PHYLOGENETIC ANALYSIS AND SNP IDENTIFICATION OF NAC GENE FAMILY IN COTTON

Jie Chen
Huazhong Agricultural University
Wuhan, Hubei Province, China
New Mexico State University
Las Cruces, NM
Jianyong Wu
Chaozhu Xing
Xihua Li
Jiwen Yu
Institute of Cotton Research, Chinese Academy of Agricultural Sciences
Anyang, Henan Province, China
Mingzhou Song
Jinfa Zhang
New Mexico State University
Las Cruces, NM

Abstract

NAC (NAM, ATAF, and CUC) transcription factors (*NAC*-TFs), as one of the largest TF families, play an important role in plant development, and are involved in multiple biotic and abiotic responses. In this study, four recently sequenced cotton genomes were analyzed to identify TF-coding genes for a phylogenetic and comparative structural genomic analysis. Single nucleotide polymorphisms (SNPs) are being further identified for *NAC*-TFs. Genome-wide gene expression studies were performed to reveal their regulations associated with ovule and fiber development and male fertility restoration in cotton. Of a total of 5,022 (7.13%), 4,815 (6.27%), 2,532 (6.31%) and 2,639 (7.04%) genes coding for transcription factors (TFs) predicted in *G. hirsutum*, *G. barbadense*, *G. arboreum* and *G. raimondii*, respectively, 306, 283, 150 and 153 belong to the *NAC* family. Of the 306 predicted *NAC* TFs from *G. hirsutum* TM-1, 135 were expressed (FPKM $\geq$ 1) in flowering buds (3 mm in length), while 107 were expressed in ovules at -3, 0 or 3 DPA and 81 in fibers at 10 DPA, and 67 were commonly expressed in the three organs. Moreover, two *NAC*-TFs, *Gh_A09G1677* and *Gh_A13G0828*, were found commonly differentially expressed in 0 DPA ovules from a comparison of WT/*fl* and NMGA-105/NMGA-062. They also shared similar expression patterns in ovule and fiber samples. The detailed expression analyses and SNP discovery of *NAC*-TFs will facilitate a better understanding of their roles in these plant developmental pathways.

Introduction

Transcription factors (TFs) regulate gene expressions via interacting with regulatory sequences located in promoter regions, and they are classified into ~60 families based on their DNA-binding domains (Jin et al., 2017). Among them, NAC (NAM, ATAF, and CUC) transcription factors (*NAC*-TFs), as one of the largest TF families, play an important role in plant development, and are also involved in multiple biotic and abiotic responses (Shao et al., 2015; Wang et al., 2011). For example, *NAC*-TFs, along with MYB genes, participated in secondary cell wall biosynthesis, thus affecting multiple plant developmental processes like root and fiber growth (Nakano et al., 2015). In rice, *ONAC122* and *ONAC131* can resist the infection against *Magnaporthe grisea* (Sun et al., 2013); and transgenic rice plants overexpressing *OsNAC6* exhibited a tolerance to blast disease (Nakashima et al., 2007). In soybean, 38 *NAC*-TFs were involved in response to drought (Le et al., 2011). 32 *NAC*-TFs responded to at least two kinds of stress treatments in *Chrysanthemum lavandulifolium* (Huang et al., 2012). In recent years, genome-wide analyses of NAC family were conducted in multiple species such as *Arabidopsis* (Ooka et al., 2003), *Vitis vinifera* (Wang et al., 2013), *Populus trichocarpa* (Hu et al., 2010), *Gossypium raimondii* (Shang et al., 2013), *Musa acuminata* (Cenci et al., 2014).

Cotton, serving as a primary fiber resource as well as an important oil-seed, is an economically significant crop around the world. The identification and functional studies of NAC family in cotton were conducted before (Shang et al., 2013; Shah et al., 2013), in which a comprehensive study regarding phylogeny, chromosomal location, gene structure, conserved motifs, and expression profiling was conducted (Shang et al., 2013), and ten stress-responsive NAC genes (*GhNAC8–GhNAC17*) were isolated (Shah et al., 2013). The whole genome sequencing of the ancestral diploid cotton species *G. raimondii* (Wang et al., 2012) and *G. arboreum* (Li et al., 2014), and their tetroploids *G.*

hirsutum (Li et al., 2015) and *G. barbadense* (Liu et al., 2015) has provided an excellent opportunity to perform a genome-wide analysis on NAC gene family, which was already conducted in *G. raimondii* (Shang et al., 2013) and in the two diploid species of cotton (Shang et al., 2016). In this study, all the four genomic information were included in the genome-wide discovery and analysis for *NAC*-TFs, which made it possible to identify single nucleotide polymorphisms (SNPs) among homologous and homeologous *NAC*-TFs under a full-view. Moreover, RNA-seq data from multiple spatial organs were also involved to evaluate possible roles of *NAC*-TFs in participating and affecting these developmental pathways.

Materials and Methods

Identification of TF genes

The genome sequence information for the four cotton species (ancestral diploids *G. arboreum*, and *G. raimondii*, and their tetraploids *G. hirsutum*- TM-1 and *G. barbadense*- Xinhai 21) were downloaded online (https://www.cottongen.org, and http://database.chgc.sh.cn/cotton). TF genes were then predicted (http://planttfdb.cbi.pku.edu.cn) using the peptide sequences. Multiple peptide sequences corresponding to the same DNA fragment were considered redundant and only counted once.

Further analysis of *NAC* family

Sequences were aligned using ClustalX version 2.1 (Fig. 1), which was used to generate a phylogenetic tree by the FastTree software and visualized through the Figtree version 1.4.2. A SNP discovery was conducted based on manual alignments using MEGA7 software on coding sequences (CDS).

Expression profiling of *NAC* family

In this study, the following RNA-seq were performed in conducting the expression analyses of the *NAC* family:

1) Flowering buds during meiosis (3 mm in length) from three isogenic lines: a CMS line (A), a maintainer line (B), and a restorer line (R).
2) Ovules at 0 and 3 DPA (days post-anthesis) and fibers at 10 DPA from two BILs (backcross inbred lines): NMGA-062 (long fiber) and NMGA-105 (short fiber).
3) Ovules at -3 and 0 DPA from Xuzhou 142 (WT) and its fiberless and fuzzless mutant (*fl*).

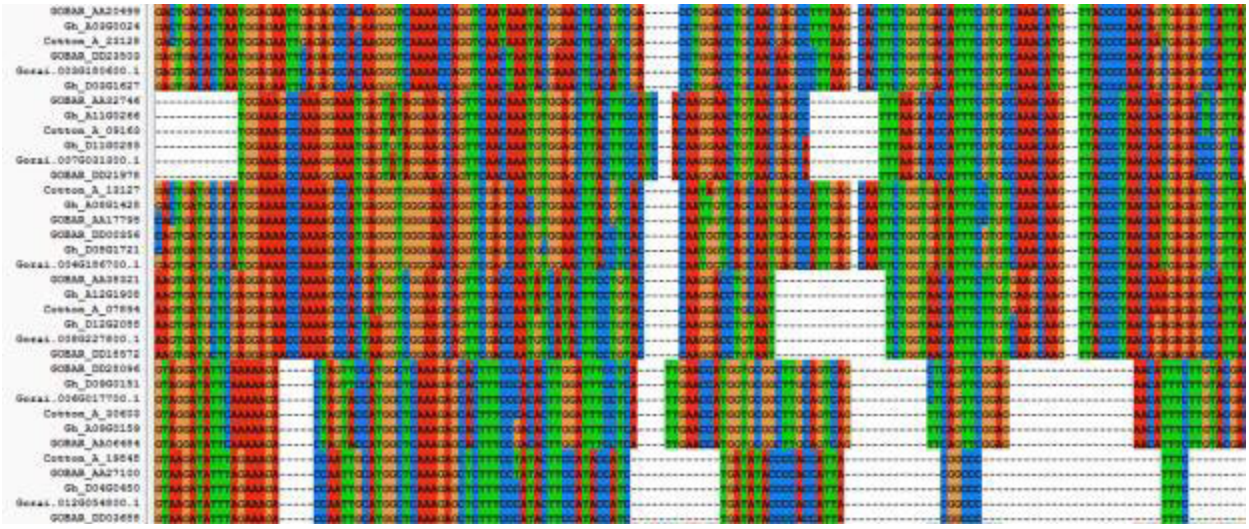

Fig. 1. Partial results of ClustalX alignment output for *NAC* family.
The input sequences were automatically by ClustalX based on alignment.

Results and Discussion

TF discovery and SNP development

By gathering the genomic information, around 40,000 genes were found in the ancestral diploid cotton *G. arboreum* and *G. raimondii*, while 70,000 genes were identifiedbin their tetraploids *G. hirsutum*- TM-1 and *G. barbadense*- Xinhai 21. Among these genes, approximately 7% of them (~2,500 for diploids and ~5,000 for tetraploids) were

predicted as TF genes (Table 1). The numbers of TFs were reflective of the origin of the tetraploids in that the total gene numbers in tetraploid cottons were approximately equal to the sum of the two diploids.

A phylogenetic tree for *NAC*-TFs was generated (Fig. 2), in which genes from the four cotton genomes were distributed evenly, which makes it suitable for subsequent sequence variation analyses. Consequently, a total of 995 SNPs were identified in the *NAC* family among the four cotton genomes, among which 195 had amino acid changes between the two sequenced tetraploid cotton genomes (i.e., *G. hirsutum* TM-1 and *G. barbadense* Xinhai 21) including 48 amino acid changes in polarities (Table 2). Those sequence variation sites with amino acid changes in polarities were most likely to cause protein (gene translation product) structural change and thus were to be verified and studied in higher priority.

Table 1. 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	77,358	4,851	6.27
G. arboreum	40,134	2,532	6.31
G. raimondii	37,505	2,639	7.04

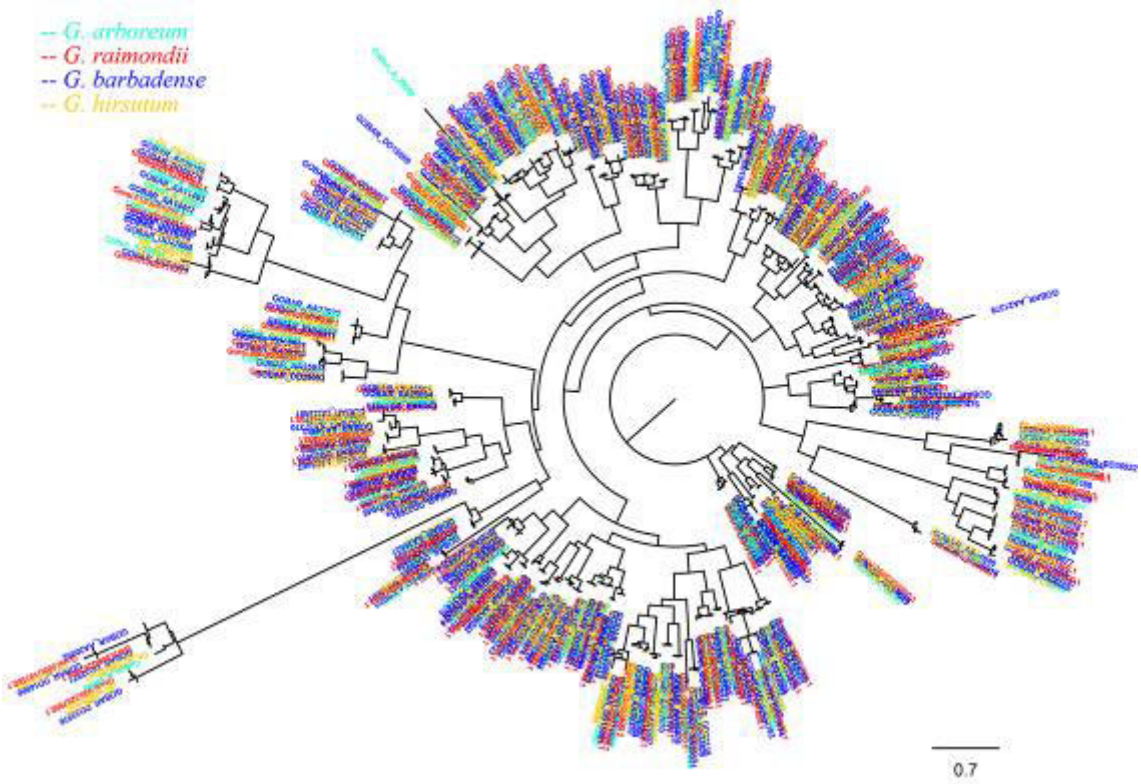

Fig. 2. A phylogenetic tree of NAC gene family in cotton.
Sequences from different cotton genomes were represented by different colors, as displayed in the figure.

Table 2. Statistics of the sequence variations in NAC family between diploid (2x) and tetraploid (4x), and between tetraploid genotypes.

	Sequence variation		
	All	2x vs. 4x	4x vs. 4x
All	995	693	302
Missense mutation			195
Polarity change			48

Expression profiling of *NAC* family

Among the 5,022 predicted TFs in Upland cotton, 306 belong to the *NAC* family, in which 135 were expressed (FPKM ≥ 1) in flowering buds (during meiosis, 3 mm in length), while 107 were expressed in ovules at -3, 0 or 3 DPA and 81 in fibers at 10 DPA, and 67 were commonly expressed in the three organs (Fig. 3A). When considering differentially expressed genes in ovules (Fig. 3B) and in the three isogenic lines (Fig. 3C), around ten or less of *NAC*-TFs were differentially expressed (DE). These DE *NAC*-TFs may be involved in the diversity of corresponding developmental pathways.

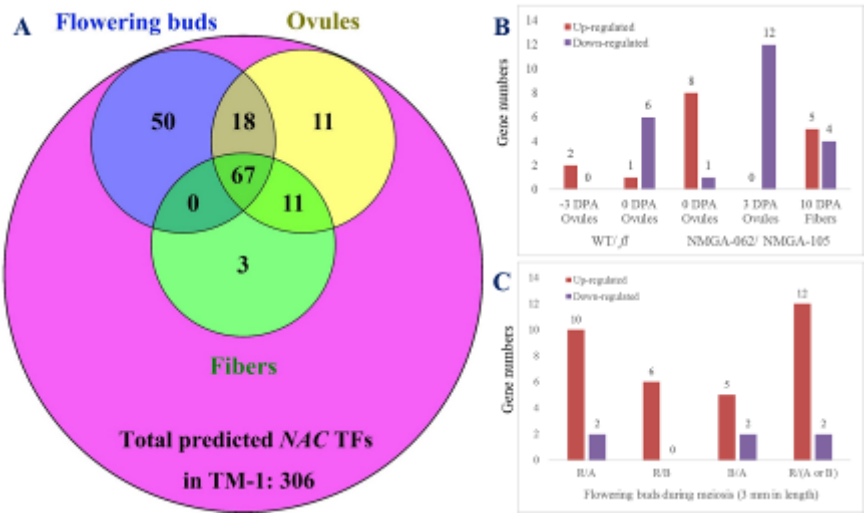

Fig. 3. Overall expression profiling of *NAC* TF genes.

A: Numbers of expressed *NAC* TFs (FPKM ≥ 1) in flowering buds (3 mm in length), ovules (BILs at 0 and 3 DPA and Xuzhou 142 at -3 and 0 DPA) and fibers (BILs at 10 DPA); and B: Differentially expressed genes (DEGs) ($|\log_2 FC| \geq 1$) in ovules; C: DEGs among the three isogenic lines.

When it came to the specific information of the 135 expressed (FPKM ≥ 1) *NAC*-TFs in the three isogenic lines (dA, the sterile line; dB, the maintainer line; and dR, the restorer line), eight, six and seven of them were specifically expressed (FPKM ≥ 1) in line A, B, and R, respectively, while one, five and five of them were commonly expressed in lines A/B, B/R, and A/R, whereas a large portion (103 of 135) of *NAC*-TFs were commonly expressed in all the three isogenic lines (Fig. 4A). In addition, only seven, six and 12 *NAC*-TFs were differentially expressed ($|\log_2 FC| \geq 1$) between comparisons of lines A/B, B/R, and A/R, respectively, and none of them were differentially expressed among the three lines (Fig. 4B), which were further detailed in Table 3. The relationship between the slight portion of specifically expressed *NAC*-TFs (8, 6, and 7 in A, B, and R lines, respectively) and the differentially expressed ones (7, 6, and 12 in comparisons A/B, B/R, and A/R, respectively) would be the most inspiring work to unveil the role of *NAC*-TFs in participating in possible male sterility pathways. The large portion (103 of 135) of *NAC*-TFs may be vital for essential plant developments.

As to the detailed information in ovules, 70 of 107 *NAC*-TFs were commonly expressed (FPKM ≥ 1) in different spatiotemporal ovule samples (line WT at -3 and 0 DPA; line *fl* at -3 and 0 DPA; line NMGA-105 at 0 and 3 DPA; and line NMGA-062 at 0 and 3 DPA, Fig. 5A), while 65 of 97 *NAC*-TFs were commonly expressed in 0 DPA of the four lines (Fig. 5B). It appeared that more *NAC* genes were expressed in lines NMGA-105/NMGA-062 than in Xuzhou 142 lines (WT/*fl*). Similar to the three isogenic lines, most of the *NAC*-TFs were commonly expressed in ovules from the two time-points of the two lines (70 of 107 in spatiotemporal ovules and 65 of 97 in 0 DPA ovules of the four lines), indicating the necessity of *NACs* maintaining basic plant growth.

Moreover, two of the *NAC*-TFs, *Gh_A09G1677* and *Gh_A13G0828*, were found commonly differentially expressed in 0 DPA ovules from lines WT/*fl* and lines NMGA-105/NMGA-062 (Fig. 6A), while *Gh_A13G0828* was expressed

consistently lower than *Gh_A09G1677* (Fig. 6B, C1 to C4). They also shared similar expression trends within lines *fl* (C1), WT (C2), NMGA-105 (C3) and NMGA-062 (C4), and at each time-point (Fig. 6B). These two genes were presumably important to fiber initiation in 0 DPA ovules.

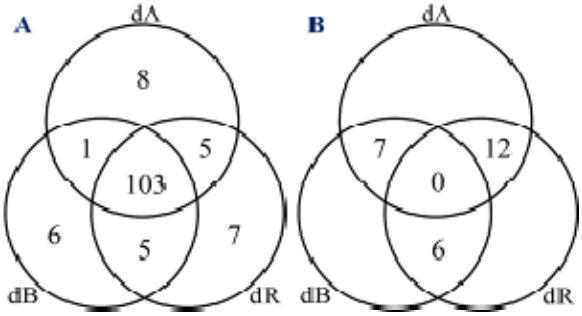

Fig. 4. Expression analysis of *NAC* TFs in isogenic lines.
A: Expressed (FPKM ≥ 1) genes in three isogenic lines, i.e., dA, a CMS line; dB, a maintainer line; and dR, a restorer line. B: Differentially expressed ($|\log_2 FC| \geq 1$) genes among the three isogenic lines. Numbers in the intersections indicated genes that were differentially expressed between the two lines under comparison.

Table 3. Differentially expressed *NAC* TFs among A, B and R lines.

Gene_id	Chr.	Length (aa)	$\log_2$FC(dB/dA)	Up/down	$\log_2$FC(dR/dB)	Up/down	$\log_2$FC(dR/dA)	Up/down
Gh_A01G0267	A01	320	-1.63	down	-1.14	-	-2.78	down
Gh_A05G2928	A05	307	1.44	up	0.7	-	2.14	up
Gh_A05G3310	A05	277	1.48	up	0.2	-	1.68	up
Gh_A07G1811	A07	404	-0.35	-	1.12	up	0.77	-
Gh_A09G0913	A09	229	1.16	up	-0.14	-	1.02	up
Gh_A11G2168	A11	383	-0.31	-	1.39	up	1.08	up
Gh_A12G1505	A12	282	-1.76	-	3.2	up	1.47	up
Gh_D01G0278	D01	321	-2.64	down	0.86	-	-1.79	down
Gh_D01G0514	D01	347	0.55	-	0.84	-	1.39	up
Gh_D02G1769	D02	276	0.23	-	0.86	-	1.09	up
Gh_D04G0293	D04	277	1.68	up	-0.03	-	1.65	up
Gh_D09G0943	D09	229	1.06	up	-0.34	-	0.72	-
Gh_D11G0743	D11	319	-1.19	-	2.91	up	1.74	up
Gh_D11G2469	D11	358	-0.86	-	2.17	up	1.31	up
Gh_D12G2767	D12	282	-0.77	-	1.96	up	1.19	-

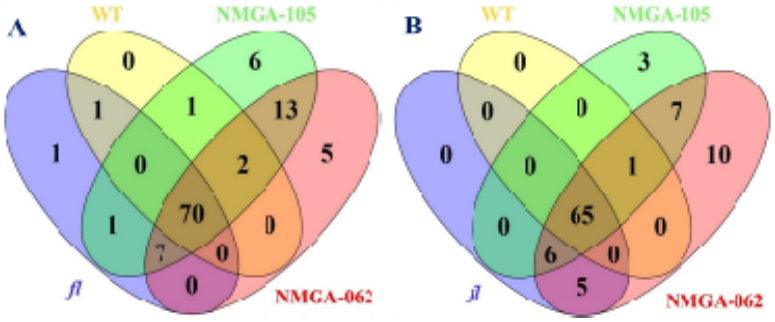

Fig. 5. Numbers of expressed *NAC*-TFs in ovules of four cotton genotypes.
Genes were counted with FPKM ≥ 1. Overall expressions in ovules from lines *fl* (-3 and 0 DPA), WT (-3 and 0

DPA), NMGA-105 (0 and 3 DPA) and NMGA-062 (0 and 3 DPA) were displayed in Fig. 5A, while specific expressions in ovules at 0 DPA from the four lines were shown in Fig. 5B.

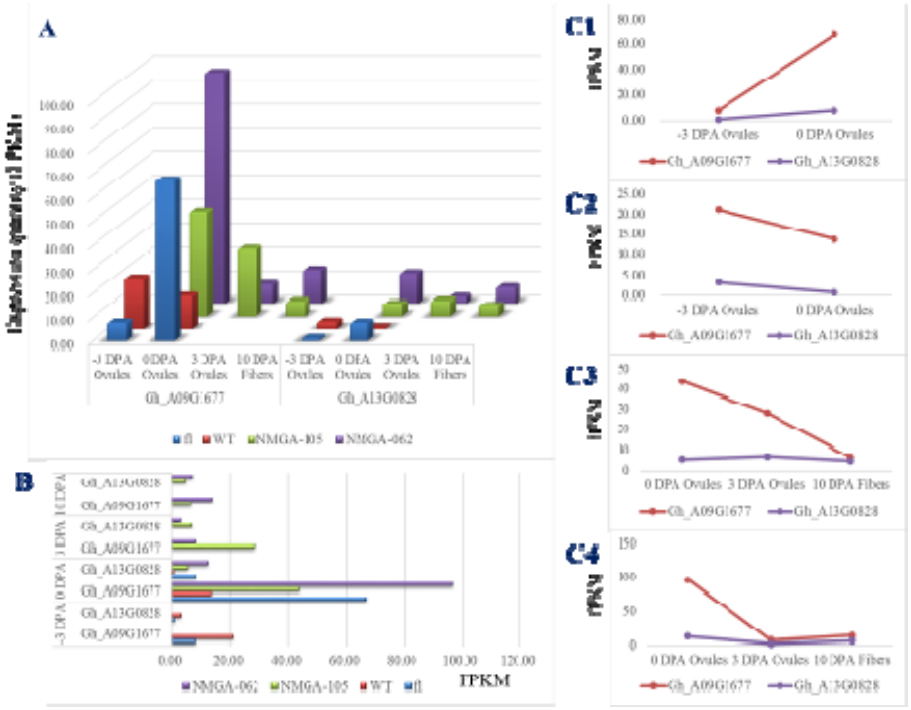

Fig. 6. Expression details of *Gh_A09G1677* and *Gh_A13G0828*.
The two genes shared similar expression trends within lines *fl* (C1), WT (C2), NMGA-105 (C3) and NMGA-062 (C4), and at each timing (B). The overall expression levels were shown in Fig. 6A, and *Gh_A13G0828* was expressed consistently lower than *Gh_A09G1677* (Fig. 6B, C1 to C4).

Summary

In this study, a total of 5,022 (7.13%), 4,815 (6.27%), 2,532 (6.31%) and 2,639 (7.04%) genes coding for transcription factors (TFs) were predicted in *G. hirsutum*, *G. barbadense*, *G. arboreum* and *G. raimondii*, respectively, among which 306, 283, 150 and 153 belong to the *NAC* family. A phylogenetic tree was generated and the SNP discovery was conducted for the *NAC* family, which provided a solid base for downstream analyses. Of the 306 predicted *NAC*-TFs in *G. hirsutum*, 135 were expressed (FPKM ≥ 1) in flowering buds (3 mm in length), while 107 were expressed in ovules at -3, 0 or 3 DPA and 81 in fibers at 10 DPA, and 67 were commonly expressed in the three organs. The majority of the expressed *NAC* TFs (103 of 135) were commonly expressed in the three isogenic A, B and R lines, while most of the differentially expressed genes (12 of 14) in the restorer R line were up-regulated. Similarly, most of *NAC*-TFs were commonly expressed in ovules (70 of 107 in spatiotemporal ovules and 65 of 97 in 0 DPA ovules), while more *NAC*-TFs were expressed in the two BILs (NMGA-105 and NMGA-062) than in Xuzhou 142 WT/*fl* ovules; and two (*Gh_A09G1677* and *Gh_A13G0828*) shared similar expression diversity and trends. The results will facilitate us in understanding the possible roles of *NAC*-TFs in participating in the sterility pathways, and in affecting the fiber developmental processes when more detailed experiments are performed based on the improved knowledge.

Acknowledgements

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

References

Cenci, A., Guignon, V., Roux, N., et al. 2014. Genomic analysis of NAC transcription factors in banana (*Musa acuminata*) and definition of NAC orthologous groups for monocots and dicots. Plant Mol. Biol. 85:63-80.

Hu, R., Qi, G., Kong, Y., et al. 2010. Comprehensive analysis of NAC domain transcription factor gene family in *Populus trichocarpa*. BMC Plant Biol. 10:145.

Huang, H., Wang, Y., Wang, S., et al. 2012. Transcriptome-wide survey and expression analysis of stress-responsive *NAC* genes in *Chrysanthemum lavandulifolium*. Plant Sci. 193:18-27.

Jin, J.P., Tian, F., Yang, D.C., et al. 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 45:D1040-D1045.

Le, D.T., Nishiyama, R.I.E., Watanabe, Y., et al. 2011. Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. DNA Res. 2011:dsr015.

Li, F., Fan, G., Lu, C., et al. 2015. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat. Biotech. 33:524-530.

Li, F., Fan, G., Wang, K., et al. 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat. Genet. 46:567-572.

Liu, X., Zhao, B., Zheng, H.J., et al. 2015. *Gossypium barbadense* genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Sci. Rep. 5.

Nakano, Y., Yamaguchi, M., Endo, H., et al. 2015. NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. Front. Plant Sci. 6:288.

Nakashima, K., Tran, L.S., VanNguyen, D., et al. 2007. Functional analysis of a NAC-type transcription factor *OsNAC6* involved in abiotic and biotic stress-responsive gene expression in rice. Plant J. 51:617-630.

Ooka, H., Satoh, K., Doi, K., et al. 2003. Comprehensive analysis of NAC family genes in Oryza sativa and Arabidopsis thaliana. DNA Res. 10:239-247.

Shah, S.T., Pang, C., Fan, S., et al. 2013. Isolation and expression profiling of *GhNAC* transcription factor genes in cotton (*Gossypium hirsutum* L.) during leaf senescence and in response to stresses. Gene. 531:220-234.

Shang, H., Li, W., Zou, C., et al. 2013. Analyses of the NAC Transcription factor gene family in *Gossypium raimondii* Ulbr.: chromosomal location, structure, phylogeny, and expression patterns. J. Integr. Plant Biol. 55:663-676.

Shang, H., Wang, Z., Zou, C., et al. 2016. Comprehensive analysis of NAC transcription factors in diploid *Gossypium*: sequence conservation and expression analysis uncover their roles during fiber development. Sci. China Life. Sci. 59:142-153.

Shao, H., Wang, H., and Tang, X. 2015. NAC transcription factors in plant multiple abiotic stress responses: progress and prospects. Front. Plant Sci. 6.

Sun, L., Zhang, H., Li, D., et al. 2013. Functions of rice NAC transcriptional factors, *ONAC122* and *ONAC131*, in defense responses against *Magnaporthe grisea*. Plant Mol. Biol. 81:41-56.

Wang, H., Zhao, Q., Chen, F., et al. 2011. NAC domain function and transcriptional control of a secondary cell wall master switch. Plant J. 68:1104-1114.

Wang, K., Wang, Z., Li, F., et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. Nat. Genet. 44:1098-1103.

Wang, N., Zheng, Y., Xin, H., et al. 2013. Comprehensive analysis of NAC domain transcription factor gene family in *Vitis vinifera*. Plant Cell. Rep. 32:61-75.