Gene 680 (2019) 72–83

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Review

Genome-wide identification and analysis of the evolution and expression patterns of the GATA transcription factors in three species of *Gossypium* genus

Zhen Zhang[a,b,1], Xianyan Zou[a,1], Zhen Huang[b,1], Senmiao Fan[a], Ge Qun[a], Aiying Liu[a], Juwu Gong[a], Junwen Li[a], Wankui Gong[a], Yuzhen Shi[a], Liqiang Fan[a], Zhibin Zhang[a], Ruixian Liu[a], Xiao Jiang[a], Kang Lei[a], Haihong Shang[a,*], Aixia Xu[b,*], Youlu Yuan[a,*]

[a] State Key Laboratory of Cotton Biology, Key Laboratory of Biological and Genetic Breeding of Cotton, Ministry of Agriculture, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, Henan, China
[b] State Key Laboratory of Crop Stress Biology for Arid Areas, College of Agronomy, Northwest A&F University, Yangling, China

ARTICLE INFO

Keywords:
Gossypium
GATA TFs
Gene and protein structure
Phylogeny
Expression pattern

ABSTRACT

GATA transcription factors (TFs), which bind to DNA in regulatory regions, are involved in cell differentiation and possess a type-IV zinc finger and a DNA-binding domain. *GATA* genes have been characterized in plant species such as *Arabidopsis thaliana*, *Oryza sativa*, and *Glycine max*, and their functions have been elucidated in *A. thaliana*. Although many *Gossypium* quantitative trait loci for fiber quality harbor *GATA* TFs, *GATA* genes have not yet been characterized in cotton. In this study, we identified 179 *GATA* genes from the genomes of three *Gossypium* species. We analyzed the phylogenetic relationships, chromosomal distribution, gene structure, expression pattern, and predicted promoters of all 179 *Gossypium GATA* genes (46 in *G. raimondii*, 46 in *G. arboreum*, and 87 in *G. hirsutum*). Phylogenetic analysis grouped the 179 *GATA* genes into four subfamilies. Domain analysis revealed that *GATA* domains in subfamilies I, II, and III were located near the C-terminal, whereas those in subfamily IV were adjacent to the N-terminal. RNA-seq and (Real-time PCR) qRT-PCR revealed that 39.1% (34/87) of *GATA* genes were expressed in growing plant tissues in *G. hirsutum*, but only 12.6% (11/87) were expressed during fiber development. In addition, 45.7% (21/46) and 26.1% (12/46) of *GATA* genes were expressed in *G. arboreum* and *G. raimondii*, respectively. Our results may be useful for elucidating the evolution, expression patterns, and functional divergence of *GATA* genes in *Gossypium*.

1. Introduction

GATA transcription factors (TFs) constitute a protein family characterized by the presence of one or two highly conserved type-IV zinc fingers (C-X_2-C-X_{17-20}-C-X_2-C) and a DNA-binding domain recognizing the DNA consensus sequence (A/T)*GATA*(A/G) (Lowry and Atchley, 2000). *GATA* TFs are widely distributed in fungi, animals, and plants (Patient and McGhee, 2002a). Although most *GATA* domains harbor a class-IV zinc-finger motif of the form C-X_2-C-X_{17-20}-C-X_2-C followed by a basic region (Reyes et al., 2004), this structure differs among kingdoms. In the animal kingdom, the *GATA* domain harbors two zinc-finger motifs with the structure C-X_2-C-X_{17}-C-X_2-C, but only the C-terminal finger is related to DNA binding (Lowry and Atchley, 2000; Patient and McGhee, 2002a). In fungi, the *GATA* domain contains a single C-X_2-C-X_{17}-C-X_2-C or C-X_2-C-X_{18}-C-X_2-C motif (Teakle and Gilmartin, 1998; Scazzocchio, 2000). In plants, most *GATA* domains include a single C-X_2-C-X_{18}-C-X_2-C motif, but some harbor more than two zinc-finger motifs or zinc-finger loops of 20 residues (Reyes et al., 2004).

The functions of *GATA* TFs have been studied in fungi, animals, and

Abbreviations: QTL, quantitative trait locus; FPKM, fragments per kilobase of exon per million fragments mapped; SNP, single nucleotide polymorphisms; qRT-PCR, quantitative real-time PCR; DPA, days post anthesis; SRA, sequence read archive; NCBI, National Center for Biotechnology Information; GSDS, Gene Structure Display Server; CDD, conserved domain database; MEME, multiple em for motif elicitation; bp, base pairs; aa, amino acid; d_N, nonsynonymous substitution rate; d_S, synonymous substitution rate; KEGG, kyoto encyclopedia of genes and genomes; KOG, eukaryotic orthologous groups; GO, gene ontology

* Corresponding authors.
E-mail addresses: shanghaihong@caas.cn (H. Shang), 799170439@qq.com (A. Xu), youluyuan@caas.cn (Y. Yuan).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.gene.2018.09.039
Received 10 May 2018; Received in revised form 28 August 2018; Accepted 21 September 2018
Available online 22 September 2018
0378-1119/

plants. In fungi, *GATA* TFs are related to light induction, siderophore biosynthesis, regulation of nitrogen metabolism, and mating-type switching (Scazzocchio, 2000). *GATA* TFs in animals are involved in the regulation of differentiation and development, including cell-fate specification and the control of cell proliferation and movement (Patient and McGhee, 2002b). In plants, *GATA* TFs are related to RNA polymerase II regulatory region sequence-specific DNA binding, RNA polymerase II TF binding, chromatin binding and TF activity, sequence-specific DNA binding, transcriptional activator activity, RNA polymerase II transcription regulatory region sequence-specific binding, and zinc ion binding (Giuliano et al., 1988; Buzby et al., 1990; Lam et al., 1990; Carre and Kay, 1995; Teakle and Kay, 1995; Manfield et al., 2007).

GATA TFs are expressed at several stages of plant growth and development, especially during light-mediated processes such as flowering, maturation, embryo development, and petal differentiation and expansion (Borello et al., 1993; Terzaghi and Cashmore, 1995; Luo et al., 2010). The first *GATA* TF gene identified in plants was *NTL1* in tobacco (*Nicotiana tabacum*) (Daniel-Vedele and Caboche, 1993). In *Arabidopsis thaliana*, 30 *GATA* TFs have been identified and characterized (Reyes et al., 2004). In one study, *GNC* (*GATA*, Nitrate-inducible, Carbon-metabolism involved) and *GNL* (*GNC*-like) were demonstrated to play key roles in the control of greening, flowering time, and senescence (Richter et al., 2013). In addition, 28 *GATA* loci have been identified in *Oryza sativa* (Reyes et al., 2004), and 64 *GATA* TFs have in detected in the soybean genome (*Glycine max*) (C. Zhang et al., 2015a), thus further implying that plant *GATA* TFs play crucial roles in plant growth and development.

As a source of natural fiber, *Gossypium* is one of the most important cash crops. High-throughput single-nucleotide polymorphisms (SNPs) have been developed for upland cotton (*G. hirsutum*) in previous studies and mostly used for linkage genetic map construction and quantitative trait locus (QTL) mapping (Wang et al., 2015; Qi et al., 2017; Zhang et al., 2017). In previous research, we identified several candidate genes corresponding to *GATA* TFs in stable QTLs, such as *GhGATA40* in *qFS-chr13-1*, *GhGATA52* in *qFS-chr19-1*, and *GATA60* and *GATA61* in *qFS-chr25-1* (Zhang et al., 2017). These genes have also been detected in stable QTLs such as *qBW-chr25-5* and *qBW-chr13-4* (Zhang et al., 2016), *qFM-chr13-1* (Jamshed et al., 2016), and *qFS-chr25-4*, *qFL-chr25-3*, and *qFM-chr25-7* (Z. Zhang et al., 2015b), which suggests that *GATA* TFs contribute to cotton plant growth and development and the formation of fiber quality traits. To explore the evolution and expression patterns of *GATA* genes in *Gossypium* species, we therefore identified *GATA* genes and constructed a phylogenetic tree. We used RNA-seq data and qRT-PCR to demonstrate and validate the expression levels of *GATA* genes at each cotton developmental stage. In addition, we analyzed the chromosomal distribution, gene structure, conserved motifs, synteny, promoters, and selection pressures of *GATA* genes in *G. arboreum*, *G. raimondii*, and *G. hirsutum*.

2. Materials and methods

2.1. Plant materials

Plants of *G. hirsutum* 'TM-1' were grown under standard field conditions in Anyang, Henan Province, China. Flower buds were tagged, and the first day of flowering was recorded as 0 DPA. Cotton fiber for qRT-PCR experiments was collected in the morning, with three biological replicates, every 5 days from 0 to 30 DPA. All fiber samples were frozen in liquid nitrogen and stored at −80 °C.

2.2. RNA isolation and cDNA synthesis

Total RNA was extracted from cotton fiber using a RNA Prep Polysaccharides and Polyphenolics-Rich Pure Plant kit (Tiangen, Beijing, China) following the manufacturer's instructions. The extracted RNAs were quantified on a Nanodrop 2000 spectrophotometer and examined on 1% agarose gels. The A260/280 ratio of each RNA sample ranged from 1.8 to 2.1. cDNA was synthesized from the total RNA of each sample using a PrimeScript RT Perfect Real Time reagent kit (Takara, Dalian, China).

2.3. Transcriptome and qRT-PCR analyses

RNA-seq data were downloaded from the Sequence Read Archive (SRA) of the NCBI (https://www.ncbi.nlm.nih.gov/) from different developmental stages of *G. raimondii* (SRA048621; Wang et al., 2012a), *G. arboreum* (SRA150181; Li et al., 2014b), and *G. hirsutum* (SRP049330; Tuttle et al., 2015) and from different tissues of *G. hirsutum* (PRJNA248163; T. Zhang et al., 2015d). The SRA data were converted to FASTQ format using the SRA Toolkit with the parameter –split-3. Reference genome sequences of *G. hirsutum* (T. Zhang et al., 2015c), *G. arboreum* (Li et al., 2014a), and *G. raimondii* (Wang et al., 2012b) were downloaded and used for library construction with Bowtie2 (Langmead and Salzberg, 2012). The FASTQ data were filtered using fastx-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) based on a quality score criterion of Q20 and were then inspected with the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Fragments per kilobase of exon per million fragments mapped (FPKM) values of isoform genes were calculated by Cufflinks (Trapnell et al., 2010) from BAM files produced by Tophat2 using the parameters – library-type and fr-understand (Kim et al., 2013). *GATA* TF genes with FPKM values > 10 during at least one developmental stage were considered to be expressed genes. The expression data were analyzed and then transformed using the expression $\log_2(FPKM + 1)$. Gene expression patterns were visualized with heatmaps generated using v3.3.0 of R (https://CRAN.R-project.org/package=pheatmap).

GhGATA gene-specific primers for qRT-PCR were designed using Primer-BLAST (Ye et al., 2012) (Additional file 1). The *GhHistone3* (AF024716) gene was used as an internal control to normalize *GhGATA* gene expression (Xu et al., 2004). All qRT-PCR samples were amplified on a Roche 480 PCR system (Roche, Diagnostics, Penzberg, Germany). Gene expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method based on three biological replicates and three independent PCR runs (Livak and Schmittgen, 2001).

2.4. Identification of GATA genes and multiple sequence alignment

To identify members of the *GATA* TF family, HMMER 3.0 (Finn et al., 2011) was used along with profile hidden Markov model (PF00320.26) to search for protein sequences in published genomes of the following five species: *A. thaliana* (Cao et al., 2011), *O. sativa* (Goff et al., 2002), *G. raimondii* (Paterson et al., 2012), *G. arboreum* (Li et al., 2014b), and *G. hirsutum* (T. Zhang et al., 2015d). Sequences were verified in Blastp using *AtGATA* sequences as queries and an *e*-value cutoff of 1×10^{-5}. After removing redundant sequences, a multiple sequence alignment was produced with ClustalW (Thompson et al., 2002). Taxon labels and their corresponding gene IDs are shown in Additional file 2.

2.5. Phylogenetic tree construction

The multiple sequence alignment was used to construct a phylogenetic tree by the neighbor-joining method based on *p*-distances in MEGA6.06 (Tamura et al., 2013) with 1000 bootstrap replicates. To examine the evolution of cotton *GATA* genes, another multiple sequence alignment consisting of 179 cotton sequences was used to construct a second phylogenetic tree by the same method.

2.6. Chromosomal location, exon/intron structure, and motif analyses

The chromosomal distribution of *GATA* genes was determined