Transcription Factors in Cotton

Jie Chen[1], Mingzhou Song[2], and Jinfa Zhang[1]

(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.

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" and *G. barbadense* "Xinhai 21") were downloaded online (https://www.cottongen.org, and http://database.chgc.sh.cn/cotton). TF genes were then predicted based on http://planttfdb.cbi.pku.edu.cn (Table 1) using the peptide sequences. Multiple peptide sequences corresponding to the same DNA fragment were considered redundant and only counted once.**

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.**

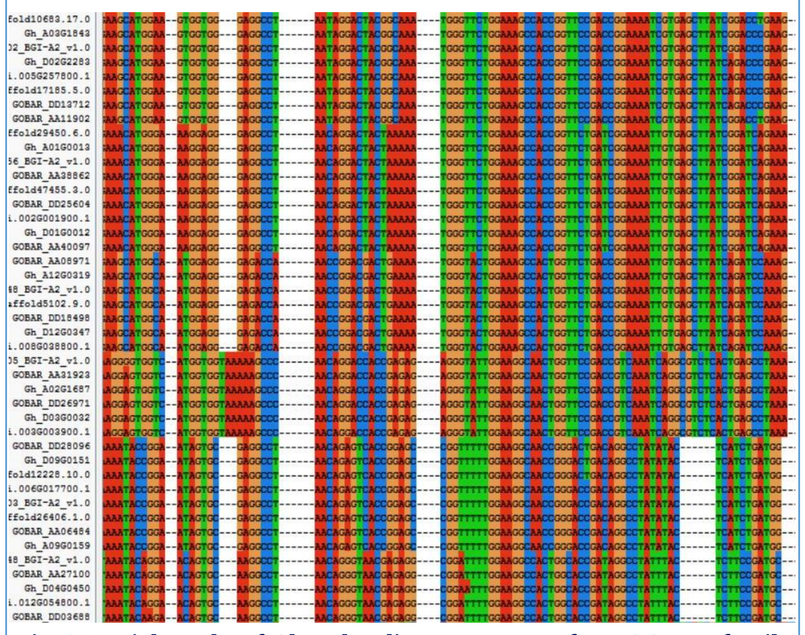

Fig. 1 Partial results of ClustalX alignment output for *NAC* gene family
The input sequences were automatically sorted by ClustalX based on alignment.

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

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
G. raimondii	37,505	2,639	7.04

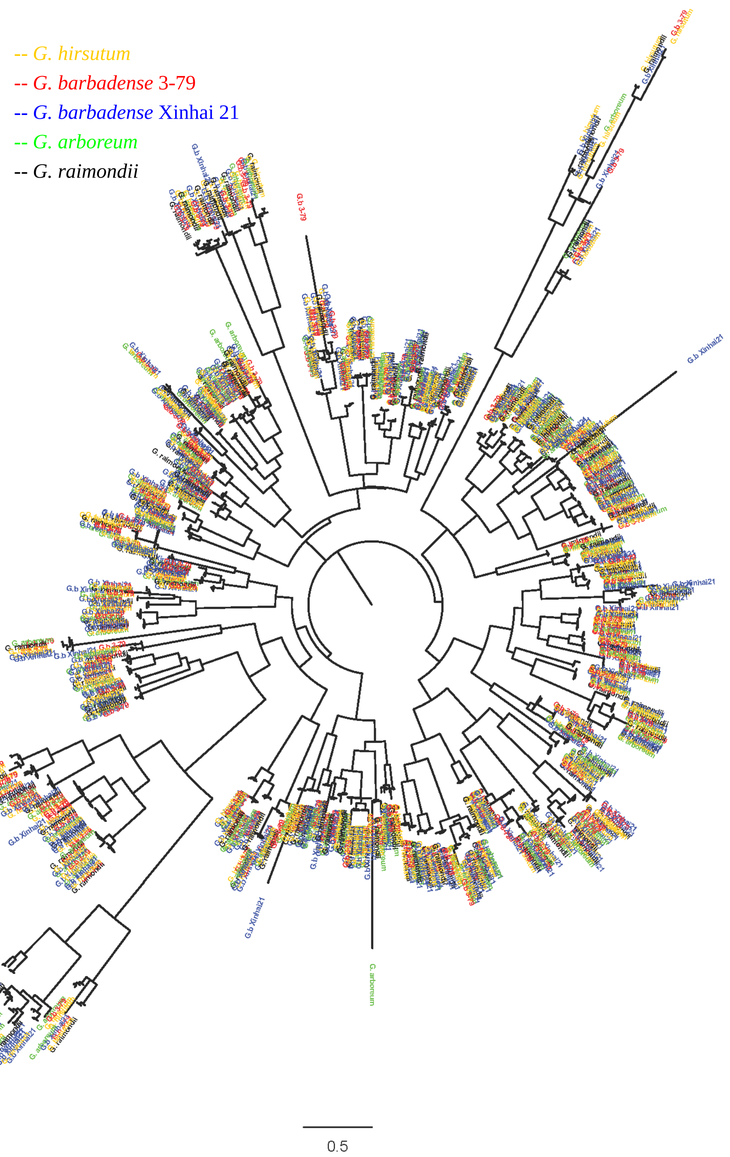

Fig. 2 A phylogenetic tree of the *NAC* gene family in cotton
Sequences from different cotton genomes were displayed by different colors, as explained in the figure.

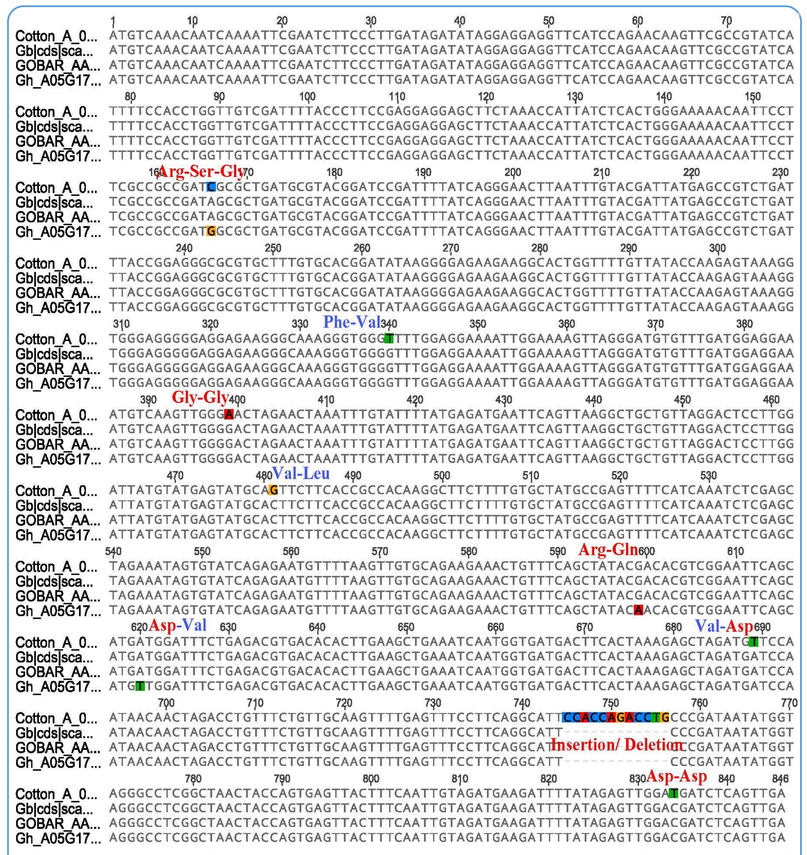

Fig. 3 An example of one homologous gene sequence alignment and sequence variations
Sequence variations were marked by colored background and the change of corresponding amino acids were labeled above, in which the polar and non-polar amino acids were displayed by red and blue fonts, respectively.

Table 3 Statistics of the sequence variations in *NAC* family between diploid (2x) and tetraploid (4x), between tetraploid *G. hirsutum* (Gh) and *G. barbadense* (Gb) and between the two Gb genotypes

	Sequence variation			
	All	2x vs. 4x	4x 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 sub-genome) 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.**

References

Jin J, Zhang H, Kong L, et al. PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Research 2013: gkt1016.

Larkin M A, Blackshields G, Brown N P, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007, 23: 2947-2948.

Acknowledgements

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