



Comparative genomic analysis of the PKS genes in five species and expression analysis in upland cotton

Xueqiang Su^{*}, Xu Sun^{*}, Xi Cheng, Yanan Wang, Muhammad Abdullah, Manli Li, Dahui Li, Junshan Gao, Yongping Cai and Yi Lin

School of Life Science, Anhui Agricultural University, Hefei, China

^{*}These authors contributed equally to this work.

ABSTRACT

Plant type III polyketide synthase (PKS) can catalyse the formation of a series of secondary metabolites with different structures and different biological functions; the enzyme plays an important role in plant growth, development and resistance to stress. At present, the PKS gene has been identified and studied in a variety of plants. Here, we identified 11 PKS genes from upland cotton (*Gossypium hirsutum*) and compared them with 41 PKS genes in *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*. According to the phylogenetic tree, a total of 52 PKS genes can be divided into four subfamilies (I–IV). The analysis of gene structures and conserved motifs revealed that most of the PKS genes were composed of two exons and one intron and there are two characteristic conserved domains (Chal_sti_synt_N and Chal_sti_synt_C) of the PKS gene family. In our study of the five species, gene duplication was found in addition to *Arabidopsis thaliana* and we determined that purifying selection has been of great significance in maintaining the function of PKS gene family. From qRT-PCR analysis and a combination of the role of the accumulation of proanthocyanidins (PAs) in brown cotton fibers, we concluded that five PKS genes are candidate genes involved in brown cotton fiber pigment synthesis. These results are important for the further study of brown cotton PKS genes. It not only reveals the relationship between PKS gene family and pigment in brown cotton, but also creates conditions for improving the quality of brown cotton fiber.

Submitted 29 May 2017
Accepted 10 October 2017
Published 30 October 2017

Corresponding authors
Yongping Cai,
1806149539@QQ.COM,
ypcaiah@163.com
Yi Lin, linyi992547404@163.com

Academic editor
Sheila McCormick

Additional Information and
Declarations can be found on
page 19

DOI 10.7717/peerj.3974

© Copyright
2017 Su et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Agricultural Science, Bioinformatics, Genomics, Plant Science

Keywords Upland cotton, Polyketide synthase, Procyanidins, Gene expression, Genome-wide analysis

INTRODUCTION

Plant polyketone compounds are secondary metabolites having a cyclic structure with an oxygen atom bound to the carbon ring. This group includes phenols, stilbene and flavonoid compounds (*Abe & Morita, 2010*). Owing to the complexity and variety of the pathways and mechanisms of biosynthesis, the number of polyketone compounds is very large and their molecular structures are complex. This complexity results in the compounds having prominent and varied biological activities (*Austin & Noel, 2002*). The biosynthesis of this group has a common mechanism that includes the enzyme polyketide synthase (PKS). According to the structure of the protein, PKS can be divided into PKS I, II and III

(Funa *et al.*, 1999). PKS I and PKS II only exist in microorganisms. Each form has many functional modules and monofunctional subunits (Xie *et al.*, 2016). The PKS III gene family exists mainly in the plant kingdom, but some occur in a few species of bacteria. PKS III gene family members can catalyse plant secondary metabolites having various structures, biological activities and chalcone synthase (CHS) backbones. Examples of such metabolites include chalcone, stilbene, benzophenone, acridone, phloroglucinol, resorcinol and pyrone (Austin & Noel, 2002). These secondary metabolites play important roles in the colouring of plant organs, safeguarding from pesticides and prevention of UV irradiation damage (Li *et al.*, 2016).

The type III PKS gene family is divided into chalcone synthase (CHS) and chalcone synthase-like protein (CHSL) subfamilies. Chalcone synthase is the core enzyme of the PKS III gene family and is the first key enzyme for the plant flavonoid synthesis pathway and the rate-limiting enzyme (Martinez-Perez *et al.*, 2014). The PKS III gene family also includes a series of gene duplications and functional differentiation derived from the class of CHS-like proteins (CHSL) (Eom & Hyun, 2016). CHSL protein is far from the biosynthesis of PAs, The main role is to help plants adapt to changes in the environment, especially in response to fungal invasion (Han *et al.*, 2014). The CHSL of the PKS III gene family include 2-pyrone synthase cloned from *Gerbera hybrida* (Helariutta *et al.*, 1995), acridone synthase cloned from (Junghanns *et al.*, 1995), benzalacetone synthase cloned from *Rheum palmatum* (Abe *et al.*, 2001) and stilbene synthase cloned from *Pinus sylvestris* (Schanz, Schröder & Schröder, 1992). Because of the evolution from a common ancestor, PKS III gene family members have a high degree of homology between the structure and catalytic mechanisms and are very similar. For example, their proteins are essentially homodimers consisting of 40–45 kDa subunits and their active sites have a catalytic triad that is composed of Cys-His-Asn. The functional differences of CHSL and CHS lie in the preference towards different substrates when catalytic reactions occur, changes in the malonyl-CoA number of condensation and different cyclic ways of production (Schröder, 2000).

The first PKS gene was reported in 1983 in a study of *PcCHS* in *Petroselinum crispum* and was shown to be involved in the biosynthesis of flavonoids (Reimold *et al.*, 1983). The study of the PKS III gene family continues today. Chalcone synthase (CHS) is by far the most thoroughly studied type III polyketide synthase. CHS catalyses the first step in the synthesis of flavonoids and CHS is responsible for catalysing the reaction of 1 molecule of 4-benzoyl-CoA with 3 molecules of malonyl-CoA to form chalcone (Burbulis & Winkel-Shirley, 1999), the precursor of many flavonoid compounds. The enzymes chalcone isomerase (CHI), flavanone 3-hydroxylase (F3H), flavonoid 3'-hydroxylase (F3'H), dihydroflavone-4-reductase (DFR) and other enzymes have a common catalytic role in the formation of a variety of flavonoids (Feng *et al.*, 2013). Currently, the cloning and functional analysis of CHS have been reported for many species, e.g., *Oryza sativa* (Hu *et al.*, 2017), *Hypericum monogynum* (Jepson *et al.*, 2014), *Gerbera hybrida* (Helariutta *et al.*, 1995), *Petunia hybrida* (Koes *et al.*, 1989), *Malus domestica* (Dare *et al.*, 2013) and *Glycine max* (Tuteja *et al.*, 2004).

Study of the PKS III gene family in the important cash crop cotton has yet to be conducted. Cotton is an important fiber crop, but it is also used for oil ([Cui et al., 2017](#)), drugs ([Stipanovic et al., 2005](#)) and other purposes. Naturally colored cotton can be divided into two categories: brown cotton and green cotton. It can synthesize and accumulate pigment to make mature fibers with varied colours during fiber development ([Yuan et al., 2012](#)). At present, the application and cultivation of a wide range of naturally coloured cotton varieties produce mainly brown cotton and green cotton. Brown cotton fiber pigments are more stable than those of green cotton; this, combined with its high yield, has led to brown cotton becoming the dominant colour of natural cotton varieties ([Qian et al., 2015](#)). Brown cotton is widely favoured for its commercial value and application characteristics, including the lack of need for dyeing, its anti-static electricity properties, ultraviolet resistance and good flame retardance ([Hinchliffe et al., 2016](#)). Brown cotton flavonoids are also closely related to resistance to pests and diseases; increasing the flavonoid content can increase plant resistance to insects and thus brown cotton has been widely favoured with increasing commercial value and application prospects ([Fan et al., 2016](#)). However, brown cotton fibers do have some problems; these include poor pigment stability, uneven pigment distribution and poor fiber quality ([Hua et al., 2007](#)). These problems can restrict the market value of brown cotton. To solve these problems, we focused on the synthesis of brown cotton pigment to improve the quality of brown cotton at the molecular level. At present, many studies have shown that brown cotton pigment is mainly composed of PAs ([Gao et al., 2016](#)). In addition, high quality varieties rich in procyanidins are reported in many species, these breeds not only have high commercial value but also help to improve our understanding of flavonoid metabolic pathways precious resources. For example: black rice since ancient times is a very precious ingredient, the color of this grain deepened is due to the accumulation of PAs in rice ([Oikawa et al., 2015](#)); *Solanum tuberosum* has high intensity of coloring and high nutritional value of food, which are due to *Solanum tuberosum* rich in PAs ([Gras et al., 2017](#)); what we used to know is that corn are orange particles, but purple corn is more resistant to storage than orange corn and has higher nutritional value ([Luna-Vital et al., 2017](#)). These varieties are all rich in PAs, PAs metabolism is an important branch of flavonoid metabolism and thus the PKS III gene family plays an important role in the synthesis of PAs. Thus it can be seen that the study of the PKS gene family is very important not only in brown cotton, but also has a very important significance in many species. The study of PKS gene family can not only help us to better understand the metabolic pathway of flavonoids but also can produce huge commercial value.

Although the whole genome of upland cotton (*Gossypium hirsutum*) has been sequenced ([Li et al., 2015](#)), the whole genome identification and analysis of the type III polyketide synthase family in terrestrial cotton have not yet been reported. The relationship between PKS genes and fiber quality in brown cotton remains unknown. In the present study, we screened the PKS family in upland cotton and analysed the characteristics of its evolution, gene structure, conserved motifs and duplication events. The study species for comparison of the PKS III gene family included *Populus tremula*, *Arabidopsis thaliana*, *Vitis vinifera* and *Malus domestica*. *Arabidopsis thaliana* is a widely used research plant and its synthesis of flavonoids is more thoroughly understood, while the other three species are rich in

flavonoids. Therefore, the choice of these four species for comparison with the upland cotton can help us better understand terrestrial cotton flavonoid metabolism. According to the analysis of promoter cis-acting elements and the expression patterns of PKS genes in upland cotton, the candidate PKS genes relating to the brown cotton fiber pigment were identified, which provides an important theoretical foundation and genetic resource for improving the uneven distribution, poor stability and fiber quality of natural brown cotton. At the same time, we further analysed the expression patterns of PKS family members and discussed their relationship with the changes in PAs at different developmental stages to determine the PKS candidate genes associated with brown cotton fiber pigment. These results will provide an important theoretical basis for improving the uneven distribution and poor stability of natural brown cotton pigment.

MATERIALS AND METHODS

Plant materials

Brown cotton plants used line Zongcaixuan No. 1 (brown fiber line) in the experiment were grown in an agricultural park (Hefei, Anhui, China). This brown cotton line belongs to tetraploid upland cotton. In July 2016, 50 brown cotton plants with good growth characteristics were selected at the blooming stage. We began collecting cotton bolls after 3, 6, 9, 12, 15, 18 and 21 days after flowering (DAF). The experimental materials were frozen in liquid nitrogen and quickly transferred to the laboratory refrigerator.

Identification and collection of PKS proteins

In our study, the genomic data of *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana* were downloaded from the Phytozome database (Hu et al., 2016) (<https://phytozome.jgi.doe.gov/pz/portal.html>). DNATOOLS software was used to establish a local database of the amino acid sequences (Curran & Tvedebrink, 2013), including the whole genomes of *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*. The sequences in TBlastN (E -value = 0.001) were queried according to the two conservative domains Chal_sti_synt_N and Chal_sti_synt_C (Han et al., 2016) and compared with the established local database sequences of *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*. Preliminary PKS candidate gene sequences were screened out. The PKS candidate gene sequences obtained by BLAST were tested for whether they contained the two conserved Chal_sti_synt_N and Chal_sti_synt_C domains using Pfam (Bateman et al., 2004) (<http://pfam.xfam.org/>) and SMART (Letunic, Doerks & Bork, 2012) (<http://smart.embl-heidelberg.de/>) online software. Multiple sequence alignment and repeat sequence removal were analysed using the ClustalW tool of the MEGA 7.0 software (Kumar, Stecher & Tamura, 2016). The molecular weight of the PKS protein was predicted using the ExPASy Proteomics Server software (Artimo et al., 2012) (<http://web.expasy.org/protparam/>). WoLFPSORT (Horton et al., 2007) (<http://www.genscript.com/wolf-psort.html>) was used to predict the PKS protein subcellular localization.

Phylogenetic analysis

Protein sequence alignment was performed using the Clustal X program (Des Higgins, Dublin, Ireland). The phylogenetic tree was built using the Neighbour-Joining (N-J) method with 1,000 bootstraps and MEGA 7.0 (Kumar, Stecher & Tamura, 2016). The *GhPKS* genes were classified according to the phylogenetic relationships. If two different species of genes are located in the phylogenetic tree at the same node and the sequence similarity is more than 80%, we consider two of these are orthologous genes (Van der Heijden et al., 2007).

Gene structural and conserved motif analysis

The map of the PKS gene structure including *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana* was displayed using Gene Structure Server (Guo et al., 2007) (<http://gsds.cbi.pku.edu.cn>). The motifs of PKS genes in *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana* were analysed using MEME online analysis software (Bailey et al., 2015) (http://meme.sdsc.edu/meme4_3_0/intro.html). The specific parameters were as follows: the motif number was 20 and the minimum and maximum widths were 6 and 200, respectively. The motif annotations were obtained from the SMART and Pfam databases.

Chromosomal location and gene duplication

Chromosome starting position and other relevant information concerning the PKS genes were obtained from the public genome database of *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*. The chromosome physical locations of the PKS genes of all five species were obtained using MapInspect (Niu et al., 2016) (<http://mapinspect.software.informer.com>) software. The gene is located on the same chromosome, separated from the 200 kb and more than 80% similarity gene called tandem duplication; whereas genes that duplicated genes on different chromosomes and more than 80% similarity gene called fragment duplication (Long & Thornton, 2001). Non-synonymous (Ka) and synonymous (Ks) sites were calculated using the DnasP v5.0 software (Librado & Rozas, 2009). Sliding window analysis was also performed using the DnasP v5.0 software; the parameters were as follows: window size, 150 bp; step size, 9 bp.

Upland cotton PKS gene promoter cis-acting element analysis

The promoter sequence of each PKS gene was obtained from the genome database for *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*; this includes the DNA sequence of the initiation codon (ATG) located 1,500 bp upstream of each PKS gene. We used the online software Plantcare (Rombauts et al., 1999) (<http://bioinformatics.psb.ugent.be/webtools/plantcarere/html/>) to analyse the promoter region cis-acting elements.

RNA extraction and qRT-PCR

In this study, 11 PKS genes of upland cotton were quantitatively analysed by real-time fluorescence. Cotton bolls at 3 DAF, 6 DAF, 9 DAF, 12 DAF, 15 DAF, 18 DAF, 21 DAF were collected and RNA was extracted using the Tiangen (Beijing, China) plant RNA

extraction kit. Reverse transcription was performed using a PrimeScript™ RT reagent kit with gDNA Eraser (Takara, Tokyo, Japan) and each reaction used 1 µg of RNA. The specific primers for the PKS gene of upland cotton (Table S1) were designed using Beacon Designer 7 software and the internal reference gene used UBQ7 (Table S1). The qRT-PCR system consisted of 20 µL: 10 µL of SYBR® Premix Ex Taq™ II (2×) (Takara), 2 µL of cDNA and 0.8 µL of GhPKS-F and GhPKS-R. The reaction procedure was 40 cycles of 50 °C for 2 min, 95 °C for 30 s, 95 °C for 5 s and 60 °C for 20 s, followed by 72 °C for 10 min; the experiment was repeated three times. Finally, we used $2^{-\Delta\Delta C_t}$ for the calculation of relative expression (Livak & Schmittgen, 2001).

Determination of proanthocyanidin content in brown cotton fibers

The fibers of brown cotton bolls at 3 DAF, 6 DAF, 9 DAF, 12 DAF, 15 DAF, 18 DAF, 21 DAF were stripped, extracted with 80% methanol and subjected to ultrasonic extraction for 30 min. After centrifuging for 15 min, the resulting supernatant was analysed for soluble PAs. A methanol solution containing 1% HCl was added to the precipitate and the solution was placed in a 6 °C water bath for 1 h; after centrifugation for 15 min, the supernatant contained the insoluble PAs. The content of PAs was determined by the method of n-butanol-hydrochloric acid: 400 µL of procyanidin extract was added to 1.5 mL of n-butanol (containing 5% hydrochloric acid) in a boiling water bath for 20 min, after which the absorbance read at 550 nm (Ikegami et al., 2009).

RESULTS

Identification and evolutionary analysis using five genomes

Two kinds of plant PKS III genes conserved domains, Chal_sti_synt_N and Chal_sti_synt_C, were obtained from the Pfam protein database using a hidden Markov model. The two conserved domains have the respective molecular functions of transacylation and transferase. Sequences from TblastN (E -value = 0.001) were compared to the genome database of *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana* using Chal_sti_synt_N and Chal_sti_synt_C. A total of 52 PKS genes were identified (Table S2), including 11 in *Gossypium hirsutum* (GhPKS1-GhPKS11), 14 in *Populus tremula* (PtPKS1-PtPKS14), 13 in *Vitis vinifera* (VvPKS1-VvPKS13), 10 in *Malus domestica* (MdPKS1-MdPKS10) and 4 in *Arabidopsis thaliana* (AtPKS1-AtPKS4). In addition to the small number of PKS genes in *Arabidopsis thaliana*, the number of PKS genes in other species was not very different. To clarify the evolutionary relationships between the 11 PKS genes and PKS genes in four other cultivars, we constructed a phylogenetic tree for a total of 52 PKS genes (Fig. 1). According to the phylogenetic tree nodes, the 52 PKS genes can be divided into four subfamilies: I, II, III and IV. The number of subfamily I members was 17, followed by subfamily IV (12), subfamily III (11) and the lowest number of members was in the subfamily II (10). PtPKS5 and PtPKS7 separated into a class. Among the four subfamilies, the subfamilies I, IV included all five species and each species provided at least one PKS gene. Subfamily III contained three species (*Gossypium hirsutum*, *Populus tremula* and *Malus domestica*), while subfamily II only consisted of *Vitis vinifera*. It is noteworthy that subfamily I includes an

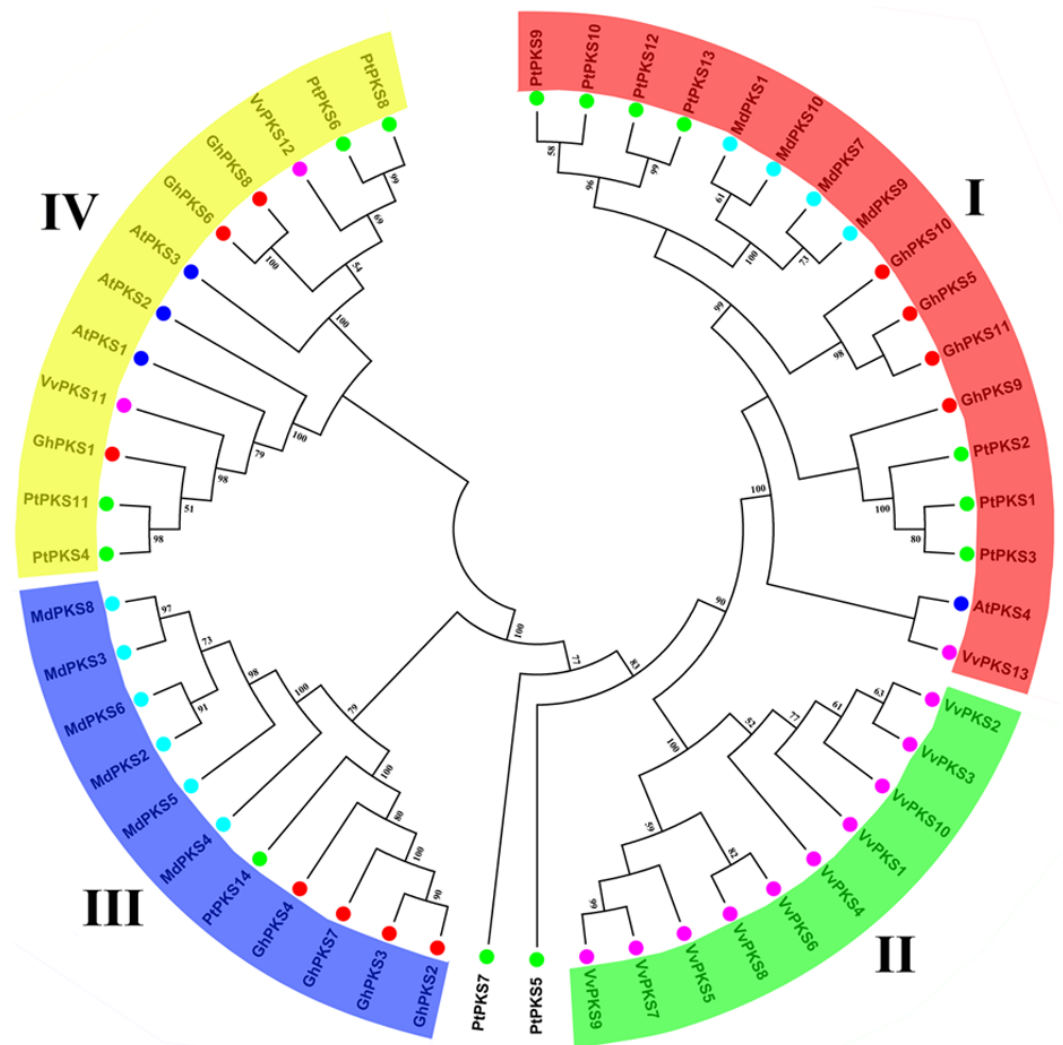


Figure 1 Phylogenetic analysis of PKS genes in upland cotton (*Gossypium hirsutum*), *Populus tremula*, *Vitis vinifera*, *Malus domestica*, and *Arabidopsis thaliana*. The PKS gene of each species is represented by a different colour: red indicates upland cotton; green represents *Populus tremula*; purple represents *Vitis vinifera*; pale blue represents *Malus domestica*; and the deep blue indicates *Arabidopsis thaliana*. According to the phylogenetic tree nodes, the PKS genes were divided into four subfamilies (*PtPKS5* and *PtPKS7* were placed separately into a class). Specific gene names are listed in Table S2.

Full-size [DOI: 10.7717/peerj.3974/fig-1](https://doi.org/10.7717/peerj.3974/fig-1)

Arabidopsis thaliana PKS gene (*AtPKS4*) (Owens et al., 2008). This gene is a CHS gene that has been reported in *Arabidopsis thaliana*. *Arabidopsis thaliana* plants were treated with high-intensity light for 24 h, resulting in a 50-fold increase in chalcone synthase activity and the accumulation of large amounts of anthocyanins (Courtney-gutterson et al., 1994). The four *GhPKSs* (*GhPKS5*, *GhPKS9*, *GhPKS10*, *GhPKS11*) were present in subfamily I, which may indicate that they are closely related to the accumulation of brown cotton fiber pigments. In addition, according to the results of the phylogenetic tree, there were no orthologous genes between the five species.

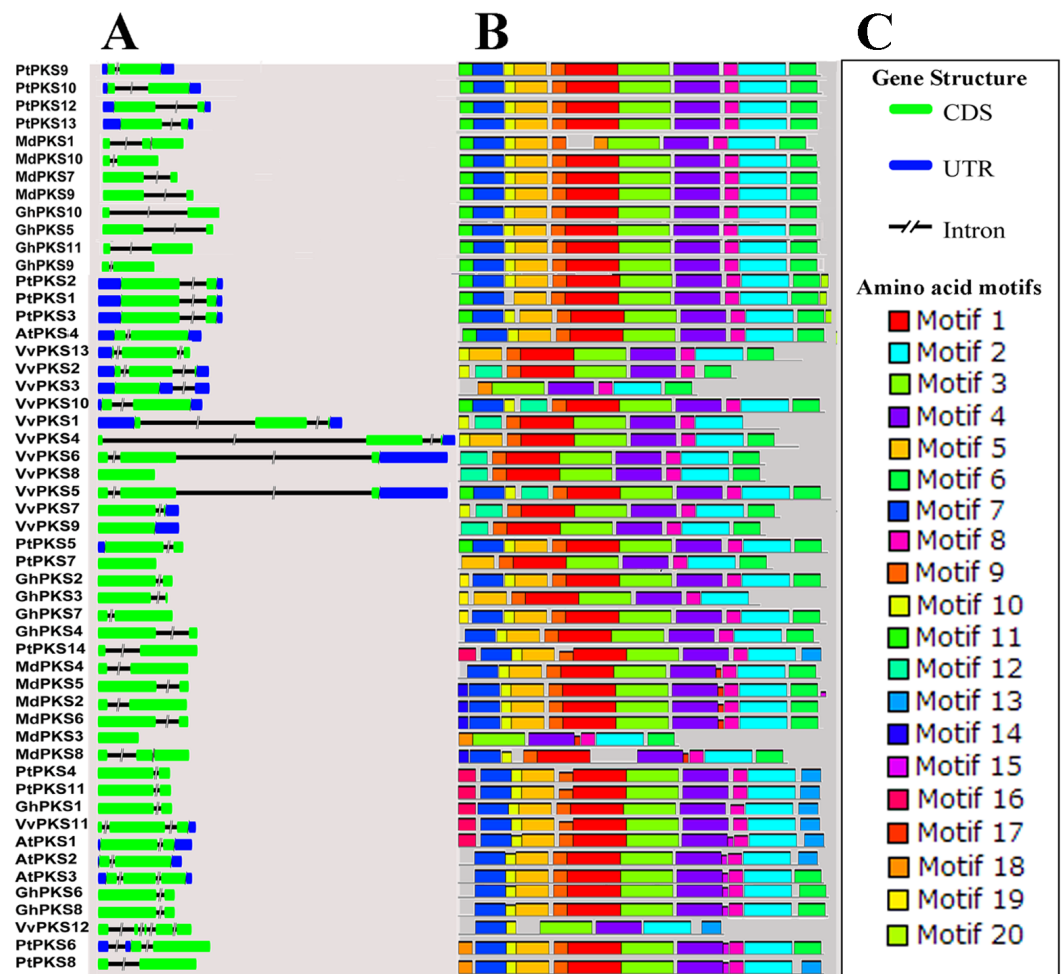


Figure 2 Exon-intron structure and motif composition of PKS genes across five plant species. (A) Gene structures of the PKS genes. (B) Distribution of MEME motifs in PKS genes. (C) Gene structure element and motif BOX serial number.

Full-size [DOI: 10.7717/peerj.3974/fig-2](https://doi.org/10.7717/peerj.3974/fig-2)

Structural and conserved motif analysis of PKS proteins

To understand the structural diversity of the PKS gene in a more comprehensive way, exon-intron pattern maps were constructed for the 52 PKS genes. As seen from the figure (Fig. 2A), there are 38 members of the 52 PKS genes consisting of two exons and one intron and as in previous reports, most of the plant PKS genes contain two exons and one intron (Durbin, Mccaig & Clegg, 2000). In the remaining 14 members, *VvPKS3* contains an exon and an intron. There are six members (*GhPKS9*, *MdPKS3*, *PtPKS7*, *VvPKS6*, *VvPKS8*, *VvPKS9*) with no introns. The remaining seven members (*AtPKS3*, *MdPKS8*, *VvPKS2*, *VvPKS4*, *VvPKS5*, *VvPKS11* and *VvPKS13*) are composed of three exons and two introns. *VvPKS12* has the largest number with five exons and four introns. There were no UTR regions found in the 23 PKS genes of *Gossypium hirsutum* and *Malus domestica*, while 73% of the members of the *Populus tremula*, *Vitis vinifera* and *Arabidopsis thaliana* group

had at least one UTR region. The results indicated that the structures of these genes were more complex. All the above results show that the PKS gene family has a diverse genetic structure, which helps to explain the divergence of PKS gene family members. To clarify the structures of the PKS genes, we attempted to gain a better understanding of the conserved motifs of these genes; we thus identified 20 conserved motifs (6–200 amino acid residue widths) using the MEME software (Table S3). The probability of occurrence of motifs 1–10 in upland cotton is more than 65%; we refer to this set as “General Motifs”. The remaining motifs 11–20 we refer to as “Specific Motifs” (Fig. 3) (Cao et al., 2016). Among the 20 motifs (Fig. 2B) we found that motifs 1, 3, 5, 7 and 12 encode a Chal_sti_synt_N conservative domain. Motifs 2, 4, 6 and 13 encode a Chal_sti_synt_C conservative domain. In upland cotton, in spite of *GhPKS3* lacking motifs 6, 7 and *GhPKS1* lacking motif 6. Almost all PKS family members included motifs 1, 2, 3, 4, 5, 6 and 7. However, in the other four species, this lack of motifs containing the Chal_sti_synt_N and Chal_sti_synt_C domains is more pronounced. For example, *Populus tremula* *PtPKS4*, 8 and 11 lack motif 6; *Malus domestica* *MdPKS8* lacks motifs 3, 5 and 7; in *Vitis vinifera* motifs 5 and 7 are present in only 3 and 4 members, respectively. In addition, motif 12 did not appear in 42 PKS proteins of *Gossypium hirsutum*, *Populus tremula*, *Malus domestica* and *Arabidopsis thaliana*, but motif 12 appeared only in two of the PKS proteins of *Vitis vinifera* (*VvPKS5*, *VvPKS10*). The frequency of motif 13 is also very low, with a total of only seven PKS family members. In the phylogenetic tree, the nearest members of each subfamily have similar motif combinations. Example combinations include *MdPKS7*, 9, *VvPKS6*, 8 and *PtPKS4*, 11. In addition, there are some proteins belonging to a subfamily with unique motifs. For example motif 15 is unique to subfamily IV and motif 17 only appears in the subfamily III. These subfamily-specific motifs play a very important role in the subfamily PKS proteins regarding function.

Comparison of GhPKS protein sequences with those of other plants

We identified and compared 11 sequences of PKS protein in upland cotton with the sequences of *Oryza sativa* chalcone synthase (*OsCHS*), *Arabidopsis thaliana* chalcone synthase (*AtCHS*) and *Medicago sativa* chalcone synthase (*MsCHS*) to clarify the functional divergence of PKS III gene family members. The results are shown in the figure (Fig. 4). The blue box and the red font in the figure represent the conservative amino acid residues and the sequence of the red regions shows a very high degree of conservation. The black wavy lines and arrows represent the α -helix and the β -sheet, respectively. The purple five-pointed star represents the catalytic triad (Cys-His-Asn) and the active amino acids (Thr, Phe, Gly, Ser) in the catalytically active central cavity are expressed as green or black triangles. When the plant PKS III enzyme catalyses the polyketone reaction, the starting substrate is first bound at the Cys in the catalytic triplet, followed by decarboxylation of the malonyl-CoA and the occurrence of the substrate condensation reaction so that the polyketone chain is continuously extended (Jez, Bowman & Noel, 2002). The final intermediate product undergoes a series of complex cyclization reactions that ultimately form the final product (Abe et al., 2001). Active amino acids located in the catalytically active central chamber can adjust the type of reaction-starting substrate and the length of

General Motifs	<i>Gossypium hirsutum</i> (11)	<i>Populus tremula</i> (14)	<i>Vitis vinifera</i> (13)	<i>Malus domestica</i> (10)	<i>Arabidopsis thaliana</i> (4)	Colour Key(%)	
Motif 1	11	14	11	10	4		10%
Motif 2	11	14	12	10	4		20%
Motif 3	11	14	13	9	4		30%
Motif 4	11	14	13	10	4		40%
Motif 5	11	14	3	8	4		50%
Motif 6	11	11	10	10	2		60%
Motif 7	10	14	4	9	4		70%
Motif 8	11	14	12	10	4		80%
Motif 9	11	14	11	9	4		90%
Motif 10	10	12	8	9	4		100%
Specific Motifs							
Motif 11	5	8	2	4	1		
Motif 12	0	0	2	0	0		
Motif 13	1	3	1	0	2		
Motif 14	0	0	0	4	0		
Motif 15	2	2	0	0	3		
Motif 16	1	2	1	0	1		
Motif 17	0	0	0	6	0		
Motif 18	0	2	1	2	0		
Motif 19	3	0	0	0	0		
Motif 20	0	2	0	0	0		

Figure 3 Distribution of motifs in PKS proteins from *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*. Colour key: the depth of colour indicates the percentage of motifs in the species.

Full-size  DOI: 10.7717/peerj.3974/fig-3

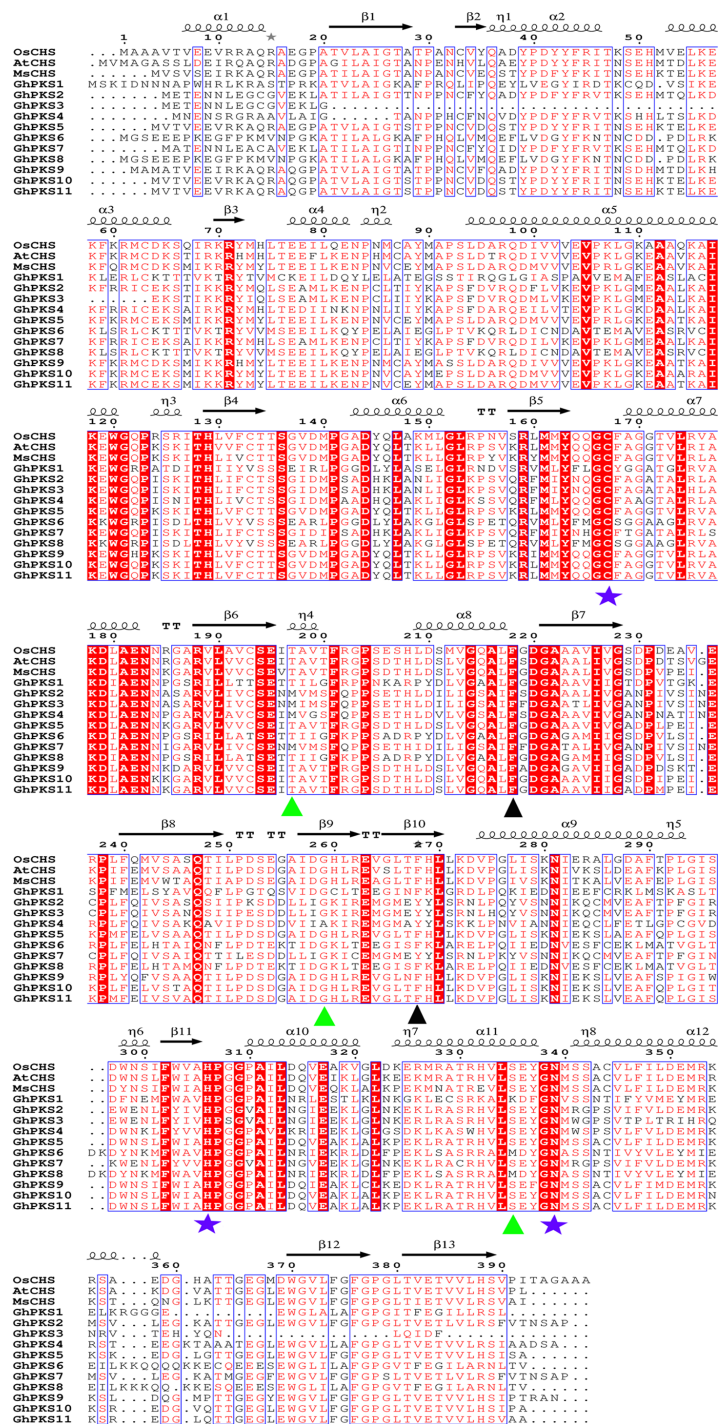


Figure 4 Sequence alignment of *GhPKSs* against the other plant species. The first line represents the secondary structure of *Oryza sativa* CHS. The blue box and the red font in the figure represent the conservative amino acid residues, and the sequence of the red regions shows a very high degree of conservation. The black wavy lines and arrows represent α -helices and β -sheet, respectively. The purple five-pointed star represents the catalytic triad, and the active amino acids are expressed in green or black triangles. *OsCHS*, *Oryza sativa* chalcone synthase (4350636); *AtCHS*, *Arabidopsis thaliana* chalcone synthase (AAB35812.1); *MsCHS*, *Medicago sativa* chalcone synthase (P30074).

Full-size DOI: 10.7717/peerj.3974/fig-4

the polyketone chain by adjusting the size of the catalytically active central chamber space (Jez, Bowman & Noel, 2002). The Cys-His-Asn catalysed triplets inherited from keto acyl synthase III (KASIII) (Austin & Noel, 2002) are highly conserved in each sequence in 11 PKS proteins of upland cotton. However, more amino acid substitutions occur at the four active amino acid positions. Thr at *GhPKS2*, 3, 4, 7 is replaced by a Met. Ser at *GhPKS1* is replaced by Lys and in *GhPKS6*, 8 is replaced by Met at the same position. The active amino acid Phe has two sites in the catalytically active central cavity and is closely related to the decarboxylation reaction of malonyl-CoA, which is represented by a black triangle in the figure. The first Phe active site was highly conserved in all upland cotton PKS proteins, but at the second Phe active site, Phe at *GhPKS2*, 3, 4, 7 was replaced by Tyr. The active amino acids Thr, Gly and Ser can regulate the specificity of the reaction substrate as well as the product. In the upland cotton PKS protein, amino acid substitution occurs in active amino acids Thr, Gly and Ser in multiple protein sequences; this phenomenon may be closely related to PKS III gene family functional diversity.

Chromosomal localization and gene duplication

To identify the distribution of PKS genes on the chromosome of each species and in the gene cluster, simultaneously to confirm the type of gene duplication events in upland cotton. We mapped the 52 PKS genes in five species (*Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*) to identify the chromosomal distribution of these PKS genes (Fig. S1). In our study, the PKS genes in the other four species were unevenly distributed on the chromosomes except for the distribution of the PKS gene in the *Vitis vinifera*, which was more concentrated on chromosome 16. In upland cotton, the PKS gene distribution was A2_chr6 (1), A2_chr8 (2), A2_chr9 (1), At_chr11 (1), Dt_chr8 (1), Dt_chr10 (1) and Dt_chr11 (4). In *Populus tremula*, the 14 PKS genes were distributed on chromosomes 1, 2, 3, 4, 5, 9 and 12. In *Malus domestica*, we found that the PKS genes were distributed on chromosomes 2, 9, 14, 15 and that *MdPKS1* was not mapped to any chromosome. The PKS genes in *Arabidopsis thaliana* are distributed on chromosomes 1, 4 and 5. However, in *Vitis vinifera*, 10 PKS genes were distributed on chromosome 16 and the remaining 3 PKS genes were distributed on chromosomes 3, 14 and 15. In the evolution of genes, most gene family expansion is due to the phenomenon of gene duplication, including tandem duplication and fragment duplication. To clarify how the PKS gene family was amplified, we examined the duplication of the PKS genes in five species (*Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*). Among the 52 PKS genes, we identified 10 gene duplication events in *Gossypium hirsutum* (2), *Populus tremula* (3), *Vitis vinifera* (3) and *Malus domestica* (2); in *Arabidopsis thaliana*, no gene duplication events were found. Five pairs of duplicated genes belonged to tandem duplication and five pairs of duplicated genes belonged to fragment duplication (Table 1). After analysing the gene duplication events of the PKS III gene family in five species, we calculated the Ka, Ks and Ka/Ks ratios of the eleven gene duplication events to explore the effects of these genes on the evolutionary processes (Table 1). In general, $Ka/Ks < 1$ represents negative selection or purification selection, $Ka/Ks > 1$ represents positive selection and $Ka/Ks = 1$ indicates neutral selection

Table 1 Ka/Ks analysis for the duplicated PKS paralogues from upland cotton, *Populus tremula*, *Vitis vinifera*, and *Malus domestica*. The chromosomal localization results are shown in Fig. S1, and the sliding window analysis results are shown in Fig. S2.

Duplicated pairs	Ka	Ks	Ka/Ks	Purifying selection	Duplicated type
<i>GhPKS5-GhPKS11</i>	0.0159	0.9533	0.017	Yes	Segmental
<i>GhPKS6-GhPKS8</i>	0.0033	0.0601	0.055	Yes	Segmental
<i>PtPKS6-PtPKS8</i>	0.0387	0.3075	0.126	Yes	Segmental
<i>PtPKS4-PtPKS11</i>	0.0475	0.3185	0.149	Yes	Segmental
<i>PtPKS12-PtPKS13</i>	0.0081	0.1357	0.060	Yes	Tandem
<i>MdPKS2-MdPKS6</i>	0.009	0.0291	0.309	Yes	Segmental
<i>MdPKS7-MdPKS9</i>	0.0068	0.3129	0.022	Yes	Tandem
<i>VvPKS1-VvPKS4</i>	0.0807	0.4019	0.201	Yes	Tandem
<i>VvPKS6-VvPKS8</i>	0.0094	0.0699	0.134	Yes	Tandem
<i>VvPKS7-VvPKS9</i>	0.0053	0.0213	0.249	Yes	Tandem

(Bitocchi et al., 2017). In our study, the Ka/Ks values of the 10 pairs of duplicated genes were less than 0.309. The results indicated that in these five species, the PKS III gene family was expanded due to gene duplication events and these repeated genes that undergo gene duplication experience strong purifying selection. Sometimes positive selection may be masked by strong negative selection. To identify positive selection of PKS loci in the occurrence of gene duplication events, we also performed a sliding window analysis of two pairs of duplicated genes in upland cotton (Fig. S2). There was never more than one repeat locus in the upland cotton, indicating that there was no positive selection for the two pairs of duplicated genes.

Analysis of cis-acting elements in the promoter of the PKS gene in upland cotton

To clarify the characteristics of the promoters of PKS genes in upland cotton, we analysed the cis-acting elements of 11 PKS gene promoters in upland cotton (promoter length = 1,500 bp) (Table S4). Strong light can regulate the expression of the PKS gene and there are many cis-acting elements in the promoter regions of PKS genes in upland cotton, e.g., Box4 (ATTAAT), SP1 (CC(G/A)CCC), CATT-Motifs (GCATTC) and many G-Boxes (CACGTT). It has been reported that *Arabidopsis thaliana* CHS genes were regulated by MYB transcription factors (Chezem & Clay, 2016). In our study, cis-acting elements associated with MYB transcription factors were also found in the promoter region of PKS genes in upland cotton, e.g., MBS (CGGTCA) and MRE (AACCTAA). This suggests that the expression of PKS genes may be regulated by MYB transcription factors. In addition, there are some cis-acting elements related to various life activities, TC-rich repeats (GTTTTCTTAC) associated with defence and stress, anaerobic induction of ARE (TGGTTT) and CGTCA-motifs (CGTCA) related to methyl jasmonate reactions. The specific cis-acting elements, concrete sequences and functions are shown in Table S5.

Expression characteristics of the PKS gene in upland cotton

The function and expression patterns of genes are closely related (Zhang *et al.*, 2014). To explore the expression patterns of PKS genes in upland cotton, we studied the expression patterns of 11 PKS genes in upland cotton at different stages of cotton fiber development, including 3 DAF, 6 DAF, 12 DAF, 15 DAF, 9 DAF, 18 DAF, 21 DAF and different plant parts, including roots, stems, leaves, fiber (cotton fiber development represented by 6 DAF). *GhPKS8* is a special gene in the 11 upland cotton PKS genes because no expression was detected in any tissue at any stage of cotton fiber development. The other 10 PKS genes were detected in all tissues and at different stages of cotton fiber development (Fig. 5). We found that *GhPKS1* is present at a higher level of transcription in the roots. *GhPKS2*, 3 and 7 showed a high expression level in the stems, while the expression levels in roots, leaves and fiber were low. *GhPKS4* showed high levels of transcription in all tissues of upland cotton. *GhPKS6* was highly expressed in the leaves, while the expression of *GhPKS5*, 9, 10 and 11 in cotton fiber was significantly higher than that in the other three plant tissues. The results of expression patterns of the 11 PKS genes in different tissues of upland cotton showed that *GhPKS5*, 9, 10 and 11 were mainly expressed in upland cotton fibers. We analysed the expression patterns of 11 PKS genes in upland cotton at different stages of cotton fiber development. The results showed that 11 PKS genes had multiple expression patterns. *GhPKS1*, 6 and 10 showed a gradual increase in transcription level from 3 DAF–15 DAF and the transcriptional level began to decrease after 15 DAF. *GhPKS2*, 7 had higher transcription levels at the later stages of fiber development and *GhPKS3*, 9 reached their highest levels at 12 DAF. *GhPKS4*, 5, 9 and 11 showed the highest amounts of transcriptional accumulation in the early stages of cotton fiber development. In brown cotton fibers, PAs are the main precursors of pigment. We also studied the accumulation of PAs in the fibers of brown cotton at different developmental stages (Fig. 6). The determination of PAs showed that both soluble and insoluble PAs had mainly accumulated before 15 DAF, after which its content gradually decreased; these results were consistent with the previously reported results (Li *et al.*, 2012). Interestingly, *GhPKS4*, 5, 9, 11 had a higher level of transcription at the early stages of cotton fiber development; the amount of expression then decreased gradually, which is consistent with the rule of accumulation of PAs in brown cotton fibers.

DISCUSSION

The plant PKS III gene family, which only exists in the plant kingdom, is associated with a variety of plant life activities (Shimizu, Ogata & Goto, 2017). The PKS III gene family is not very large and PKS III gene family members have been identified or cloned in several species. For instance, 14 PKS genes have been identified in *Zea mays* (Han *et al.*, 2016), 12 PKS genes have been isolated and sequenced in *Petunia hybrida* (Koes *et al.*, 1989) and 27 PKS genes have been reported in *Oryza sativa* (Hu *et al.*, 2017), which is the species with the largest number of PKS genes reported to date. In this study, we identified 11 PKS genes from upland cotton and compared these with PKS genes in *Populus tremula* (14), *Vitis vinifera* (13), *Malus domestica* (10) and *Arabidopsis thaliana* (4). The 52 PKS genes were divided into four subfamilies, I, II, III and IV, according to the phylogenetic tree nodes.

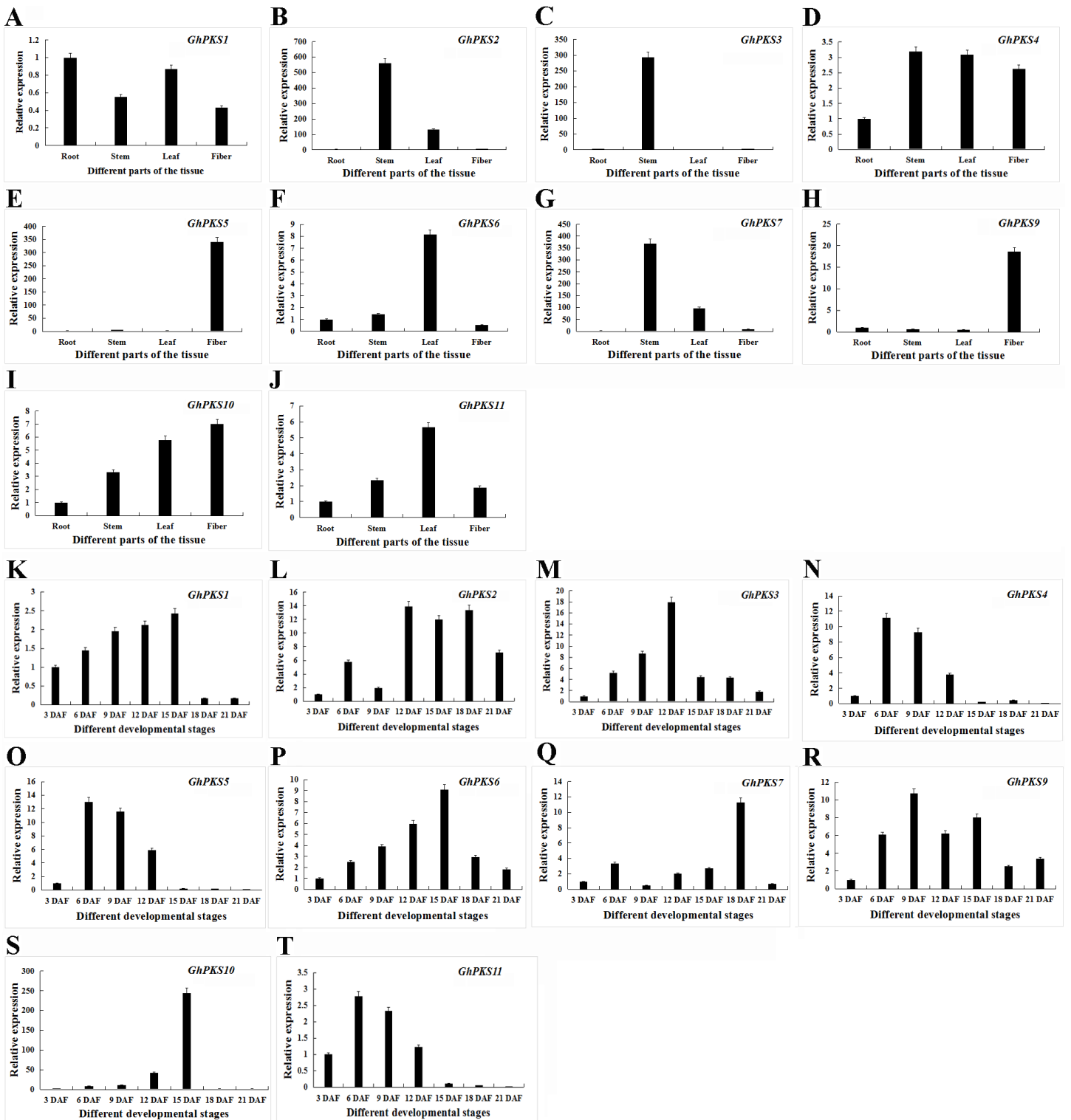


Figure 5 Expression patterns of PKS genes of upland cotton in different tissues and brown cotton fibers at different growth stages. (A–J) Expression patterns of PKS genes in upland cotton in different tissues. (K–T) Expression patterns of PKS genes in upland cotton at different growth stages of cotton fibers.

Full-size DOI: 10.7717/peerj.3974/fig-5

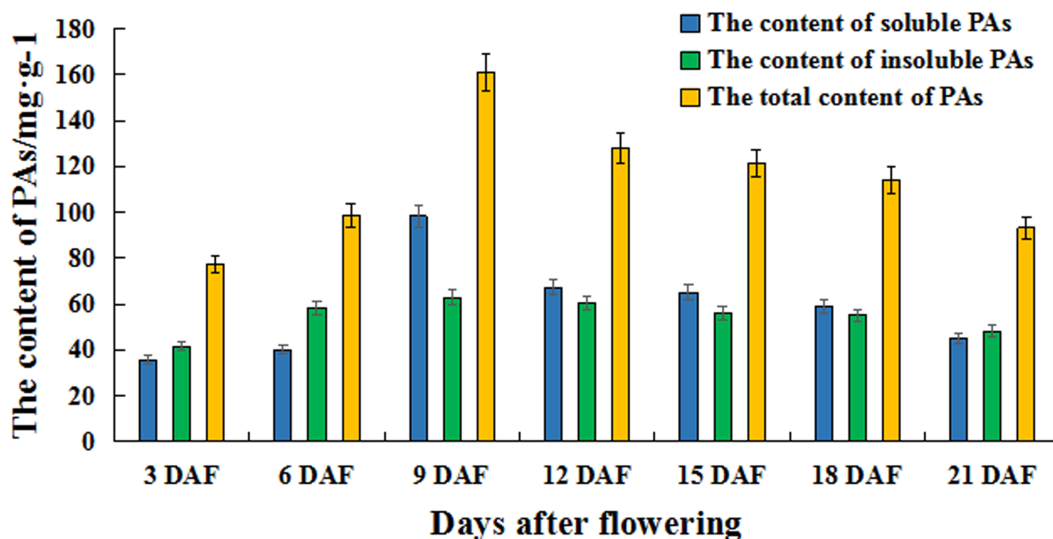


Figure 6 The content of PAs at different fiber development stages in brown cotton. The contents of soluble proanthocyanidins, insoluble proanthocyanidins and total proanthocyanidins are expressed as different colours.

Full-size DOI: [10.7717/peerj.3974/fig-6](https://doi.org/10.7717/peerj.3974/fig-6)

Previous researchers have suggested that most of the CHS genes consist of two exons and one intron (Durbin, Mccaig & Clegg, 2000) and the diversity of gene structures is important for the evolution of gene families (Swarbreck et al., 2008). According to our study, 72% of the 52 PKS genes consisted of two exons and one intron. However, some genes also had different compositions. For example, *VvPKS12* consists of five exons and four introns. Six PKS genes including *GhPKS9* had no introns and seven PKS genes had three exons and two introns.

We identified 20 conservative motifs using MEME software (Bailey et al., 2015). Among these 20 motifs, motifs 1, 3, 5, 7 encoded a Chal_sti_synt_N conservative domain and motifs 2, 4, 6, 13 encoded a Chal_sti_synt_C conservative domain. All 52 PKS genes with motifs encoding these two conserved domains demonstrate that the PKS III gene family has been highly conserved during evolution. These two conserved domains are associated respectively with acyl transfer activity and transferase activity (Götz et al., 2008), which indicates that these genes function in catalysing the formation of polyketone compounds. We found that the PKS genes in the same subfamily had similar motif compositions, e.g., *MdPKS7*, 9, *VvPKS6*, 8 and *PtPKS4*, 11. At the same time, there were some subfamily-specific motifs. The diversity of gene structure and conserved motif distribution may help to explain the functional dispersion of PKS gene family members.

The plant PKS III enzyme protein-specific catalytic triad composed of Cys-His-Asn could be traced back to the earliest ancestors of KAS III (Austin & Noel, 2002), which was considered to be important for the maintenance of PKS III gene family functions. Therefore, using BLAST, we queried the protein sequences of the 11 upland cotton PKS genes and the *AtCHS* and *MsCHS* protein sequences with the reported secondary structure of *OsCHS* as a template (Consortium et al., 2003). The results showed that the Cys-His-Asn

catalytic triad was highly conserved in all *GhPKS* sequences. However, there were more amino acid substitutions in active amino acids in the catalytic active site. For example, the first Phe site was highly conserved in all *GhPKS*s in the two Phe sites that are closely related to the binding of various CoA, while more amino acid substitutions appear in the second Phe site. At the same time, the three active amino acids (Thr, Gly, Ser), which are responsible for the regulation of the substrate and the length of the polyketide chain, have also been replaced by other amino acids. This suggested that the catalytic triad of the *GhPKS* protein was highly conserved in the process of gene evolution, whereas the active amino acids were not highly conserved. Therefore, we speculated that the diversity of amino acids at the active amino acid sites was the main cause of the functional dispersion of the PKS gene family.

Chromosomal localization analysis showed that the distribution of PKS genes in five species in our study was irregular. The PKS genes were more concentrated on chromosome 16 except for the PKS genes in *Vitis vinifera*. The rest of the PKS genes were scattered on multiple chromosomes, which is consistent with previous studies ([Han et al., 2016](#)). Subsequently, we found 10 pairs of duplicated genes in the five species: two pairs in upland cotton, three pairs in *Populus tremula*, three pairs in *Vitis vinifera* and two pairs in *Malus domestica*. No duplicated genes were found in *Arabidopsis thaliana*. Among the 10 pairs of duplicated genes, five of the duplicated genes in the *Vitis vinifera*, *Malus domestica* and *Populus tremula* were from tandem duplication and the other five pairs of duplicated genes were derived from segmental duplication. It has been reported that there are seven pairs of duplicated genes in the 27 PKS genes of *Oryza sativa*, but only one pair of duplicated genes was the result of segmental duplication ([Hu et al., 2017](#)). In *Zea mays*, there were two pairs of duplicated genes in the 14 PKS genes and these were from segmental duplication ([Han et al., 2016](#)). The PKS gene family in *Oryza sativa* has many duplicated genes and there are two types of gene duplication in *Oryza sativa*, tandem duplication and fragment duplication, which also explains why the number of PKS genes in *Oryza sativa* is greater than that of other species. We speculated that there were two kinds of duplication modes in the process of PKS gene duplication in terrestrial plants: tandem duplication and fragment duplication. However, it is unknown whether the duplications were mainly in the form of tandem duplication or segmental duplication, which have varied tendencies in different plants. It is generally believed that tandem duplication contributes to the generation of new genes and fragment duplication leads to slower evolution of the gene family ([Cao et al., 2016](#)). In upland cotton, the duplications of the PKS gene carried out in the form of segmental duplication indicated that the evolution of the PKS gene family was slow. The analysis of the Ka/Ks values of the 10 repeat genes showed that the Ka/Ks values of the 10 duplicated gene pairs were less than 0.309, which indicated that these replicates had undergone strong purification selection after duplication was complete. This was for a factor in maintaining the PKS gene family.

In *Arabidopsis thaliana*, *AtCHS* is regulated by a variety of MYB transcription factors such as *AtMYB11*, 58, 63, 111 and other transcription factors that can activate *AtCHS* transcription ([Chezem & Clay, 2016](#)). Furthermore, *Arabidopsis thaliana* treated with high-intensity light for 24 h resulted in a 50-fold increase in the activity of chalcone synthase

and a large amount of anthocyanin accumulation (*Courtney-gutterson et al., 1994*). In this study, the analysis of the cis-elements in the promoter regions of these 11 PKS genes of upland cotton showed that the regions contained many elements related to light regulation and MYB transcription factor binding. Therefore, we believe that upland cotton PKS genes may be regulated by light and MYB transcription factors. The expression patterns of PKS genes of upland cotton in different tissues and cotton fiber development were studied by qRT-PCR. *GhPKS1* showed a higher transcription level in the roots; *GhPKS2, 3, 7* showed a high expression level in the stem; *GhPKS5, 9, 10, 11* were mainly expressed in the fibers. The accumulation of PAs in brown cotton fibers occurred mainly before stage 15 DAF of cotton fiber development (*Li et al., 2012*). The expression of *GhPKS4, 5, 9, 11* was higher in the early stages of cotton fiber development and PAs in the brown cotton fibers gradually accumulated as their expression increased. The procyanidin content then decreased as the amount of expression also gradually decreased. Previous studies have shown that the PKS gene encodes a key enzyme in the flavonoid biosynthetic pathway as the first rate-limiting enzyme (*Martinez-Perez et al., 2014*). The precursor material of the pigment in the brown cotton fiber is PAs, which are flavonoids (*Liu et al., 2016*). The expression trend of *GhPKS4, 5, 9, 11* was consistent with the trend of the accumulation of PAs in brown cotton fibers; therefore, we speculate that *GhPKS4, 5, 9, 11* may be involved in brown cotton fiber pigment biosynthesis.

CONCLUSION

In this study, we identified 11 PKS genes from upland cotton and compared them with analogous genes from *Populus tremula*, *Arabidopsis thaliana*, *Vitis vinifera* and *Malus domestica*; there were 41 PKS genes with respect to phylogeny, gene structure, conserved motifs and selection pressure. According to the constructed phylogenetic tree, the 52 total PKS genes were divided into four subfamilies. Most of the PKS genes were composed of two exons and one intron. The PKS genes in the same subfamily had similar gene structure and conserved motifs. At the same time, our research on structure showed that gene duplication has been the main driving force of the expansion of the PKS III gene family, but there is a kind of species-specificity concerning fragment duplication vs. tandem duplication. The results of the Ka/Ks ratio analysis showed that purification selection has been important in maintaining the function of the PKS III gene family. According to the analysis of cis-acting elements of PKS promoters in upland cotton, the PKS gene may be regulated by MYB transcription factors and light. The analysis of qRT-PCR and the accumulation of PAs in brown cotton fibers suggest that *GhPKS4, 5, 9* and *11* may be involved in the accumulation of PAs in brown cotton fibers.

ACKNOWLEDGEMENTS

The authors are deeply grateful to Prof. Yongping Cai and Prof. Yi Lin, who provided the sample used in the study and very effective direction. The authors also thank Prof. JunShan Gao, Prof. Dahui Li, Dr. Yanan Wang, Dr. Xi Cheng and Dr. Muhammad Abdullah for providing valuable suggestions and comments.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by grants from the National Natural Science Foundation of China (No. 31640068), the Cotton Industry Innovation System of Anhui Province, the Natural Science Foundation of the Higher Education Institutions of Anhui Province (No. KJ2016A840 and KJ2014A078) and Postgraduate Student's Innovation Fund of Anhui Agricultural University (No. 2017yjs-32). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

National Natural Science Foundation of China: 31640068.

Cotton Industry Innovation System of Anhui Province.

Natural Science Foundation of the Higher Education Institutions of Anhui Province: KJ2016A840, KJ2014A078.

Postgraduate Student's Innovation Fund of Anhui Agricultural University: 2017yjs-32.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Xueqiang Su conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables.
- Xu Sun performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables.
- Xi Cheng performed the experiments, wrote the paper, prepared figures and/or tables.
- Yanan Wang performed the experiments, prepared figures and/or tables.
- Muhammad Abdullah wrote the paper.
- Manli Li performed the experiments, wrote the paper.
- Dahui Li and Junshan Gao performed the experiments, reviewed drafts of the paper.
- Yongping Cai and Yi Lin conceived and designed the experiments, reviewed drafts of the paper.

Data Availability

The following information was supplied regarding data availability:

The raw data has been provided as a [Supplemental File](#). The following information is available at <https://phytozome.jgi.doe.gov/pz/portal.html>: The *Gossypium hirsutum* genome (version1_0.fa), annotation information (version1_0.gff), coding sequences (version1_0.cds.fa) and protein sequences (version1_0.pep.fa); the *Populus tremula* protein sequences (version3_0.pep.fa); annotation information (version3_0.gene.gff3); the *Vitis vinifera* protein sequences (version12x.pep.fa), annotation information (version12X.gene.gff3); the *Malus domestica* protein sequences (version1_0.pep.fa),

annotation information (version1_0.gene.gff3); the *Arabidopsis thaliana* protein sequences (version167_TAIR10.pep.fa), annotation information (version167_TAIR10.gene.gff3).

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.3974#supplemental-information>.

REFERENCES

- Abe I, Morita H. 2010.** Structure and function of the chalcone synthase super family of plant type III polyketide synthases. *Natural Product Reports* **27(6)**:809–838 DOI [10.1039/b909988n](https://doi.org/10.1039/b909988n).
- Abe I, Takahashi Y, Morita H, Noguchi H. 2001.** Benzalacetone synthase. A novel polyketide synthase that plays a crucial role in the biosynthesis of phenylbutanones in *Rheum palmatum*. *European Journal of Biochemistry* **268(11)**:3354–3359 DOI [10.1046/j.1432-1327.2001.02255.x](https://doi.org/10.1046/j.1432-1327.2001.02255.x).
- Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, Castro E, Duvaud E, Flegel V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D, Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H. 2012.** ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research* **40**:597–603 DOI [10.1093/nar/gks400](https://doi.org/10.1093/nar/gks400).
- Austin MB, Noel JP. 2002.** The chalcone synthase superfamily of type III polyketide synthases. *Natural Product Reports* **20(1)**:79–110 DOI [10.1039/B100917F](https://doi.org/10.1039/B100917F).
- Bailey TL, Johnson J, Grant CE, Noble WS. 2015.** The MEME suite. *Nucleic Acids Research* **43**:W39–W49 DOI [10.1093/nar/gkv416](https://doi.org/10.1093/nar/gkv416).
- Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme YC, Eddy SR. 2004.** The pfam protein families database. *Nucleic Acids Research* **32**:263–266 DOI [10.1093/nar/gkh150](https://doi.org/10.1093/nar/gkh150).
- Bitocchi E, Rau D, Benazzo A, Bellucci E, Goretti D, Biagetti E, Panziera A, Laidò G, Rodriguez M, Gioia T, Attene G, McClean P, Lee RK, Jackson SA, Bertorelle G, Papa R. 2017.** High level of nonsynonymous changes in common bean suggests that selection under domestication increased functional diversity at target traits. *Frontiers in Plant Science* **7**:2005–2020 DOI [10.3389/fpls.2016.02005](https://doi.org/10.3389/fpls.2016.02005).
- Burbulis IE, Winkel-Shirley B. 1999.** Interactions among enzymes of the Arabidopsis flavonoid biosynthetic pathway. *Proceedings of the National Academy of Sciences of the United States of America* **96(22)**:12929–12934 DOI [10.1073/pnas.96.22.12929](https://doi.org/10.1073/pnas.96.22.12929).
- Cao YP, Han YH, Meng DD, Li DH, Jin Q, Lin Y, Cai YP. 2016.** Structural, evolutionary, and functional analysis of the class III peroxidase gene family in Chinese Pear (*Pyrus bretschneideri*). *Frontiers in Plant Science* **7**:1874–1886 DOI [10.3389/fpls.2016.01874](https://doi.org/10.3389/fpls.2016.01874).
- Chezem WR, Clay NK. 2016.** Regulation of plant secondary metabolism and associated specialized cell development by MYBs and bHLHs. *Phytochemistry* **131**:26–43 DOI [10.1016/j.phytochem.2016.08.006](https://doi.org/10.1016/j.phytochem.2016.08.006).

- Consortium FLC, Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E, Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y, Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H, Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu R, Ueda M, Matsubara K, RIKEN, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I, Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K, Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301(5631):376–379 DOI 10.1126/science.1081288.
- Courtney-gutterson N, Napoli C, Lemieux C, Morgan A, Firoozabady E, Robinson KE. 1994. Modification of flower color in florist's chrysanthemum: production of a white-flowering variety through molecular genetics. *Nature Biotechnology* 12(3):268–271 DOI 10.1038/nbt0394-268.
- Cui YP, Liu ZJ, Zhao YP, Wang YM, Huang Y, Li L, Wu H, Xu SX, Hua JP. 2017. Overexpression of heteromeric GhACCase subunits enhanced oil accumulation in upland cotton. *Plant Molecular Biology Reporter* 35(2):287–297 DOI 10.1007/s11105-016-1022-y.
- Curran JM, Tvedebrink T. 2013. DNAtools: tools for empirical testing of DNA match probabilities. R package. Available at <https://cran.r-project.org/package=DNAtools>.
- Dare AP, Tomes S, Jones M, McGhie TK, Stevenson DE, Johnson RA, Greenwood DR, Hellens RP. 2013. Phenotypic changes associated with RNA interference silencing of chalcone synthase in apple (*Malus × domestica*). *Plant Journal* 74(3):398–410 DOI 10.1111/tpj.12140.
- Durbin ML, Mccaig B, Clegg MT. 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. *Plant Molecular Biology* 42(1):79–92 DOI 10.1023/A:1006375904820.
- Eom SH, Hyun TK. 2016. Genome-wide identification and transcriptional expression analysis of chalcone synthase in flax (*Linum usitatissimum*, L.). *Gene Reports* 5:51–56 DOI 10.1016/j.genrep.2016.08.011.
- Fan X, Fan B, Wang Y, Yang W. 2016. Anthocyanin accumulation enhanced in Lc-transgenic cotton under light and increased resistance to bollworm. *Plant Biotechnology Reports* 10(1):1–11 DOI 10.1007/s11816-015-0382-3.
- Feng H, Tian X, Liu Y, Zhang X, Jones BJ, Sun Y, Sun J. 2013. Analysis of flavonoids and the flavonoid structural genes in brown fiber of upland cotton. *PLOS ONE* 8(3):e58820 DOI 10.1371/journal.pone.0058820.
- Funa N, Ohnishi Y, Fujii I, Shibuya M, Ebizuka Y, Horinouchi S. 1999. A new pathway for polyketide synthesis in microorganisms. *Nature* 400:897–899 DOI 10.1038/23748.
- Gao JS, Nan W, Shen ZL, Lv K, Qian SH, Guo N, Sun X, Cai YP, Lin Y. 2016. Molecular cloning, expression analysis and subcellular localization of a Transparent Testa

- 12, ortholog in brown cotton (*Gossypium hirsutum*, L.). *Gene* **576**:763–769 DOI [10.1016/j.gene.2015.11.002](https://doi.org/10.1016/j.gene.2015.11.002).
- Götz S, Garcíagómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008.** High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36**(10):3420–3435 DOI [10.1093/nar/gkn176](https://doi.org/10.1093/nar/gkn176).
- Gras CC, Nemetz N, Carle R, Schweiggert RM. 2017.** Anthocyanins from purple sweet potato (*Ipomoea batatas* (L.) Lam.) and their color modulation by the addition of phenolic acids and food-grade phenolic plant extracts. *Food Chemistry* **235**(11):265–274 DOI [10.1016/j.foodchem.2017.04.169](https://doi.org/10.1016/j.foodchem.2017.04.169).
- Guo AY, Zhu QH, Chen X, Luo JC. 2007.** GSDS: a gene structure display server. *Yi Chuan* **29**(8):1023–1026 DOI [10.1360/yc-007-1023](https://doi.org/10.1360/yc-007-1023).
- Han Y, Ding T, Su B, Jiang H. 2016.** Genome-wide identification, characterization and expression analysis of the chalcone synthase family in maize. *International Journal of Molecular Sciences* **17**(2):161–176 DOI [10.3390/ijms17020161](https://doi.org/10.3390/ijms17020161).
- Han Y, Zhao W, Wang Z, Zhu J, Liu Q. 2014.** Molecular evolution and sequence divergence of plant chalcone synthase and chalcone synthase-Like genes. *Genetica* **142**(3):215–225 DOI [10.1007/s10709-014-9768-3](https://doi.org/10.1007/s10709-014-9768-3).
- Helariutta Y, Elomaa P, Kotilainen M, Griesbach RJ, Schröder J, Teeri TH. 1995.** Chalcone synthase-like genes active during corolla development are differentially expressed and encode enzymes with different catalytic properties in *Gerbera hybrida* (Asteraceae). *Plant Molecular Biology* **28**:47–60 DOI [10.1007/BF00042037](https://doi.org/10.1007/BF00042037).
- Hinchliffe DJ, Condon BD, Thyssen G, Naoumkina M, Madison CA, Reynolds M, Delhom CD, Fang DD, Li P, McCarty J. 2016.** The GhTT2_A07 gene is linked to the brown colour and natural flame retardancy phenotypes of Lc1 cotton (*Gossypium hirsutum* L.) fibers. *Journal of Experimental Botany* **67**(18):5461–5471 DOI [10.1093/jxb/erw312](https://doi.org/10.1093/jxb/erw312).
- Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007.** WoLF PSORT: protein localization predictor. *Nucleic Acids Research* **35**:W585–W587 DOI [10.1093/nar/gkm259](https://doi.org/10.1093/nar/gkm259).
- Hu L, He H, Zhu C, Peng X, Fu J, He X, Chen X, Ouyang L, Bian J, Liu S. 2017.** Genome-wide identification and phylogenetic analysis of the chalcone synthase gene family in rice. *Journal of Plant Research* **130**(1):1–11 DOI [10.1007/s10265-016-0885-1](https://doi.org/10.1007/s10265-016-0885-1).
- Hu W, Yang H, Yan Y, Wei Y, Tie W, Ding Z, Zuo J, Peng M, Li K. 2016.** Genome-wide characterization and analysis of bZIP transcription factor gene family related to abiotic stress in cassava. *Scientific Reports* **6**:22783 DOI [10.1038/srep22783](https://doi.org/10.1038/srep22783).
- Hua SJ, Wang XD, Yuan SN, Shao MY, Zhao XQ, Zhu SJ, Jiang LX. 2007.** Characterization of pigmentation and cellulose synthesis in colored cotton fibers. *Crop Science* **47**(4):1540–1546 DOI [10.2135/cropsci2006.12.0835](https://doi.org/10.2135/cropsci2006.12.0835).
- Ikegami A, Akagi T, Potter D, Yamada M, Sato A, Yonemori K, Kitajima A, Inoue K. 2009.** Molecular identification of 1-Cys peroxiredoxin and anthocyanidin/flavonol 3-O-galactosyltransferase from proanthocyanidin-rich young fruits of persimmon (*Diospyros kaki* Thunb). *Planta* **230**(4):841–855 DOI [10.1007/s00425-009-0989-0](https://doi.org/10.1007/s00425-009-0989-0).

- Jepson C, Karppinen K, Daku RM, Sterenberg BT, Suh DY. 2014. Hypericum perforatum hydroxyalkylpyrone synthase involved in sporopollenin biosynthesis—phylogeny, site-directed mutagenesis, and expression in nonanther tissues. *Febs Journal* **281**(17):3855–3868 DOI [10.1111/febs.12920](https://doi.org/10.1111/febs.12920).
- Jez JM, Bowman ME, Noel JP. 2002. Expanding the biosynthetic repertoire of plant type III polyketide synthases by altering starter molecule specificity. *Proceedings of the National Academy of Sciences of the United States of America* **99**(8):5319–5324 DOI [10.1073/pnas.082590499](https://doi.org/10.1073/pnas.082590499).
- Junghanns KT, Kneusel RE, Baumert A, Maier W, Gröger D, Matern U. 1995. Molecular cloning and heterologous expression of acridone synthase from elicited *Ruta graveolens* L. cell suspension cultures. *Plant Molecular Biology* **27**(4):681–692 DOI [10.1007/BF00020222](https://doi.org/10.1007/BF00020222).
- Koes RE, Spelt CE, Van den Elzen PJ, Mol JN. 1989. Cloning and molecular characterization of the chalcone synthase multigene family of *Petunia hybrida*. *Gene* **81**(2):245–257 DOI [10.1016/0378-1119\(89\)90185-6](https://doi.org/10.1016/0378-1119(89)90185-6).
- Kumar S, Stecher G, Tamura K. 2016. MEGA 7.0: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology & Evolution* **33**(7):1870–1874 DOI [10.1093/molbev/msw054](https://doi.org/10.1093/molbev/msw054).
- Letunic I, Doerks T, Bork P. 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research* **40**:302–305 DOI [10.1093/nar/gkr931](https://doi.org/10.1093/nar/gkr931).
- Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G, Wang K, Xu X, Yu JZ, Zhu Y, Yu S. 2015. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nature Biotechnology* **33**(5):524–530 DOI [10.1038/nbt.3208](https://doi.org/10.1038/nbt.3208).
- Li H, Liang J, Chen H, Ding G, Ma B, He N. 2016. Evolutionary and functional analysis of mulberry type III polyketide synthases. *BMC Genomics* **17**(1):540–558 DOI [10.1186/s12864-016-2843-7](https://doi.org/10.1186/s12864-016-2843-7).
- Li TC, Fan HH, Li ZP, Wei J, Lin Y, Cai YP. 2012. The accumulation of pigment in fiber related to proanthocyanidins synthesis for brown cotton. *Acta Physiologiae Plantarum* **34**(2):813–818 DOI [10.1007/s11738-011-0858-x](https://doi.org/10.1007/s11738-011-0858-x).
- Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**(11):1451–1452 DOI [10.1093/bioinformatics/btp187](https://doi.org/10.1093/bioinformatics/btp187).
- Liu C, Wang X, Shulaev V, Dixon RA. 2016. A role for leucoanthocyanidin reductase in the extension of proanthocyanidins. *Nature Plants* **2**:16182 DOI [10.1038/nplants.2016.182](https://doi.org/10.1038/nplants.2016.182).
- Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{(-Delta Delta C(T))} Method. *Methods* **25**(4):402–408 DOI [10.1006/meth.2001.1262](https://doi.org/10.1006/meth.2001.1262).

- Long M, Thornton K. 2001.** Gene duplication and evolution. *Science* 293(5535):1551–1155 DOI [10.1126/science.293.5535.1551a](https://doi.org/10.1126/science.293.5535.1551a).
- Luna-Vital D, Li Q, West L, West M, Gonzalez de Mejia E. 2017.** Anthocyanin condensed forms do not affect color or chemical stability of purple corn pericarp extracts stored under different pHs. *Food Chemistry* 232:639–647 DOI [10.1016/j.foodchem.2017.03.169](https://doi.org/10.1016/j.foodchem.2017.03.169).
- Martinez-Perez C, Ward C, Cook G, Mullen P, McPhail D, Harrison DJ, Langdon SP. 2014.** Novel flavonoids as anti-cancer agents: mechanisms of action and promise for their potential application in breast cancer. *Biochemical Society Transactions* 42(4):1017–1023 DOI [10.1042/BST20140073](https://doi.org/10.1042/BST20140073).
- Niu E, Cai C, Zheng Y, Shang X, Fang L, Guo W. 2016.** Genome-wide analysis of CrRLK1L, gene family in *Gossypium*, and identification of candidate CrRLK1L, genes related to fiber development. *Molecular Genetics & Genomics* 291(3):1137–1154 DOI [10.1007/s00438-016-1169-0](https://doi.org/10.1007/s00438-016-1169-0).
- Oikawa T, Maeda H, Oguchi T, Yamaguchi T, Tanabe N, Ebana K, Yano M, Ebitani T, Izawa T. 2015.** The birth of a black rice gene and its local spread by introgression. *Plant Cell* 27(9):2401–2414 DOI [10.1105/tpc.15.00310](https://doi.org/10.1105/tpc.15.00310).
- Owens DK, Alerding AB, Crosby KC, Bandara AB, Westwood JH, Winkel BS. 2008.** Functional analysis of a predicted flavonol synthase gene family in Arabidopsis. *Plant Physiology* 147(3):1046–1061 DOI [10.1104/pp.108.117457](https://doi.org/10.1104/pp.108.117457).
- Qian SH, Hong L, Xu M, Cai YP, Lin Y, Gao JS. 2015.** Cellulose synthesis in coloured cotton. *ScienceAsia* 41(3):180–186 DOI [10.2306/scienceasia1513-1874.2015.41.180](https://doi.org/10.2306/scienceasia1513-1874.2015.41.180).
- Reimold U, Kröger M, Kreuzaler F, Hahlbrock K. 1983.** Coding and 3' non-coding nucleotide sequence of chalcone synthase mRNA and assignment of amino acid sequence of the enzyme. *Embo Journal* 2(10):1801–1805.
- Rombauts S, Déhais P, Van Montagu M, Rouzé P. 1999.** PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Research* 27(1):295–296 DOI [10.1093/nar/27.1.295](https://doi.org/10.1093/nar/27.1.295).
- Schanz S, Schröder G, Schröder J. 1992.** Stilbene synthase from Scots pine (*Pinus sylvestris*). *Febs Letters* 313(1):71–74 DOI [10.1016/0014-5793\(92\)81187-Q](https://doi.org/10.1016/0014-5793(92)81187-Q).
- Schröder J. 2000.** The family of chalcone synthase-related proteins: functional diversity and evolution. *Recent Advances in Phytochemistry* 34:55–89 DOI [10.1016/S0079-9920\(00\)80004-0](https://doi.org/10.1016/S0079-9920(00)80004-0).
- Shimizu Y, Ogata H, Goto S. 2017.** Type III polyketide synthases: functional classification and phylogenomics. *ChemBioChem* 18:50–65 DOI [10.1002/cbic.201600522](https://doi.org/10.1002/cbic.201600522).
- Stipanovic RD, Puckhaber LS, Bell AA, Percival AE, Jacobs J. 2005.** Occurrence of (+)- and (–)- gossypol in wild species of cotton and in *Gossypium hirsutum* Var. marie-galante (Watt) Hutchinson. *Journal of Agricultural and Food Chemistry* 53(16):6266–6271 DOI [10.1021/jf050702d](https://doi.org/10.1021/jf050702d).
- Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E. 2008.** The Arabidopsis Information Resource (TAIR): gene

structure and function annotation. *Nucleic Acids Research* **36**:D1009–D1014
[DOI 10.1093/nar/gkm965](https://doi.org/10.1093/nar/gkm965).

Tuteja JH, Clough SJ, Chan WC, Vodkin LO. 2004. Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in glycine max. *Plant Cell* **16**(4):819–835 [DOI 10.1105/tpc.021352](https://doi.org/10.1105/tpc.021352).

Van der Heijden RT, Snel B, Van Noort V, Huynen MA. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* **8**:83 [DOI 10.1186/1471-2105-8-83](https://doi.org/10.1186/1471-2105-8-83).

Xie L, Liu P, Zhu Z, Zhang S, Zhang S, Li F, Zhang H, Li G, Wei Y, Sun R. 2016. Phylogeny and expression analyses reveal important roles for plant PKS III family during the conquest of land by plants and angiosperm diversification. *Front Plant Sci* **7**:1312–1327 [DOI 10.3389/fpls.2016.01312](https://doi.org/10.3389/fpls.2016.01312).

Yuan S, Hua SJ, Malik W, Bibi N, Wang XD. 2012. Physiological and biochemical dissection of fiber development in colored cotton. *Euphytica* **187**(2):215–226 [DOI 10.1007/s10681-012-0653-9](https://doi.org/10.1007/s10681-012-0653-9).

Zhang XW, Xiong HR, Liu AL, Zhou XY, Peng Y, Li ZX, Luo GY, Tian XR, Chen XB. 2014. Microarray data uncover the genome-wide gene expression patterns in response to heat stress in rice post-meiosis panicle. *Journal of Plant Biology* **57**(6):327–336 [DOI 10.1007/s12374-014-0177-z](https://doi.org/10.1007/s12374-014-0177-z).