Gene 551 (2014) 206–213

Contents lists available at ScienceDirect

Gene

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

Characterization of the global transcriptome for cotton (*Gossypium hirsutum* L.) anther and development of SSR marker

Xianwen Zhang [a], Zhenwei Ye [a], Tiankang Wang [b], Hairong Xiong [c], Xiaoling Yuan [a], Zhigang Zhang [d], Youlu Yuan [b,*], Zhi Liu [a,*]

[a] College of Bioscience and Biotechnology, Hunan Agricultural University, Changsha 410128, China
[b] State Key Laboratory of Cotton Biology, Key Laboratory of Biological and Genetic Breeding of Cotton, The Ministry of Agriculture, Institute of Cotton Research, The Chinese Academy of Agricultural Sciences, Anyang 455004, China
[c] Key Laboratory for Crop Germplasm Innovation and Utilization of Hunan Province, Hunan Agricultural University, Changsha 410128, China
[d] Cotton Sciences Research Institute of Hunan Province, Changde 415101, China

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form 26 August 2014
Accepted 29 August 2014
Available online 29 August 2014

Keywords:
Cotton
Transcriptome
Transcription factor
SSR
Polymorphism

ABSTRACT

Cotton is an important fiber plant, and it's attractive to elucidate the molecular mechanism of anther development due to the close relationship between the anther fertility and boll-setting, and also fiber yield. In the present paper, 47.2 million paired-end reads with average length of 82.87 bp from the anthers of TM-1 (*Gossypium hirsutum* L.), a genetic standard line, were generated through transcriptome sequencing, and 210,965 unigenes of more than 100 bp were obtained. BLAST, KEGG, COG, and GO analyses showed that the genes were enriched in the processes of transcription, translation, and post-translation as well as hormone signal transduction, the transcription factor families, and cell wall-related genes mainly participating in cell expansion and carbohydrate metabolism. Further analysis identified 11,153 potential SSRs. A suit of 5122 primer pair sequences were designed, and 82 of 300 randomly selected primer pairs produced reproducible amplicons that were polymorphic among 22 cotton accessions from *G. hirsutum*, *Gossypium barbadense* and *Gossypium arboreum*. The UPGMA clustering analysis further confirmed high quality and effectiveness of these novel SSR markers. The present study provided insights into the transcriptome profile of the cotton and established a public information platform for functional genomics and molecular breeding.

1. Introduction

Cotton (*Gossypium* spp.) is one of the most economically important crops due to its fiber used as the principal natural source for the textile industry worldwide. The *Gossypium* genus contains 5 tetraploid (AD_1 to AD_5, $2n = 4\times$) and over 45 diploid ($2n = 2\times$) species (where n is the number of chromosomes in the gamete of an individual). There are four cultivated cotton species, two diploids from Africa–Asia, *Gossypium herbaceum* L. (Gher, A_1 genome) and *Gossypium arboreum* L. (Ga, A_2 genome), and two tetraploids from Americas, *Gossypium hirsutum* L. (Gh, AD_1 genome) and *Gossypium barbadense* L. (Gb, AD_2 genome). At present, *G. hirsutum* is the most widely cultivated cotton species, accounting for more than 90% of the world cotton production (http://en.wikipedia.org/wiki/Gossypium_hirsutum). Apart from its economic value, cotton is also an excellent model system for studying polyploidization, cell elongation and cell wall biosynthesis (Al-Ghazi et al., 2009; Paterson et al., 2012; Qin and Zhu, 2011). Recently, a number of genome resources have been developed from the genus *Gossypium* including the construction of high-density tetraploid cotton genetic linkage maps (Guo et al., 2008; Yu et al., 2011). It's more attractive that a draft genome of a diploid cotton putative *G. raimondii*, a putative D-genome parent, was created using a whole-genome shortgun strategy (K. Wang et al., 2012). However, it's an urgent task to provide more genome-wide information about tetraploid cotton.

The development of functional pollen and its releasing at appropriate stage to maximize pollination and fertilization are critical for plant reproduction, the creation of genetic diversity and biological productivity. These processes require cooperative interactions between gametophytic and sporophytic tissues within anther (Wilson and Zhang, 2009; Zhang and Wilson, 2009). Based on molecular studies, large numbers of genes related to pollen and anther development have been identified, and anther and pollen development pathway has been elucidated, especially in *Arabidopsis* and rice (Wilson and Zhang, 2009; Wilson et al., 2011). Stamen initiation is controlled by the homeotic genes *APETALA3* (*AP3*), *PISTILLATA* (*PI*), and *AGAMOUS*

Abbreviations: BLAST, Basic Local Alignment Search Tool; KEGG, Kyoto Encyclopedia of Genes and Genomes database; COG, Clusters of Orthologous Groups; GO, gene ontology; PCR, Polymerase Chain Reaction; SSR, Simple Sequence Repeat; TF, transcription factor; EST, Expressed Sequence Tag; UPGMA, Unweighted Pair Group Method with Arithmetic Mean.

* Corresponding authors.
E-mail addresses: yylcri@126.com (Y. Yuan), tigerzhiliu@gmail.com (Z. Liu).

http://dx.doi.org/10.1016/j.gene.2014.08.058
0378-1119/

(*AG*), with the primordia forming as a tetrad of archesporial cells. Then *AG* induces microsporogenesis via activation of *NOZZLE/SPOROCYTELESS* (*NZZ/SPL*) (Ito et al., 2004), and regulates stamen development at least in part by controlling the transcription of a catalytic enzyme of the lipid-derived phytohormone jasmonic acid (JA), DEFECTIVE IN ANTHER DEHISCENCE1 (Ito et al., 2007). The transcription factors *JAGGED* (*JAG*) and *NUBBIN* (*NAB*) are also involved in the process of defining stamen structure (Dinneny et al., 2006). A transcriptome analysis successfully selected from hundreds of transcripts several transcripts encoding potential proteins for lipid exine synthesis during early anther development in rice (Huang et al., 2009).

The latest research on cotton (*G. hirsutum*) anthers of the wild type (WT) and the genetic male sterility (GMS) mutant (in the WT background) in three stages of meiosis, tetrad, and uninucleate microspore using digital gene expression (DGE) method identified many genes specific to anther development (Wei et al., 2013). Generation of ESTs from shoot apexes, squares, and flowers in upland cotton (*G. hirsutum*) forms a valuable foundation for gene expression profiling analysis, functional analysis of newly discovered genes, genetic linkage, and quantitative trait loci analysis (Lai et al., 2011).

Though the basic mechanisms of pollen and anther development could be cross-referenced, each species has its own peculiarity. Furthermore, most of these studies were carried out on self-pollinated and cross-pollinated plants. Pollen and anther development in Upland cotton, an often cross-pollinated crop, may somewhat differ from these other species.

Herein, in the present study, we generated the transcriptome profiling of anthers from Upland cotton TM-1 (*G. hirsutum* L.), a genetic standard line, and identified some transcription factor genes and some genes involved in cell wall formation. A total of 4376 potential SSRs were characterized, and further PCR verification confirmed the significant polymorphism of 82 SSR loci. The objective of the present study is to get a close understanding of the molecular mechanism of cotton anther development, and establish a sound foundation for functional genomics, comparative genomics analysis, molecular breeding and new gene cloning in cotton based on high throughput sequence.

2. Materials and methods

2.1. Plant materials

Plants of upland cotton, the genetic standard TM-1 (*G. hirsutum* L.), were grown in an experimental field under standard field conditions in 2012. The anther at the stages of sporogenous cell division (bud length < 3.0 mm), pollen mother cell meiosis (bud length = 3.1–4.5 mm), uninucleate microspore (bud length = 4.6–12 mm) and mature pollen (bud length > 12 mm) were harvested according to Deng et al. (2010), respectively, frozen in liquid nitrogen immediately, and stored at −80 °C for use.

2.2. RNA isolation and sequencing

For Illumina sequencing, the total RNA of each sample was isolated using Trizol (Invitrogen, Carlsbad, CA) and further purified with the RNeasy Plant Mini Kit (Qiagen, Valencia, CA). RNA quality was verified using a 2100 Bioanalyzer RNA Nanochip (Agilent, Santa Clara, CA) and all samples had RNA Integrity Number (RIN) value more than 8.5. Then RNA was quantified using NanoDrop ND-1000 Spectrophotometer (Nano-Drop, Wilmington, DE).

The RNAs from four samples were mixed with equal amount, and then Illumina sequencing using the Solexa platform was performed according to the manufacturer's instructions (Illumina, San Diego, CA) in Shanghai Quanmai Bio-Technology Co., Ltd. Then, the stringent filter process on the raw sequencing data was carried out with the criteria that quality threshold is 20 (error rate 1%) to remove low-quality data, and length threshold is 35 bp to rule out the ambiguous N-containing sequences. Therefore, 47.2 million sequencing reads with 82-mer length. The sequencing data are deposited in NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra) with accession number SRP041153.

2.3. De Novo assembly and analysis of illumina reads

The samples were assembled with SOAPdenovo (Li et al., 2010). The reads were first combined to form longer fragments, i.e., contigs. The EST sequences were downloaded from cotton database (http://gossypium.info/, cotton46a), and further mapping was performed using Bowtie and Velvet software to obtain non-redundant unigenes that were as long as possible.

The resulting unigenes were determined by performing BLASTX (V2.2.14) (Altschul et al., 1997) searches against protein databases, with the priority order of NR (non-redundant protein sequences in NCBI), Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes database (KEGG) (V56.0, Oct. 1, 2010) (Kanehisa et al., 2008), and COG with the criteria of E-value ≤ 1e−3. Based on the results of the protein database annotation, Blast2GO (Conesa et al., 2005) was employed to obtain the functional classification of the unigenes based on Gene Ontology (GO) terms. WEGO software (Ye et al., 2006) was used to perform the GO functional classification for all unigenes. Further BLASTx against *Arabidopsis thaliana* Transcription Factors database (PlantTFDB), Cell Wall Navigator (CWN) protein database and MAIZEWALL database, respectively, were performed. The expectation (E)-value cutoff was set at 1E−5.

2.4. SSR mining and confirmation

MISA (http://pgrc.ipk-gatersleben.de/misa/), a Perl script, was used to identify microsatellites (SSR, Simple Sequence Repeat) in the unigenes identified in the study. The parameters for the SSR search were defined as follows: the size of motifs was two to six nucleotides, and the minimum repeat unit was defined as six for dinucleotides, and five for trinucleotides to hexanucleotides.

Primer Premier 6.0 (PREMIER Biosoft International, Palo Alto, CA) was used to design PCR primers in the flanking regions of the SSRs. The criteria of the primer design were as follows: primer length of 18–24 bp, GC content between 40–65%, and melting temperature between 50–65 °C. The expected product size was between 100 bp and 350 bp with no secondary structures. The redundancy analysis of these primers for all SSRs obtained in this study was carried out using SSRD software (W. Wang et al., 2012) after comparing with primers from CMD (http://www.cottonmarker.org/). Among all the non-redundant designed primers, 300 primer pairs were randomly selected to evaluate their application and polymorphisms in 22 cotton accessions that include TM-1, CCRI36, zhong221, zhongR014121, xinluzao24, CCRI19, CCRI35, Acala3080, NM970513, 7235, TAM91D-3, CCRI60, 0–153, 9708, 177 and 48 from *G. hirsutum*, hai1, 7124, P62ne10, Pima_S6 and 3–79 from *G. barbadense*, and shixiya1 from *G. arboreum*.

For SSR confirmation using PCR, the genomic DNA was extracted with 1 g young cotton leaf according to the previous method (Paterson et al., 1993). Each 10 μL PCR reaction mixture contained 1 μL 10× PCR buffer, 0.5 μL 10 mM dNTP, 0.5 μL forward primer (10 μM), 0.5 μL reverse primer (10 μM), 1.2 μL DNA (30 ng/μL), 0.1 μL Taq DNA polymerase (5 U/μL) and 6.2 μL ddH_2O. DNA amplification was programmed at 94 °C for 5 min for initial denaturation, then 30 cycles at 94 °C (30 s)/58 °C (45 s)/72 °C (45 s). The final extension step was 2 min at 72 °C. Each PCR product was analyzed using 8% native PAGE (polyacrylamide gel electrophoresis), and SSR marker analysis was done according to the previous reports (Zhang et al., 2000, 2002). The similarity and clustering analysis on 22 cotton accessions was performed using NTSYS-pc 2.20 with UPGMA (Unweighted Pair Group Method with Arithmetic Mean) (Rohlf, 2005), and finally the systematic tree was produced by MEGA4 software (Tamura et al., 2007). The number of alleles (*Na*), expected