
Gh A05G17	TTTTCCACCTGGT	<i>ITGTCGATTTT</i>	ACCCTTCCG/	AGGAGGAGCT	TCTAAACCA	ΓΤΑΤCTCACT	GGGAAAAACAA	ATT
	164 rg-Ser-	G1170	180	190	200	210	220	
O-H A 0	164 rg-Ser-	COCTONTOCO	TACCCATCC	- A T T T T A T C A	LCCCAACTTA	ATTTCTACCA	TTATCACCCC	гсти
	TEGEEGEEGAT	JCGCTGATGCG	TACGGATCCC		GGGAACTTAA	ATTIGIACOA	TATGAGECGI	
Gb cds sca	TCGCCGCCGATAG	CGCIGAIGCG	TACGGATCC		GGGAACTIAA	ATTIGTACGA	ITATGAGCCGI	
GOBAR AA	TCGCCGCCGATAG	GCGCTGATGCG	TACGGATCC	GATTTTTATCA	AGGGAACTTA/	ATTTGTACGA	TTATGAGCCG	ГСТО
Gh A05G17	TCGCCGCCGAT <mark>G</mark> G	GCGCTGATGCG	TACGGATCC	GATTTTATCA	GGGAACTTA/	ATTTGTACGA	TTATGAGCCG1	ГСТО
		100 T 100 T					200	
	240	250	260	270	280	290	340	
	TTACCGGAGGGCG	CGIGCIIIGI	GCACGGATA	AAGGGGAGA	AGAAGGCAC	IGGIIIGII	ATACCAAGAGI	ΙΑΑΑ
Gb cds sca	TTACCGGAGGGCG	SCGTGCTTTGT	GCACGGATAT	FAAGGGGAGA	AGAAGGCACT	rggttttgtt	ATACCAAGAGI	ΓΑΑΑ
GOBAR AA	TTACCGGAGGGGG	SCGTGCTTTGT	GCACGGATA 7	FAAGGGGAG	AGAAGGCACT	<i>FGGTTTTGTT</i>	ATACCAAGAG1	ГААА
Gh A05G17	TTACCGGAGGGCG							
			The second se					
	310 320		Phe-Val340	27.03			370 38	
Cotton A 0	TGGGAGGGGGGGGG							
Gblcdslsca	TGGGAGGGGGAGG	JAGAAGGGCAA	AGGGTGGGGT	<i>ITTGGAGGA</i>	AATTGGAAA/	AGTTAGGGAT	GTGTTTGATGO	GAG
GOBAR AA	TGGGAGGGGGAGG	GAGAAGGGCAA	AGGGTGGGG	TTGGAGGA	AATTGGAAA	AGTTAGGGAT	GTGTTTGATGO	GAGO
· 영상 전자 · 영상 전 전 전 사회 · 영상 · 전 · 영상 · · · · · · · · · · · · · · · · ·	TGGGAGGGGGGAGG							
GII_A05G17		Colorado						
	390 Gly-Gl		410		430		450	46
Cotton A 0	ATGTCAAGTTGGG	ACTAGAACT	AAATTTGTA	<i>ITTTATGAGA</i>	TGAATTCAG	FTAAGGCTGC	TGTTAGGACTO	
Ghicdeleca	ATGTCAAGTTGGG	GACTAGAACT	AAATTTGTA	ETTTATGAGA	TGAATTCAG	FTAAGGCTGC	TGTTAGGACTO	стт
COBAD AA	ATGTCAAGTTGGG	CACTAGAACT	AAATTTGTAT	TTTATGAGA	TGAATTCAG	TTAAGGCTGC	TGTTAGGACTO	стт
Gn_A05G17	ATGTCAAGTTGGG						IGIIAGGACIC	
	470		<mark>eu</mark> 490			520	530	
Cotton A 0	ATTATGTATGAGT	FATGCAGTTCT	TCACCGCCAC	CAAGGCTTCT	TTTGTGCTA	<i>FGCCGAGTTT</i>	TCATCAAATCI	FCG A
Gblcdslsca	ATTATGTATGAGT	FATGCACTTCT	TCACCGCCA	CAAGGCTTCT	TTTGTGCTA	FGCCGAGTTT	TCATCAAATCI	ECGA
CORAD AA	ATTATGTATGAGT	LATGCACTTCT	TCACCGCCA	AAGGCTTCT	TTTGTGCTAT	FGCCGAGTTT	ΤΓΑΤΓΑΑΑΤΓΙ	FCGA
OUDAR AA	ATTATCTATCACT	TATOCACTICI	TCACCOCCA		TTTGTGCTA	TOCCOACTTT	TCATCAAATCI	
Gn_A05G17	ATTATGTATGAGT	AIGCACIICI	ICACCGCCA		IIIGIGCIA	IGCCGAGITI	ICAICAAAICI	ICGA
	540 550	560	570	580	590	Arg-Glneo	0 610	
Cotton A 0	TAGAAATAGTGTA	ATCAGAGAATG		STGCAGAAGA	9 1112.7 AC 3.2 HO 1111 W 2. DO 121 MIN 1		The state of the second s	TTC
	TAGAAATAGTGTA							
GUBAK AA	TAGAAATAGTGTA	ATCAGAGAAIG	TITTAAGII(JIGCAGAAGA	AACIGIIICA	AGCIATACGA	CACGICGGAAI	
	TAGAAATAGTGTA							
	620 Asp-Val 6 ATGATGGATTTCT	630 6	40	650	660	670	680 Val-As	D690
Cotton A 0	ATGATGGATTTCT	IGAGACGTGAC	ACACTTGAAG	CTGAAATCA	ATGGTGATG	ACTTCACTAA	AGAGCTAGATO	TI
	ATGATGGATTTCT	TCACACCTCAC	ACACTTCAA	CTCAAATCA	ATCCTCATC	ACTTCACTAA	ACACCTACATO	и п
Gb cds sca	AIGAIGGAIIICI	GAGACGIGAC.	ACACTIGAA	JUI GAAATUA	AIGGIGAIGA	ACTICACIAA	AGAGCIAGATO	JAIU
GOBAR_AA	ATGATGGATTTCT	FGAGACGTGAC	ACACTTGAA	JCTGAAATCA	ATGGTGATG	ΑCTTCACTAA	AGAGCTAGATO	SATC
Gh A05G17	ATGTTGGATTTCT	FGAGACGTGAC	ACACTTGAAG	SCTGAAATCA	ATGGTGATG/	ΑCTTCACTAA	AGAGCTAGATO	GAT
A CARDON AND A DECEMBER OF A CONTRACT AND A DECEMBER.								
Cotton A O	700 ATAACAACTAGAC	CTCTTTCTCT	TCCAACTTT	ICACTTTCCT		CACCACACE	TCCCCCATAAT	
	ATAACAACTAGAC		TOCAAGITI		TCAGGCATI		CCCGATAAT	
Gb cds sca	ATAACAACTAGAC	LCIGTTTCTGT	I GCAAGTTT7	GAGTTTCCT	ICAGGCATT		CCCGATAAT	ATC
GOBAR AA	ATAACAACTAGAC	CTGTTTCTGT	TGCAAGTTT	<i>IGAGTTTCCT</i>	TCAGGCAT	sertion/Del	CIONC CGATAAT	ΓΑΤ
Gh_A05G17	ATAACAACTAGAC							
							Asp-Asp 840	
<u>.</u>	780	790	800	810	820			
	AGGGCCTCGGCTA							
Gblcdsisca	AGGGCCTCGGCTA	ACTACCAGTG	AGTTACTTT	CAATTGTAGA	TGAAGATTT	FATAGAGTTG	GACGATCTCAG	GTTC
	AGGGCCTCGGCTA							
	AGGGCCTCGGCTA							
	AUGULTEGULTA	ACTACKOID	AUTACITIC	ALL OLAGA		AIAGAGIIG	UNCONICICAC	arre
		•				1	•	
	1	!						/
FIQ. J A	n example	e of one	homolo	gous ge	ene seau	ience al	ignment	ar
U	An example		homolo	gous ge	ene sequ	ience al	ignment	ar

Objectives

To identify and map the TF genes in cultivated tetraploid cotton species (*Gossypium hirsutum* and *G. barbadense*) and their ancestral diploid species (*G. arboreum* and *G. raimondii*); and

To identify and develop SNP markers for gene and QTL mapping and other studies in cotton.

Materials & Methods

Identification of TF genes: The genomic information for the five genomes of four cotton species (ancestral diploids G. arboreum and G. raimondii, and their tetraploids G. hirsutum, G. barbadense "3-79" G. barbadense "Xinhai 21") were downloaded online and (<u>https://www.cottongen.org</u>, and <u>http://database.chgc.sh.cn/cotton</u>). TF predicted then based genes were on <u>http://planttfdb.cbi.pku.edu.cn</u> (Table 1) using the peptide sequences. Multiple peptide sequences corresponding to the same DNA fragment were considered redundant and only counted once.

Results

Table 2 Statistics of the predicted TFs in cotton

Gossypium species	Total no. genes	No. TFs	TF %
G. hirsutum	70,478	5,022	7.13
G. barbadense 3-79	80,876	4,910	6.07
G. barbadense Xinhai 21	77,358	4,851	6.27
G. arboreum	40,134	2,532	6.31
<u>G. raimondii</u>	37,505	2,639	7.04

-- G. hirsutum
-- G. barbadense 3-79
-- G. barbadense Xinhai 21
-- G. arboreum
-- G. raimondii

sequence variations

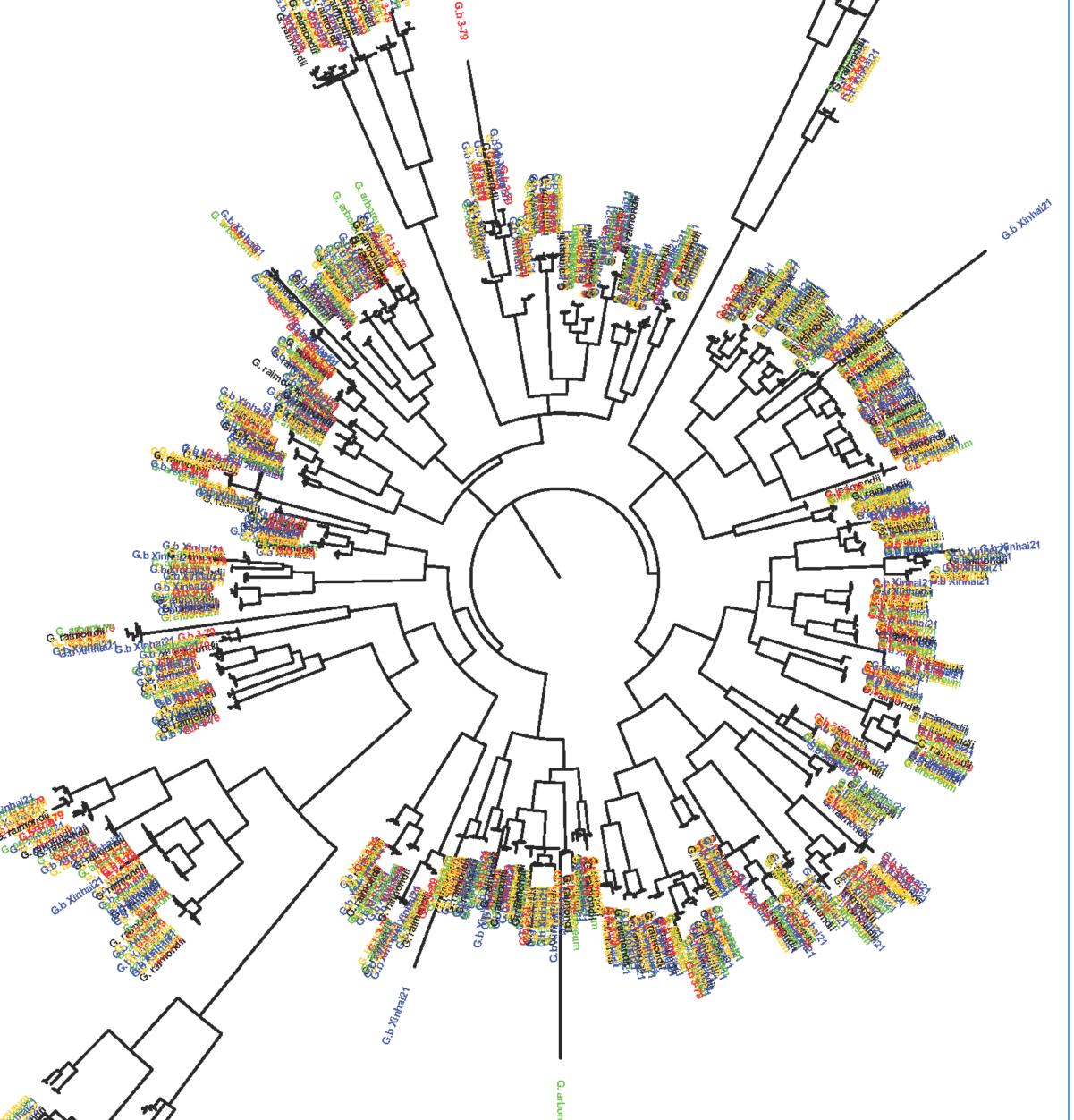
Sequence variations were marked by colored background and the change of corresponding amino acids were labeled above, in which the polar and non-polar amino acids were displayed by red and blue fonts, respectively.

Table 3 Statistics of the sequence variations in *NAC* family between diploid (2x) and tetraploid (4x), between tetraploid *G*. *hirsutum* (Gh) and *C* hashed area (Ch) and between the two Ch constructs

Sequence alignment: All the DNA sequences for each TF family were aligned using the ClustalX version 2.1 to distinguish homologous sequences in the groups (Fig. 1).

Further analysis of *NAC* family: Coding sequences (CDS) for each gene from the previously divided same homologous sequences group were manually aligned using MEGA software version 7.0.18 for SNP discovery. The amino acids at individual sequence variation sites were translated. A phylogenetic tree was obtained from the ClustalX align file by the FastTree software, which was further visualized through the Figtree version 1.4.2.

Eold10683.17.0	AAGCATGGAAGTGGTGGGAGGCCTAATAGGACTACGGCAAATGGGTTCTGGAAAGCCACCGGTTCCGACCGGAAAATCGTGAGCTTATCGGACCTGAAG-
Gh_A03G1843	AAGCATCGAAGTCCTCCCACCCCTAATACCACTACCCCAAATCCCTTCCCAAACCCACCC
02_BGI-A2_v1.0	AAGCATCCAAGTCCTCCCACCCCTAATACCACTACCCCAAATCCCTTCTCCAAACCCACCC
Gh_D02G2283	AAGCATGGAAGTGGTGGGAGGCCTAATAGGACTACGGCAAATGGGTTCTGGAAAGCCACCGGTTCCGACCGGAAAATCGTGAGCCTTATCAGACCCGAAG
1.005G257800.1	AAGCATGGAAGTGGTGGGAGGCCTAATAGGACTACGGCAAATGGGTTCTGGAAAGCCACCGGTTCCGACCGGAAAATCGTGAGCTTATCAGACCCGAAG
ffold17185.5.0	AAGCATGGAAGTCGTCGGACCCCTAATAGCACTACCCCCAAATCCGTTCTCGAAAGCCACCCCCCCCCC
GOBAR_DD13712	AAGCATGGAAGTGGTGGGAGGCCTAATAGGACTACGGCAAATGGGTTCTGGAAAGCCACCGGTTCCCACCGGAAAATCGTGAGCCTTATCAGACCCGAAG
GOBAR_AA11902	AAGCATGGAAGTGGTGGGAGGCCTAATAGGACTACGGCAAATGGGTTCTGGAAAGCCACCGGTTCCGACCGGAAAATCGTGAGCTTATCGGACCTGAAG
Efold29450.6.0	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCTTATCGGATCAGAAA
Gh_A01G0013	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCTTATCGGATCAGAAA
56_BGI-A2_v1.0	BAAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCCTTATCGGATCAGAAA
GOBAR_AA38862	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCCTTATCGGATCAGAAA
Efold47455.3.0	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCCTTATCGGATCAGAAA
GOBAR_DD25604	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCCTTATCGGATCAGAAA
i.002G001900.1	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCCTTATCGGATCAGAAA
Gh_D01G0012	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCTTATCGGATCAGAAA
GOBAR_AA40097	AAACATGGGAAAGGAGGGAGGCCTAACAGGACTACTAAAAATGGGTTCTGGAAAGCCACCGGTTCTGATCGGAAAATTGTGAGCCTTATCGGATCAGAAA
GOBAR_AA08971	AAGCATGGCAATGGAGGGAGACCAAACCGGACGACTGAAAATGGGTACTGGAAAGCCACTGGTTCTCACCGGAAAATTGTGAGCTTATCAGATCCAAAG
Gh_A12G0319	AAGCATGOCAATGGAGGGAGACCAAACCGGACGACTGAAAATGGGTACTGGAAAGCCACTGGTTCTGACCGGAAAATTGTGAGCTTATCAGATCCAAAG
48_BGI-A2_v1.0	AAGCATGOCAATGCAGGGAGACCAAACCGGACGACTGAAAATGGGTACTGGAAAGCCACTGGTTCTGACCGGAAAATTGTGAGCTTATCAGATCCAAAG-
affold5102.9.0	AAGCATGOCAATGGAGGGAGACCAAACCGGACGACTGAAAATGGGTACTGGAAAGCCACTGGTTCTGACCGGAAAATTGTGAGCTTATCAGATCCAAAG
GOBAR_DD18498	AAGCATGGCAATGGAGGGAGACCAAACCGGACGACTGAAAATGGGTACTGGAAAGCCACTGGTTCTGACCGGAAAATTGTGAGCTTATCAGATCCAAAG
Gh_D12G0347	AAGCATGGCAATGGAGGGAGACCAAACCGGACGACTGAAAATGGGTACTGGAAAGCCACTGGTTCTGACCGGAAAATTGTGAGCTTATCAGATCCAAAG
i.008G038800.1	AAGCATGGCAATGGAGGCAGACCAAACCGGACGACTGAAAATGGGTACTGGAAAGCCACTGGTTCTGACCGGAAAATTGTGAGCTTATCAGATCCAAAG
05_BGI-A2_v1.0	AGGGGTGGTCATGGTGGTAAAAAGCCCAACAGGACCACCGAGAGAGGGTATTGGAAGGCAACTGGTTCCGACCGTCAAATCAGGCGTCTCACTGAGCCTAAA-
GOBAR AA31923	ACCACTCGTC ATCCTCCTCAAAAACCCCCAACACCACCCACCCAC



and G. barbadense (Gb) and between the two Gb genotypes

	Sequence variation			
	All $2x vs. 4x$		4x v	vs. 4x
	All	$\Delta X VS. 4X$	Gh-Gb	Gb1-Gb2
All	1240	736	361	129
Missense mutation			208	54
Polarity change			56 16	

Conclusions

♦ A total of 5,022 (7.13%), 4,910 (6.07%), 4,815 (6.27%), 2,532 (6.31%) and 2,639 (7.04%) genes coding for transcription factors were predicted in *G. hirsutum*, *G. barbadense* cv. 3-79, *G. barbadense* cv. Xinhai 21, *G. arboreum* and *G. raimondii*, respectively, among which 306, 245, 283, 150 and 153 belong to the NAC family.

◆In the phylogenetic tree for NAC family, many clades consist of more than 3 sequences in each sub-genome (i.e., A or D subgenome) and are therefore suitable for the downstream sequence variation analysis.

A total of 1,240 SNPs were identified in the *NAC* family among the five cotton genomes, among which 262 had amino acid changes among the three sequenced tetraploid cotton genomes including 72 amino acid changes in polarities.

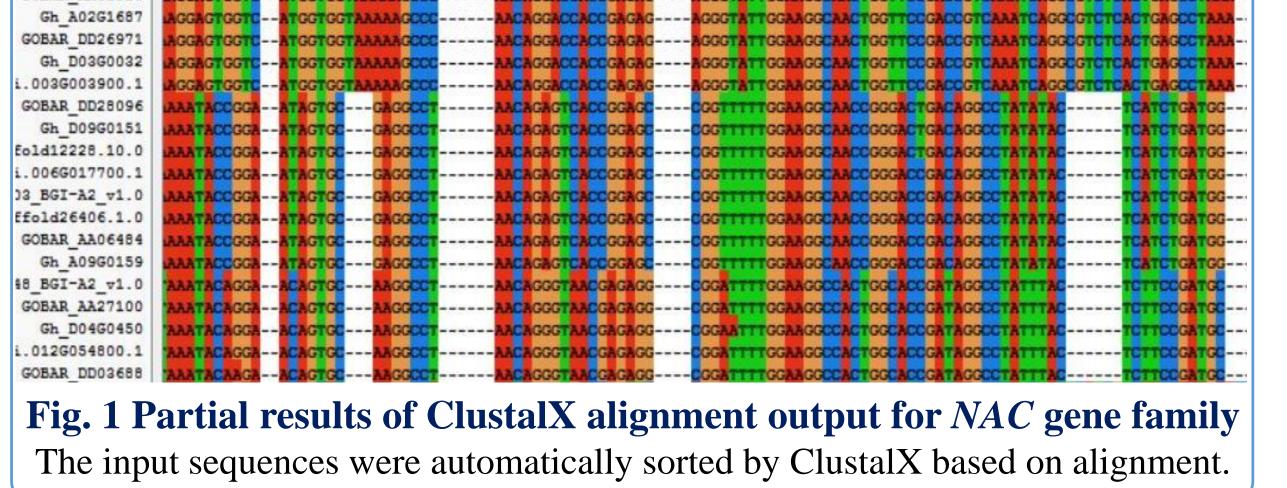
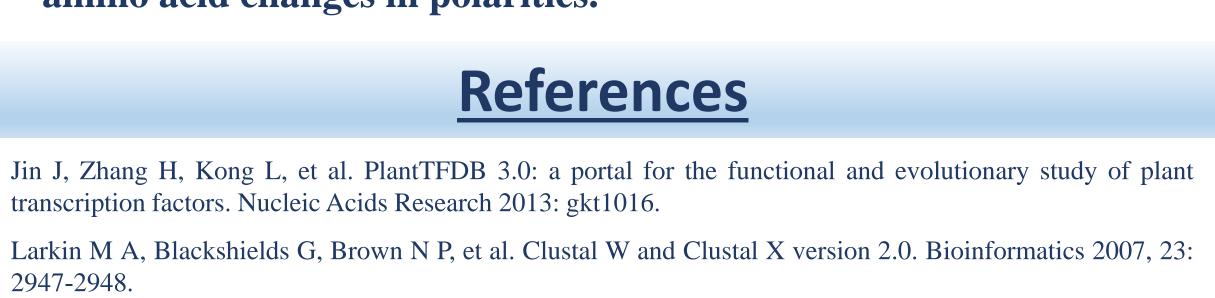


Fig. 2 A phylogenetic tree of the *NAC* **gene family in cotton** Sequences from different cotton genomes were displayed by different colors, as explained in the figure.

0.5



Acknowledgements

Mr. Zhihua Pei in assisting with the bioinformatics work and the graduate students in the cotton lab of NMSU for their help.