Turkish Journal of Agriculture and Forestry

Volume 46 | Number 5 Article 11

1-1-2022

# In silico analysis of the regulatory gene families for proanthocyanidins biosynthesis in the genus Gossypium L.

ALEKSANDRA MIKHAILOVA

KSENIA STRYGINA

ELENA KHLESTKINA

Turkish Journal of Agriculture and Forestry

http://journals.tubitak.gov.tr/agriculture/

Research Article

Turk J Agric For
(2022) 46: 743-761

doi:10.55730/1300-011X.3039

# In silico analysis of the regulatory gene families for proanthocyanidins biosynthesis in the genus *Gossypium* L.

**Aleksandra MIKHAILOVA**[1,2,*], **Ksenia STRYGINA**[2], **Elena KHLESTKINA**[2]

[1]Department of Genetics and Biotechnology, Faculty of Biology, Saint-Petersburg State University, St. Petersburg, Russia
[2]Federal Research Center N. I. Vavilov All-Russian Institute of Plant Genetic Resources, St. Petersburg, Russia

**Received:** 09.09.2020 • **Accepted/Published Online:** 23.02.2022 • **Final Version:** 03.10.2022

**Abstract:** The flavonoid pigments provide cotton cultivars with various shades of the naturally colored fiber. Despite of the deep knowledge about the flavonoids biosynthesis pathway in cotton, the genetic features underlying its control are not fully understood. Particularly, data on regulatory genes for the proanthocyanidins (PAs) branch are incomplete for the genus *Gossypium* L. Here we report results of comprehensive in silico analysis of the *R2R3-Myb*, *bHLH-Myc* and *WD40* regulatory gene families, involved in the biosynthesis of flavonoid pigment PAs in cotton. For the first time, we identified regulatory genes *R2R3-Myb*, *bHLH-Myc* and *WD40*, including paralogous and homoeologous copies, among all sequenced diploid and allotetraploid cotton species: *G. hirsutum* L., *G. barbadence* L., *G. raimondii* L. and *G. arboreum* L. All duplication events occurred in the genome of the common diploid ancestor of the genus *Gossypium* and were supported by a negative evolution ($Ka/Ks < 1$). However, the *GhTTG1/GhTTG3* genes are proved to be the most conservative and predominantly supported by selection, while *GhTT2/GhMYB10* and *GhTT8* are more variable and suitable for further analysis. The thorough study has suggested that all genes, with the exception of *GhTT8-A2*, possibly involved in the PAs biosynthesis. Therefore, information about structural organization and functionality of the all *MBW* genes in cotton genome is represented here. Revealing of the allelic diversities among identified genes, associating with contrast phenotype, may be useful for marker-assisted selection or genome editing for creation of naturally colored cotton cultivars with agriculturally valuable traits.

**Key words:** Cotton, brown cotton fiber, flavonoid biosynthesis, MBW complex, proanthocyanidins, in silico

## 1. Introduction

Flavonoids and the other phenolic compounds are known to be an important plant secondary metabolites (Falcone Ferreyra et al., 2012; Panche et al., 2016; Kumar et al., 2018). Flavonoids are an indispensable compounds which, on the one hand, provide color of flowers, extra common seeds, and fruits and participate in many important biochemical processes as protectors against infections, pathogens and UV-radiation from the another hand (Winkel-Shirley, 2002; Taylor and Grotewold, 2005; Kumar et al., 2018).

The chemical structure of flavonoids represents benzo-γ-pyrone derivatives with a common C6-C3-C6 carbon backbone and consists of the oxygen-containing C-ring between two aromatic A- and B- benzene rings. Depending on degree of oxidation of C-ring and its linking position with B-ring, following subclasses are divided into: flavones, flavanones, flavonols, isoflavones, flavanols and anthocyanidins (Figure 1) (Panche et al., 2016). The proanthocyanidins (PAs), also known as condensed tannins and consisting of polymer flavonols units, often stand out into the seventh subclass. The PAs formation begins with flavan-3-ol and leucoanthocyanidin monomer dimerization, which is followed by leucoanthocyanidin units condensation during phenylpropanoid branch of the flavonoid biosynthesis pathway (Figure 1) (Tanner et al., 2003; Zhao and Dixon, 2009; Xu et al., 2018).

Different environmental factors, such as solar irradiance, extreme temperatures, drought, nitrogen deficiency, sucrose and phytohormones (jasmonate, abscisic acid, cytokines), positively influence on PAs biosynthesis and its accumulation (Christie et al., 1994; Lea et al., 2007; Shan et al., 2009). It was reported that metabolites of the starch and sucrose biosynthesis pathways are proposed to be precursors for formation of condensed tannins, which is a main candidate responsible for brown cotton fiber pigmentation (Peng et al., 2020).

Pigmentation of cotton (the genus *Gossypium L.*) fiber, stem and leaves is observed due to PAs and anthocyanins accumulation in the large central cell vacuole (Harland, 1932; Lane and Schuster, 1981; Debeaujon et al., 2001; Xiao et al., 2007; Feng et al., 2013; Sun et al., 2019). Early studies concerning genetic aspects of colored cotton fiber

* Correspondence: a.mikhailova@vir.nw.ru

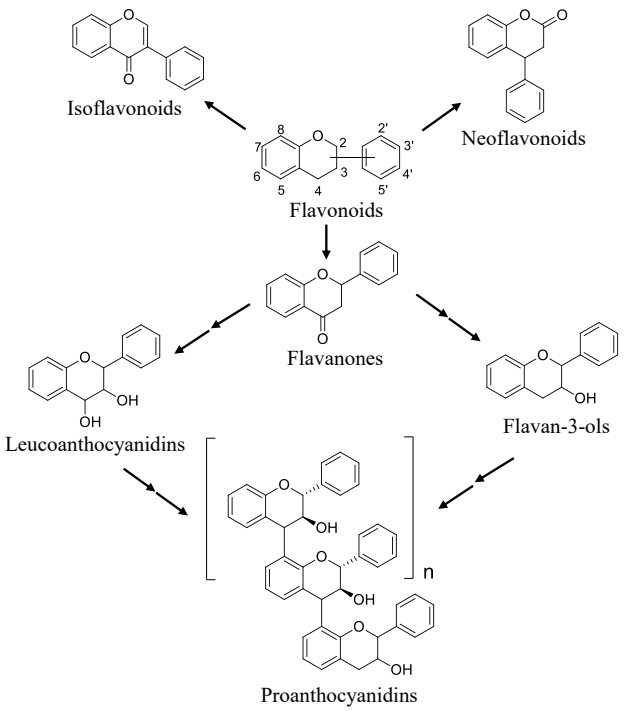



**Figure 1.** Classification of flavonoids based on C-ring radical varieties.

phenotype began with Harland (1932). He established that brown color is linked with two independent loci, *Lc1* and *Lc2*; *Lc1* was mapped on the long arm of chromosome 7 (Endrizzi and Taylor, 1968). Later, additional four loci (*Lc3-Lc6*) associated with different shades of brown fiber were revealed (Kohel, 1985). Hinchliffe (2016) linked the *GhTT2_A07* gene (also known as *GhTT2*/*GhTT2-3A*/ *GhMYB36*, chromosome 7A) with the *Lc1* phenotype (see below).

It is widely known that a raw of structural genes such as *Chalcone synthase* (*CHS*), *Anthocyanidin reductase* (*ANR*) and *Leucoanthocyanidin reductase* (*LAR*) are responsible for the PAs accumulation (Peng et al., 2012; Li et al., 2013; Feng et al., 2014; Gao et al., 2019). The high expression level of structural genes involved in flavonoid biosynthesis: *Phenylalanine ammonia-lyase* (*PAL*), *Cinnamate 4-hydroxylase* (*C4H*), *Flavonoid 3',5'-hydroxylase* (*F3'5'H*), *Flavonol synthase* (*FLS*), *Dihydroflavonol 4-reductase* (*DFR*) and *LAR* positively correlates with the pigment content and upregulates during colored cotton fiber development (Peng et al., 2020; Tang et al., 2021). In addition to this, transgenic RNAi *G. hirsutum* L. plants (*GhCHSi*, *GhANRi* and *GhLARi*) demonstrate reduced fiber coloration (Gao et al., 2019). The myeloblastosis family with two imperfect amino acid sequence repeats (R2R3-Myb), basic helix-loop-helix type Avian virus oncogene Myelocytomatosis (MYC) (bHLH-Myc) and tryptophan-aspartic acid (W-D) dipeptide repeats (WD40) transcription factors (TFs) form together the ternary MBW complex, which regulates the transcriptional activity of mentioned structural genes. There are a raw of studies underlying the importance of MBW protein complex for the flavonoid biosynthesis and trichome pattern formation pathways (Werber and Weisshaar, 2001; Dubos et al., 2010; Li, 2014; Xu et al., 2014; Xu et al., 2015; Sun et al., 2016; Lloyd et al., 2017; Doroshkov et al., 2019; Huang et al., 2019; Stracke et al., 2020). A specific combination of R2R3-Myb and bHLH-Myc proteins is stabilized by WD40 A specific combination of R2R3-Myb and bHLHMyc proteins stabilized by WD40 repeat proteins. This MBW complex mediates anthocyanidins and PAs biosynthesis pathways as well as trichome hair pattering in higher plants (Koes et al., 2005; Li, 2014; Lloyd et al., 2017). For instance, products of the *R2R3-Myb* family genes: *PRODUCTION OF ANTHOCYANIN PIGMENT 1* (*PAP1*), *PAP2*, *MYB113* and *MYB114* of *Arabidopsis thaliana (*L.*)* Heynh interact with *TRANSPARENT TESTA 8*, *GL3* GLABROUS 3 (GL3) and *EGL3* ENHANCER OF GLABRA 3 (EGL3) (*bHLH* family) and WD40 proteins together activate anthocyanin biosynthesis (Gonzalez et al., 2008; Guo et al., 2014; Liang et al., 2020). In *A. thaliana* PRODUCTS of the *R2R3-Myb* gene *TRANSPARENT TESTA 2* (*AtTT2*), *bHLH-Myc* gene (*AtTT8*) and the *WD40* gene *TRANSPARENT TESTA GLABRA 1 (AtTTG1)* positively regulate the PAs biosynthesis in seed coat endothelium via activating expression of *ANR,* encoding the enzyme in this pathway

(Devic et al., 1999; Walker et al., 1999; Nesi et al., 2000; Nesi et al., 2001; Maier et al., 2013; Li, 2014).

The known *MBW* genes: control expression of the PAs biosynthesis structural genes *GhMYB10* (*R2R3-Myb*), *GhTT2*/*GhTT2-3A*/*GhMYB36* (*R2R3-Myb*), *GhTT8*/*GhbHLH130D* (*bHLH-Myc*), *GhTTG1* (*WD40*) and *GhTTG3* (*WD40*) were found out earlier in the *G. hirsutum* genome (Humphries et al., 2005; Hinchliffe et al., 2016; Lu et al., 2017; Yan et al., 2018; Mikhailova et al., 2019).

For the first time, in this study we identified paralogous and homoeologous copies of the *R2R3-Myb* (*TT2-1*, *TT2-2*, *TT2-3*, *MYB10* genes; 24 in total), *bHLH-Myc* (*TT8-1*, *TT8-2* genes; 12 in total) and *WD40* (*TTG1, TTG3* genes; 11 in total) regulatory genes in allotetraploid (*G. hirsutum* (2n = 4x = 52, the genome $(AD)_1$), *G. barbadense* L. (2n = 4x = 52, the genome $(AD)_2$)) and diploid (*G. raimondii* L. (2n = 2x = 26, the genome $D_5$), *G. arboreum* L. (2n = 2x = 26, the genome $A_2$)) cotton species. Analysis of genetic similarity displayed that all duplications occurred in the genome of the common diploid ancestor of the genus *Gossypium* about 27.46–29.98 MYA for *R2R3-Myb*, 32.74 MYA for *bHLH-Myc* and 45.19 MYA for the *WD40* genes. Conservative domain motifs and structural organization of the regulatory gene copies, including a comprehensive analysis concerning promoter regions and evolutionary development features of detected copies, are described in details. Moreover, comparative characteristics of the 3D protein structures of the *R2R3-Myb*, *bHLH-Myc* and *WD40* genes are given here.

The aim of this work was to reveal and characterize the regulatory *R2R3-Myb*, *bHLH-Myc* and *WD40* genes instead gene copies. Revealing of different allelic variants of gene, associating with phenotype, may be useful for development of genetic markers for the next-generation breeding of naturally colored cotton cultivars by marker-assisted selection or gene editing.

## 2. Materials and methods

### 2.1. Identification of the *R2R3-Myb, bHLH-Myc* and *WD40* gene copies

The search for *R2R3-Myb*, *bHLH-Myc* and *WD40* families' members was based on *GhTT2* (CottonFGD: Gh_A07G2341), *GhMYB10* (CottonFGD: Gohir.A06G075700), *GhTT8* (also known as *GhbHLH130D*; CottonFGD: Gh_D11G1273), *GhTTG1* (GenBank: AF530907) and *GhTTG3* (GenBank: AF530911) sequences in the genomes of diploid *G. raimondii* ($D_5$), *G. arboreum* ($A_2$) and tetraploid *G. hirsutum* $(AD)_1$, *G. barbadence* $(AD)_2$ cotton species using BLASTN algorithm in databases CottonFGD (https://cottonfgd:), CottonGene (https://www.cottongen:) and NCBI (https://www.ncbi.
nlm.nih.gov/) databases. (Yu et al., 2014; Zhu et al., 2017). Prediction of *cis*-acting regulatory DNA elements in promoter region was carried out with New PLACE database (https://www.dna.affrc.go.jp/PLACE/?action=newplace) (Higo et al., 1999).

### 2.2. Cluster analysis and evolutionary studies

Multiple alignments of nucleotide sequences were conducted with ClustalW (Kumar et al., 2018). Cluster analysis was performed using MEGA-X and UPGMA algorithm with 1000 bootstrap replications and the maximum composite likelihood method as a substitution model (Felsenstein, 1985; Saitou and Nei, 1987; Kumar et al., 2018). The *R2R3-Myb* genes of *A. thaliana AtPAP1* (GenBank: AT1G56650), *AtPAP2* (GenBank: AT1G66390), *AtMYB113* (GenBank: AT1G66370), *AtMYB114* (GenBank: AT1G66380), *AtMYB115* (as outgroup; GenBank: AT5G40360), the *bHLH-Myc* genes *AtTT8* (GenBank: AJ277509), *AtGL3* (GenBank: AT5G41315), *AtEGL3* (GenBank: AT1G63650), *AtMYC2* (as outgroup; GenBank: AT1G32640) and the *WD40* genes *AtTTG1* (GenBank: AT5G24520), *AtMSI3* (as outgroup; GenBank: AT4G35050) were used for the phylogenetic trees reconstruction. The number of nonsynonymous substitutions per nonsynonymous sites (*Ka*), synonymous substitutions per synonymous sites (*Ks*), *Ka*/*Ks* ratio and divergence time were calculated. To apply the molecular clocks model, the divergence time of the *Malvaceae* Juss. and *Brassicaceae* Burnett families (68–96 MYA based on Abdurakhmonov (2010)) was used for calibration.

### 2.3. In silico analysis of predicted amino acid sequences

MultAlin was applied for multiple amino acid sequences alignment (Kumar et al., 2018). Conservative domains were identified using Pfam database (http://pfam.xfam:). Protein similarity determined with LALIGN (https://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=lalign&pgm=lal), the templates with the largest degree of homology for AtTT2 (PDB: 6KKS), AtTT8 (PDB: 4RS9, 5GNJ), and AtTTG1 (PDB: 2HES) protein sequences were found in RSCB PDB database using BLASTp search (https://www.rcsb:). These templates were applied to predict 3D protein models belonging to the *R2R3-Myb*, *bHLH-Myc* and *WD40* families. Presumable 3D protein structures were visualized using the SWISS-MODEL software (https://swissmodel.expasy:interactive) (Waterhouse et al., 2018). PDBeFold (https://www.ebi.ac.uk/msd-srv/ssm/) was used to carry out the multiple comparison of 3D alignment of protein structures. HOPE server (http://www.cmbi.ru.nl/hope) provides 3D protein structures modelling by available databases so that it allows to describe physicochemical features for the wild-type and mutant proteins (Venselaar et al., 2010). Input data included GhTT8-A2 protein sequence and a single mutation Arginine into Serine at position 495, while output gave out changes affected by this mutation.

## 3. Results

### 3.1. Identification and structural organization of duplicated regulatory gene copies in diploid and allotetraploid cotton species

***R2R3-Myb.*** Search for homologous copies in the genomes of diploid *G. arboreum* ($A_2$), *G. raimondi* ($D_5$)

and allotetraploid *G. barbadence* ((AD)2) cotton species, deposited in databases CottonFGD and CottonGene databases was conducted using BLASTN algorithm, based on known *G. hirsutum* L. ($(AD)_1$) *GhTT2* (chromosome 7; CottonFGD: Gh_A07G2341) and *GhMYB10* (chromosome 6; CottonFGD: Gohir.A06G075700) gene sequences.

Three pairs of the *TT2* paralogous genes tandemly locate in chromosomes 7A and 7D found in the genomes of all species, while *GrTT2* of *G. raimondii,* on the authority of the CottonFGD database, locate in chromosome 1D. All identified genes were designated as *TT2-A1*, *TT2-A2*, *TT2-A3* or *TT2-D1*, *TT2-D2*, *TT2-D3* according to the rules of designation of homoeologous copies and based on the *TT2* gene name in *A. thaliana*.

Single *GhMYB10* gene copy was detected in the A- and D-genomes among all considered species. The *GhMYB10* genes located in chromosome 6, however in the D-genome of *G. raimondii GhMYB10* is in chromosome 10. All found sequences are listed in Table 1.

***bHLH-Myc.*** The known *GhTT8* gene (chromosome 11D; CottonFGD: Gh_D11G1273) was used for identification of the *bHLH-Myc* gene copies in the A- and D-genomes of *G. hirsutum*. *GhTT8* is known to be orthologue of *A. thaliana AtTT8*, therefore *TT8* is assigned for all considered in the framework of this study *Gossypium* species.

Two *GhTT8* homoeologous gene pairs in chromosomes 11A (*TT8-A1*), 11D (*TT8-D1*), 8A (*TT8-A2*) and 8D (*TT8-D2*) in the *G. hirsutum* and *G. barbadense* genomes were revealed. In the A-genome of *G. arboretum* two *GaTT8* paralogous gene copies were detected in chromosomes 11A (*GaTT8-A1*) and 8A (*GaTT8-A2*). However, regarding to the CottonFGD data, two paralogous copies of the *GrTT8* gene were found in the D-genome of *G. raimondii* and locate in chromosomes 7D (*GrTT8-D1*) and 4D (*GrTT8-D2*) (Table 1).

***WD40.*** The homologues genes in *Gossypium* species were found based on sequences of the *WD40* coding genes *GhTTG1* (GenBank: AF530907) and *GhTTG3* (GenBank: AF530911) of *G. hirsutum* using the BLASTN algorithm. *TTG1* and *TTG3* designation was accepted for all identified genes regarding with orthologous gene in the *A. thaliana* genome.

Homoeologous pairs of the *GhTTG1* genes (*GhTTG-A1* and *GhTTG-D1*) and the *GhTTG3* genes (*GhTTG-A3* and *GhTTG-D3*) were found in chromosomes 8A and 8D and in chromosomes 5A and 4D of *G. hirsutum*, respectively. *G. barbadense* has two homoeologous copies of *GbTTG-A1* and *GbTTG-D1* in chromosome 8A and 8D of the A-genome, respectively, and D-genome, but only one *GbTTG-A3* gene was observed in the A-genome. The *GaTTG-A1* and *GaTTG-A3* genes were found in chromosomes 8A and 4A in the *G. arboretum* genome. In *G. raimondii* two *GrTTG-D1* and *GrTTG-D3* homologous genes locate in chromosomes 4D and 12D, respectively (Table 1).

### 3.2. Structural organization

The exon-intron organization for all *MBW* genes was observed for *G. hirsutum* species.

We established that all identified *Myb* gene copies have the R2 and R3 motifs (Pfamseq database accession number: PF00249) and consist of three exons and two introns (Figure 2). Critical amino acid residues of the R2- and R3-Myb domains, according to Bedon (2007), are indicated by asterisks (Supplementary file 1). It was found that genome assembling of the Joint Genome Institute, Walnut Creek, California (JGI) differs from the Zhejiang University, Hangzhou, China (ZJU) and the Holy Angel University (HAU) in the CottonFGD database. The *GhMYB10-A1* and *GhMYB10-D1* gene sequences have 24 amino acids insertion, according to the ZJU data (Supplementary file 2). *GhTT2* contains the 20 amino acids deletion in the Myb domain region (Supplementary file 2). Therefore, the data of the JGI assembly were used for further analysis.

In addition to this, some replacements: Leu18Trp, Gly50Cys and Lys51Arg in GhTT2-A2, Gly65Cys and Lys66Arg in GhTT2-D2, Pro42Ser in GhTT2-A3, Pro39Ser and His69Arg in GhTT2-D3 were observed.

The all *TT8* cotton genes consist of eight exons and seven introns with two conservative region as the Helix-Loop-Helix DNA-binding domain (HLH; Pfamseq: PF00010) and the N-terminal region of the MYB and MYC TFs families (bHLH-Myc_N, Pfamseq: PF14215) (Figure 2). These conserved motifs were turned up among all observed *MBW* genes using Pfam database. Conservative amino acids are pointed as asterisks in accordance with earlier reported data (Atchley et al., 2003; Toledo-Ortiz et al., 2003) (Supplementary files 3 and 4). Analysis of the amino acid sequences of the identified TT8 proteins showed the substitution of arginine to serine at position 13 (Arg13Ser) in the Myc2 domain of GhTT8-A2, as well as valine to isoleucine (Val52Ile) and arginine to lysine (Arg53Lys) replacements among all identified genes in the HLH region (Myc2) (Supplementary file 4).

Two the WD40 conservative motifs (Pfamseq: PF00400), located in a single exon of all detected *WD40* genes in the *G. hirsutum* genome, were revealed (Figure 2). There are no critical substitutions noticed in this region.

### 3.3. Analysis of the promoter regions

We searched for *cis*-acting regulatory DNA elements in approximately 1000 bp regions upstream to the ATG start codon of the *R2R3-Myb, bHLH-Myc* and *WD40* genes of *G. hirsutum* (Figure 3). We observed strongly conserved sequences, such as CAAT-box and GATA-box for both the *R2R3-Myb* and *bHLH-Myc* genes. The group of the core TATA-box promoters, including TATABOX2, TATABOX3,

**Table 1**. Copies of the *GhTT2*, *GhMYB10* (*R2R3-MYB* family), *GhTT8* (*bHLH-MYC* family), *GhTTG1, GhTTG3* (*WD40* family) genes, which were found out in the Cotton Functional Genomics Database among allotetraploid and diploid genome cotton species.

| Gene | Species | Chromosome | Gene | CottonGen CDS | CottonGen genome |
|---|---|---|---|---|---|
| ***R2R3-Myb*** | | | | | |
| *TT2-1* | *Gossypium hirsutum* ((AD)1) | 7A | *GhTT2-A1* | Gohir.A07G020200 | Gh-JGI-A07 |
| | | 7D | *GhTT2-D1* | Gohir.D07G019900 | Gh-JGI-D07 |
| | *Gossypium barbadense* ((AD)2) | 7A | *GbTT2-A1* | GB_A07G0201 | Gb-ZJU-A07 |
| | | 7D | *GbTT2-D1* | GB_D07G0208 | Gb-ZJU-D07 |
| | *Gossypium arboreum* $A_2$ | 7A | *GaTT2-A1* | Ga07G0216 | Ga-CRI-Chr07 |
| | *Gossypium raimondii* $D_5$ | 1D | *GrTT2-D1* | Gorai.001G020600 | Gr-JGI-Chr01 |
| *TT2-2* | *G. hirsutum* | 7A | *GhTT2-A2* | Gohir.A07G020000 | Gh-JGI-A08 |
| | | 7D | *GhTT2-D2* | Gohir.D07G019800 | Gh-JGI-D08 |
| | *G. barbadense* | 7A | *GbTT2-A2* | GB_A07G0199 | Gb-ZJU-A07 |
| | | 7D | *GbTT2-D2* | GB_D07G0207 | Gb-ZJU-D07 |
| | *G. arboreum* | 7A | *GaTT2-A2* | Ga07G0215 | Ga-CRI-Chr07 |
| | *G. raimondii* | 1D | *GrTT2-D2* | Gorai.001G020500 | Gr-JGI-Chr01 |
| *TT2-3* | *G. hirsutum* | 7A | *GhTT2-A3* | Gohir.A07G019900 | - |
| | | 7D | *GhTT2-D3* | Gohir.D07G019700 | Gh-JGI-D05 |
| | *G. barbadense* | 7A | *GbTT2-A3* | GB_A07G0198 | Gb-ZJU-A07 |
| | | 7D | *GbTT2-A3* | GB_D07G0206 | Gb-ZJU-D07 |
| | *G. arboreum* | 7A | *GaTT2-A3* | Ga07G0214 | Ga-CRI-Chr07 |
| | *G. raimondii* | 1D | *GrTT2-D3* | Gorai.001G020400 | Gr-JGI-Chr01 |
| *MYB10* | *G. hirsutum* | 6A | *GhMYB10-A1* | Ghir_A06G007950 | Ghir-HAU-A06 |
| | | 6D | *GhMYB10-D1* | Ghir_D06G008190 | Ghir-HAU-D06 |
| | *G. barbadense* | 6A | *GbMYB10-A1* | Gbar_A06G007930 | Gbar-HAU-A06 |
| | | 6D | *GbMYB10-D1* | Gbar_D06G008240 | Gbar-HAU-D06 |
| | *G. arboreum* | 6A | *GaMYB10-A1* | Ga06G0803 | Ga-CRI-Chr06 |
| | *G. raimondii* | 10D | *GrMYB10-D1* | Gorai.010G087200 | Gr-JGI-Chr10 |
| ***bHLH-Myc*** | | | | | |
| *TT8-1* | *G. hirsutum* | 11A | *GhTT8-A1* | GH_A11G1273 | Gh-NAU-A11 |
| | | 11D | *GhTT8-D1* | GH_D11G1302 | Gh-NAU-D11 |
| | *G. barbadense* | 11A | *GbTT8-A1* | GB_A11G1279 | Gb-ZJUA11 |
| | | 11D | *GbTT8-A1* | GB_D11G1316 | Gb-ZJU-D11 |
| | *G. arboreum* | 11A | *GaTT8-A1* | Ga11G2761 | Ga-CRI-Chr11 |
| | *G. raimondii* | 7D | *GrTT8-D1* | Gorai.007G136400 | Gr-JGI-Chr07 |
| *TT8-2* | *G. hirsutum* | 8A | *GhTT8-A2* | GH_A08G2121 | Gh-NAU-A08 |
| | | 8D | *GhTT8-A2* | GH_D08G2138 | Gh-NAU-A08 |
| | *G. barbadense* | 8A | *GbTT8-A2* | GB_A08G2229 | Gb-ZJU-A08 |
| | | 8D | *GbTT8-D2* | GB_D08G2219 | Gb-ZJU-D08 |
| | *G. arboreum* | 8A | *GaTT8-A2* | Ga08G2187 | Ga-CRI-Chr08 |
| | *G. raimondii* | 4D | *GrTT8-D2* | Gorai.004G215900 | Gr-JGI-Chr04 |
| ***WD40*** | | | | | |
| *TTG1* | *G. hirsutum* | 8A | *GhTTG-A1* | Gh_A08G0926 | Gh-JGI-A08 |
| | | 8D | *GhTTG-D1* | Gh_D08G1130 | Gh-JGI-D08 |
| | *G. barbadense* | 8A | *GbTTG-A1* | Gbar_A08G012290 | Gbar-HAU-A08 |
| | | 8D | *GbTTG-D1* | Gbar_D08G011870 | Gbar-HAU-D08 |
| | *G. arboreum* | 8A | *GaTTG-A1* | Ga08G1234 | Ga-CRI-Chr08 |
| | *G. raimondii* | 4D | *GrTTG-D1* | Gorai.004G125200 | Gr-JGI-Chr04 |
| *TTG3* | *G. hirsutum* | 5A | *GhTTG-A3* | Gohir.A05G415900 | Gh-JGI-A05 |
| | | 4D | *GhTTG-D3* | Gohir.D04G000300 | Gh-JGI-D04 |
| | *G. barbadense* | 5A | *GbTTG-A3* | Gbar_A05G041970 | Gbar-HAU-A05 |
| | | - | - | - | - |
| | *G. arboreum* | 4A | *GaTTG-A3* | Ga04G2136 | Ga-CRI-Chr04 |
| | *G. raimondii* | 12D | *GrTTG-D3* | Gorai.012G001200 | Gr-JGI-Chr12 |

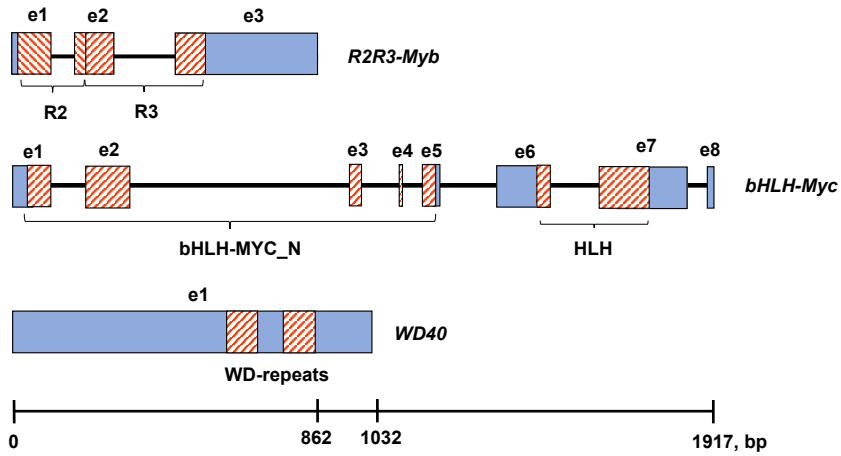



**Figure 2.** Structure of flavonoid biosynthesis regulatory genes in cotton with shading of the main domains.

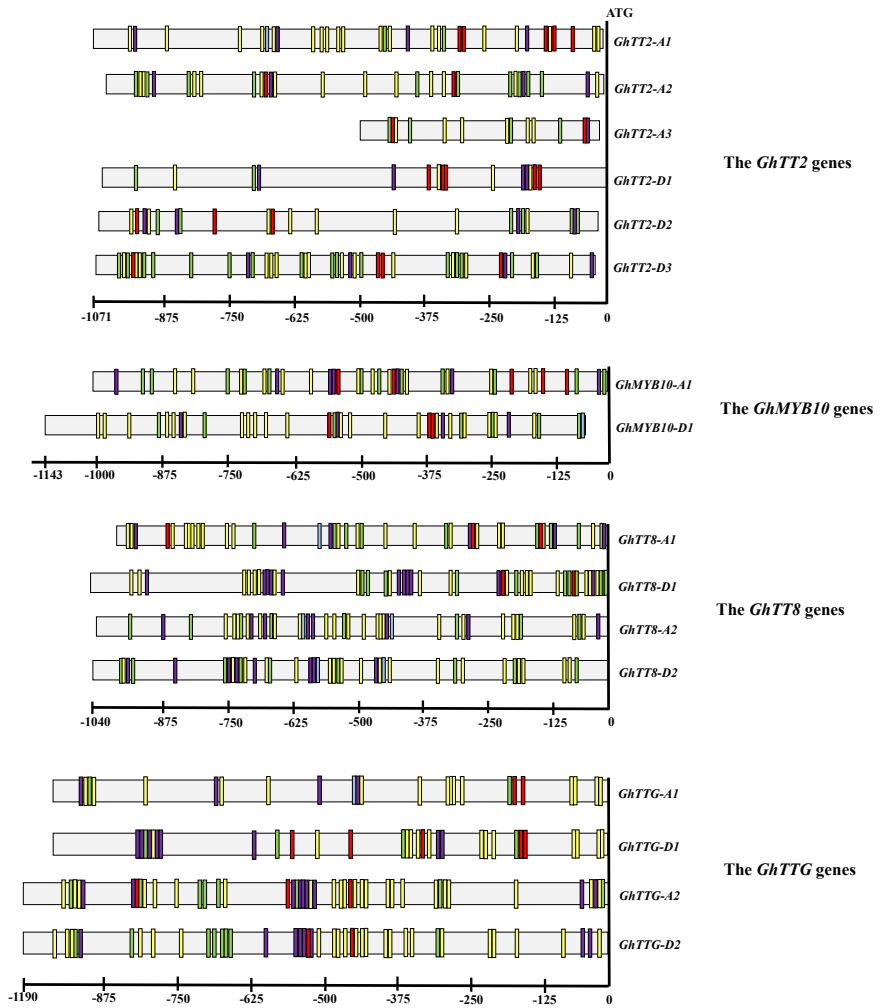



**Figure 3.** Pattern of cis-acting relulatory elements on the 5'-upstream region of the *GhTT2*, *GhMYB10*, *GhTT8* and *GhTTG* genes family copies in the A- and D-genomes of cotton. Green – hormone-responsive site (jasmonate, auxin); yellow – light-responsive site; red – Myc factors binding motif; purple – Myb factors binding motif; blue – low temperature responsive element.

TATABOX4, TATABOXOSPAL and TATABOX5, locates on the 5'-upstream region of the *R2R3-Myb*, *bHLH-Myc* and *WD40* genes.

Moreover, we detected the GT1 and GATA cis-regulatory elements of many light-regulated genes. In addition, we found the conservative light-responsive motifs, T-box and I-box, on the 5'-upstream regions not only in the *R2R3-Myb* genes, but also in *bHLH-Myc* and *WD40* (Figure 3).

Group of the sequences over-represented in light-repressed promoters (SORLREPs), such as SORLREP3AT, SORLIP5AT, SORLIP1AT and SORLIP2AT, represented in *A. thaliana*, also were found among promoter region of cotton genes (Jiao et al., 2005). Moreover, the GT-1 binding sites, including GT1CONSENSUS and GATABOX, known to be the *cis*-regulatory elements of many light-regulated genes, were observed (Liu et al., 2011). We found some initiator elements (INRNTPSADB, IBOXCORE, PRECONSCRHSP70A and TBOXATGAPB), which are conservative light-responsive motifs, situating not only in 5'-upstream region of the *Myb* genes, but also in *Myc* and *WD40*. The HDZIP2ATATHB2 is a binding site of the *A. thaliana* homeobox gene (ATHB-2), which also is exposed under light control (Ohgishi et al., 2001). The *GhTT2-D3* and *GhMYB10-A1* genes contain this motif in the 5'-upstream promoter region.

Among considering gene copies the hormone- and pathogen-responsible motifs were revealed. For instance, we found ATHB6COREAT and GATABOX motifs; the 5'-upstream sequence of the *GhTT2-D3, GhTT8-A1* and *GhTTG-D1* genes have GCCCORE (5'- GCCGCC -3') motif, which is known to be involved in regulation of jasmonate responses (Himmelbach, 2002).

It is interesting to note that promoter regions of *MBW* genes include many Myb recognizing sites (AGMOTIFNTMYB2, MYB1AT, MYBATRD22, MYBCORE, MYBST1) (Hudson and Quail, 2003). MYB2CONSENSUSAT and MYBPZM motifs were identified only in the *GhTT2-D2, GhTT8-A2 GhTT8-D1*, *GhTT8-D2, GhTTG-D1*, *GhTTG-A2* and *GhTTG-D2* genes. Among all *TT2* and *MYB10* gene copies in *G. hirsutum* we identified MYB core and AC-rich regions (CAAT- and CCAAT-boxes), which are usually bound by Myb factors in *A. thaliana*.

The bHLH-Myc transcription factors was reported to be bind with the E/G-box motifs in *Arabidopsis* genes (Ishida et al., 2007). There are MYCCONSENSUSAT, MYCATERD1 and MYCATRD22 motifs in promoter region of *GhTT8-A1*, *GhTT8-D1* and *TTG1*.

Transcription factors of the ternary MBW complex through regulation of flavonoid biosynthesis genes are known to determine the tissue-specific accumulation of anthocyanins and PAs in various plant tissues (Ludwig et al., 1989; Reddy et al., 1995; Quattrocchio et al., 1998; Xu et al., 2013). In this study we observed the binding sites, required for tissue-specific gene expression, such as NTBBF1ARROLB, S2FSORPL21, SORLIP1AT and RHERPATEXPA7 (Lagrange et al., 1997; Jiao et al., 2005).

**3.4. Evolutionary analysis**

We calculated the ratio of the number of nonsynonymous substitutions per nonsynonymous sites (*Ka*) to the number of synonymous substitutions per synonymous sites (*Ks*) for the annotated *R2R3-Myb*, *bHLH-Myc* and *WD40* cotton genes. The obtained values are listed in Table 2. The meanings of *Ka* and *Ks* for *R2R3-Myb* ranged between 0.129–0.375 and 0.497–0.975, for *bHLH-Myc* – 0.127–0.157 and 0.174–0.203, for *WD40* – 0.037–0.052 and 0.360–0.446, respectively. *Ka/Ks* value reaches 0.413 for *R2R3-Myb,* 0.674 for *bHLH-Myc* and 0.104 in case the *WD40* gene family.

Genetic relationship between 24 identified *R2R3-Myb* genes of *Gossypium* in comparison with the five *A. thaliana R2R3-Myb* genes involved in PA biosynthesis (*AtMYB113* (GenBank: AT1G66370), *AtMYB114* GenBank: (AT1G66380), *AtPAP1* (GenBank: AT1G56650), *AtPAP2* (GenBank: AT1G66390), *AtTT2* (GenBank: AJ299452) was established using full-length nucleotide coding sequence and the UPGMA phylogenetic analysis algorithm. The *AtMYB115* gene (GenBank: AT5G40360) was used as outgroup. The results of the analysis demonstrate existence of several clades (Figure 4). The cotton *TT2* genes show similarity to the *A. thaliana R2R3-Myb* family genes. The sequences of the *AtMYB114*, *AtPAP2*, *AtPAP1* and *AtMYB113* genes formed an independent paraphyletic group. The cotton *MYB10* genes clustered jointly into one common branch.

In accordance with the phylogenetic analysis, two duplications of the *TT2* paralogous copies occurred about 29.98 and 27.46 million years ago (MYA). Divergence of the homoeologous copies in the A- and D-genomes was taking place during the period from 5.16 to 3.28 MYA. The *MYB10* cotton genes constitute the isolative group, which separated from the *TT2* cotton genes about 82.00 MYA (Figure 4).

The twelve identified *Gossypium bHLH* genes and the four in *A. thaliana* (*AtTT8* (GenBank: AT4G09820), *AtGL3* (GenBank: AT5G41315), *AtEGL3* (GenBank: AT1G63650) and *AtMYC2* (GenBank: AT1G32640)) genes were also analyzed. In accordance with phylogenetic analysis, *AtTT8* is the evolutionarily closest to the *Gossypium TT8* genes and form together a common clade (Figure 5). Considering results of divergence genes time, the duplication of the *TT8* cotton genes dated about 32.74 MYA with formation of two subclades. Division of the A- and D-genomes took place about 2.05–3.20 MYA.

**Table 2.** Number of *Ka*, *Ks* and *Ka/Ks* ratio for predicted cotton gene copies related to the *R2R3-Myb*, *bHLH-Myc* and *WD40* families.

| ***R2R3-Myb*** | ***Ka*** | ***Ks*** | ***Ka/Ks*** |
|---|---|---|---|
| *GhTT2-A1* | 0.129 | 0.501 | 0.258 |
| *GhTT2-D1* | 0.213 | 0.513 | 0.415 |
| *GhTT2-A2* | 0.233 | 0.546 | 0.427 |
| *GhTT2-D2* | 0.228 | 0.547 | 0.416 |
| *GhTT2-A3* | 0.244 | 0.548 | 0.445 |
| *GhTT2-D3* | 0.248 | 0.547 | 0.453 |
| *GrTT2-D1* | 0.215 | 0.497 | 0.433 |
| *GrTT2-D2* | 0.231 | 0.513 | 0.451 |
| *GrTT2-D3* | 0,247 | 0,520 | 0,475 |
| *GaTT2-A1* | 0.211 | 0.498 | 0.423 |
| *GaTT2-A2* | 0.230 | 0.534 | 0.430 |
| *GaTT2-A3* | 0.242 | 0.543 | 0.445 |
| *GbTT2-A1* | 0.211 | 0.501 | 0.421 |
| *GbTT2-D1* | 0.212 | 0.500 | 0.424 |
| *GbTT2-A2* | 0.233 | 0.546 | 0.427 |
| *GbTT2-D2* | 0.228 | 0.520 | 0.437 |
| *GbTT2-A3* | 0.244 | 0.548 | 0.445 |
| *GbTT2-D3* | 0.248 | 0.561 | 0.442 |
| *GbMYB10-A1* | 0.375 | 0.975 | 0.385 |
| *GbMYB10-D1* | 0.369 | 0.955 | 0.386 |
| *GrMYB10-D1* | 0.375 | 0.919 | 0.408 |
| *GhMYB10-A1* | 0.375 | 0.975 | 0.385 |
| *GhMYB10-D1* | 0.370 | 0.956 | 0.387 |
| *GaMYB10-A1* | 0.373 | 0.971 | 0.384 |
| **Average meaning** | 0.262 | 0.635 | 0.413 |
| ***bHLH-Myc*** | ***Ka*** | ***Ks*** | ***Ka/Ks*** |
| *GaTT8-A1* | 0.127 | 0.201 | 0.634 |
| *GaTT8-A2* | 0.131 | 0.192 | 0.683 |
| *GbTT8-D1* | 0.129 | 0.202 | 0.639 |
| *GbTT8-A1* | 0.157 | 0.198 | 0.792 |
| *GbTT8-A2* | 0.130 | 0.195 | 0.670 |
| *GbTT8-D2* | 0.131 | 0.201 | 0.651 |
| *GhTT8-D1* | 0.129 | 0.202 | 0.640 |
| *GhTT8-A1* | 0.129 | 0.174 | 0.741 |
| *GhTT8-A2* | 0.131 | 0.194 | 0.674 |
| *GhTT8-D2* | 0.130 | 0.203 | 0.641 |
| *GrTT8-D1* | 0.130 | 0.198 | 0.659 |
| *GrTT8-D2* | 0.131 | 0.196 | 0.667 |
| **Average meaning** | 0.132 | 0.196 | 0.674 |
| ***WD40*** | ***Ka*** | ***Ks*** | ***Ka/Ks*** |
| *GrTTG-D1* | 0.040 | 0.363 | 0.110 |
| *GrTTG-D3* | 0.043 | 0.433 | 0.099 |
| *GhTTG-D1* | 0.040 | 0.360 | 0.111 |
| *GhTTG-D3* | 0.043 | 0.433 | 0.099 |
| *GhTTG-A1* | 0.037 | 0.386 | 0.096 |
| *GhTTG-A3* | 0.043 | 0.433 | 0.099 |
| *GbTTG-D1* | 0.042 | 0.360 | 0.117 |
| *GbTTG-A1* | 0.037 | 0.386 | 0.095 |
| *GbTTG-A3* | 0.043 | 0.433 | 0.099 |
| *GaTTG-A1* | 0.037 | 0.385 | 0.097 |
| *GaTTG-A3* | 0.052 | 0.446 | 0.117 |
| **Average meaning** | 0.042 | 0.402 | 0.104 |

Phylogenetic analysis of the eleven *WD40* genes of *Gossypium* in comparison with the *AtTTG1* (GenBank: AT5G24520) and *AtMSI3* (GenBank: AT4G35050) gene sequences of *Arabidopsis* was also done. The cotton *TTG1* and *TTG3* genes were demonstrated to be evolutionarily close to the *AtTTG1* gene sequence (Figure 6). The cotton *TTG1* and *TTG3* homologous gene subclades split up about 45.19 MYA. The homoeologous genes in the A- and D-genomes diverged 2.34–3.92 MYA.

### 3.5. Prediction of 3D structures for R2R3-Myb, bHLH-Myc and WD40 proteins

The 6KKS template, deposited in the RSCB PDB, was used to predict 3D structures of all revealed R2R3-Myb proteins in the *G. hirsutum* genome. It corresponds to the *AtMYB66* gene product (R2R3-Myb family; GenBank: AT5G14750) in *A. thaliana* and has responsibility for regulation of root hair pattern formation (Lee and Schiefelbein, 1999).

Applied template covers R2- and R3-Myb encoding regions of 100 amino acids in length. Amino acid sequences identity to the template for all identified genes arises about 60% percent (Figure 7; Supplementary file 6). Total root-mean-square deviation (RMSD) value of atomic positions between 3D structures of AtTT2 protein sequence with the 6KKS template compiles 0.06 Å; the average RMSD score among all structures constitutes 0.08 Å, that point on the high structural similarity of coding sequences. Amino acid sequences similarity is the highest between homoeologous copies (more than 96%), while among all revealed copies with AtTT2 it achieves more than 80%. Sequence identity between GhTT2 and GhMYB10 is 86.5% in average (Supplementary file 7).

The Myc2, Myc3 and Myc4 TFs correspond to the bHLH class proteins and play an important role as mediators in jasmonate signaling reactions in *A. thaliana* (Zhang et al., 2015; Lian et al., 2017). The Myc2 and Myc3 coding domains were observed among regulatory genes in the *G. hirsutum* genome, so that to predict 3D protein models of related sequences, the 4RS9 (relates to the bHLH-Myc_N superfamily, Myc3 TF, GenBank: AT5G46760) and the 5GNJ (the HLH superfamily, Myc2 TF, GenBank: AT1G32640) templates were chosen (Zhang et al., 2015; Lian et al., 2017).

The 5GNJ template covers the conservative Myc2 coding region and constitutes approximately 50% identity with amino acid sequences of all detected genes in *G. hirsutum* (Supplementary file 6). The predicted conformation of the Myc2 TF expectedly folds into helix-loop-helix structure (Figure 8). 3D structures of Myc3 protein among all considered *GhTT8* copies in the *G. hirsutum* genome compile the amino acid sequence identity with the 4RS9 template in range 30%–40%. The identity percent between amino acid sequence of the *AtTT8* gene product and the *GhTT8* gene copies achieves 63.9% in average (Supplementary file 7).

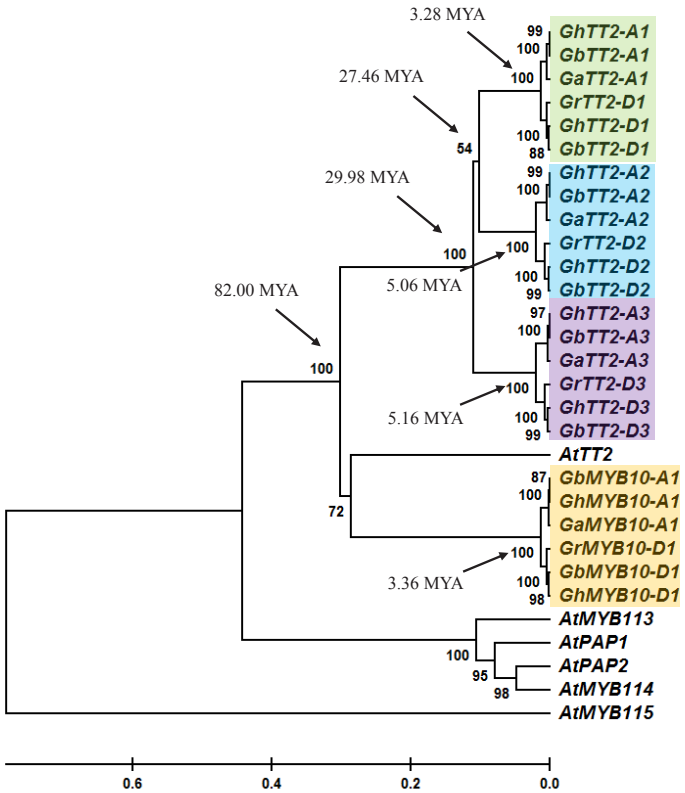



**Figure 4.** The phylogenetic relationship between *Gossypium TT2, MYB10* and *A. thaliana TT2* genes. Cluster analysis of the nucleotide sequences was conducted using the UPGMA algorithm with 1000 bootstrap trials**.** GenBank accession numbers for outgroup sequences: *AtPAP1* (AT1G56650), *AtPAP2* (AT1G66390), *AtMYB113* (AT1G66370), *AtMYB114* (AT1G66380), *AtMYB115* (AT5G40360). Cotton sequences identifiers are given in Table 1.

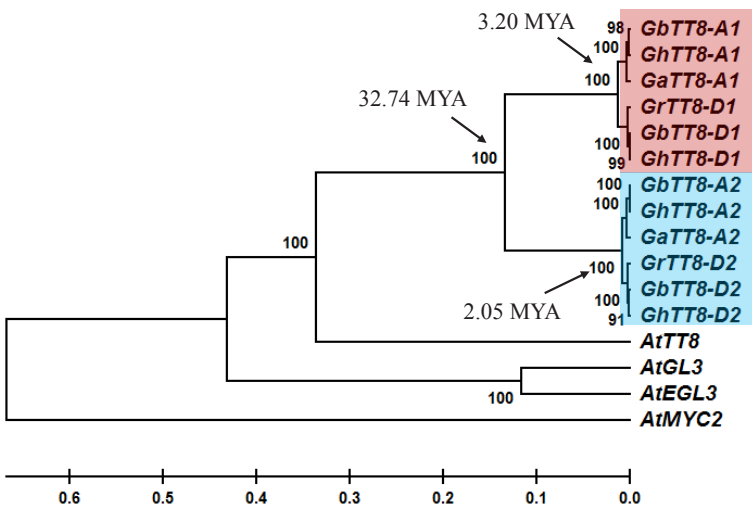



**Figure 5.** Phylogenetic relationship between *Gossypium* and *A. thaliana TT8* genes. Cluster analysis of the nucleotide sequences was conducted using the UPGMA algorithm with 1000 bootstrap trials**.** GenBank accession numbers for outgroup sequences: *AtTT8* (AJ277509), *AtGL3* (AT5G41315), *AtEGL3* (AT1G63650), *AtMYC2* (AT1G32640). Cotton sequences identifiers are given in Table 1.

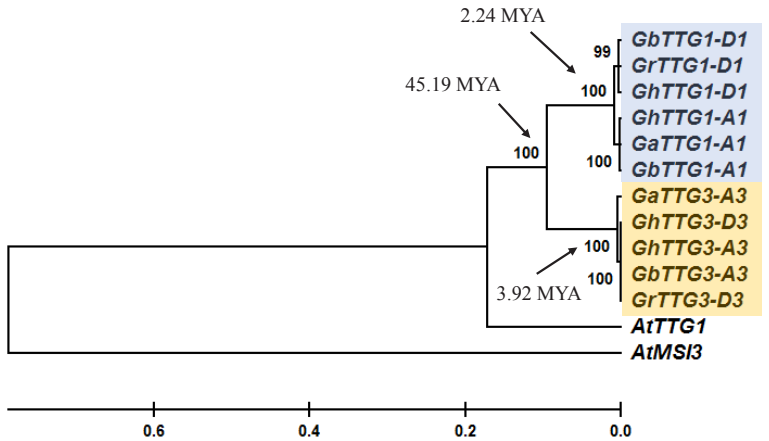



**Figure 6.** Phylogenetic relationship between cotton *TTG1*, *TTG3* and *A. thaliana AtTTG1* genes. The cluster analysis of nucleotide sequences was performed using the UPGMA algorithm with 1000 bootstrap trials**.** GenBank accession numbers for outgroup sequences: *AtTTG1* (AT5G24520), *AtMSI3* (AT4G35050). Cotton sequences identifiers are given in Table 1.

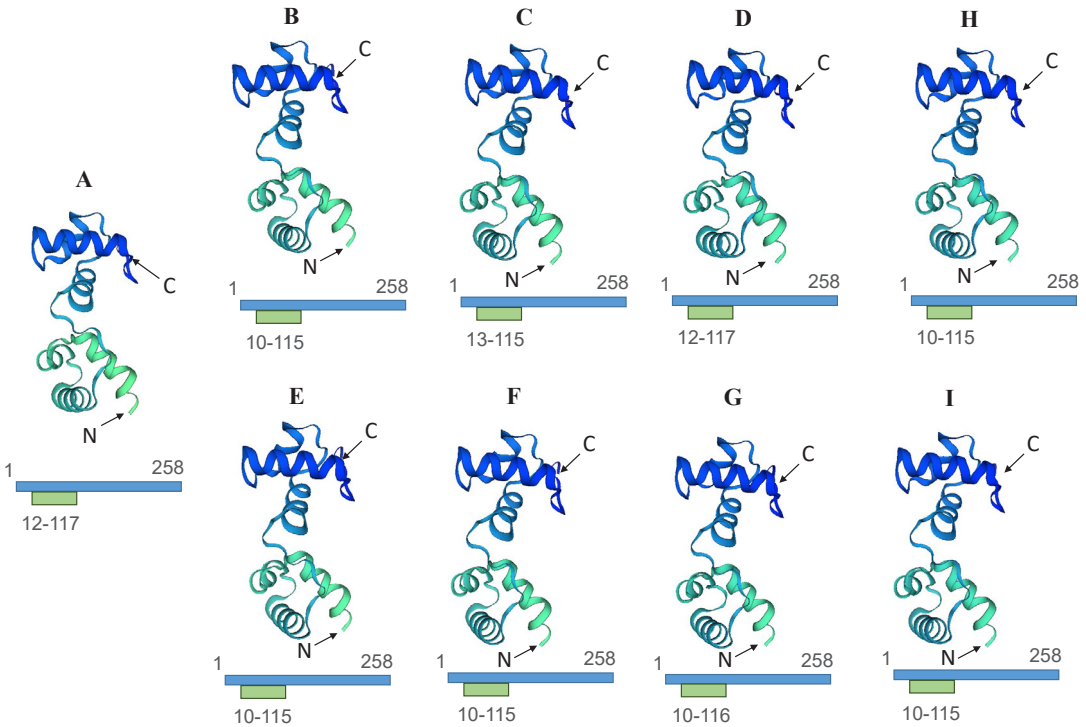



**Figure 7.** Predicted 3D-structures of *A. thaliana* and *G. hirsutum* gene products belonging to *R2R3-Myb* family (6KKS templates from PDB). A – AtTT2, B – GhTT2-A1, C – GhTT2-A2, D – GhTT2-A3, E – GhTT2-D1, F – GhTT2-D2, G – GhTT2-D3, H – GhMYB10-A1, I – GhMYB10-D1. Blue line shows the length of investigating protein sequence as well as green its coverage by the applied template.

The FASTA format of the AtTTG amino acid sequence was used for BLASTp search in NCBI to retrieve the 2HES template from the RSCB PDB database. The template relates with Cia1 protein in yeast (*Saccharomyces cerevisiae* Meyen ex E.C.Hansen), which takes part in cytosolic iron-sulfur (Fe/S) protein assembly (CIA) process in eukaryotes (Srinivasan et al., 2007). Template covers almost all WD40 coding region except for around the first C-end 15 amino

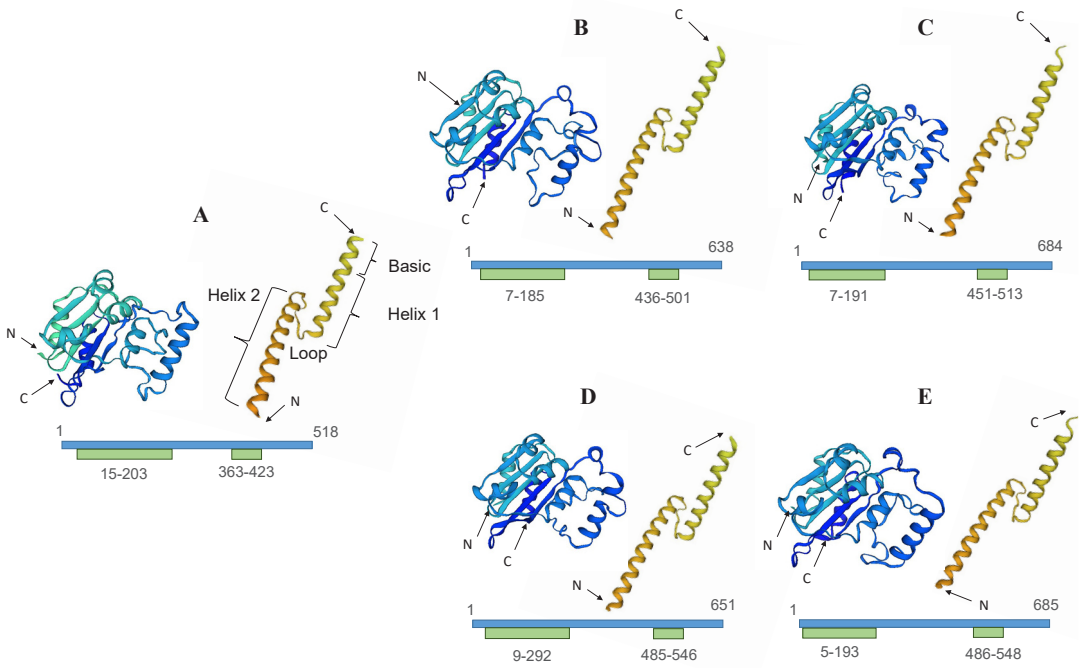



**Figure 8.** Predicted 3D-structures of *A. thaliana* and *G. hirsutum* gene products related to bHLH-Myc family (4RS9 and 5GNJ templates from PDB). A – AtTT8, B – GhTT8-A1, C – GhTT8-A2, D – GhTT8-D1, E – GhTT8-D2. Blue line shows the length of investigating protein sequence as well as green its coverage by the applied templates.

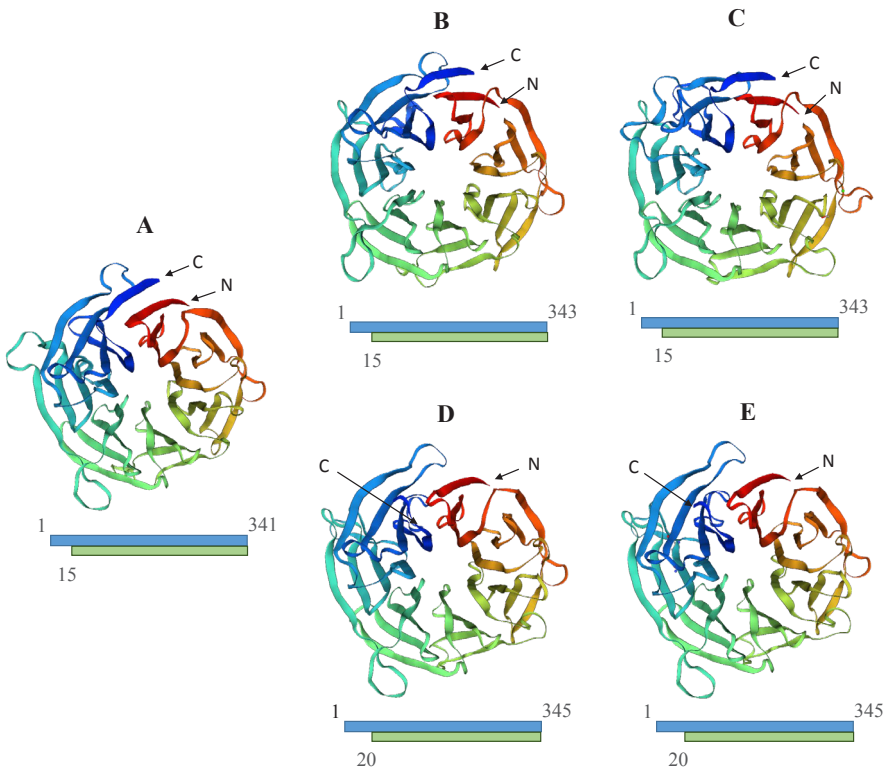



**Figure 9.** Predicted 3D-structures of *A. thaliana* and *G. hirsutum* gene products belonging to WD40 family. A – AtTTG1, B – GhTTG1-A1, C – GhTTG1-D1, D – GhTTG3-A3, E – GhTTG3-D3. Blue line shows the length of investigating protein sequence as well as green its coverage by the applied template.

acids (Figure 9; Supplementary file 6). The 3D structure of crystallized Cia1 protein represents β-propeller with seven WD40 repeats as blades. The predicted *AtTTG1* and the *GhTTG1* gene copies products fold into analogical conformations. The percent of amino acid identity between AtTTG1 and GhTTG1/GhTTG3 compiles 91% in average, while between GhTTG1 and GhTTG3 gene products it arises more than 93% (Supplementary file 7).

## 4.Discussion

Naturally colored cotton is an important agriculture crop for the textile industry due to its hypoallergenic and eco-friendly properties (Semizer-Cuming et al., 2015; Kochetkova, 2018; Abbasi, 2019; Günaydina et al., 2020). PA accumulation into cell vacuole during the flavonoids biosynthesis pathway provides various brown shades of cotton fiber (Xiao et al., 2007; Li et al., 2013; Feng et al., 2014; Li et al., 2020; Peng et al., 2020). Structural genes of the PAs biosynthesis pathway are controlled by TFs related to the R2R3-Myb, bHLH-Myc and WD40 families, that form together the ternary MBW complex (Koes et al., 2005; Li 2014; Xu, et al., 2015). However, to date only two *R2R3-Myb* genes (*GhTT2-A1*, also known as *GhMYB36* and *GhMYB10*), one *bHLH-Myc* gene (*GhTT8-D1*, known as *GhbHLH130D*) and two *WD40* genes in the *G.hirsutum* (*GhTTG1* and *GhTTG3*) genome were established to be involved in PAs accumulation and responsible for anthocyanin biosynthesis regulation in cotton (Humphries et al., 2005; Hinchliffe et al., 2016; Lu et al., 2017; Yan et al., 2018; Mikhailova et al., 2019). Another two genes related the *WD40* family, *GhTTG2* and *GhTTG4,* were reported to consist separate group and be closer to ATAN11-A protein (Light-Regulated WD1 (LWD1) protein controls circadian period length and photoperiodic flowering; GenBank: AT1G12910) in *A. thaliana* (Humphries et al., 2005). For the first time this work represents detailed in silico study of an unidentified earlier 24 *R2R3-Myb*, 12 *bHLH-Myc* and 11 *WD40* candidate gene copies with their divergence time period in the common ancient of the genus *Gossypium* as well as features of the structural and functional organization of found copies.

### 4.1 In silico structural and functional analysis

***R2R3-Myb***. The *Myb* genes represent the biggest family in higher plants, however only *R2R3-Myb* gene products play an important role in the secondary metabolism and the PA biosynthesis (Nesi, 2001). There are several types of the *Myb*-binding domains, which are classified according to the number of amino acid repeats (R), namely one 1R, two 2R (R2R3-Myb), three 3R (1R2R3R-Myb) and four 4R (the mixed R1/R2 type). Every repeat consists of about 53 amino acid residues and folds into helix-turn-helix structure. The third helix fragment has binding sites for interaction with a major groove of DNA.

We showed that the all identified *R2R3-Myb* genes in the genus *Gossypium* contain the R2 and R3 Myb DNA-binding domains (Pfam: PF00249) with 44-47 amino acid residues (Supplementary file 1). The conserved functionally significant regions in the tertiary protein structure of predicted 3D models have similar conformation (Figure 7). R2/R3-Myb domains contain 46 critical amino acids in *A. thaliana* that comparable with the length of R2 and R3 motifs in all identified genes.

In R2R3-Myb repeats tryptophan residues are known to play a crucial role in hydrophobic core formation for proper 3D helix-turn-helix structure folding to effectively recognize DNA specific sequences (Stracke et al., 2001). These critical residues locate at positions: Trp19, Trp39, Trp59, Trp91, Trp110 in R2-/R3-Myb-repeats under multiple amino acid sequences alignment of AtTT2 (Supplementary file 1). The R3 domain of the R2R3-Myb TF contains region with 20 amino acids residues from glutamic acid to arginine, which is essential for bHLH protein interaction (Zimmermann et al., 2004; Sun, et al., 2016). In the present study all identified R2R3-Myb proteins kept critical tryptophan residues and conserved motifs for bHLH binding unchangeable.

TFs, belonging to the MYB family, is a part of MBW complex determining the specific promoter activation at both different proteins transcriptional and posttranscriptional levels (Cao et al., 2020). Consequently, TT2 regulates PA biosynthesis, while PAP4 controls anthocyanins accumulation (Nesi et al., 2001; Gonzalez et al., 2008; Heppel et al., 2013; Li, 2014). It was reported that a single glycine to arginine exchange at position 39 in the R2 domain along with simultaneously exchange of a four amino acid in the R3-repeat convert the target specificity between TT2 and PAP4 activation (Heppel et al., 2013). Multiple alignment of the R2R3-Myb amino acid sequences revealed several single point mutations in amino acid sequences of GhTT2-A2, GhTT2-D2, GhTT2-A3 and GhTT2-D3 proteins. Glycine at 39 position in R2 motif was observed with exception of GhTT2-A2 and GhTT2-D2 (Supplementary file 1). Moreover, we revealed replacements Asp4Arg, Gly41Cys, Lys42Arg in GhTT2-A2 as well as Asn4Lys, Gly41Cys, Lys42Arg for GhTT2-D2; Ile15Leu, Pro32Ser for GhTT2-A3 and GhTT2-D3 in conservative R2 motif.

These amino acid replacements do not relate to the critical residues listed above, so the protein function has probably kept. Despite of some differences in a primary amino acid sequences, a tertiary one kept conservative (Figures 7). Moreover, the mutated residues are not critical in accordance with the HOPE server prediction; therefore, we assumed that these genes maintained in a functional state.

***bHLH-Myc***. The bHLH-type Myc TFs were predicted to form the amphipathic two α-helices connected by an

intervening loop (Murre et al., 1994). The helix motif of the longest N-terminal HLH basic region (Myc2) has basic amino acids residues that serve as a crucial site for DNA-binding; the Myc3 C-terminal bHLH motif is responsible for dimerization (Ferré-D'Amaré et al., 1994). The conservative bHLH-Myc_N and HLH DNA-binding domains are equal to 175 and 53 amino acids in AtTT8 and GhTT8. (Supplementary file 3). The high structural similarities among all predicted 3D models of GhTT8 with AtTT8 were proven based on low RMSD values (Figure 8). However, there are single point mutations of GhTT8 amino acid sequences that are needed to be described. The arginine to serine substitution at position 13 (Arg13Ser) in the N-terminal HLH basic region of GhTT8-A2 protein sequence was observed. There is a difference in charge and in size between the wild-type and mutant amino acid, so that the positive charge of the wild-type residue will be lost, this can cause loss of noncovalent interactions with other molecules or residues. Moreover, the wild-type and mutant amino acids differ in size; the mutant residue is smaller, then initial one, that may lead to disappearance of interactions. Due to the diaminocarbene, functional group of arginine introduces more hydrophobic residue at this position, then the oxyaminocarbene of Serine, loss of hydrogen bonds or disturb correct folding may be occurred. Moreover, Arg13 locates in functionally significant region, there Arg10 and Glu13 with Arg11 and Arg12 residues interact with nucleotide bases and DNA backbone, respectively. Considering this observation, substitution may disrupt the Myc 2 domain and leads to incorrect folding of the whole protein.

Moreover, the Val52Ile and Arg53Lys replacements of the critical amino acids were fixed among all uncovered copies. Probably, pointed differences do not have critical influence on the protein function. This statement is supported by earlier study, where the *GhTT8-D1* gene (*GhbHLH130D*) was demonstrated to take part in PA biosynthesis (Yan et al., 2018).

***WD40***. The *WD40* encoding domain comprises of about 40 amino acids. N-terminus starts with glycine-histidine (GH) dipeptide, as well as tryptophan-aspartate (WD) dipeptide is usually at the C–terminal end. WD40 proteins consist of at least seven repeated motifs and curl up into four stranded β-sheets or blades. Further the blades aggregate together with formation of a β-propeller structure (Villanueva et al., 2016).

Based on multiple alignment of four WD repeats, the conservative motif without critical amino acid substitutions were revealed among all identified copies (Supplementary file 5). Predicted 3D models of WD40 proteins fold into 7-bladed β-propeller conformation in both cotton and *Arabidopsis* (Figure 9). The lowest RMSD scores between 3D structures of AtTTG and cotton TTG1, TTG3 proteins indicate the high mutual similarity. Due to the high homology of gene product sequences, preservation of conserved domains and great structural similarity detected gene copies possibly involved in the phenylpropanoid biosynthesis pathway.

**4.2. Structure of the MBW genes promoters**

As a part of our study, we considered *cis*-acting regulatory DNA elements in the region approximately 1000 bp upstream to the ATG start translation codon. The vast majority among found regulatory motifs are light- and phytohormones inducible (Figure 3). Besides these motifs, we observed strongly conserved CAAT and CCAAT motifs, TATA- and GATA-boxes as well as transcription the initiation sites among the *R2R3-Myb, bHLH-Myc*, and *WD40* genes. It is believed that the combination of such core promoters is characteristically for highly expressed genes (Sawant et al., 1999; Roy and Singer, 2015; Haberle and Stark, 2018).

As was reported earlier, the *TT8* gene transcription is activated by binding of the TT2 TF with the specific sites in DNA promoter sequence, inducing the positive expression of PAs structural genes as mutual synergetic effect (Baudry et al., 2004; Baudry et al., 2006; Xu et al., 2013). Namely, plant R2R3-Myb TFs recognize AC-rich and Myb-core *cis*-acting regulatory elements (Prouse and Campbell, 2012; Kelemen et al., 2015).

The identified *TT2, MYB10* and *TTG* gene copies, except for *GhTT8-A2* and *GhTT8-D2*, have numerous binding sites for the Myb and Myc factors, therefore the majority of the identified gene copies can be involved in transcription activating processes (Figure 3).

Here, we revealed the raw of *cis*-acting regulatory DNA-elements define tissue-specific gene expression depends on biotic (phytohormones, sucrose, pathogen infections) and abiotic factors (extremely temperature, highlight induction). Since promoter regions of all investigated gene families characterize a various pattern of *cis*-acting elements, the identified copies possibly diverged during their separation from common donor species.

MBW complex is known to regulate many vital processes in higher plants (Johnson et al., 2002; Wang et al., 2004; Gonzalez et al., 2009; Appelhagen et al., 2011; Qi et al., 2011). Examples of preservation and functional specialization of duplicated copies are typical for the flavonoid biosynthesis genes in a higher plants (Strygina and Khlestkina, 2019a, 2019b; Strygina et al., 2019; Vikhorev et al., 2019). Taking into account this fact, including observed the high variability of promoter region, we assumed that identified duplicated genes control not only PA biosynthesis pathway but also potentially acquire other functions because of subfunctionalization.

**4.3 Evolutionary analysis of the *R2R3-Myb, bHLH-Myc* and *WD40* genes**

Early hexaploidization among eudicot ancestors took place about 115-146 MYA and followed by division of the

genus *Gossypium* into eight groups with genomes from A to G and K (Strygina et al., 2020). The D-genome diploids were included in the New World clade and diverged from the A-genome diploids (the African-Asian clade) around 5–10 MYA. Duplications in *G. herbaceum* ($A_1$) and *G. arboreum* ($A_2$), native to Africa, and *G. raimondii* ($D_2$), origin in Mexica, ancestors occurred 13-20 MYA during Miocene period. The ancient progenitor similar to *G. herbaceum* ($A_1$), *G. arboretum* ($A_2$) and another progenitor similar to *G. raimondii* ($D_2$) hybridized with formation of allotetraploid species with the AADD genome due to transoceanic resettlement approximately 1.5 MYA (Hu et al., 2019). Later these species spread from the North and the South America to the West part of the Pacific Ocean with sharing into seven known species: *G. tomentosum* Nutt. ex Seem $(AD)_3$, *G. mustelinum* Miers ex G. Watt $(AD)_4$, *G. darwinii* G. Watt $(AD)_5$, *G. hirsutum, G. barbadense, G. ekmanianum* Wittmack $(AD)_6$ and *G. stephensii* J. Gallagher et al. $(AD)_7$ (Wendel et al., 2009; Strygina et al., 2020).

To establish evolutionary relationships between Gossypium and *A. thaliana* orthologue genes, we constructed phylogenetic tree with regarding to the results of full size nucleotide sequences alignment of the *R2R3-Myb*, *bHLH-Myc* and *WD40* family genes. Despite on the fact that the *TT2* and *MYB10* genes coding proteins with similar functions, they correspond to the different types of *TT2*-like *MYB* genes. Thus, the *TT2* and *MYB10* genes form two separate subclades. The *TT2* nucleotide sequences separated into the biggest subclade, consisting of the three groups. Moreover, every group can be divided into several subgroups in obedient to homoeologous copies in the A- and D-genomes with a high bootstrap values from 86 to 100.

According to the phylogenetic analysis, paralogous and homoeologous cotton *TT8* genes, related to the *bHLH-Myc* family, formed two clades approximately 32.74 MYA in the common ancestor. The *TTG1* and *TTG3* genes (*WD40* family) divided out into distinct evolutionary group.

Two duplication events between the *TT2* gene copies occurred much later, in comparison with genes belonged to the *bHLH-Myc* and *WD40* families. Divergence among the *TT2* (27.46 and 29.98 MYA), *TT8* (32.74 MYA) and *TTG* genes (45.19 MYA) in the genus *Gossypium* took place in one common ancestor before formation of the three *Gossypium* clades (7.5–10 MYA): the New World (the D-genome species), the Australian (the C-, G-, and K-genomes) and the African-Asian (the A-, B-, E- and F-genomes) and before polyploidization (Wendel et al., 2009) (Figures 4–6). Homoeologous genes in the A- and D-genomes separated from the common donor about 3.28–5.16 MYA for *TT2*, 2.05–3.20 MYA for *TT8* and 2.24–3.92 MYA for *WD40*. Therefore, duplication events after polyploidization did not take place.

In accordance with obtained data, the *G. raimondii* genome has been exposed by a multiple rearrangement events. We put forward several reasons to explain this observation. Possible mistakes may be associated with the low quality of genome assembling as well as too high evolution rate could lead to inaccurate data concerning nucleotide sequences and its chromosome localization.

To determine the rate of evolution, we calculate the *Ka/Ks* ratio. As known, the *Ka* displays the number of occurred changes normalized to an all possible, whereas *Ks* does not regard any functional and phenotypic improvements. *Ks* shows the background rate of "silent evolution" (Hu and Banzhaf, 2008). All obtained *Ka/Ks* values satisfy *Ka/Ks* < 1 condition for considered gene families (Table 2). Therefore, the rate of protein fixing is slow and these genes exist under "negative selection". The lowest *Ka/Ks* ratio was observed for the *WD40* genes (0.104). These genes are predominantly supported by stabilizing selection, therefore are excluded from candidate responsible for phenotypic variability.

## 5. Conclusion

In silico analysis is intended to be one of the most essential stage in row investigations devoted to plant breeding, genome editing and molecular genetics. It is important to reveal candidate genes and predict their potential functions to improve the quality and credibility of the future genetics studies.

We characterized highly homologous regulatory genes *R2R3-Myb*, *bHLH-Myc* and *WD40* in the *Gossypium* genomes, including diploid and allotetraploid cotton species. The *TTG3* gene absents in the D-genome of *G. barbadence*, while *G. raimondii* genes were subjected by significant rearrangements during evolution process. According to the phylogenetic analysis, all duplication events of *R2R3-Mybs*, *bHLH-Mycs* and *WD40s* occurred in the common diploid ancestor of the genus *Gossypium* before formation of the allotetraploid species. The *GhTTG1/GhTTG3* (*WD40*) genes were demonstrated to be the most conservative and predominantly supported by selection, while *GhTT2/GhMYB10* (*R2R3-Myb*) and *GhTT8* (*bHLH-Myc*) are more variable. Numerous light, extreme temperature and phytohormones sensitive *cis*-acting regulatory DNA elements, as well as specific binding motifs in promoter regions of regulatory genes, were revealed. Variable pattern of the *cis*-regulatory elements between homoeologous genes point on their possible specialization. Under multiple alignment of HLH domain sequences it turned out that GhTT8-A2 carries Arg13Ser replacement leading to loss-of-function mutation in a highly conserved position in the Myc2 domain. Therefore, we assumed that MBW genes among considering families, except for *GhTT8-A2*, can be associated with the PA biosynthesis pathway.

To summarize, we have revealed 24 *R2R3-Myb,* 12 *bHLH-Myc* and 11 *WD40* gene copies in cotton genome. We singled out more suitable among them for further allelic diversity screening among genotypes of cotton with contrast colored fiber by a comprehensive evolutionary and functionally analysis of identified genes. Obtained results will be necessary for the in-depth genetic studies and can be potentially appropriated for creation of naturally colored cotton cultivars using marker-assisted selection or gene editing.

**Acknowledgment**
The work was carried out with the support of the Ministry of Science and Higher Education of the Russian Federation in the framework of Agreement No. 075-15-2020-911 of 16.11.2020 on providing a grant in the form of subsidies from the federal budget for the implementation of state support for the creation and development of the world-class scientific center "Agrotechnologies for the future".

## References

Abbasi SM (2019). The future of organic colored cotton. Vlakna a Textil 26 (4): 13–18.

Abdurakhmonov IY, Buriev ZT, Logan-Young CJ, Abdukarimov A, Pepper AE (2010). Duplication, divergence and persistence in the phytochrome photoreceptor gene family of cottons (*Gossypium* Spp.). BMC Plant Biology 10 (1): 119. doi: 10.1186/1471-2229-10-119

Appelhagen I, Lu GH, Huep G, Schmelzer E, Weisshaar B et al. (2011). TRANSPARENT TESTA1 interacts with R2R3-MYB factors and affects early and late steps of flavonoid biosynthesis in the endothelium of *Arabidopsis thaliana* seeds. The Plant Journal 67 (3): 406–419. doi: 10.1111/j.1365-313X.2011.04603.x

Atchley WR, Terhalle W, Dress A (1999). positional dependence, cliques and predictive motifs in the BHLH protein domain. Journal of Molecular Evolution 48 (5): 501–516. doi: 10.1007/PL00006494

Baker SS, Wilhelm KS, Thomashow MF (1994). The 5'-region of *Arabidopsis thaliana* Cor15a Has *cis*-acting elements that confer cold-, drought- and ABA-regulated gene expression. Plant Molecular Biology 24 (5): 701–713. doi: 10.1007/BF00029852

Baudry A, Caboche M, Lepiniec L (2006). TT8 controls its own expression in a feedback regulation involving TTG1 and homologous MYB and BHLH factors, allowing a strong and cell-specific accumulation of flavonoids in *Arabidopsis thaliana*. The Plant Journal 46 (5): 768–779. doi: 10.1111/j.1365-313X.2006.02733.x

Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B et al. (2004). *TT2*, *TT8*, and *TTG1* synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. The Plant Journal 39 (3): 366–380. doi: 10.1111/j.1365-313X.2004.02138.x

Bedon F, Grima-Pettenati J, Mackay J (2007). Conifer R2R3-MYB Transcription factors: sequence analyses and gene expression in wood-forming tissues of white spruce (*Picea glauca*). BMC Plant Biology 7 (1): 17. doi: 10.1186/1471-2229-7-17

Cao Y, Li K, Li Y, Zhao X, Wang L (2020). MYB transcription factors as regulators of secondary metabolism in plants. Biology 9 (3): 61. doi: 10.3390/biology9030061

Christie PJ, Alfenito MR, Walbot V (1994). Impact of low-temperature stress on general phenylpropanoid and anthocyanin pathways: enhancement of transcript abundance and anthocyanin pigmentation in Maize seedlings. Planta 194 (4): 541–549. doi: 10.1007/BF00714468

Corpet F (1988). Multiple Sequence alignment with hierarchical clustering. Nucleic Acids Research 16 (22): 10881–10890. doi: 10.1093/nar/16.22.10881

Debeaujon I, Peeters AJM, Leon-Kloosterziel KM, Koornneef M (2001). The *TRANSPARENT TESTA12* gene of Arabidopsis encodes a multidrug secondary transporter-like protein required for flavonoid sequestration in vacuoles of the seed coat endothelium. The Plant Cell 13 (4): 853. doi: 10.2307/3871345

Devic M, Guilleminot J, Debeaujon I, Bechtold N, Bensaude E et al. (1999). The *BANYULS* gene encodes a DFR-like protein and is a marker of early seed coat development. The Plant Journal 19 (4): 387–398. doi: 10.1046/j.1365-313X.1999.00529.x

Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C (2010). MYB transcription factors in Arabidopsis. Trends Plant Sci 15 (10): 573–581. doi: 10.1016/j.tplants.2010.06.005

Endrizzi JE, Taylor T (1968). Cytogenetic studies of N *Lc1 Yg2 R2* marker genes and chromosome deficiences in cotton. Genetical Research 12 (03): 295. doi: 10.1017/S0016672300011885

Falcone Ferreyra ML, Rius SP, Casati P (2012). Flavonoids: Biosynthesis, biological functions, and biotechnological applications. Frontiers in Plant Science 28 (3):222. doi: 10.3389/fpls.2012.00222

Felsenstein J (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39 (4): 783. doi: 10.2307/2408678

Feng H, Li Y, Wang S, Zhang L, Liu Y et al. (2014). Molecular analysis of proanthocyanidins related to pigmentation in brown cotton fibre (*Gossypium hirsutum* L.). Journal of Experimental Botany 65 (20): 5759–5769. doi: 10.1093/jxb/eru286

Feng H, Tian X, Liu Y, Li Y, Zhang X et al. (2013). Analysis of flavonoids and the flavonoid structural genes in brown fiber of upland cotton. Edited by Jinfa Zhang. PLoS ONE 8 (3): e58820. doi: 10.1371/journal.pone.0058820

Gao J, Shen L, Yuan J, Zheng H, Su Q et al. (2019). Functional analysis of *GhCHS*, *GhANR* and *GhLAR* in colored fiber formation of *Gossypium hirsutum* L. BMC Plant Biology 19 (1): 455. doi: 10.1186/s12870-019-2065-7

Gonzalez A, Zhao M, Leavitt JM, Lloyd AM (2008). Regulation of the Anthocyanin biosynthetic pathway by the TTG1/BHLH/Myb transcriptional complex in Arabidopsis seedlings. The Plant Journal 53 (5): 814–27. doi: 10.1111/j.1365-313X.2007.03373.x

Gonzalez A, Mendenhall J, Huo Y, Lloyd A (2009). TTG1 Complex MYBs, MYB5 and TT2, Control outer seed coat differentiation. Developmental Biology 325 (2): 412–21. doi: 10.1016/j.ydbio.2008.10.005

Günaydina GK, Palamutcu S, Soydan AS, Yavas A, Avinc O et al. (2020). Evaluation of fiber, yarn, and woven fabric properties of naturally colored and white Turkish organic cotton. The Journal of The Textile Institute 111 (10): 1436–53. doi: 10.1080/00405000.2019.1702611

Guo N, Cheng F, Wu J, Liu B, Zheng S et al. (2014). Anthocyanin biosynthetic genes in *Brassica rap*a. BMC Genomics 15 (1): 426. doi: 10.1186/1471-2164-15-426

Haberle V, Stark A (2018). Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation. Nature Reviews Molecular Cell Biology 19 (10): 621–37. doi: 10.1038/s41580-018-0028-8

Harlan SC (1932). The Genetics of Cotton. Journal of Genetics 25 (3): 261–70. doi: 10.1007/BF02984590

Heim MA (2003). The Basic Helix-Loop-Helix Transcription Factor Family in Plants: A Genome-Wide Study of Protein Structure and Functional Diversity. Molecular Biology and Evolution 20 (5): 735–47. doi: 10.1093/molbev/msg088

Heppel SC, Jaffé FW, Takos AM, Schellmann S, Rausch T et al. (2013). Identification of key amino acids for the evolution of promoter target specificity of anthocyanin and proanthocyanidin regulating MYB factors. Plant Molecular Biology 82 (4–5): 457–71. doi: 10.1007/s11103-013-0074-8

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999). Plant *cis*-acting regulatory DNA elements (PLACE) database. Nucleic Acids Res 27 (1): 297-300. doi: 10.1093/nar/27.1.297

Himmelbach A, Hoffmann T, Leube M, Höhener B, Grill E (2002). Homeodomain protein ATHB6 is a target of the protein phosphatase ABI1 and regulates hormone responses in Arabidopsis. EMBO J 21 (12): 3029-3038. doi: 10.1093/emboj/cdf316

Hinchliffe DJ, Condon BD, Thyssen G, Naoumkina M, Madison CA et al. (2016). The *GhTT2_A07* gene is linked to the brown colour and natural flame retardancy phenotypes of *Lc1* cotton (*Gossypium hirsutum* L.) fibres. J Exp Bot. 67 (18): 5461-5471. doi: 10.1093/jxb/erw312

Hu T, Banzhaf W (2008). Nonsynonymous to Synonymous Substitution Ratio ka/ks: Measurement for Rate of Evolution in Evolutionary Computation. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Dortmund, Germany. pp. 448–457. doi: 10.1007/978-3-540-87700-4_45

Hu Y, Chen J, Fang L, Zhang Z, Ma W et al. (2019). *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. Nat Genet. 51 (4): 739-748. doi: 10.1038/s41588-019-0371-5

Huang J, Guo Y, Sun Q, Zeng W, Li J et al. (2019). Genome-Wide Identification of R2R3-MYB Transcription Factors Regulating Secondary Cell Wall Thickening in Cotton Fiber Development. Plant Cell Physiol 60 (3): 687-701. doi: 10.1093/pcp/pcy238

Hudson ME, Quail PH (2003). Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. Plant Physiol 133 (4): 1605-1616. doi: 10.1104/pp.103.030437

Humphries JA, Walker AR, Timmis JN, Orford SJ (2005). Two WD-repeat genes from cotton are functional homologues of the *Arabidopsis thaliana* TRANSPARENT TESTA GLABRA1 (TTG1) gene. Plant Mol Biol. 57 (1): 67-81. doi: 10.1007/s11103-004-6768-1

Ishida T, Hattori S, Sano R, Inoue K, Shirano Y et al. (2007). Arabidopsis TRANSPARENT TESTA GLABRA2 is directly regulated by R2R3 MYB transcription factors and is involved in regulation of GLABRA2 transcription in epidermal differentiation. Plant Cell. 19 (8): 2531-2543. doi: 10.1105/tpc.107.052274

Jiao Y, Ma L, Strickland E, Deng XW (2005). Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and Arabidopsis. Plant Cell. 17 (12): 3239-3256. doi: 10.1105/tpc.105.035840

Johnson CS, Kolevski B, Smyth DR (2002). TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. Plant Cell. 14 (6): 1359-1375. doi: 10.1105/tpc.001404

Kelemen Z, Sebastian A, Xu W, Grain D, Salsac F et al. (2015). Analysis of the DNA-Binding Activities of the Arabidopsis R2R3-MYB Transcription Factor Family by One-Hybrid Experiments in Yeast. PLoS One. 10 (10): e0141044. doi: 10.1371/journal.pone.0141044

Kochetkova OV (2018). Formalization and analysis of technological processes of primary processing of cotton-raw maternal. Proceedings of Nizhnevolzhskiy Agrouniversity Complex: Science and Higher Vocational Education 3 (51): 291–300. doi: 10.32786/2071-9485-2018-01-291-300

Koes R, Verweij W, Quattrocchio F (2005). Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. Trends Plant Sci. 10 (5): 236-242. doi: 10.1016/j.tplants.2005.03.002

Kohel RJ (1985). Genetic analysis of fiber color variants in cotton 1. Crop Science 25 (5): 793–797. doi: 10.2135/cropsci1985.0011183X0025000500017x

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 35 (6): 1547-1549. doi: 10.1093/molbev/msy096

Kumar V, Suman U, Rubal, Yadav SK (2018). Flavonoid secondary metabolite: biosynthesis and role in growth and development in plants. In: Yadav SK, Kumar V, Singh SP (editors). Recent trends and techniques in plant metabolic engineering. Singapore, Republic of Singapore: Springer Singapore, pp. 19–45. doi: 10.1007/978-981-13-2251-8_2

Lagrange T, Gauvin S, Yeo HJ, Mache R (1997). S2F, a leaf-specific trans-acting factor, binds to a novel *cis*-acting element and differentially activates the *Rpl21* gene. The Plant Cell 9 (8): 1469-1479. doi: 10.2307/3870396

Lane HC, Schuster MF (1981). Condensed tannins of cotton leaves. Phytochemistry 20 (3): 425–427. doi: 10.1016/S0031-9422(00)84158-8

Lea US, Slimestad R, Smedvig P, Lillo C (2007). Nitrogen deficiency enhances expression of specific MYB and bHLH transcription factors and accumulation of end products in the flavonoid pathway. Planta 225 (5): 1245-1253. doi: 10.1007/s00425-006-0414-x

Lee MM, Schiefelbein J (1999). WEREWOLF, a MYB-Related protein in Arabidopsis, is a position-dependent regulator of epidermal cell patterning. Cell 99 (5): 473–483. doi: 10.1016/S0092-8674(00)81536-6

Li S (2014). Transcriptional control of flavonoid biosynthesis: fine-tuning of the MYB-bHLH-WD40 (MBW) complex. Plant Signal Behav 9 (1): e27522. doi: 10.4161/psb.27522

Li YJ, Zhang XY, Wang FX, Yang CL, Liu F et al. (2013). A comparative proteomic analysis provides insights into pigment biosynthesis in brown color fiber. Proteomics 78 (1): 374-388. doi: 10.1016/j.jprot.2012.10.005

Li Z, Qian SU, Mingqi XU, Jiaqi YOU, Khan AQ et al. (2020). Phenylpropanoid metabolism and pigmentation show divergent patterns between brown color and green color cottons as revealed by metabolic and gene expression analyses. Journal of Cotton Research 3 (1): 27. doi: 10.1186/s42397-020-00069-x

Lian TF, Xu YP, Li LF, Su XD (2017). Crystal structure of tetrameric arabidopsis MYC2 reveals the mechanism of enhanced interaction with DNA. Cell Rep. 19 (7): 1334-1342. doi: 10.1016/j.celrep.2017.04.057

Liu F, Chang XJ, Ye Y, Xie WB, Wu P et al. (2011). Comprehensive sequence and whole-life-cycle expression profile analysis of the phosphate transporter gene family in rice. Mol Plant. 4 (6): 1105-1122. doi: 10.1093/mp/ssr058

Lu N, Roldan M, Dixon RA (2017). Characterization of two TT2-type MYB transcription factors regulating proanthocyanidin biosynthesis in tetraploid cotton, *Gossypium hirsutum*. Planta 246 (2): 323-335. doi: 10.1007/s00425-017-2682-z

Ludwig SR, Habera LF, Dellaporta SL, Wessler SR (1989). *Lc*, a member of the maize *R* gene family responsible for tissue-specific anthocyanin production, encodes a protein similar to transcriptional activators and contains the myc-homology region. Proc Natl Acad Sci USA. 86 (18): 7092-7096. doi: 10.1073/pnas.86.18.7092

Mikhailova A, Strygina K, Khlestkina E (2019). The genes determining synthesis of pigments in cotton. Biological Communications 64 (2): 133–45. doi: 10.21638/spbu03.2019.205

Mol J, Grotewold E, Koes R (1998). How genes paint flowers and seeds. Trends in Plant Science 3 (6): 212–217. doi: 10.1016/S1360-1385(98)01242-4

Murre C, Bain G, Dijk MA, Engel I, Furnari BA et al. (1994). Structure and function of helix-loop-helix proteins. Biochimica Et Biophysica Acta (BBA) - Gene Structure and Expression 1218 (2): 129–135. doi: 10.1016/0167-4781(94)90001-9

Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M et al. (2000). The *TT8* gene encodes a basic helix-loop-helix domain protein required for expression of *DFR* and *BAN* genes in Arabidopsis siliques. Plant Cell 12 (10): 1863-1878. doi: 10.1105/tpc.12.10.1863

Nesi N, Jond C, Debeaujon I, Caboche M, Lepiniec L (2001). The Arabidopsis *TT2* gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. Plant Cell 13 (9): 2099-2114. doi: 10.1105/tpc.010098

Ohgishi M, Oka A, Morelli G, Ruberti I, Aoyama T (2001). Negative autoregulation of the Arabidopsis homeobox gene *ATHB-2*. Plant J 25 (4): 389-398. doi: 10.1046/j.1365-313x.2001.00966.x

Panche AN, Diwan AD, Chandra SR (2016). Flavonoids: an overview. J Nutr Sci 5:e47. doi: 10.1017/jns.2016.41

Peng Z, Gao Q, Luo C, Gong W, Tang S et al. (2020). Flavonoid biosynthetic and starch and sucrose metabolic pathways are involved in the pigmentation of naturally brown-colored cotton fibers. Industrial Crops and Products 158:113045. doi: 10.1016/j.indcrop.2020.113045

Prouse MB, Campbell MM (2012). The interaction between MYB proteins and their target DNA binding sites. Biochim Biophys Acta 1819 (1): 67-77. doi: 10.1016/j.bbagrm.2011.10.010

Qi T, Song S, Ren Q, Wu D, Huang H et al. (2011). The Jasmonate-ZIM-domain proteins interact with the WD-Repeat/bHLH/MYB complexes to regulate jasmonate-mediated anthocyanin accumulation and trichome initiation in *Arabidopsis thaliana*. Plant Cell 23 (5): 1795-1814. doi: 10.1105/tpc.111.083261

Quattrocchio F, Wing JF, van der Woude K, Mol JN, Koes R (1998). Analysis of bHLH and MYB domain proteins: species-specific regulatory differences are caused by divergent evolution of target anthocyanin genes. Plant J 13 (4): 475-488. doi: 10.1046/j.1365-313x.1998.00046.x

Reddy VS, Dash S, Reddy AR (1995). Anthocyanin pathway in rice (*Oryza sativa* L.): identification of a mutant showing dominant inhibition of anthocyanins in leaf and accumulation of proanthocyanidins in pericarp. Theor Appl Genet 91 (2): 301-312. doi: 10.1007/BF00220892

Roy AL, Singer DS (2015). Core promoters in transcription: old problem, new insights. Trends Biochem Sci 40 (3): 165-171. doi: 10.1016/j.tibs.2015.01.007

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4 (4): 406-425. doi: 10.1093/oxfordjournals.molbev.a040454

Sawant SV, Singh PK, Gupta SK, Madnala R, Tuli R (1999). Conserved nucleotide sequences in highly expressed genes in plants. Journal of Genetics 78 (2): 123–131. doi: 10.1007/BF02924562

Semizer-Cuming D, Altan F, Akdemir H, Tosun M, Gurel A et al. (2015). QTL analysis of fiber color and fiber quality in naturally green colored cotton (*Gossypium hirsutum* L.). Turkish Journal Of Field Crops 20 (1): 49–58. doi: 10.17557/.94527

Shan X, Zhang Y, Peng W, Wang Z, Xie D (2009). Molecular mechanism for jasmonate-induction of anthocyanin accumulation in Arabidopsis. J Exp Bot 60 (13): 3849-3860. doi: 10.1093/jxb/erp223

Srinivasan V, Netz DJ, Webert H, Mascarenhas J, Pierik AJ et al. (2007). Structure of the yeast WD40 domain protein Cia1, a component acting late in iron-sulfur protein biogenesis. Structure 15 (10): 1246-1257. doi: 10.1016/j.str.2007.08.009

Stracke R, Werber M, Weisshaar B (2001). The R2R3-MYB Gene family in *Arabidopsis thaliana*. Current Opinion in Plant Biology 4 (5): 447–456. doi: 10.1016/S1369-5266(00)00199-0

Strygina K, Khlestkina E, Podolnaya L (2020). Cotton genome evolution and features of its structural and functional organization. Biological Communications 65 (1). doi: 10.21638/spbu03.2020.102

Strygina KV, Khlestkina EK (2019a). Myc-like transcriptional factors in wheat: structural and functional organization of the subfamily I members. BMC Plant Biol 19 (Suppl 1):50. doi: 10.1186/s12870-019-1639-8

Strygina KV, Khlestkina EK (2019b). Structural and functional divergence of the Mpc1 genes in wheat and barley. BMC Evol Biol 19 (Suppl 1): 45. doi: 10.1186/s12862-019-1378-3

Strygina KV, Kochetov AV, Khlestkina EK (2019). Genetic control of anthocyanin pigmentation of potato tissues. BMC Genet 20 (Suppl 1):27. doi: 10.1186/s12863-019-0728-x

Sun SS, Gugger PF, Wang QF, Chen JM (2016). Identification of a *R2R3-MYB* gene regulating anthocyanin biosynthesis and relationships between its variation and flower color difference in lotus (Nelumbo Adans.). PeerJ 4: e2369. doi: 10.7717/peerj.2369

Sun S, Xiong XP, Zhu Q, Li YJ, Sun J (2019). Transcriptome sequencing and metabolome analysis reveal genes involved in pigmentation of green-colored cotton fibers. Int J Mol Sci 20 (19): 4838. doi: 10.3390/ijms20194838

Tang Z, Fan Y, Zhang L, Zheng C, Chen A et al. (2021). Quantitative metabolome and transcriptome analysis reveals complex regulatory pathway underlying photoinduced fiber color formation in cotton. Gene 767: 145180. doi: 10.1016/j.gene.2020.145180

Tanner GJ, Francki KT, Abrahams S, Watson JM, Larkin PJ et al. (2003). Proanthocyanidin biosynthesis in plants. Purification of legume leucoanthocyanidin reductase and molecular cloning of its cDNA. J Biol Chem 278 (34): 31647-31656. doi: 10.1074/jbc.M302783200

Taylor LP, Grotewold E (2005). Flavonoids as developmental regulators. Curr Opin Plant Biol 8 (3): 317-233. doi: 10.1016/j.pbi.2005.03.005

Toledo-Ortiz G, Huq E, Quail PH (2003). The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell 15 (8): 1749-1770. doi: 10.1105/tpc.013839

Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G (2010). Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 11: 548. doi: 10.1186/1471-2105-11-548

Vikhorev AV, Strygina KV, Khlestkina EK (2019). Duplicated flavonoid 3'-hydroxylase and flavonoid 3', 5'-hydroxylase genes in barley genome. PeerJ 7: e6266. doi: 10.7717/peerj.6266

Villanueva MA, Islas-Flores T, Ullah H (2016). Editorial: signaling through WD-repeat proteins in plants. Front Plant Sci 7: 1157. doi: 10.3389/fpls.2016.01157

Walker AR, Davison PA, Bolognesi-Winfield AC, James CM, Srinivasan N et al. (1999). The TRANSPARENT TESTA GLABRA1 locus, which regulates trichome differentiation and anthocyanin biosynthesis in Arabidopsis, encodes a WD40 repeat protein. Plant Cell 11 (7): 1337-1350. doi: 10.1105/tpc.11.7.1337

Wang S, Wang JW, Yu N, Li CH, Luo B et al. (2004). Control of plant trichome development by a cotton fiber *MYB* gene. Plant Cell 16 (9): 2323-2334. doi: 10.1105/tpc.104.024844

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 46 (W1): W296-W303. doi: 10.1093/nar/gky427

Wendel JF, Brubaker C, Alvarez I, Cronn R, McD. Stewart J (2009). Evolution and natural history of the cotton genus. In: Paterson AH (editor). Genetics and Genomics of Cotton. New York, USA: Springer US, pp. 3–22. doi: 10.1007/978-0-387-70810-2_1

Winkel-Shirley B (2002). Biosynthesis of Flavonoids and effects of stress. Current Opinion in Plant Biology 5 (3): 218–223. doi: 10.1016/S1369-5266(02)00256-X

Xiao YH, Zhang ZS, Yin MH, Luo M, Li XB et al. (2007). Cotton flavonoid structural genes related to the pigmentation in brown fibers. Biochem Biophys Res Commun 358 (1): 73-78. doi: 10.1016/j.bbrc.2007.04.084

Xu L, Shen ZL, Chen W, Si GY, Meng Y et al. (2019). Phylogenetic analysis of upland cotton MATE gene family reveals a conserved subfamily involved in transport of proanthocyanidins. Mol Biol Rep 46 (1): 161-175. doi: 10.1007/s11033-018-4457-4

Xu W, Dubos C, Lepiniec L (2015). Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. Trends Plant Sci 20 (3): 176-185. doi: 10.1016/j.tplants.2014.12.001

Xu W, Grain D, Bobet S, Le Gourrierec J, Thévenin J et al. (2014). Complexity and robustness of the flavonoid transcriptional regulatory network revealed by comprehensive analyses of MYB-bHLH-WDR complexes and their targets in Arabidopsis seed. New Phytol 202 (1): 132-144. doi: 10.1111/nph.12620

Yan Q, Wang Y, Li Q, Zhang Z, Ding H et al. (2018). Up-regulation of *GhTT2-3A* in cotton fibres during secondary wall thickening results in brown fibres with improved quality. Plant Biotechnol J 16 (10): 1735-1747. doi: 10.1111/pbi.12910

Yu J, Jung S, Cheng CH, Ficklin SP, Lee T et al. (2014). CottonGen: a genomics, genetics and breeding database for cotton research. Nucleic Acids Res 42 (Database issue): D1229-1236. doi: 10.1093/nar/gkt1064

Zhang F, Yao J, Ke J, Zhang L, Lam VQ et al. (2015). Structural basis of JAZ repression of MYC transcription factors in jasmonate signalling. Nature 525 (7568): 269-273. doi: 10.1038/nature14661

Zhao J, Dixon RA (2009). MATE transporters facilitate vacuolar uptake of epicatechin 3'-O-glucoside for proanthocyanidin biosynthesis in *Medicago truncatula* and *Arabidopsis*. Plant Cell 21 (8): 2323-2340. doi: 10.1105/tpc.109.067819

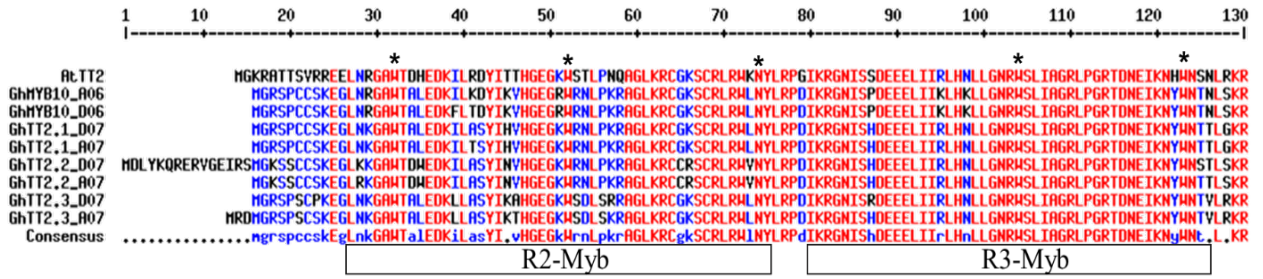
Zhu T, Liang C, Meng Z, Sun G, Meng Z et al. (2017). CottonFGD: an integrated functional genomics database for cotton. BMC Plant Biol 17 (1): 101. doi: 10.1186/s12870-017-1039-x

Zimmermann IM, Heim MA, Weisshaar B, Uhrig JF (2004). Comprehensive identification of Arabidopsis thaliana MYB transcription factors interacting with R/B-like BHLH proteins. Plant J 40 (1): 22-34. doi: 10.1111/j.1365-313X.2004.02183.x

**Supplementary material**

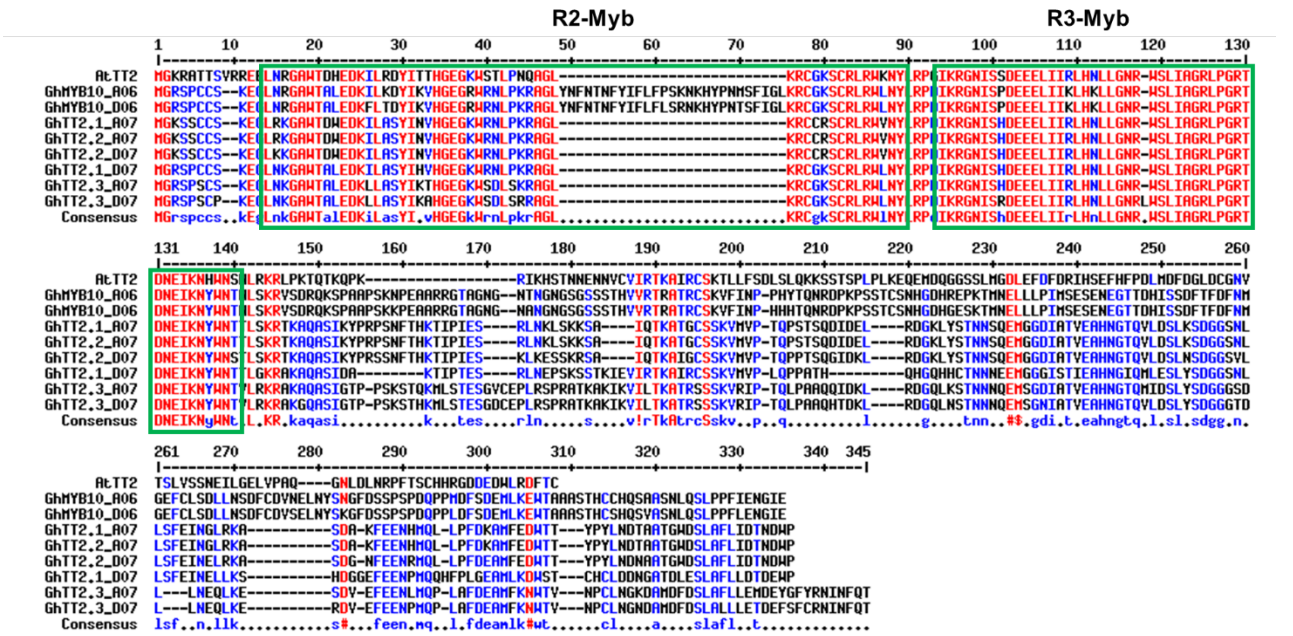
**Supplementary file 1**

Multiple amino acid sequences alignment of the R2-/R3-Myb regions in the *AtTT2* and *GhTT2* genes; asterisks point on critical tryptophan residues (Trp19, Trp39, Trp59, Trp91, Trp110).

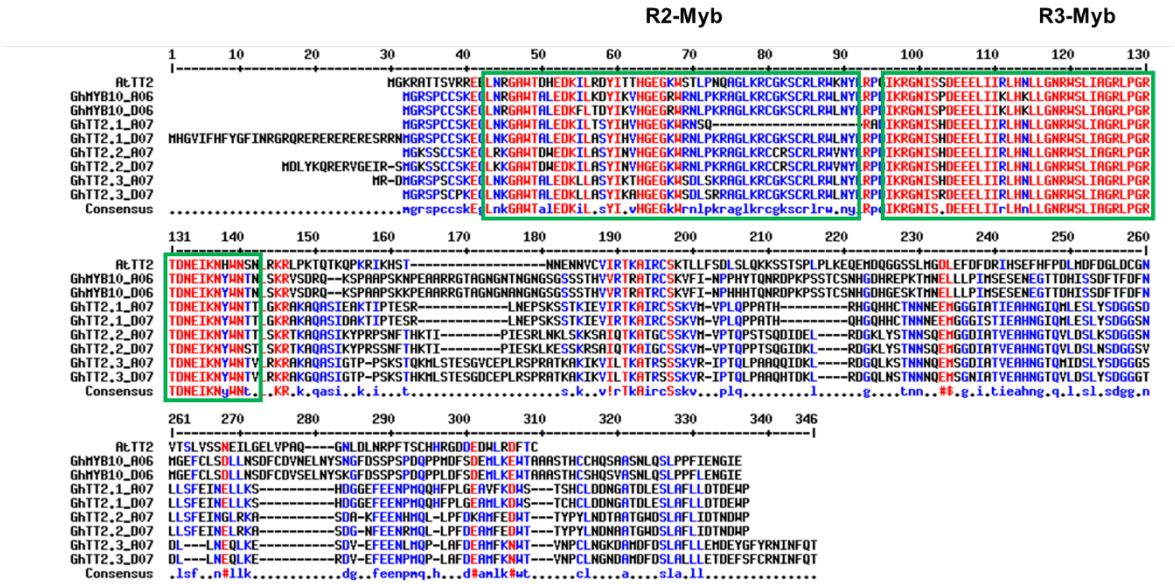



**Supplementary file 2**

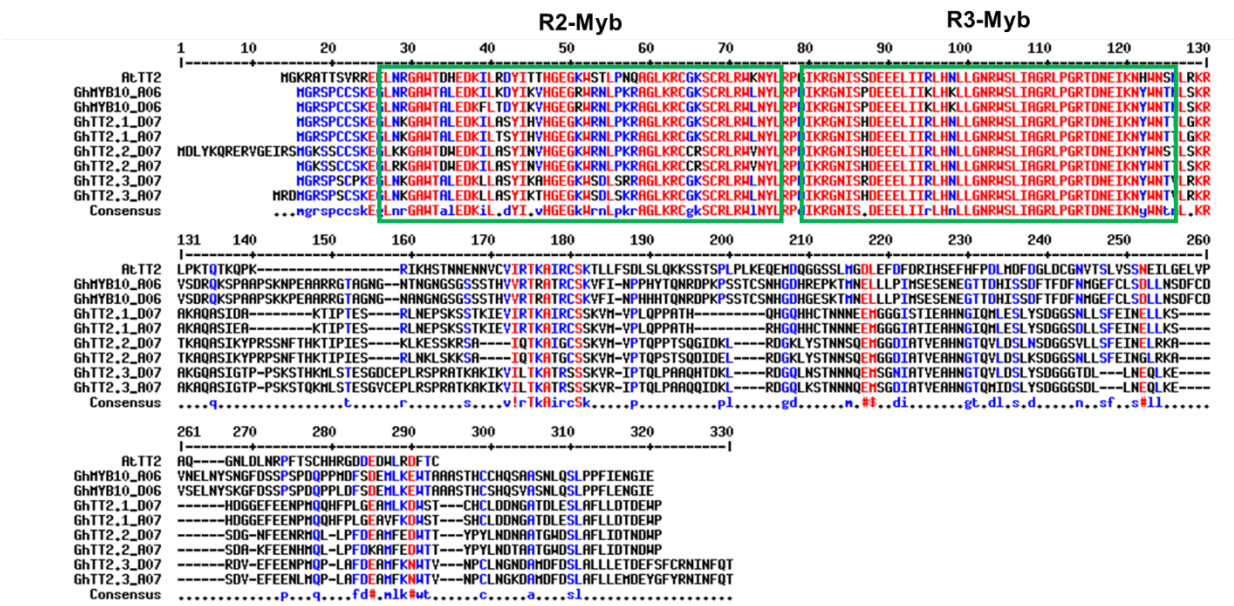
Multiple amino acid sequences alignment of the *AtTT2* gene and its orthologues *GhTT2-3A* and *GhMYB10* in the *G. hirsutum* genome. The degree of similarity is decreasing from red to black color (assembled by ZJU). Conserved R3-Myb and R2-Myb DNA-binding motifs are highlighted with a green frame.



Multiple amino acid sequences alignment of the *AtTT2* gene and its orthologues *GhTT2-3A* and *GhMYB10* in the *G. hirsutum* genome. The degree of similarity is decreasing from red to black color (assembled by HAU). Conserved R3-Myb and R2-Myb DNA-binding motifs are highlighted with a green frame.

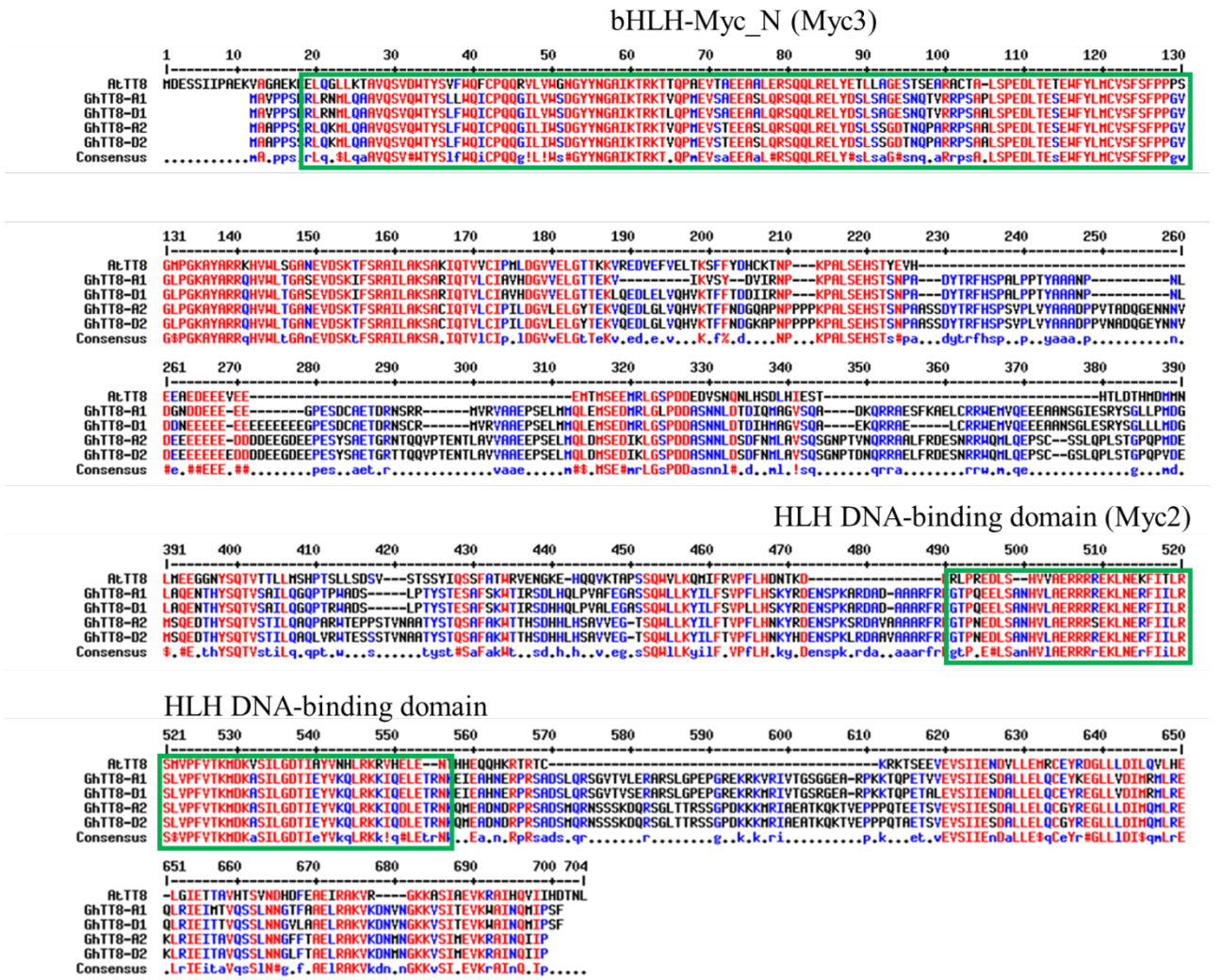



Multiple amino acid sequences alignment of the *AtTT2* gene and its orthologues *GhTT2-3A* and *GhMYB10* in the *G. hirsutum* genome. The degree of similarity is decreasing from red to black color (assembled by JGI). Conserved R2-Myb and R3-Myb DNA-binding motifs are highlighted with green frame.



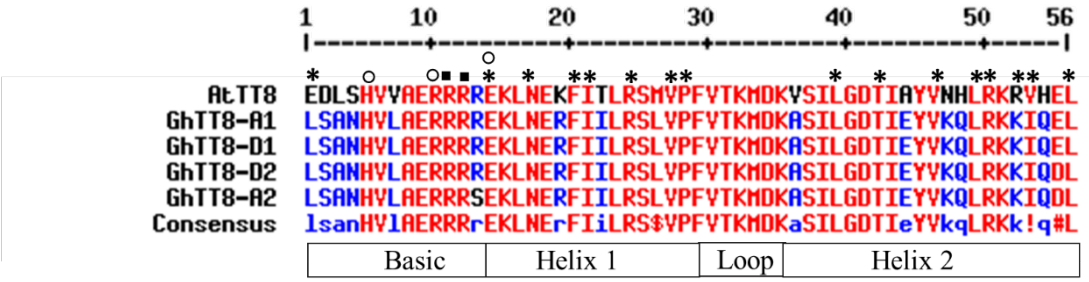
**Supplementary file 3**

Multiple amino acid sequences alignment of the *AtTT8* gene and its orthologue *GhTT8* in the *G. hirsutum* genome (assembled by NAU). The conserved Myc3, Myc2 and HLH DNA-binding domain motifs are highlighted with green frame.



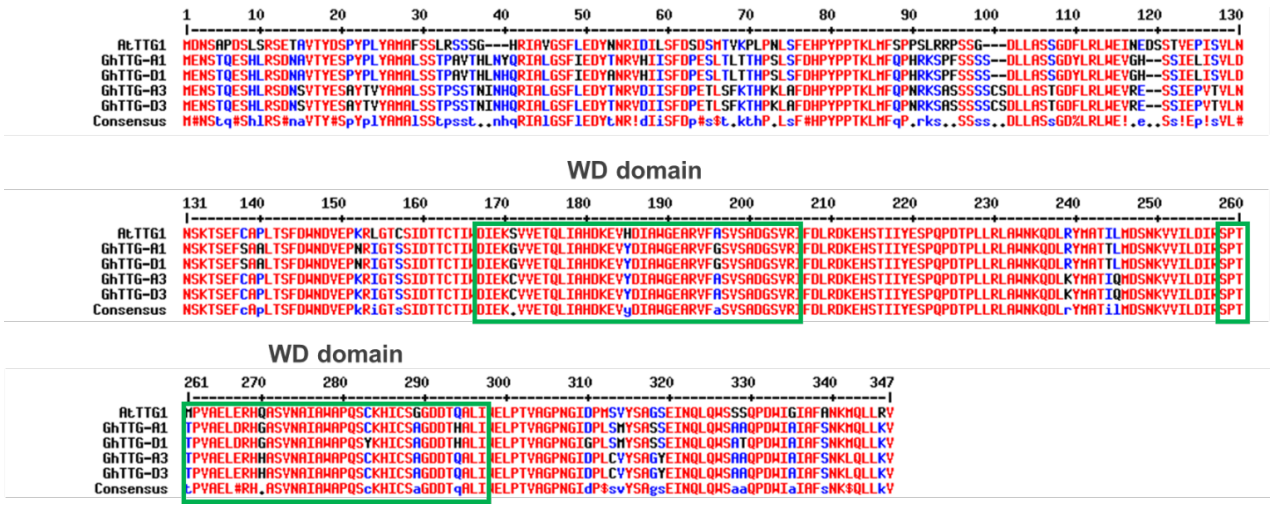
**Supplementary file 4**

Multiple amino acid sequences alignment of bHLH regions in the *AtTT8* and *GhTT8* genes; asterisks indicate the critical amino acids residues (Arg13, Glu14, Phe20, Ile21, Arg24, Val27, Pro28, Leu39, Thr42, Tyr45, Val46, Leu49, Arg50, Arg52, Val53 and Leu56), circle and square - amino acids that interact with nucleotide bases (His5, Arg10, Glu14) and DNA backbone (Arg11, Arg12), respectively (isolated by searching in the Pfam database).




**Supplementary file 5**

Multiple amino acid sequences alignment of the *AtTTG1* gene and its orthologues *GhTTG1* and *GhTTG3* in the *G. hirsutum* genome (assembled by JGI). Conserved WD-repeats motifs are highlighted with a green frame.



**Supplementary file 6.**
Identity percent of the predicted protein model (the *AtTT2, AtTT8, AtTTG1* genes and its orthologues in the *G. hirsutum* genome) in comparison with the applied template.

| **The gene name (allele)** | **Identity percent** |
|---|---|
| *AtTT2* (A) | 60.18% |
| *GhTT2-A1* (B) | 61.61% |
| *GhTT2-A2* (C) | 66.02% |
| *GhTT2-A3* (D) | 62.16% |
| *GhTT2-D1* (E) | 61.61% |
| *GhTT2-D2* (F) | 60.71% |
| *GhTT2-D3* (G) | 62.16% |
| *GhMYB10-A1* (H) | 62.83% |
| *GhMYB10-D1* (I) | 62.50% |

The 6KKS template citation: Wang, B. et al., Structural insights into target DNA recognition by R2R3-MYB transcription factors. Nucleic Acids Res. (2020).

| **The gene name (allele)** | **Identity percent** |
|---|---|
| *AtTT8* (A) | 28.83%<br>52.46% |
| *GhTT8-A1* (B) | 39.02%<br>51.52% |
| *GhTT8-D1* (D) | 37.42%<br>52.31% |
| *GhTT8-A2* (C) | 36.57%<br>50.00% |
| *GhTT8-D2* (E) | 35.00%<br>51.52% |

The 4RS9 template citation: Zhang, F. et al., Structural basis of JAZ repression of MYC transcription factors in jasmonate signalling. Nature (2015).
The 5GNJ template citation: Lian, T.F. et al., Crystal Structure of Tetrameric Arabidopsis MYC2 Reveals the Mechanism of Enhanced Interaction with DNA. Cell Rep (2017).

| **The gene name (allele)** | **Identity percent** |
|---|---|
| *AtTTG1* (A) | 20.57% |
| *GhTTG-A1* (B) | 18.97% |
| *GhTTG-D1* (C) | 19.29% |
| *GhTTG-A3* (D) | 19.57% |
| *GhTTG-D3* (E) | 19.57% |

The 2HES template citation: Liu, B.H. et al., Targeting cancer addiction for SALL4 by shifting its transcriptome with a pharmacologic peptide. Proc. Natl. Acad. Sci. U.S.A. (2018).

**Supplementary file 7.**
Protein sequence identity of *A. thaliana AtTT2, AtTT8, AtTTG1* and *G. hirsutum GhTT2*, *GhMYB10*, *GhTT8*, *GhTTG1*, including their predicted copies in the A- and D-genomes of *G.hirsutum* are presented.

| | ***AtTTG1*** | ***GhTTG1-A1*** | ***GhTTG1-D1*** | ***GhTTG1-A3*** | ***GhTTG1-D3*** |
|---|---|---|---|---|---|
| ***AtTTG1*** | - | | | | |
| ***GhTTG1-A1*** | 91.3% | - | | | |
| ***GhTTG1-D1*** | 90.5% | 99.4% | - | | |
| ***GhTTG1-A3*** | 90.8% | 94.5% | 93.9% | - | |
| ***GhTTG1-D3*** | 90.8% | 94.5% | 93.9% | 54.9% | - |

| | ***AtTT2*** | ***GhTT2-A1*** | ***GhTT2-A2*** | ***GhTT2-A3*** | ***GhTT2-D1*** | ***GhTT2-D2*** | ***GhTT2-D3*** | ***GhMYB10-A1*** | ***GhMYB10-D1*** |
|---|---|---|---|---|---|---|---|---|---|
| ***AtTT2*** | - | | | | | | | | |
| ***GhTT2-A1*** | 82.5% | - | | | | | | | |
| ***GhTT2-A2*** | 80.6% | 91.3% | - | | | | | | |
| ***GhTT2-A3*** | 83.5 % | 91.3% | 86.4% | - | | | | | |
| ***GhTT2-D1*** | 82.5% | 99.0% | 92.2% | 92.2% | - | | | | |
| ***GhTT2-D2*** | 81.6% | 90.3% | 98.1% | 85.4% | 91.3% | - | | | |
| ***GhTT2-D3*** | 81.6 % | 88.3% | 83.5% | 96.1% | 89.3% | 82.5% | - | | |
| ***GhMYB10-A1*** | 82.5% | 90.3% | 85.4% | 86.4% | 90.3% | 84.5% | 84.5% | - | |
| ***GhMYB10-D1*** | 81.6% | 90.3% | 84.5% | 86.4% | 89.3% | 83.5% | 84.5% | 98.1% | - |

| | ***AtTT8*** | ***GhTT8-A1*** | ***GhTT8-D1*** | ***GhTT8-A2*** | ***GhTT8-D2*** |
|---|---|---|---|---|---|
| ***AtTT8*** | - | | | | |
| ***GhTT8-A1*** | 65% | - | | | |
| ***GhTT8-D1*** | 66.4% | 94.8% | - | | |
| ***GhTT8-A2*** | 62.2% | 81.2% | 84.6% | - | |
| ***GhTT8-D2*** | 62.3% | 81.3% | 84.9% | 98.5% | - |