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*, Yi Lin*

School of Life Science, Anhui Agricultural University, No. 130, Changjiang West Road, Hefei 230036, China;

# This authors have contributed equally to this work.

***Corresponding author**: linyi1957@126.com (Yi Lin);

Co-Corresponding author: ypcaiah@163.com (Yongping Cai);

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

26   pigment synthesis. These results are important for the further study of brown cotton PKS genes.

27   It not only reveals the relationship between PKS gene family and pigment in brown cotton, but

28   also creates conditions for improving the quality of brown cotton fiber.

INTRODUCTION

31   Plant polyketone compounds are secondary metabolites having a cyclic structure with an

32   oxygen atom bound to the carbon ring. This group includes phenols, stilbene and flavonoid

33   compounds (Abe & Morita, 2010). Owing to the complexity and variety of the pathways and

34   mechanisms of biosynthesis, the number of polyketone compounds is very large and their

35   molecular structures are complex. This complexity results in the compounds having prominent

36   and varied biological activities (Austin & Noel, 2002). The biosynthesis of this group has a

37   common mechanism that includes the enzyme polyketide synthase (PKS). According to the

38   structure of the protein, PKS can be divided into PKS I, II and III (Funa et al., 1999). PKS I and

39   PKS II only exist in microorganisms. Each form has many functional modules and

40   monofunctional subunits (Xie et al., 2016). The PKS III gene family exists mainly in the plant

41   kingdom, but some occur in a few species of bacteria. PKS III gene family members can catalyse

42   plant secondary metabolites having various structures, biological activities and chalcone

43   synthase (CHS) backbones. Examples of such metabolites include chalcone, stilbene,

44   benzophenone, acridone, phloroglucinol, resorcinol and pyrone (Austin & Noel, 2002). These

45   secondary metabolites play important roles in the colouring of plant organs, safeguarding from

46   pesticides and prevention of UV irradiation damage (Li et al., 2016).

47   The type III PKS gene family is divided into chalcone synthase (CHS) and chalcone

48   synthase-like protein (CHSL) subfamilies. Chalcone synthase is the core enzyme of the PKS III

49   gene family and is the first key enzyme for the plant flavonoid synthesis pathway and the rate-

50   limiting enzyme (Martinez-Perez et al., 2014). The PKS III gene family also includes a series of

51   gene duplications and functional differentiation derived from the class of CHS-like proteins

52 (CHSL) (Eom & Hyun, 2016). CHSL protein is far from the biosynthesis of PAs, The main role

53 is to help plants adapt to changes in the environment, especially in response to fungal invasion

54 (Han et al., 2014). The CHSL of the PKS III gene family include 2-pyrone synthase cloned from

55 *Gerbera hybrida* (Helariutta et al., 1995), acridone synthase cloned from (Junghanns et al., 1995),

56 benzalacetone synthase cloned from *Rheum palmatum* (Abe et al., 2001) and stilbene synthase

57 cloned from *Pinus sylvestris* (Schanz et al., 1992). Because of the evolution from a common

58 ancestor, PKS III gene family members have a high degree of homology between the structure

59 and catalytic mechanisms and are very similar. For example, their proteins are essentially

60 homodimers consisting of 40-45 kDa subunits and their active sites have a catalytic triad that is

61 composed of Cys-His-Asn. The functional differences of CHSL and CHS lie in the preference

62 towards different substrates when catalytic reactions occur, changes in the malonyl-CoA number

63 of condensation and different cyclic ways of production (Schröder, 2000).

64 The first PKS gene was reported in 1983 in a study of *PcCHS* in *Petroselinum crispum* and

65 was shown to be involved in the biosynthesis of flavonoids (Reimold et al., 1983). The study of

66 the PKS III gene family continues today. Chalcone synthase (CHS) is by far the most thoroughly

67 studied type III polyketide synthase. CHS catalyses the first step in the synthesis of flavonoids

68 and CHS is responsible for catalysing the reaction of 1 molecule of 4-benzoyl-CoA with 3

69 molecules of malonyl-CoA to form chalcone (Burbulis & Winkel-Shirley, 1999), the precursor

70 of many flavonoid compounds. The enzymes chalcone isomerase (CHI), flavanone 3-

71 hydroxylase (F3H), flavonoid 3'-hydroxylase (F3'H), dihydroflavone-4-reductase (DFR) and

72 other enzymes have a common catalytic role in the formation of a variety of flavonoids (Feng et

73 al., 2013). Currently, the cloning and functional analysis of CHS have been reported for many

74 species, e.g., *Oryza sativa* (Hu et al., 2017), *Hypericum monogynum* (Jepson et al., 2014),

75 *Gerbera hybrida* (Helariutta et al., 1995), *Petunia hybrida* (Koes et al., 1989), *Malus domestica*

76 (Dare et al., 2013) and *Glycine max* (Tuteja et al., 2004).

77 Study of the PKS III gene family in the important cash crop cotton has yet to be conducted.

78 Cotton is an important fiber crop, but it is also used for oil (Cui et al., 2017), drugs (Stipanovic et

79   al., 2005) and other purposes. Naturally colored cotton can be divided into two categories: brown

80   cotton and green cotton. It can synthesize and accumulate pigment to make mature fibers with

81   varied colours during fiber development (Yuan et al., 2012). At present, the application and

82   cultivation of a wide range of naturally coloured cotton varieties produce mainly brown cotton

83   and green cotton. Brown cotton fiber pigments are more stable than those of green cotton; this,

84   combined with its high yield, has led to brown cotton becoming the dominant colour of natural

85   cotton varieties (Qian et al., 2015). Brown cotton is widely favoured for its commercial value

86   and application characteristics, including the lack of need for dyeing, its anti-static electricity

87   properties, ultraviolet resistance and good flame retardance (Hinchliffe et al., 2016). Brown

88   cotton flavonoids are also closely related to resistance to pests and diseases; increasing the

89   flavonoid content can increase plant resistance to insects and thus brown cotton has been widely

90   favoured with increasing commercial value and application prospects (Fan et al., 2016).

91   However, brown cotton fibers do have some problems; these include poor pigment stability,

92   uneven pigment distribution and poor fiber quality (Hua et al., 2007). These problems can

93   restrict the market value of brown cotton. To solve these problems, we focused on the synthesis

94   of brown cotton pigment to improve the quality of brown cotton at the molecular level. At

95   present, many studies have shown that brown cotton pigment is mainly composed of PAs (Gao et

96   al., 2016). In addition, high quality varieties rich in procyanidins are reported in many species,

97   these breeds not only have high commercial value but also help to improve our understanding of

98   flavonoid metabolic pathways precious resources. For example: black rice since ancient times is

99   a very precious ingredients, the color of this grain deepened is due to the accumulation of PAs in

100 rice (Oikawa et al., 2015); *Solanum tuberdsm* is with high intensity of coloring and high

101 nutritional value of food; which are due to *Solanum tuberdsm* rich in PAs (Gras et al., 2017);

102 what we used to know corn is orange particles, but purple corn is more resistant to storage than

103 orange corn and has higher nutritional value (Luna-Vital et al., 2017). These varieties are all rich

104 in PAs, PAs metabolism is an important branch of flavonoid metabolism and thus the PKS III

105 gene family plays an important role in the synthesis of PAs. Thus it can be seen the study of the

106 PKS gene family is very important not only in brown cotton, but also has a very important

107 significance in many species. The study of PKS gene family can not only help us to better

108 understand the metabolic pathway of flavonoids but also can produce huge commercial value.

109   Although the whole genome of upland cotton (*Gossypium hirsutum*) has been sequenced (Li

110 et al., 2015), the whole genome identification and analysis of the type III polyketide synthase

111 family in terrestrial cotton have not yet been reported. The relationship between PKS genes and

112 fiber quality in brown cotton remains unknown. In the present study, we screened the PKS

113 family in upland cotton and analysed the characteristics of its evolution, gene structure,

114 conserved motifs and duplication events. The study species for comparison of the PKS III gene

115 family included *Populus tremula*, *Arabidopsis thaliana*, *Vitis vinifera* and *Malus domestica*.

116 *Arabidopsis thaliana* is a widely used research plant and its synthesis of flavonoids is more

117 thoroughly understood, while the other three species are rich in flavonoids. Therefore, the choice

118 of these four species for comparison with the upland cotton can help us better understand

119 terrestrial cotton flavonoid metabolism. According to the analysis of promoter cis-acting

120 elements and the expression patterns of PKS genes in upland cotton, the candidate PKS genes

121 relating to the brown cotton fiber pigment were identified, which provides an important

122 theoretical foundation and genetic resource for improving the uneven distribution, poor stability

123 and fiber quality of natural brown cotton. At the same time, we further analysed the expression

124 patterns of PKS family members and discussed their relationship with the changes in PAs at

125 different developmental stages to determine the PKS candidate genes associated with brown

126 cotton fiber pigment. These results will provide an important theoretical basis for improving the

127 uneven distribution and poor stability of natural brown cotton pigment.

128

MATERIALS AND METHODS

129

Plant materials

130

131   Brown cotton plants used line Zongcaixuan No. 1 (brown fiber line) in the experiment were

132   grown in an agricultural park (Hefei, Anhui, China). This brown cotton line belongs to tetraploid

133   upland cotton. In July 2016, 50 brown cotton plants with good growth characteristics were

134   selected at the blooming stage. We began collecting cotton bolls after 3, 6, 9, 12, 15, 18 and 21

135   days after flowering (DAF). The experimental materials were frozen in liquid nitrogen and

136   quickly transferred to the laboratory refrigerator.

137

138   **Identification and collection of PKS proteins**

139   In our study, the genomic data of *Gossypium hirsutum, Populus tremula*, *Vitis vinifera*,

140   *Malus domestica* and *Arabidopsis thaliana* were downloaded from the Phytozome database (Hu

141   at al., 2016) (https://phytozome.jgi.doe.gov/pz/portal.html). DNATOOLS software was used to

142   establish a local database of the amino acid sequences (Curran & Tvedebrink, 2013), including

143   the whole genomes of *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica*

144   and *Arabidopsis thaliana*. The sequences in TBlastN (E-value=0.001) were queried according to

145   the two conservative domains Chal_sti_synt_N and Chal_sti_synt_C (Han et al., 2016) and

146   compared with the established local database sequences of *Gossypium hirsutum*, *Populus*

147   *tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*. Preliminary PKS candidate

148   gene sequences were screened out. The PKS candidate gene sequences obtained by BLAST were

149   tested for whether they contained the two conserved Chal_sti_synt_N and Chal_sti_synt_C

150   domains using Pfam (Bateman et al., 2004) (http://pfam.xfam.org/) and SMART (Letunic et al.,

151   2012) (http://smart.embl-heidelberg.de/) online software. Multiple sequence alignment and

152   repeat sequence removal were analysed using the ClustalW tool of the MEGA 7.0 software

153   (Kumar et al., 2016). The molecular weight of the PKS protein was predicted using the ExPASy

154   Proteomics Server software (Artimo et al., 2012) (http://web.expasy.org/protparam/).

155   WoLFPSORT (Horton et al., 2006) (http://www.g enscript.com/wolf-psort.html) was used to

156   predict the PKS protein subcellular localization.

157

Phylogenetic analysis

159   Protein sequence alignment was performed using the Clustal X program (Des Higgins,
160   DUB, Ireland). The phylogenetic tree was built using the Neighbour-Joining (N-J) method with
161   1000 bootstraps and MEGA 7.0 (Kumar et al., 2016). The *GhPKS* genes were classified
162   according to the phylogenetic relationships. Two different species of genes are located in the
163   phylogenetic tree at the same node and the sequence similarity is more than 80%, we consider
164   two of these are orthologous genes (van der Heijden RT et al., 2007).

165

Gene structural and conserved motif analysis

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

175

Chromosomal location and gene duplication

177   Chromosome starting position and other relevant information concerning the PKS genes
178   were obtained from the public genome database of *Gossypium hirsutum*, *Populus tremula*, *Vitis*
179   *vinifera*, *Malus domestica* and *Arabidopsis thaliana*. The chromosome physical locations of the
180   PKS genes of all five species were obtained using MapInspect (Niu et al., 2016)
181   (http://mapinspect.software.informer.com) software. The gene is located on the same

182 chromosome, separated from the 200 kb and more than 80% similarity gene called tandem

183 duplication; whereas genes that duplicated genes on different chromosomes and more than 80%

184 similarity gene called fragment duplication (Long & Thornton, 2001). Non-synonymous (Ka)

185 and synonymous (Ks) sites were calculated using the DnasP v5.0 software (Librado & Rozas,

186 2009). Sliding window analysis was also performed using the DnasP v5.0 software; the

187 parameters were as follows: window size, 150 bp; step size, 9 bp.

188

189 **Upland cotton PKS gene promoter cis-acting element analysis**

190   The promoter sequence of each PKS gene was obtained from the genome database for

191 *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis thaliana*;

192 this includes the DNA sequence of the initiation codon (ATG) located 1500 bp upstream of each

193 PKS gene. We used the online software Plantcare (Rombauts et al., 1999)

194 (http://bioinformatics.psb.ugent.be/webtools/plantcarere/html/) to analyse the promoter region

195 cis-acting elements.

196

197 **RNA extraction and qRT-PCR**

198   In this study, 11 PKS genes of upland cotton were quantitatively analysed by real-time

199 fluorescence. Cotton bolls at 3 DAF, 6 DAF, 9 DAF, 12 DAF, 15 DAF, 18 DAF, 21 DAF were

200 collected and RNA was extracted using the Tiangen (Beijing, China) plant RNA extraction kit.

201 Reverse transcription was performed using a PrimeScript™ RT reagent kit with gDNA Eraser

202 (Takara, Japan) and each reaction used 1 μg of RNA. The specific primers for the PKS gene of

203 upland cotton (Table S1) were designed using Beacon Designer 7 software and the internal

204 reference gene used UBQ7 (Table S1). The qRT-PCR system consisted of 20 μL: 10 μL of

205 SYBR® Premix Ex Taq™ II (2×) (Takara, Japan), 2 μL of cDNA and 0.8 μL of GhPKS-F and

206 GhPKS-R. The reaction procedure was 40 cycles of 50°C for 2 min, 95°C for 30 s, 95°C for 5 s

207   and 60°C for 20 s, followed by 72°C for 10 min; the experiment was repeated three times.

208   Finally, we used $2^{-\Delta\Delta Ct}$ for the calculation of relative expression (Livak & Schmittgen, 2001).

209

Determination of proanthocyanidin content in brown cotton fibers

211   The fibers of brown cotton bolls at 3 DAF, 6 DAF, 9 DAF, 12 DAF, 15 DAF, 18 DAF, 21

212   DAF were stripped, extracted with 80% methanol and subjected to ultrasonic extraction for 30

213   min. After centrifuging for 15 min, the resulting supernatant was analysed for soluble PAs. A

214   methanol solution containing 1% HCl was added to the precipitate and the solution was placed in

215   a 6°C water bath for 1h; after centrifugation for 15 min, the supernatant contained the insoluble

216   PAs. The content of PAs was determined by the method of n-butanol-hydrochloric acid: 400 μL

217   of procyanidin extract was added to 1.5 mL of n-butanol (containing 5% hydrochloric acid) in a

218   boiling water bath for 20 min, after which the absorbance read at 550 nm (Ikegami et al., 2009).

219

RESULTS

Identification and evolutionary analysis using five genomes

222   Two kinds of plant PKS III genes conserved domains, Chal_sti_synt_N and

223   Chal_sti_synt_C, were obtained from the Pfam protein database using a hidden Markov model.

224   The two conserved domains have the respective molecular functions of transacylation and

225   transferase. Sequences from TBlastN (E-value = 0.001) were compared to the genome database

226   of *Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and *Arabidopsis*

227   *thaliana* using Chal_sti_synt_N and Chal_sti_synt_C. A total of 52 PKS genes were identified

228   (Table S2), including 11 in *Gossypium hirsutum* (*GhPKS1-GhPKS11*), 14 in *Populus tremula*

229   (*PtPKS1-PtPKS14*), 13 in *Vitis vinifera* (*VvPKS1-VvPKS13*), 10 in *Malus domestica* (*MdPKS1-*

230   *MdPKS10*) and 4 in *Arabidopsis thaliana* (*AtPKS1-AtPKS4*). In addition to the small number of

231   PKS genes in *Arabidopsis thaliana*, the number of PKS genes in other species was not very

232   different. To clarify the evolutionary relationships between the 11 PKS genes and PKS genes in

233   four other cultivars, we constructed a phylogenetic tree for a total of 52 PKS genes (Figure 1).

234   According to the phylogenetic tree nodes, the 52 PKS genes can be divided into four subfamilies:

235   I, II, III and IV. The number of subfamily I members was 17, followed by subfamily IV (12),

236   subfamily III (11) and the lowest number of members was in the subfamily II (10). *PtPKS5* and

237   *PtPKS7* separated into a class. Among the four subfamilies, the subfamilies I, IV included all

238   five species and each species provided at least one PKS gene. Subfamily III contained three

239   species (*Gossypium hirsutum*, *Populus tremula* and *Malus domestica*), while subfamily II only

240   consisted of *Vitis vinifera*. It is noteworthy that subfamily I includes an *Arabidopsis thaliana*

241   PKS gene (*AtPKS4*) (Owens et al., 2008). This gene is a CHS gene that has been reported in

242   *Arabidopsis thaliana*. *Arabidopsis thaliana* plants were treated with high-intensity light for 24

243   hours, resulting in a 50-fold increase in chalcone synthase activity and the accumulation of large

244   amounts of anthocyanins (Courtney-gutterson et al., 1994). The four *GhPKSs* (*GhPKS5*,

245   *GhPKS9*, *GhPKS10*, *GhPKS11*) were present in subfamily I, which may indicate that they are

246   closely related to the accumulation of brown cotton fiber pigments. In addition, according to the

247   results of the phylogenetic tree, there were no orthologous genes between the five species.

248

249   **Structural and conserved motif analysis of PKS proteins**

250   To understand the structural diversity of the PKS gene in a more comprehensive way, exon-

251   intron pattern maps were constructed for the 52 PKS genes. As seen from the figure (Figure 2A),

252   there are 38 members of the 52 PKS genes consisting of two exons and one intron and as in

253   previous reports, most of the plant PKS genes contain two exons and one intron (Durbin et al.,

254   2000). In the remaining 14 members, *VvPKS3* contains an exon and an intron. And there are six

255   members (*GhPKS9*, *MdPKS3*, *PtPKS7*, *VvPKS6*, *VvPKS8*, *VvPKS9*) with no introns. The

256   remaining seven members (*AtPKS3*, *MdPKS8*, *VvPKS2*, *VvPKS4*, *VvPKS5*, *VvPKS11* and

257   *VvPKS13*) are composed of three exons and two introns. *VvPKS12* has the largest number with

258   five exons and four introns. There were no UTR regions found in the 23 PKS genes of

259   *Gossypium hirsutum* and *Malus domestica*, while 73% of the members of the *Populus tremula*,

260   *Vitis vinifera* and *Arabidopsis thaliana* group had at least one UTR region. The results indicated

261   that the structures of these genes were more complex. All the above results show that the PKS

262   gene family has a diverse genetic structure, which helps to explain the divergence of PKS gene

263   family members. To clarify the structures of the PKS genes, we attempted to gain a better

264   understanding of the conserved motifs of these genes; we thus identified 20 conserved motifs (6-

265   200 amino acid residue widths) using the MEME software (Table S3). The probability of

266   occurrence of motifs 1–10 in upland cotton is more than 65%; we refer to this set as "General

267   Motifs". The remaining motifs11–20 we refer to as "Specific Motifs" (Figure 3) (Cao et al.,

268   2016). Among the 20 motifs (Figure 2B) we found that motifs 1, 3, 5, 7 and 12 encode a

269   Chal_sti_synt_N conservative domain. Motifs2, 4, 6 and 13 encode a Chal_sti_synt_C

270   conservative domain. In upland cotton, in spite of *GhPKS3* lacking motifs 6, 7 and *GhPKS1*

271   lacking motif 6. Almost all PKS family members included motifs 1, 2, 3, 4, 5, 6 and 7. However,

272   in the other four species, this lack of motifs containing the Chal_sti_synt_N and Chal_sti_synt_C

273   domains is more pronounced. For example, *Populus tremula PtPKS4*, *8* and *11* lack motif 6;

274   *Malus domestica MdPKS8* lacks motifs 3, 5 and 7; in *Vitis vinifera* motifs 5 and 7 are present in

275   only 3 and 4 members, respectively. In addition, motif 12 did not appear in 42 PKS proteins of

276   *Gossypium hirsutum*, *Populus tremula*, *Malus domestica* and *Arabidopsis thaliana*, but motif 12

277   appeared only in two of the PKS proteins of *Vitis vinifera* (*VvPKS5*, *VvPKS10*). The frequency

278   of motif 13 is also very low, with a total of only seven PKS family members. In the phylogenetic

279   tree, the nearest members of each subfamily have similar motif combinations. Example

280   combinations include *MdPKS7, 9*, *VvPKS6, 8* and *PtPKS4, 11*. In addition, there are some

281   proteins belonging to a subfamily with unique motifs. For example motif 15 is unique to

282   subfamily IV and motif 17 only appears in the subfamily III. These subfamily-specific motifs

283   play a very important role in the subfamily PKS proteins regarding function.

284

Comparison of GhPKS protein sequences with those of other plants

285

286   We identified and compared 11 sequences of PKS protein in upland cotton with the

287   sequences of *Oryza sativa* chalcone synthase (*OsCHS*), *Arabidopsis thaliana* chalcone synthase

288   (*AtCHS*) and *Medicago sativa* chalcone synthase (*MsCHS*), to clarify the functional divergence

289   of PKS III gene family members. The results are shown in the figure (Figure 4). The blue box

290   and the red font in the figure represent the conservative amino acid residues and the sequence of

291   the red regions shows a very high degree of conservation. The black wavy lines and arrows

292   represent the α-helix and the β-sheet, respectively. The purple five-pointed star represents the

293   catalytic triad (Cys-His-Asn) and the active amino acids (Thr, Phe, Gly, Ser) in the catalytically

294   active central cavity are expressed as green or black triangles. When the plant PKS III enzyme

295   catalyses the polyketone reaction, the starting substrate is first bound at the Cys in the catalytic

296   triplet, followed by decarboxylation of the malonyl-CoA and the occurrence of the substrate

297   condensation reaction so that the polyketone chain is continuously extended (Jez et al., 2002).

298   The final intermediate product undergoes a series of complex cyclization reactions that

299   ultimately form the final product (Abe et al., 2001). Active amino acids located in the

300   catalytically active central chamber can adjust the type of reaction-starting substrate and the

301   length of the polyketone chain by adjusting the size of the catalytically active central chamber

302   space (Jez et al., 2002). The Cys-His-Asn catalysed triplets inherited from keto acyl synthase III

303   (KASIII) (Austin & Noel, 2002) are highly conserved in each sequence in 11 PKS proteins of

304   upland cotton. However, more amino acid substitutions occur at the four active amino acid

305   positions. Thr at *GhPKS2, 3, 4, 7* is replaced by a Met. Ser at *GhPKS1* is replaced by Lys and in

306   *GhPKS6, 8* is replaced by Met at the same position. The active amino acid Phe has two sites in

307   the catalytically active central cavity and is closely related to the decarboxylation reaction of

308   malonyl-CoA, which is represented by a black triangle in the figure. The first Phe active site was

309   highly conserved in all upland cotton PKS proteins, but at the second Phe active site, Phe at

310   *GhPKS2, 3, 4, 7* was replaced by Tyr. The active amino acids Thr, Gly and Ser can regulate the

311   specificity of the reaction substrate as well as the product. In the upland cotton PKS protein,

312   amino acid substitution occurs in active amino acids Thr, Gly and Ser in multiple protein

313   sequences; this phenomenon may be closely related to PKS III gene family functional diversity.

314

315   **Chromosomal localization and gene duplication**

316   To identify the distribution of PKS genes on the chromosome of each species and in the

317   gene cluster, simultaneously to confirm the type of gene duplication events in upland cotton. We

318   mapped the 52 PKS genes in five species (*Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*,

319   *Malus domestica* and *Arabidopsis thaliana*) to identify the chromosomal distribution of these

320   PKS genes (Figure S1). In our study, the PKS genes in the other four species were unevenly

321   distributed on the chromosomes except for the distribution of the PKS gene in the *Vitis vinifera*,

322   which was more concentrated on chromosome 16. In upland cotton, the PKS gene distribution

323   was A2_chr6 (1), A2_chr8 (2), A2_chr9 (1), At_chr11 (1), Dt_chr8 (1), Dt_chr10 (1) and

324   Dt_chr11 (4). In *Populus tremula*, the 14 PKS genes were distributed on chromosomes 1, 2, 3, 4,

325   5, 9 and 12. In *Malus domestica*, we found that the PKS genes were distributed on chromosomes

326   2, 9, 14, 15 and that *MdPKS1* was not mapped to any chromosome. The PKS genes in

327   *Arabidopsis thaliana* are distributed on chromosomes 1, 4 and 5. However, in *Vitis vinifera*, 10

328   PKS genes were distributed on chromosome 16 and the remaining 3 PKS genes were distributed

329   on chromosomes 3, 14 and 15. In the evolution of genes, most gene family expansion is due to

330   the phenomenon of gene duplication, including tandem duplication and fragment duplication. To

331   clarify how the PKS gene family was amplified, we examined the duplication of the PKS genes

332   in five species (*Gossypium hirsutum*, *Populus tremula*, *Vitis vinifera*, *Malus domestica* and

333   *Arabidopsis thaliana*). Among the 52 PKS genes, we identified 10 gene duplication events in

334   *Gossypium hirsutum* (2), *Populus tremula* (3), *Vitis vinifera* (3) and *Malus domestica* (2); in

335   *Arabidopsis thaliana*, no gene duplication events were found. Five pairs of duplicated genes

336   belonged to tandem duplication and five pairs of duplicated genes belonged to fragment

337   duplication (Table 1). After analysing the gene duplication events of the PKS III gene family in

338   five species, we calculated the Ka, Ks and Ka/Ks ratios of the eleven gene duplication events to

339   explore the effects of these genes on the evolutionary processes (Table 1). In general, Ka/Ks<1

340   represents negative selection or purification selection, Ka/Ks>1 represents positive selection and

341   Ka/Ks=1 indicates neutral selection (Bitocchi et al., 2017). In our study, the Ka/Ks values of the

342   10 pairs of duplicated genes were less than 0.309. The results indicated that in these five species,

343   the PKS III gene family was expanded due to gene duplication events and these repeated genes

344   that undergo gene duplication experience strong purifying selection. Sometimes positive

345   selection may be masked by strong negative selection. To identify positive selection of PKS loci

346   in the occurrence of gene duplication events, we also performed a sliding window analysis of

347   two pairs of duplicated genes in upland cotton (Figure S2). There was never more than one

348   repeat locus in the upland cotton, indicating that there was no positive selection for the two pairs

349   of duplicated genes.

350

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

352   To clarify the characteristics of the promoters of PKS genes in upland cotton, we analysed

353   the cis-acting elements of 11 PKS gene promoters in upland cotton (promoter length=1500 bp)

354   (Table S4). Strong light can regulate the expression of the PKS gene and there are many cis-

355   acting elements in the promoter regions of PKS genes in upland cotton, e.g., Box4 (ATTAAT),

356   SP1 (CC(G/A)CCC), CATT-Motifs (GCATTC) and many G-Boxes (CACGTT). It has been

357   reported that *Arabidopsis thaliana* CHS genes were regulated by MYB transcription factors

358   (Chezem & Clay, 2016). In our study, cis-acting elements associated with MYB transcription

359   factors were also found in the promoter region of PKS genes in upland cotton, e.g., MBS

360   (CGGTCA) and MRE (AACCTAA). This suggests that the expression of PKS genes may be

361   regulated by MYB transcription factors. In addition, there are some cis-acting elements related to

362   various life activities, TC-rich repeats (GTTTTCTTAC) associated with defence and stress,

363   anaerobic induction of ARE (TGGTTT) and CGTCA-motifs (CGTCA) related to methyl

364   jasmonate reactions. The specific cis-acting elements, concrete sequences and functions are

365   shown in Table S5.

366

Expression characteristics of the PKS gene in upland cotton

368   The function and expression patterns of genes are closely related (Zhang et al., 2014). To

369   explore the expression patterns of PKS genes in upland cotton, we studied the expression

370   patterns of 11 PKS genes in upland cotton at different stages of cotton fiber development,

371   including 3 DAF, 6 DAF, 12 DAF, 15 DAF, 9 DAF, 18 DAF, 21 DAF and different plant parts,

372   including roots, stems, leaves, fiber (cotton fiber development represented by 6 DAF). *GhPKS8*

373   is a special gene in the 11 upland cotton PKS genes because no expression was detected in any

374   tissue at any stage of cotton fiber development. The other 10 PKS genes were detected in all

375   tissues and at different stages of cotton fiber development (Figure 5). We found that *GhPKS1* is

376   present at a higher level of transcription in the roots. *GhPKS2*, *3* and *7* showed a high expression

377   level in the stems, while the expression levels in roots, leaves and fiber were low. *GhPKS4*

378   showed high levels of transcription in all tissues of upland cotton. *GhPKS6* was highly expressed

379   in the leaves, while the expression of *GhPKS5*, *9*, *10* and *11* in cotton fiber was significantly

380   higher than that in the other three plant tissues. The results of expression patterns of the 11 PKS

381   genes in different tissues of upland cotton showed that *GhPKS5*, *9*, *10* and *11* were mainly

382   expressed in upland cotton fibers. We analysed the expression patterns of 11 PKS genes in

383   upland cotton at different stages of cotton fiber development. The results showed that 11 PKS

384   genes had multiple expression patterns. *GhPKS1*, *6* and *10* showed a gradual increase in

385   transcription level from 3 DAF–15 DAF and the transcriptional level began to decrease after 15

386   DAF. *GhPKS2, 7* had higher transcription levels at the later stages of fiber development and

387   *GhPKS3, 9* reached their highest levels at 12 DAF. *GhPKS4*, *5*, *9* and *11* showed the highest

388   amounts of transcriptional accumulation in the early stages of cotton fiber development. In

389   brown cotton fibers, PAs are the main precursors of pigment. We also studied the accumulation

390 of PAs in the fibers of brown cotton at different developmental stages (Figure 6). The

391 determination of PAs showed that both soluble and insoluble PAs had mainly accumulated

392 before 15 DAF, after which its content gradually decreased; these results were consistent with

393 the previously reported results (Li et al., 2012). Interestingly, *GhPKS4*, *5*, *9*, *11* had a higher

394 level of transcription at the early stages of cotton fiber development; the amount of expression

395 then decreased gradually, which is consistent with the rule of accumulation of PAs in brown

396 cotton fibers.

397

DISCUSSION

399 The plant PKS III gene family, which only exists in the plant kingdom, is associated with a

400 variety of plant life activities (Shimizu et al., 2017). The PKS III gene family is not very large

401 and PKS III gene family members have been identified or cloned in several species. For instance,

402 14 PKS genes have been identified in *Zea mays* (Han et al., 2016), 12 PKS genes have been

403 isolated and sequenced in *Petunia hybrida* (Koes et al., 1989) and 27 PKS genes have been

404 reported in *Oryza sativa* (Hu et al., 2017), which is the species with the largest number of PKS

405 genes reported to date. In this study, we identified 11 PKS genes from upland cotton and

406 compared these with PKS genes in *Populus tremula* (14), *Vitis vinifera* (13), *Malus domestica*

407 (10) and *Arabidopsis thaliana* (4). The 52 PKS genes were divided into four subfamilies, I, II, III

408 and IV, according to the phylogenetic tree nodes. Previous researchers have suggested that most

409 of the CHS genes consist of two exons and one intron (Durbin et al., 2000) and the diversity of

410 gene structures is important for the evolution of gene families (Swarbreck et al., 2008).

411 According to our study, 72% of the 52 PKS genes consisted of two exons and one intron.

412 However, some genes also had different compositions. For example, *VvPKS12* consists of five

413 exons and four introns. Six PKS genes including *GhPKS9* had no introns and seven PKS genes

414 had three exons and two introns.

415 We identified 20 conservative motifs using MEME software (Bailey et al., 2015). Among

416 these 20 motifs, motifs 1, 3, 5, 7 encoded a Chal_sti_synt_N conservative domain and motifs 2, 4,

417 6, 13 encoded a Chal_sti_synt_C conservative domain. All 52 PKS genes with motifs encoding

418 these two conserved domains demonstrate that the PKS III gene family has been highly

419 conserved during evolution. These two conserved domains are associated respectively with acyl

420 transfer activity and transferase activity (Götz et al., 2008), which indicates that these genes

421 function in catalysing the formation of polyketone compounds. We found that the PKS genes in

422 the same subfamily had similar motif compositions, e.g., *MdPKS7, 9*, *VvPKS6, 8* and *PtPKS4, 11*.

423 At the same time, there were some subfamily-specific motifs. The diversity of gene structure and

424 conserved motif distribution may help to explain the functional dispersion of PKS gene family

425 members.

426   The plant PKS III enzyme protein-specific catalytic triad composed of Cys-His-Asn could

427 be traced back to the earliest ancestors of KAS III (Austin & Noel, 2002), which was considered

428 to be important for the maintenance of PKS III gene family functions. Therefore, using BLAST,

429 we queried the protein sequences of the 11 upland cotton PKS genes and the *AtCHS* and *MsCHS*

430 protein sequences with the reported secondary structure of *OsCHS* as a template (Consortium et

431 al., 2003). The results showed that the Cys-His-Asn catalytic triad was highly conserved in all

432 *GhPKS* sequences. However, there were more amino acid substitutions in active amino acids in

433 the catalytic active site. For example, the first Phe site was highly conserved in all *GhPKS*s in the

434 two Phe sites that are closely related to the binding of various CoA, while more amino acid

435 substitutions appear in the second Phe site. At the same time, the three active amino acids (Thr,

436 Gly, Ser), which are responsible for the regulation of the substrate and the length of the

437 polyketide chain, have also been replaced by other amino acids. This suggested that the catalytic

438 triad of the *GhPKS* protein was highly conserved in the process of gene evolution, whereas the

439 active amino acids were not highly conserved. Therefore, we speculated that the diversity of

440 amino acids at the active amino acid sites was the main cause of the functional dispersion of the

441 PKS gene family.

442   Chromosomal localization analysis showed that the distribution of PKS genes in five

443 species in our study was irregular. The PKS genes were more concentrated on chromosome 16

444 except for the PKS genes in *Vitis vinifera*. The rest of the PKS genes were scattered on multiple

445 chromosomes, which is consistent with previous studies (Han et al., 2016). Subsequently, we

446 found 10 pairs of duplicated genes in the five species: two pairs in upland cotton, three pairs in

447 *Populus tremula*, three pairs in *Vitis vinifera* and two pairs in *Malus domestica*. No duplicated

448 genes were found in *Arabidopsis thaliana*. Among the 10 pairs of duplicated genes, 5 of the

449 duplicated genes in the *Vitis vinifera, Malus domestica* and *Populus tremula* were from tandem

450 duplication and the other 5 pairs of duplicated genes were derived from segmental duplication. It

451 has been reported that there are 7 pairs of duplicated genes in the 27 PKS genes of *Oryza sativa*,

452 but only one pair of duplicated genes was the result of segmental duplication (Hu et al., 2017). In

453 *Zea mays*, there were two pairs of duplicated genes in the 14 PKS genes and these were from

454 segmental duplication (Han et al., 2016). The PKS gene family in *Oryza sativa* has many

455 duplicated genes and there are two types of gene duplication in *Oryza sativa*, tandem duplication

456 and fragment duplication, which also explains why the number of PKS genes in *Oryza sativa* is

457 greater than that of other species. We speculated that there were two kinds of duplication modes

458 in the process of PKS gene duplication in terrestrial plants: tandem duplication and fragment

459 duplication. However, it is unknown whether the duplications were mainly in the form of tandem

460 duplication or segmental duplication, which have varied tendencies in different plants. It is

461 generally believed that tandem duplication contributes to the generation of new genes and

462 fragment duplication leads to slower evolution of the gene family (Cao et al., 2016). In upland

463 cotton, the duplications of the PKS gene carried out in the form of segmental duplication

464 indicated that the evolution of the PKS gene family was slow. The analysis of the Ka/Ks values

465 of the 10 repeat genes showed that the Ka/Ks values of the 10 duplicated gene pairs were less

466 than 0.309, which indicated that these replicates had undergone strong purification selection after

467 duplication was complete. This was for a factor in maintaining the PKS gene family.

468   In *Arabidopsis thaliana*, *AtCHS* is regulated by a variety of MYB transcription factors such

469 as *AtMYB11, 58, 63, 111* and other transcription factors that can activate *AtCHS* transcription

470   (Chezem & Clay, 2016). Furthermore, *Arabidopsis thaliana* treated with high-intensity light for

471   24 hours resulted in a 50-fold increase in the activity of chalcone synthase and a large amount of

472   anthocyanin accumulation (Courtney-gutterson et al., 1994). In this study, the analysis of the cis-

473   elements in the promoter regions of these 11 PKS genes of upland cotton showed that the regions

474   contained many elements related to light regulation and MYB transcription factor binding.

475   Therefore, we believe that upland cotton PKS genes may be regulated by light and MYB

476   transcription factors. The expression patterns of PKS genes of upland cotton in different tissues

477   and cotton fiber development were studied by qRT-PCR. *GhPKS1* showed a higher transcription

478   level in the roots; *GhPKS2, 3, 7* showed a high expression level in the stem; *GhPKS5, 9, 10, 11*

479   were mainly expressed in the fibers. The accumulation of PAs in brown cotton fibers occurred

480   mainly before stage 15 DAF of cotton fiber development (Li et al., 2012). The expression of

481   *GhPKS4, 5, 9, 11* was higher in the early stages of cotton fiber development and PAs in the

482   brown cotton fibers gradually accumulated as their expression increased. The procyanidin

483   content then decreased as the amount of expression also gradually decreased. Previous studies

484   have shown that the PKS gene encodes a key enzyme in the flavonoid biosynthetic pathway as

485   the first rate-limiting enzyme (Martinez-Perez et al., 2014). The precursor material of the

486   pigment in the brown cotton fiber is PAs, which are flavonoids (Liu et al., 2016). The expression

487   trend of *GhPKS4, 5, 9, 11* was consistent with the trend of the accumulation of PAs in brown

488   cotton fibers; therefore, we speculate that *GhPKS4, 5, 9, 11* may be involved in brown cotton

489   fiber pigment biosynthesis.

490

CONCLUSION

491

492   In this study, we identified 11 PKS genes from upland cotton and compared them with

493   analogous genes from *Populus tremula*, *Arabidopsis thaliana*, *Vitis vinifera* and *Malus*

494   *domestica*; there were 41 PKS genes with respect to phylogeny, gene structure, conserved motifs

495   and selection pressure. According to the constructed phylogenetic tree, the 52 total PKS genes

496   were divided into 4 subfamilies. Most of the PKS genes were composed of two exons and one

497   intron. The PKS genes in the same subfamily had similar gene structure and conserved motifs.

498   At the same time, our research on structure showed that gene duplication has been the main

499   driving force of the expansion of the PKS III gene family, but there is a kind of species-

500   specificity concerning fragment duplication vs. tandem duplication. The results of the Ka/Ks

501   ratio analysis showed that purification selection has been important in maintaining the function

502   of the PKS III gene family. According to the analysis of cis-acting elements of PKS promoters in

503   upland cotton, the PKS gene may be regulated by MYB transcription factors and light. The

504   analysis of qRT-PCR and the accumulation of PAs in brown cotton fibers suggest that *GhPKS4*,

505   *5*, *9* and *11* may be involved in the accumulation of PAs in brown cotton fibers.

506

ACKNOWLEDGEMENTS

508   The authors are deeply grateful to Prof. Yongping Cai and Prof. Yi Lin, who provided the

509   sample used in the study and very effective direction. The authors also thank Prof. JunShan Gao,

510   Prof. Dahui Li, Dr. Yanan Wang, Dr. Xi Cheng and Dr. Muhammad Abdullah for providing

511   valuable suggestions and comments.

512

ADDITIONAL INFORMATION AND DECLARATIONS

514   **Data Deposition**

515   The following information was supplied regarding data availability:

516   The *Gossypium hirsutum* genome (version1_0.fa), annotation information (version1_0.gff),

517   coding sequences (version1_0.cds.fa) and protein sequences (version1_0.pep.fa). The *Populus*

518   *tremula* protein sequences (version3_0.pep.fa), annotation information (version3_0.gene.gff3).

519 The *Vitis vinifera* protein sequences (version12x.pep.fa), annotation information

520 (version12X.gene.gff3). The *Malus domestica* protein sequences (version1_0.pep.fa), annotation

521 information (version1_0.gene.gff3). The *Arabidopsis thaliana* protein sequences

522 (version167_TAIR10.pep.fa), annotation information (version167_TAIR10.gene.gff3).   This

523 information are available at website (https://phytozome.jgi.doe.gov/pz/portal.html).

524

525 **Supplemental Information**

526 Supplemental Table S1: Primers used in RT-PCR.

527 Supplemental Table S2: The PKS genes identified in this study are listed.

528 Supplemental Table S3: Detailed information of the 20 motifs in the 52 PKS proteins.

529 Supplemental Table S4: Analysis of cis-acting elements of PKS gene promoter in upland cotton.

530 Supplemental Table S5: Potential cis-elements in the 5' regulatory sequences of the 11 *GhPKS*

531 genes.

532 Supplemental Figure S1: Chromosomal locations of PKS genes in five species.

533 Supplemental Figure S2: Sliding window analysis of 2 pairs of duplicated genes in upland cotton.

534

535 # REFERENCES

536 **Abe I, Morita H. 2010.** ChemInform Abstract: Structure and Function of the Chalcone Synthase

537   Super family of Plant Type III Polyketide Synthases. *Natural Product Reports* 27(6):809-838.

538 **Abe I, Takahashi Y, Morita H, Noguchi H. 2001.** Benzalacetone synthase. A novel polyketide

539   synthase that plays a crucial role in the biosynthesis of phenylbutanones in Rheum palmatum.

540   *European Journal of Biochemistry* 268(11):3354-3359.

541 **Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, Castro E, Duvaud E,   Flegel**

542 **V, Fortier A, Gasteiger E, Grosdidier A, Hernandez C, Ioannidis V, Kuznetsov D,**
543 **Liechti R, Moretti S, Mostaguir K, Redaschi N, Rossier G, Xenarios I, Stockinger H.**
544 **2012.** ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research* 40:597-603.

545 **Austin MB, Noel JP. 2002.** The chalcone synthase superfamily of type III polyketide synthases.
546 *Natural Product Reports* 20(1):79-110.

547 **Bailey TL, Johnson J, Grant CE, Noble WS. 2015.** The MEME Suite. *Nucleic Acids Research*
548 43:W39-W49.

549 **Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall**
550 **M, Moxon S, Sonnhammer EL, Studholme YC, Eddy SR. 2004.** The pfam protein
551 families database Nucleic Acids Res. 32. *Nucleic Acids Research* 32: 263-266.

552 **Bitocchi E, Rau D, Benazzo A, Bellucci E, Goretti D, Biagetti E, Panziera A, Laidò G,**
553 **Rodriguez M, Gioia T, Attene G, McClean P, Lee RK, Jackson SA, Bertorelle G, Papa**
554 **R. 2017.** High Level of Nonsynonymous Changes in Common Bean Suggests That Selection
555 under Domestication Increased Functional Diversity at Target Traits. *Frontiers in Plant*
556 *Science* 7:2005.

557 **Burbulis IE, Winkel-Shirley B. 1999.** Interactions among enzymes of the Arabidopsis
558 flavonoid biosynthetic pathway. *Proceedings of the National Academy of Sciences*
559 96(22):12929-12934.

560 **Cao YP, Han YH, Meng DD, Li DH, Jin Q, Lin Y, Cai YP. 2016.** Structural, Evolutionary,
561 and Functional Analysis of the Class III Peroxidase Gene Family in Chinese Pear (Pyrus
562 bretschneideri). *Frontiers in Plant Science* 7:1874-1886.

563 **Chezem WR, Clay NK. 2016.** Regulation of plant secondary metabolism and associated
564 specialized cell development by MYBs and bHLHs. *Phytochemistry* 131:26-43.

565 **Cui YP, Liu ZJ, Zhao YP,Wang YM, Huang Y, Li L, Wu H, Xu SX, Hua JP. 2017.**
566 Overexpression of Heteromeric GhACCase Subunits Enhanced Oil Accumulation in Upland
567 Cotton. *Plant Molecular Biology Reporter* 4(35):287-297.

568 **Curran JM, Tvedebrink T. 2013.** DNAtools: Tools for empirical testing of DNA match
569 probabilities. R package.

570 **Consortium FLC, Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N,**
571 **Yazaki J, Ishikawa M, Yamada H, Ooka H, Hotta I, Kojima K, Namiki T, Ohneda E,**
572 **Yahagi W, Suzuki K, Li CJ, Ohtsuki K, Shishiki T, Otomo Y, Murakami K, Iida Y,**
573 **Sugano S, Fujimura T, Suzuki Y, Tsunoda Y, Kurosaki T, Kodama T, Masuda H,**
574 **Kobayashi M, Xie Q, Lu M, Narikawa R, Sugiyama A, Mizuno K, Yokomizo S, Niikura**
575 **J, Ikeda R, Ishibiki J, Kawamata M, Yoshimura A, Miura J, Kusumegi T, Oka M, Ryu**

576   **R, Ueda M, Matsubara K, RIKEN, Kawai J, Carninci P, Adachi J, Aizawa K, Arakawa**
577   **T, Fukuda S, Hara A, Hashizume W, Hayatsu N, Imotani K, Ishii Y, Itoh M, Kagawa I,**
578   **Kondo S, Konno H, Miyazaki A, Osato N, Ota Y, Saito R, Sasaki D, Sato K, Shibata K,**
579   **Shinagawa A, Shiraki T, Yoshino M, Hayashizaki Y, Yasunishi A. 2003.** Collection,
580   mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*
581   301(5631):376.

582   **Courtney-gutterson N, Napoli C, Lemieux C, Morgan A, Firoozabady E, Robinson KE.**
583   **1994.** Modification of flower color in florist's chrysanthemum: production of a white-
584   flowering variety through molecular genetics. *Bio/technology* 12(3):268-271.

585   **Dare AP, Tomes S, Jones M, McGhie TK, Stevenson DE, Johnson RA, Greenwood DR,**
586   **Hellens RP. 2013.** Phenotypic changes associated with RNA interference silencing of
587   chalcone synthase in apple (Malus×domestica). *Plant Journal* 74(3):398-410.

588   **Durbin ML, Mccaig B, Clegg MT. 2000.** Molecular evolution of the chalcone synthase
589   multigene family in the morning glory genome. *Plant Molecular Biology* 42(1):79-92.

590   **Eom SH, Hyun TK. 2016.** Genome-wide identification and transcriptional expression analysis
591   of chalcone synthase in flax (Linum usitatissimum, L.). *Gene Reports* 5:51-56.

592   **Fan X, Fan B, Wang Y, Yang W. 2016.** Anthocyanin accumulation enhanced in Lc-transgenic
593   cotton under light and increased resistance to bollworm. *Plant Biotechnology Reports*
594   10(1):1-11.

595   **Feng H, Tian X, Liu Y, Zhang X, Jones BJ, Sun Y, Sun J. 2013.** Analysis of Flavonoids and
596   the Flavonoid Structural Genes in Brown Fiber of upland cotton. *Plos One* 8(3):e58820.

597   **Funa N, Ohnishi Y, Fujii I, Shibuya M, Ebizuka Y, Horinouchi S. 1999.** A new pathway
598   forpolyketide synthesis in microo rganisms. *Nature* 400:897-899.

599   **Gao JS, Nan W, Shen ZL, Lv K, Qian SH, Guo N, Sun X, Cai YP, Lin Y. 2016.** Molecular
600   cloning, expression analysis and subcellular localization of a Transparent Testa 12, ortholog
601   in brown cotton (Gossypium hirsutum, L.). *Gene* 576:763-769.

602   **Götz S, Garcíagómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón**
603   **M, Dopazo J, Conesa A. 2008.** High-throughput functional annotation and data mining with
604   the Blast2GO suite. *Nucleic Acids Research* 36(10):3420-3435.

605   **Guo AY, Zhu QH, Chen X, Luo JC. 2007.** [GSDS: a gene structure display server]. *Yi Chuan*
606   29(8):1023-1026.

607   **Gras CC, Nemetz N, Carle R, Schweiggert RM. 2017.** Anthocyanins from purple sweet potato

608 (Ipomoea batatas (L.) Lam.) and their color modulation by the addition of phenolic acids and
609 food-grade phenolic plant extracts. *Food Chemistry*, 235(11):265-274.

610 **Li H, Liang J, Chen H, Ding G, Ma B, He N. 2016.** Evolutionary and functional analysis of
611 mulberry type III polyketide synthases. *BMC Genomics* 17(1):540-558.

612 **Han Y, Ding T, Su B, Jiang H. 2016.** Genome-Wide Identification, Characterization and
613 Expression Analysis of the Chalcone Synthase Family in Maize. *International Journal of*
614 *Molecular Sciences* 17(2):161-176.

615 **Han Y, Zhao W, Wang Z, Zhu J, Liu Q. 2014.** Molecular evolution and sequence divergence
616 of plant chalcone synthase and chalcone synthase-Like genes. *Genetica* 142(3):215-225.

617 **Helariutta Y, Elommaa P, Kotilainen M, Griesbach RJ, Schröder J, Teeri TH. 1995.**
618 Chalconesynthase-like genes activeduringcorolla development are differentially expressed
619 and encode enzymes with differentcataly tic properties in Gerbera hybrida (Asteraceae).
620 *Plant Mol Boil* 28:47-60.

621 **Hinchliffe DJ, Condon BD, Thyssen G, Naoumkina M, Madison CA, Reynolds M,**
622 **Delhom CD, Fang DD, Li P, McCarty J. 2016.** The GhTT2_A07 gene is linked to the
623 brown colour and natural flame retardancy phenotypes of Lc1 cotton (Gossypium hirsutum
624 L.) fibers. *Journal of Experimental Botany* 67(18):5461-5471.

625 **Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007.**
626 WoLF PSORT: protein localization predictor. *Nucleic Acids Research* 35:W585-W587.

627 **Hu L, He H, Zhu C, Peng X, Fu J, He X, Chen X, Ouyang L, Bian J, Liu S. 2017.** Genome-
628 wide identification and phylogenetic analysis of the chalcone synthase gene family in rice.
629 *Journal of Plant Research* 130(1):1-11.

630 **Hua SJ, Wang XD, Yuan, SN, Shao MY, Zhao, XQ, Zhu SJ, Jiang, LX. 2007.**
631 Characterization of Pigmentation and Cellulose Synthesis in Colored Cotton fibers. *Crop*
632 *Science* 47(4):1540-1546.

633 **van der Heijden RT, Snel B, van Noort V, Huynen MA. 2007.** Orthology prediction at
634 scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8:83.

635 **Ikegami A, Akagi T, Potter D, Yamada M, Sato A, Yonemori K, Kitajima A, Inoue K. 2009.**
636 Molecular identification of 1-Cys peroxiredoxin and anthocyanidin/flavonol 3-O-
637 galactosyltransferase from proanthocyanidin-rich young fruits of persimmon (Diospyros kaki
638 Thunb.). *Planta* 230(4):841-855.

639 **Jepson C, Karppinen K, Daku RM, Sterenberg BT, Suh DY. 2014.** Hypericum perforatum
640 hydroxyalkylpyrone synthase involved in sporopollenin biosynthesis--phylogeny, site-

641   directed mutagenesis, and expression in nonanther tissues. *Febs Journal* 281(17):3855-3868.

642   **Jez JM, Bowman ME, Noel JP. 2002.** Expanding the biosynthetic repertoire of plant type III
643   polyketide synthases by altering starter molecule specificity. *Proceedings of the National*
644   *Academy of Sciences of the United States of America* 99(8):5319-5324.

645   **Junghanns KT, Kneusel RE, Baumert A, Maier W, Gröger D, Matern U. 1995.** Molecular
646   cloning and heterologous expression of acridone synthase from elicited Ruta graveolens L.
647   cell suspension cultures. *Plant Molecular Biology* 27(4):681-92.

648   **Koes RE, Spelt CE, van den Elzen PJ, Mol JN. 1989.** Cloning and molecular characterization
649   of the chalcone synthase multigene family of Petunia hybrida. *Gene* 81(2):245-257.

650   **Kumar S, Stecher G, Tamura K. 2016.** MEGA 7.0: Molecular Evolutionary Genetics Analysis
651   Version 7.0 for Bigger Datasets. *Molecular Biology & Evolution* 33(7):1870-1874.

652   **Letunic I, Doerks T, Bork P. 2012.** SMART 7: recent updates to the protein domain annotation
653   resource. *Nucleic Acids Research* 40:302-305.

654   **Librado P, Rozas J. 2009.** DnaSP v5: a software for comprehensive analysis of DNA
655   polymorphism data. *Bioinformatics* 25(11):1451-1452.

656   **Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X,**
657   **Huang G, Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X,**
658   **Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X,**
659   **Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G, Wang K, Xu X, Yu JZ, Zhu**
660   **Y, Yu S. 2015.** Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1)
661   provides insights into genome evolution. *Nature Biotechnology* 33(5):524-530.

662   **Li TC, Fan HH, Li ZP, Wei J, Lin Y, Cai YP. 2012.** The accumulation of pigment in fiber
663   related to proanthocyanidins synthesis for brown cotton. *Acta Physiologiae Plantarum*
664   34(2):813-818.

665   **Liu C, Wang X, Shulaev V, Dixon RA. 2016.** A role for leucoanthocyanidin reductase in the
666   extension of proanthocyanidins. *Nature Plants* 2:16182.

667   **Livak KJ, Schmittgen TD. 2001.** Analysis of relative gene expression data using real-time
668   quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25(4):402-408.

669   **Long M, Thornton K. 2001.** Gene duplication and evolution. *Science* 293(5535):1551.

670   **Luna-Vital D, Li Q, West L, West M, Gonzalez de Mejia E. 2017.** Anthocyanin condensed
671   forms do not affect color or chemical stability of purple corn pericarp extracts stored under
672   different pHs. *Food Chemistry* 232:639-647.

673 **Martinez-Perez C, Ward C, Cook G, Mullen P, McPhail D, Harrison DJ, Langdon SP.**
674   **2014.** Novel flavonoids as anti-cancer agents: mechanisms of action and promise for their
675   potential application in breast cancer. *Biochemical Society Transactions* 42(4):1017-1023.

676 **Niu E, Cai C, Zheng Y, Shang X, Fang L, Guo W. 2016.** Genome-wide analysis of CrRLK1L,
677   gene family in Gossypium, and identification of candidate CrRLK1L, genes related to fiber
678   development. *Molecular Genetics & Genomics* 291(3):1137-1154.

679 **Oikawa T, Maeda H, Oguchi T, Yamaguchi T, Tanabe N, Ebana K, Yano M, Ebitani T,**
680   **Izawa T. 2015.** The Birth of a Black Rice Gene and Its Local Spread by Introgression. *Plant*
681   *Cell* 27(9):2401-2414.

682 **Qian SH, Hong L, Xu M, Cai YP, Lin Y, Gao JS. 2015.** Cellulose synthesis in coloured cotton.
683   *Scienceasia* 41(3):180.

684 **Owens DK, Alerding AB, Crosby KC, Bandara AB, Westwood JH, Winkel BS. 2008.**
685   Functional Analysis of a Predicted Flavonol Synthase Gene Family in Arabidopsis. *Plant*
686   *Physiology* 147(3):1046-1061.

687 **Reimold U, Kröger M, Kreuzaler F, Hahlbrock K. 1983.** Coding and 3' non-coding
688   nucleotide sequence of chalcone synthase mRNA and assignment of amino acid sequence of
689   the enzyme. *Embo Journal* 2(10):1801-1805.

690 **Rombauts S, Déhais P, Van Montagu M, Rouzé P. 1999.** PlantCARE, a plant cis-acting
691   regulatory element database. *Nucleic Acids Research* 27(1):295-296.

692 **Schanz S, Schröder G, Schröder J. 1992.** Stilbene synthase from Scots pine (*Pinus sylvestris*).
693   *Febs Letters* 313(1):71-74.

694 **Schröder J. 2000.** The family of chalcone synthase-related proteins: functional diversity and
695   evolution. *Recent Advances in Phytochemistry* 34:55-89.

696 **Shimizu Y, Ogata H, Goto S. 2017.** Type III Polyketide Synthases: Functional Classification
697   and Phylogenomics. *ChemBioChem* 18:50–65.

698 **Stipanovic RD, Puckhaber LS, Bell AA, Percival AE, Jacobs J. 2005.** Occurrence of (+)- and
699   (-)- gossypol in wild species of cotton and in Gossypium hirsutum Var. marie-galante (Watt)
700   Hutchinson. *Journal of Agricultural and Food Chemistry* 8(53):6266-6271.

701 **Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D,**
702   **Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P,**
703   **Huala E. 2008.** The Arabidopsis Information Resource (TAIR): gene structure and function
704   annotation. *Nucleic Acids Research* 36:D1009-14.

705   **Tuteja JH, Clough SJ, Chan WC, Vodkin LO. 2004.** Tissue-Specific Gene Silencing
706   Mediated by a Naturally Occurring Chalcone Synthase Gene Cluster in Glycine max. *Plant*
707   *Cell* 16(4):819-835.

708   **Hu W, Yang H, Yan Y, Wei Y, Tie W, Ding Z, Zuo J, Peng M, Li K. 2016.** Genome-wide
709   characterization and analysis of bZIP transcription factor gene family related to abiotic stress
710   in cassava. *Scientific Reports* 6:22783.

711   **Xie L, Liu P, Zhu Z, Zhang S, Zhang S, Li F, Zhang H, Li G, Wei Y, Sun R. 2016.**
712   Phylogeny and Expression Analyses Reveal Important Roles for Plant PKS III Family during
713   the Conquest of Land by Plants and Angiosperm Diversification. *Front Plant Sci* 7:1312.

714   **Yuan S, Hua SJ, Malik W, Bibi, N, Wang, XD. 2012.** Physiological and biochemical
715   dissection of fiber development in colored cotton. *Euphytica* 187(2):215-226.

716   **Zhang XW, Xiong HR, Liu AL, Zhou XY, Peng Y, Li ZX, Luo GY, Tian XR, Chen XB.**
717   **2014.** Microarray data uncover the genome-wide gene expression patterns in response to heat
718   stress in rice post-meiosis panicle. *Journal of Plant Biology* 57(6):327-336.

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 4 subfamilies (*PtPKS5* and *PtPKS7* were placed separately into a class). Specific gene names are listed in Supplementary Table S2.

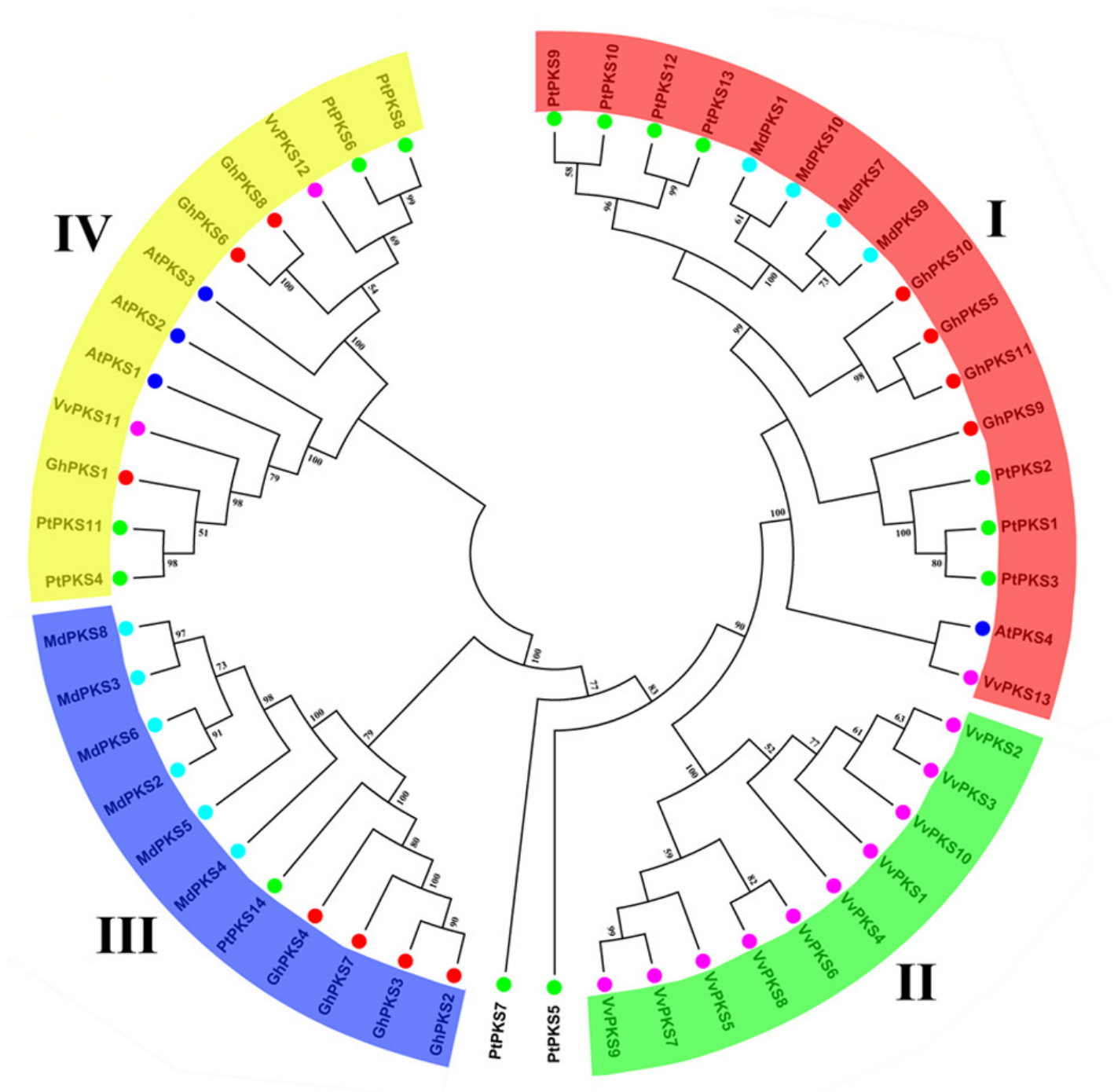



Figure 2

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.

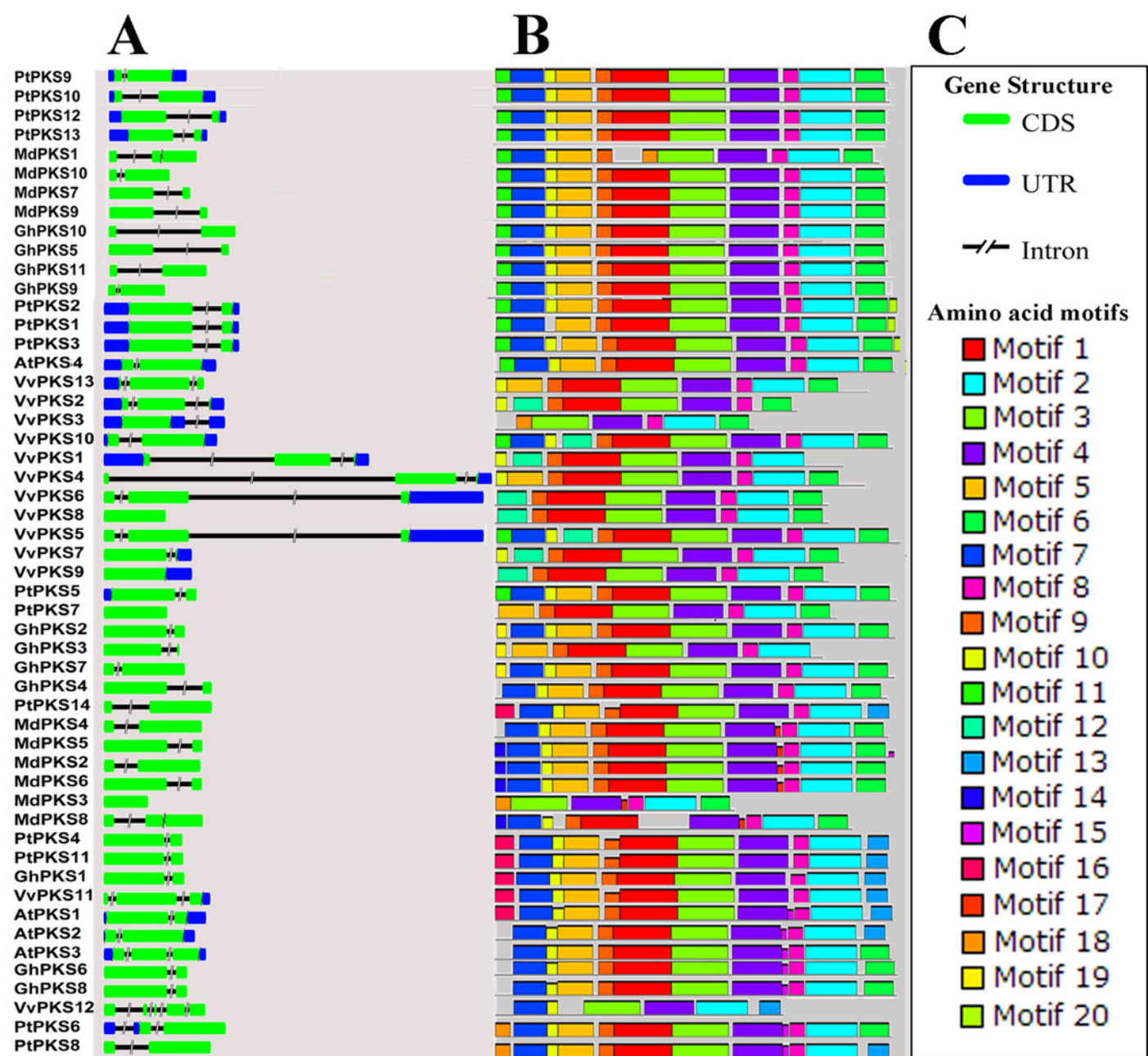



Figure 3

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.

General Motifs	Gossypium hirsutum (11)	Populus tremula (14)	Vitis vinifera (13)	Malus domestica (10)	Arabidopsis thaliana (4)
Motif 1	11	14	11	10	4
Motif 2	11	14	12	10	4
Motif 3	11	14	13	9	4
Motif 4	11	14	13	10	4
Motif 5	11	14	3	8	4
Motif 6	11	11	10	10	2
Motif 7	10	14	4	9	4
Motif 8	11	14	12	10	4
Motif 9	11	14	11	9	4
Motif 10	10	12	8	9	4

Colour Key (%)

%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

Specific Motifs	Gossypium hirsutum (11)	Populus tremula (14)	Vitis vinifera (13)	Malus domestica (10)	Arabidopsis thaliana (4)
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 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).

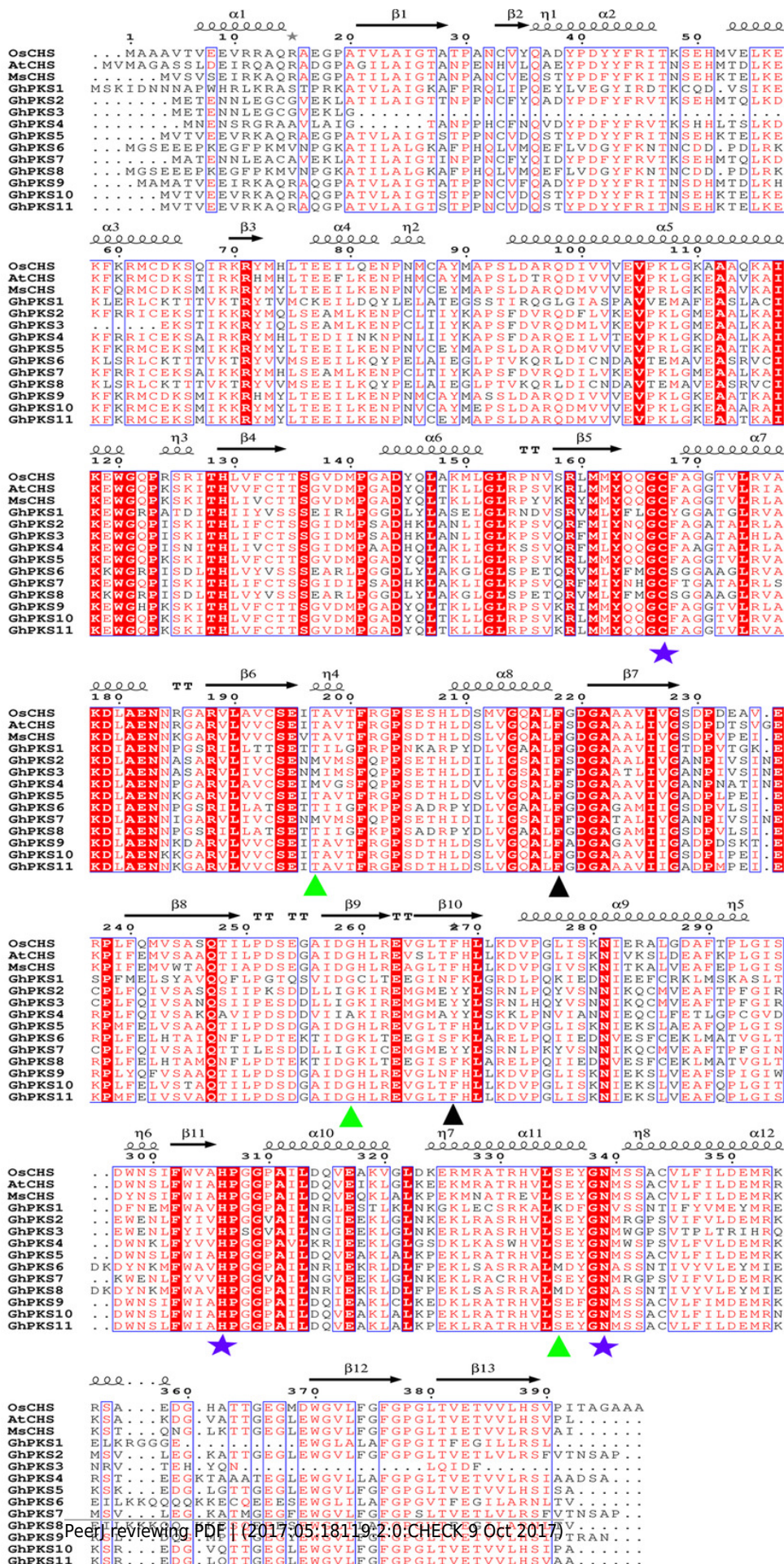



Figure 5

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.

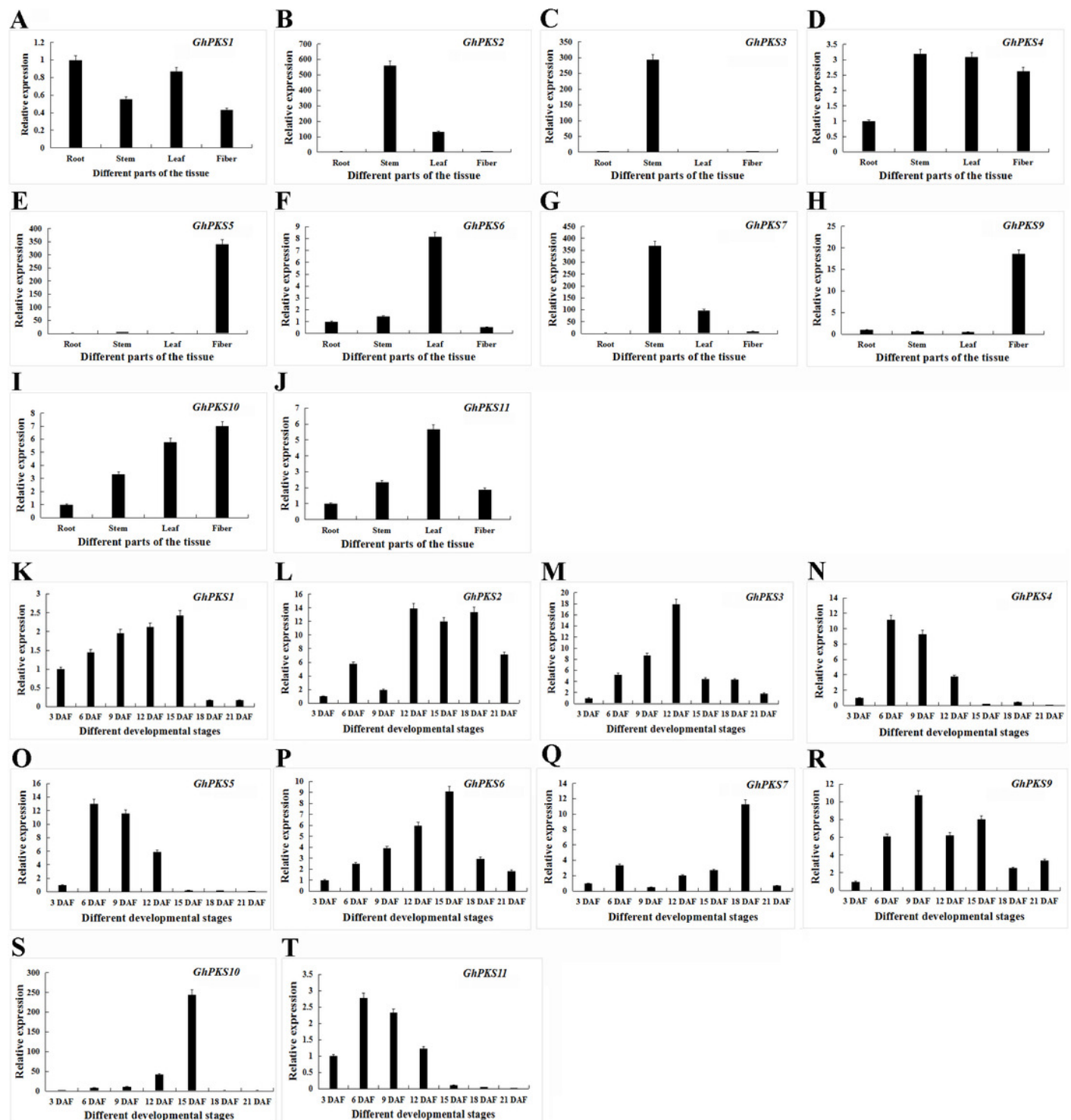
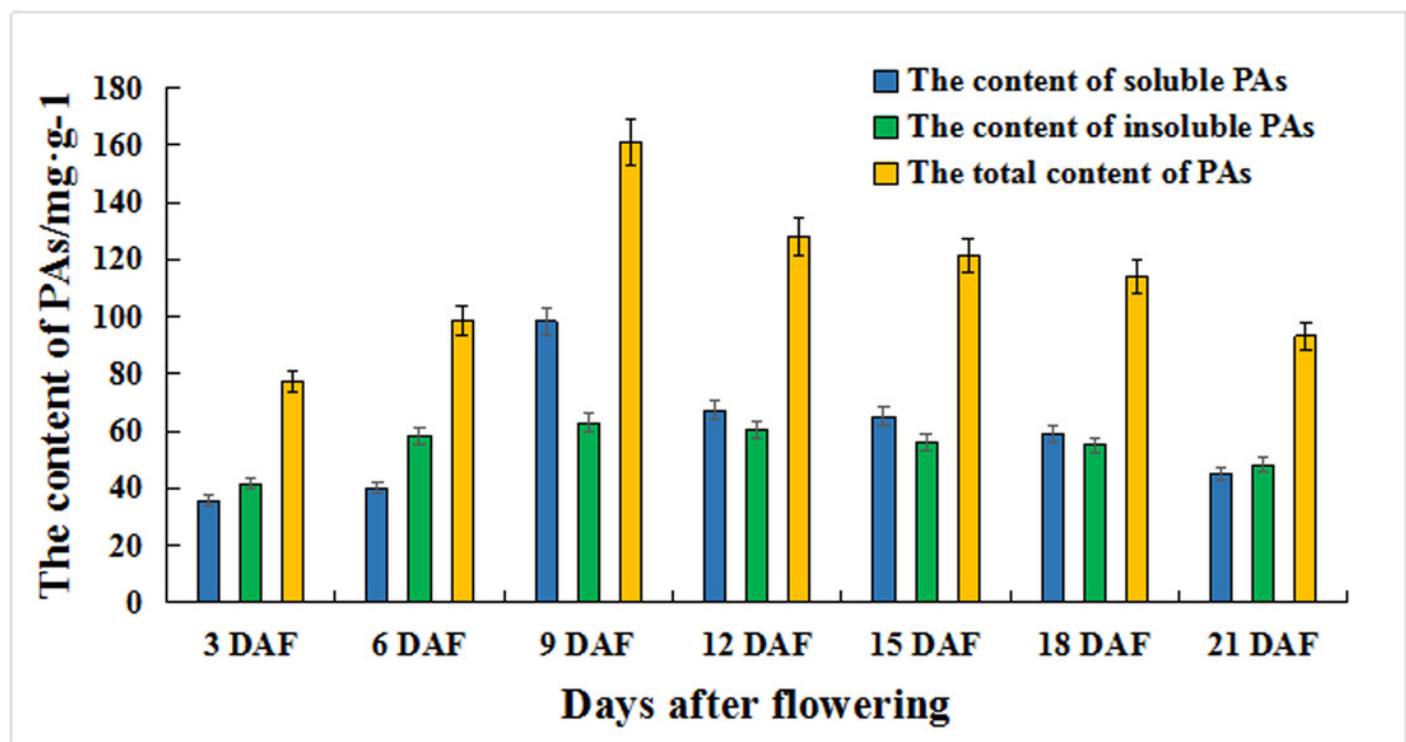



Figure 6

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.

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 Figure S1, and the sliding window analysis results are shown in Figure S2.

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