Genome-Wide identification and analysis of the AHL gene family in cotton (Gossypium)

Lanjie Zhao
Cotton Research Institute of CAAS
Youjun Lu
Anyang Institute of Technology
Wei Chen
Cotton Research institute of CAAS
Jinbo Yao
Cotton Research Institute of CAAS
Yan Li
Cotton Research Institute of CAAS
Qiulin Li
Cotton Research Institute of CAAS
Jingwen Pan
Cotton Research Institute of CAAS
Shengtao Fang
Cotton Research Institute of CAAS
Jie Sun
Shihezi University
Yongshan Zhang (13938698299@163.com)
Institute of Cotton Research of CAAS

Research article

Keywords: Cotton, AHL family, AT-Hook motif, Phylogenetics, Synteny, Expression profile, Ka/Ks

Posted Date: June 14th, 2019

DOI: https://doi.org/10.21203/rs.2.10337/v1

Version of Record: A version of this preprint was published on January 22nd, 2020. See the published version at https://doi.org/10.1186/s12864-019-6406-6.

Abstract

Background: Members of the AT-HOOK MOTIF CONTAINING NUCLEAR LOCALIZED (AHL) family are involved in various plant biological processes via protein-DNA and protein-protein interaction. However, the systematic identification and analysis of AHL gene family have yet not to be reported in cotton. Results: To investigate the potential functions of AHLs in cotton, genome-wide identification, expressions and structure analysis of the AHL gene family were performed in this study. 48, 51 and 99 AHL genes were identified from the G.raimondii, G.arboreum and G.hirsutum genome respectively. Phylogenetic trees showed that the AHLs in cotton evolved into 2 clades, Clade-A with 4-5 introns and Clade-B with intronless (excluding AHL20-2). Based on the composition of the AT-hook motif(s) and PPC/DUF 296 domain, AHL proteins were classified into three types (Type-I/-II/-III), with Type-I AHLs forming Clade-B, and the other two types together diversifying in Clade-A. The detection of synteny and collinearity showed that the AHLs expanded with specific WGD in cotton, and AHL20-2 showed the tendency of increasing intron in three different Gossypium spp. The ratios of non-synonymous (Ka) and synonymous (Ks) substitution rates of orthologous gene pairs revealed that the AHL genes of G.hirsutum had undergone through various selection pressures, purifying selection mainly in A-subgenome and positive selection mainly in D-subgenome. Examination of their expression patterns showed most of AHLs of Clade-B expressed predominantly in stem, while those of Clade-A in ovules, suggesting that the AHLs within each clade shared similar expression patterns with each other. qRT-PCR analysis further confirmed that some GhAHLs higher expression in stems and ovules. Conclusion: In this study, 48, 51 and 99 AHL genes were identified from three cotton genomes respectively. AHLs in cotton were classified into two clades by phylogenetic relationship and three type based on the composition of motif and domain. The AHLs expanded with segmental duplication, not tandem duplication. The expression profiles of GhAHLs revealed abundant differences in expression levels in various tissues and at different stages of ovules development. Our study provided significant insights into the potential functions of AHLs in regulating the growth and development in cotton.

Results

Identification and features of AHL genes in cotton

To identify the *AHL* genes, the Blastp and Hmmer search program (HMMER3.0 package) was performed against the protein databases using the *A. thaliana* AHL protein sequences as queries. After removing the redundant transcripts, the candidate *AHL* genes were confirmed using PROSITE and InterProscan 65.0 to search for the PPC and the AT-hook motifs. Finally, 12, 15, 21, 48, 51, 99 AHL genes were obtained from *P. patens*, *V. vinifera*, *T. cacao*, *G. raimondii*, *G. arboreum*, *G. hirsutum* respectively. The properties of identified *AHL*s in cotton were also analyzed by ExPASy(https://web.expasy.org/compute_pi/). The gene lengths of *AHL* genes in *G.raimondii*ranged from 684 bp to 8394 bp, which encoding polypeptides from 227 to 396 amino acid with predicted molecular weights ranging from 22.75 kD to 41.38 kDa. The theoretical pI ranged from 5.35 to 10.68 with charge from -4 to 18(Table 1). The *AHL* genes in *G.arboreum*

and *G.hirsutum* differed greatly in length (641-10972), isoelectric point (5.3-10.6), molecular weight (17.22-45.29kDa) and charge (-5-19) (Additional file 1, 2).

Phylogenetic analysis and gene structures of AHL genes

To elucidate the evolutionary relationship of the *AHL* gene family in *Gossypium*, the maximum-likelihood phylogenetic tree was reconstructed by 1000 bootstrap replicates with *AHL* proteins from *P. patens (Pp)*, *A. thaliana (At), V. vinifera (Vv)*, *T. cacao (Tc)* and *G. raimondii (Gr)*. The phylogenetic analysis showed that the *AHL*s were divided into two monophyletic clades, Clade-A and Clade-B, with 9 and 8 groups respectively (Figure 1). Each group in Clade-A (except for AHL-X, no corresponding *AHL* gene in *A. thaliana,* named AHL-X) was composed of one *VvAHL*, one *TcAHL*, different number of *AtAHL* and *GrAHL* respectively, Those in Clade-B were composed of various number of *AHL* genes from *A. thaliana (At), V. vinifera (Vv)*, *T. cacao (Tc)* and *G. raimondii (Gr)*. In *G.raimondii* Clade-A contained 22 genes including the members of *GrAHL1, GrAHL3, GrAHL5, GrAHL7, GrAHL9, GrAHL10, GrAHL13, GrAHL14, GrAHL-X1 and GrAHL-X2,* while Clade-B contained 26 members including *GrAHL15, GrAHL16, GrAHL17, GrAHL20, GrAHL22, GrAHL23, GrAHL24* and *GrAHL25*. Each group of *GrAHL* gene family usually contained 2-3 members, while the group of GrAHL17 had8 members. The Group *AHL15, AHL10* and *AHL3* showed a more rigorous evolution pattern, with only one copy left in the genomes of the 4 examined species (Figure 1, Additional file 3). This result indicated different characteristics and patterns of evolution in various group. The members of *AHLs* in *G.raimondii*, *G.arboreum* and *G.hirsutum* showed the preferably relationship of one-to-one correspondence except for *AHL17* and *AHL23*, there were 4 AHL17 members in *G. raimondii*, 6 in *G. arboreum*, and 9 in *G. hirsutum* (Additional file 4). The *AHLs* from *P. patens* evolved into two clades, suggesting an expansion of the *AHL* gene family in land plants posterior to the division between *P. patens* and the extant land plants [3].

Conservation of gene structure and motifs among AHLs in cotton

The *AHL* proteins were typically characterized by the presence of AT-hook motif(s) for binding DNA and PPC/DUF296 domain for nuclear localization and interaction with each other or themselves [4]. To investigate the presence of homologous domain sequences and the degree of conservation in the two domains, AT-hook motif(s) and PPC domain, we performed multiple sequence alignment to generate sequence logos of the two domains in cotton against the MEME website (http://meme-suite.org/tools/meme). 20 conserved motifs were predicted, and the specific amino acid sequences of each motif were also provided (Figure 2, Additional file 8).Based on the number and composition of the AT-hook motif(s) and PPC/DUF 296 domain, *AHL* proteins were classified into three types (Type-I/-II/-III), with Type-I *AHL*s forming Clade-B, and the other two types together diversifying in Clade-A. Two types of AT-hook motifs (H1 and H2) were found in the *AHL* proteins (Figure 2). Both of H1 and H2 in the AHL proteins shared the same conserved R-G-R-P core, showing the ability of bind the minor groove of AT-rich B-form DNA. The conservation of H2 with a longer core R-G-R-P-R-K-Y heptapeptides was higher than that of the H1 in cotton. H1 was found only in Clade-B, while H2 or H1 plus H2 were found in Clade-A (Figure 2). The *AHL* proteins in cacao, the closest related species of cotton, contained three types , while the AHL

proteins in grape has only two types, Type-Ⅰ and -Ⅱ. The conserved structure of *AHL9, AHL5* from 4 species contained 2 AT-hook motifs, indicating the distinct function in development. Almost all *AHL* genes in cotton (except for *Gorai.012G0247001, Ga04G1890.1, Gh_D04G0182.1 and Gh_A05G3407.1,* named *AHL-X5*) contained AT-hooking motif(s) and PPC/DUF296 domain. We considered *AHL-X5*s (Table 1, Additional file 1 and 2) in cotton as pseudogenes, because they contained the most regions of PPC/DUF296 domains although lacking the AT-hooking motifs and core sequences (motif 2), so these four genes were used as the members of *AHL* family for further analysis.

To investigate the diversity of gene structure, we performed multiple sequence alignment to generate the exons/intron pattern using the GSDS (http://gsds.cbi.pku.edu.cn/chinese.php). The structures of *AHL* genes can be divided into two types, with intronless and multiple-exon. The 26 *AHL* genes in Clade-B showed intronless in *G.raimondii*, while those in Clade-A with 5-6 exons (Figure 2). Most of the *AHL* genes in both *G. arboreum* and *G. hirsutum* presented similar exon/intron gene structure. The exception was *AHL20-2* in Clade-B, which had only one exon in *G. raimondii*, but its orthologous genes in both *G. arboreum* and *G. hirsutum* showed 4-5introns in CDS indicating the rapid evolution with intron-insertion (Additional file 5).

Chromosomal location and synteny analysis of AHL genes

A total of 48 *GrAHL* genes were unevenly mapped onto 13 chromosomes of *G. raimondii*. Each chromosome contained 3-6 *AHL* members usually. Chromosome07 contained 8 *AHL* genes, while chromosome 10 and 11 had only one *AHL*, respectively (Figure 3). The distribution of *GaAHL* and *GhAHL* genes showed similar to *G. raimondii* but some *AHL* genes in scaffolds (Additional file 1and 2).

We surveyed the collinear relationship among the orthologous *AHL* genes from *V. vinifera*, *T. cacao* and *G. raimondii* to investigate the putative clues of evolutionary events. There were 15, 21, 29 and 48 *AHL* genes in *V. vinifera, T. cacao, A. thaliana*and *G. raimondii,* respectively.Specific loss and expansion of *AHL* genes were found in four species*. AHL16* and *AHL17* were not found in *V. vinifera,* while *AHL-X* not in *A. thaliana*. Most of *AHL* genes showed one-to-one corresponding relationship in *V. vinifera and T. cacao,* while 2-4 orthologous genes were found in *G. raimondii* (Figure 3, Additional file 3). In order to investigate the pattern of gene duplication, MCScanX was used to analyze *AHL* gene family in *G. raimondii, G. arboreum* and *G.hirsutum*, the *AHL* genes in *G. raimondii* and *G. arboreum* showed correspondent relationship between those from D-subgenome, A-subgenome in *G.hirsutum* respectively(Additional file 7)*.* The result indicated the expansion of *GrAHL* gene family were with segmental duplication or Whole Genome Duplication (WGD), no tandem duplication were found (Figure 4).

Different evolution of AHL genes in A and D subgenomes of G.hirsutum

To explore the selective constrains among the orthologous AHL genes in *G. raimondii, G. arboreum* and *G.hirsutum*, we calculated Ks, Ka and the Ka/Ks ratio for the *AHL* gene pairs (Additional file 9 and 10). It is generally believed that the value of Ks was not affected by natural selection, but that of Ka was affected by natural selection. The Ka/Ks value can also explain positive selection (Ka/Ks>1), neutral

selection (Ka/Ks=1) and negative selection (Ka/Ks<1) during the evolution. In this study, 48 and 51 orthologous AHL gene paris were identified by OrthoMCL between *G. raimondii* and D-subgenome of *G.hirsutum(GrAHL/GhAHL-Dt), and* bewteen *G. arboretum* and A-subgenomeof *G.hirsutum(GaAHL/GhAHL_At),* resprectively.The distributions of Ka and Ks between each pairs were shown in Figure 5. The Ka of *GrAHL/GhAHL-Dt* ranged from 0.972745 to 1.08213, while Ks from 0.795064 to 1.08921. The Ka of *GaAHL/GhAHL_At* ranged from 0.915553 to 1.03866, while Ks from 0.899268 to 1.30387. 19 gene pairs of *GrAHL/GhAHL-Dt(39.6%)* with Ka/Ks>1 were subjected to positive selection, while 2 (AHL24-2 and AHL17-3) negative selection; 17 gene pairs of *GaAHL/GhAHL_At (33.3%)* with Ka/Ks<1 were subjected to negative selection, while only *AHL17-8* positive selection. The result suggested that the *GhAHL* genes derived from *G. raimondii* and *G. arboreum* underwent various selection directions during the evolution.

Gene expression profiles of GhAHLs

To explore the possible biological functions of *AHLs*, we inspected the expression patterns of different *AHL* genes in *G. hirsutum* based on the RNA-seq data downloaded from CottonFGD (http://www.cottonfgd.com). The *AHL* genes from *G. hirsutum* were expressed in different temporal and spatial patterns. Most *GhAHL* genes in Clade-B (such as *AHL20*, *AHL22*, *AHL23*, *AHL24)* were found strongly up-regulated expression in the stem, but extremely lowly in fiber, ovule, leaf, petal, root, stamen and pistil. Some of *GhAHL* genes in Clade-A (such as, *AHL1*, *AHL7*, *AHL9* and *AHL10)* showed an extensive expression activity in different organs, highly expressed in the fiber and ovule, suggesting a special function of these genes in the development of cotton ovule (Figure 6). Interestingly, two *AHL20-2* genes inserted by introns in *G. hirsutum* showed higher expression activity in all organs and periods than other member in Clade-B. the expression of *GrAHL* showed similar pattern in different tissues (Additional file 6). The expression result showed that the *AHL*s within each clade shared similar expression patterns with each other; however, *AHL*s in one monophyletic clade exhibited distinct expression patterns from those in the other clade.

For verification the data of RNA-seq, the qRT-PCR of six selected *AHL* genes in *G.hirsutum* was performed to analyze the expression pattern in stem, root, leaf, flower and ovule (-3,-1, 0, 1, 3, 5 PDA). The results showed that two *AHL* genes (*AHL22-1, AHL20-2*) in Clade-B displayed higher expression in stem, and lower expression in leaf. Three *AHL* genes (*AHL9-1, AHL7-1*and *AHL10*) in Clade-A expressed highly in the early development of ovule. *AHL16-1* expressed extremely lower in stem, root, leaf and flower (Figure 7). The result coincided with the data of the RNA-seq, suggesting that the data from Cottonfgd (http://www.cottonfgd.com) were reliable.

Discussion

Cotton is one of the most important economical crops worldwide, providing more than 90% of the natural fiber for textile industry. Previous research about the *AHL* gene family has been performed in *A. thaliana*, *P. patens* and other monocot and dicot plants. Up to now, no reports about *AHL* gene family were reported

in cotton, the model plants for polyploids evolution studies. In this study, we performed a comprehensive identification of *AHL* genes in *G. raimondii*, *G. arboreum*, and *G. hirsutum*, with an aim of understanding the important and diverse roles of this gene family in regulation of growth and development in plants.

Identification of AHL proteins

In our study, 48, 51 and 99 *AHL* genes were identified from the *G. raimondii*, *G. arboreum* and *G. hirsutum* genomes, respectively. According to the phylogenetic analysis and gene structure, Ga07G1158.1 were regard as the member of the group of *AHL9*, but noted as *AHL1* in Version2 of *G. arboretum;* Gh_D11G0864.1 should be regard as the member of *AHL22*, not AHL18 in notation. The group of *AHL-X5* (Gorai.012G0247001, Ga04G1890 Gh_D04G0182 and Gh_A05G3407) in 3 cotton species showed similar structure, containing the most regions of PPC/DUF296 domains, but lack of the AT-hooking motifs, so we regarded these four genes as pseudogenes and the members of *AHL* family for further analysis. GSVIVT01013438001 in grape containing AT-hooking motif, but lack of part sequences of PPC/DUF296 domains, were also regard as the member of *AHL* gene family, different to the study of Zhao et al. [1]. 12 *AHL* genes were obtained from *P. patens,* which differ from the 10 *AHL* genes in the previous study maybe because of the annotation version of genome sequencing. Genomes of *G. hirsutum* are derived from hybridization between D-subgenome of *G. raimondii* and A-subgenome in *G. arboreum* [21-24]. The 47 of 48 *GrAHL* genes were located onto 13 chromosomes, showing one to one corresponding relationship with those of D-subgenome in *G. hirsutum*. No member of *GhAHL* was located onto the Chromsome06 in D-subgenome, while *Gh_Sca005047G03* was located on scaleffolds. Based on the synteny analysis, we speculated that *Gh_Sca005047G03* was likely located on Chromosome06 of D-subgenome. *GaAHL* genes showed better linearity relationship to those in A-subgenome, it was speculated that *Gh_Sca009301G01*, *Ga14G0362*, *Ga14G0408* and *Ga14G1507* were likely located on *Gh_A11*, *GaChr09*, *GaChr06* and *GaChr02*, respectively (Additional file 8).

The *AHL* genes are divided into Clade-A and Clade-B, but the group members of Clade-A and Clade-B were respectively 5 and 4 in land plants [1], more than those from *P. patens*, suggesting an significant expansion of the *AHL* gene family in land plants. The 48 numbers of *AHL* gene family in *G. raimondii* were more than those from other species reported in previous report or closely-related species, such as cacao (21) and grape (15) [1]. Each group in Clade-A (except for AHL-X*)* was composed of one *VvAHL*, one *TcAHL*, different number of *AtAHL* and *GrAHL* respectively. Most groups had 2-3 members in diploid cotton, while the *GrAHL17* had8 members in *G. raimondii,* 9 in *G. arboreum and 18 G. hirsutum,* indicating a different expansion of the *AHL* gene family. The synteny results showed the expansion of *AHL* family were with the WGD or segmental duplications, not tandem duplication. Related research suggested that the ancestor of *Gossypium* experienced a whole-genome duplication event after its divergence from the cocoa ancestor [21,22]. So, we speculated that the numbers of the *AHL*s in *G. raimondii* or *G. arboreum* were more than that in *V. vinifera* and *T. cacao* maybe due to the specific Whole-genome duplication (WGD) event in *Gossypium* ancestor after the divergence of cotton from cacao [21,23]. The *AHL* gene losses were also found in Arabidopsis, group *AHL-X* included the corresponding *AHL* genes from *G.*

raimondii, *V. vinifera* and *T. cacao*, no member were found in Arabidopsis, suggesting that the different number of each group resulted from the various gene loss.

Conservation of the *AHL* gene family

The *AHL* gene family is a plant-specific family with conserved structure of AT-hook and PPC/DUF domain. The members of *AHL* family present diversity not only in the sequence of AT-hook and PPC motifs, but also in gene length, gene structure, as well as in motif number. An analysis of sequence logo was performed for further investigating the divergence of AT-hook motif and the PPC domains in AHL proteins. AT-hook motif(s) could be distinguished by the phylogenetic relatedness of its homeodomains. Our results demonstrated that a longer core sequence R-G-R-P in *AHL* proteins in cotton, especially in type AT-hook motif, containing a more longer and conserved core R-G-R-P-R-K-Y heptapeptide. According to the AT-hook motif and PPC domain, the *AHL* proteins in cotton were divided into three types, agree with previous study [1, 3].Two types of gene structure, with intronless and multiple-exon, were found in the *AHL* genes of cotton. The 26 *GrAHL* genes in Clade-B showed intronless, while those in Clade-A with 5-6 exons. The *AHL* genes in *V. vinifera* presented another scenario, in which most of the *AHL* genes contain multiple exons except for the sole-exon gene *GSVIVT01027625*. There were some exceptions in cotton and cocoa, such as the inclusion of multiple exons in cocoa *Thecc1EG005492* and *Thecc1EG034810*, which were clustered in Clade-A. The difference of gene structure among the *AHL20-2* genes in different cotton species were showed in Figure S4, *GrAHL20-2* possessed only one exon while its orthologous members in *G. arboreum* and *G.hirsutum* contained multiple introns, suggesting a rapid evolutionary rate during the history of cotton. Furthermore, the *AHL* genes in Clade-B in *G. hirsutum* were mainly specifically expressed in stem, with no detectable expression in other organs. Two members of *AHL20-2* from A-subgenome and D-subgenome respectively, with multiple introns, expressed in various organs and tissues, suggesting that the gene structure may have some effects on gene expression pattern.

Expression profile analysis of AHL in cotton

The *AHL* genes play important roles in plant development, floral transition and response to biotic and abiotic stress [1, 4, 30]. *AHL20*, *AHL22*, *AHL23* and *AHL24* were strongly up-regulated expressed in the stem, but extremely lowly in fiber, ovule, leaf, petal, root, stamen and pistil. *AHL1*, *AHL7*, *AHL9* and *AHL10* showed an extensive expression activity in different organs, highly expressed in the fiber and ovule, suggesting a special function of these genes in the development of cotton ovule. According to the phylogenetic analysis, Group of AHL3, AHL10 and AHL15 kept one copy left in *V. vinifera*, *T. cocoa*, *A. thaliana* and *G. raimondii*, suggesting the more conserved function or vital roles in development. The gene expression patterns of *GhAHL10* and *GhAHL15* were significantly various, *GhAHL10* was significantly expressed in all detected tissues and stages, while *GhAHL15* were not detected expression in any detected tissues and stages. Compared with the homologous groups of *V. vinifera* and *T. cacao*, the members of *AHL17* were significantly expanded to 8 in *G. raimondii* and 17 in *G. hirsutum* no expression was detected in tissues and stages except *AHL17-2* and *AHL17-6*, consistent with decreased gene expression levels after gene expansion in previous reports. The expression result showed that the *AHL*s

within each clade shared similar expression patterns with each other; however, *AHL*s in one monophyletic clade exhibited distinct expression patterns from the ones in the other clade.

Conclusions

In this study, 48, 51 and 99 *AHL* genes were identified from the *G.raimondii, G.arboreum* and *G.hirsutum* genome on a genome-wide scale respectively. All of the genes showed one-to-one homology relationship among three different genomes or subgenomes in cotton. Phylogenetic analysis revealed that *AHLs* in cotton evolved into 2 clades, Clade-A with 4-5 introns and Clade-B with intronless (excluding *AHL20-2*). Based on the composition of the motif and domain, *AHL* proteins were classified into three types. Synteny analysis indicated that *AHL*s in cotton were highly homologous to those in *V. vinifera and T. cacao*. Sequence analysis showed that segmental duplications were the major driving forces of *AHL* family evolution, suggesting that *AHLs* expanded with specific WGD in cotton. The ratios of non-synonymous (Ka) and synonymous (Ks) substitution rates between orthologous gene pairs revealed that the *AHL* genes of *G.hirsutum* had undergone through various selections during evolution, purifying selection mainly in A-subgenome and positive selection mainly in D-subgenome.

The expression profiles of *AHLs* were analyzed by using RNA-seq data, and the results indicated that most of *AHLs* of Clade-B expressed predominantly in stem, while those of Clade-A in ovules, suggesting that the *AHLs* within each clade shared similar expression patterns within each other. qRT-PCR analysis further confirmed that some *GhAHLs* displayed higher expression in stems and ovules. Our study provided a reference for the further functional investigation of these selected candidate *AHL* proteins.

Materials And Methods

Identification of the AHL genes

To identify the AHL gene family in cotton, the genome sequence and annotation data of four cotton species, including *G. raimondii* [21,22], *G. arboreum*[23], *G. hirsutum*[24]and *G. barbadense* [11], were obtained from the CottonFGD (http://www.cottonfgd.org/) [25] by blastp against protein database and tblastn against genome databases using the query sequences of the 29 AHL proteins in *Arabidopsis thaliana* acquired from TAIR 15 (http://www.arabidopsis.org), the E-value cut-off was set at 1.0e-5 to ensure confidence. The *AHL* genes from *P. patens (Pp)*, *A. thaliana (At), V. vinifera (Vv)*, *T. cacao (Tc)*were retrieved from the Phytozome database v12.1 (https://phytozome.jgi.doe.gov/pz/portal.html). Redundant sequences were detected and deleted by manual. The candidate sequences were submitted to PROSITE for PPC domain (PS51742), those sequences comprised of AT-hook motif(s) and PPC domain were confirmed as *AHL* genes for further analysis. Proteinsequences of *AHL* were submitted to ExPASy (http://web.expasy.org/protparam/) to predict the molecular weights (MW) and theoretical isoelectric points (pI) and charge.

Chromosomal location and collinearity analysis

The information of the *AHLs* loci on chromosome was obtained from annotation gff3 files. The Gene Structure Display Server (http://gsds,cbi.pku.edu.cn/) was used for gene structure analysis. Conserved protein motifs of the *AHLs* were predicted by the MEME program (http://meme-suite.org/tools/meme). The parameters of MEME were optimum width, 3-50; number of repetitions, any; maximum number of motifs, 20. A schematic diagram of gene structure was redrawn by Circos. The MCscanX program was used to identify *GrAHL* duplications as previous described by Wang et al. [26]. Total 37505 proteins sequences were used by all-all BLAST with e-value<1.0e-5. All genes were classified into various types of duplications, Dispersed, Singleton, WGD or Segmental and Tandem duplications. A schematic diagram of the putative WGD or segmental duplications of *GrAHL* was constructed using the Circos, and the *AHLs* with WGD or segmental duplications were linked by lines.

Phylogenetic analysis and classification of AHL genes in cotton

For phylogenetic analysis, All *AHL* amino sequences from *P. patens (Pp)*, *A. thaliana (At)*, *V. vinifera (Vv)*, *T. cacao (Tc)* and three cotton species were aligned by ClustalX v1.83 with default parameters [27]. MEGA7.0 was used to find best model and construct the Maximum likelihood (ML) tree with bootstrap test of 1000 replicates, the model of JTT+G was selected as the best model. Neighbor-Joining (NJ) phylogenetic trees were also generated in MEGA7.0 to validate the ML phylogenetic trees [28].

Calculation of Ka/Ks of AHL genes in cotton

The orthologs of the AHL genes in *G. raimondii, G. arboreum* and *G.hirsutum* were identified by OrthoMCL[29]. The orthologous gene pairs of AHLs were aligned by codons with Muscle in MEGA7.0 software. Non-synonymous (Ka) and synonymous (Ks) substitution rates and Ka/Ks ratios of were determined by the model average (MA) and model (MS) in Kaks_Calculator 2.0 program [30] respectively.

Expression Profiles of GhAHL and GrAHL genes

For analyzing the expression profile of *GhAHL* and *GrAHL* genes in different tissues and development stages, the expression data of fragments-per-kilobase-per-million (FPKM) were retrieved from the genome-wide RNA-seq dataset in CottonFGD (http://www.cottonfgd.com/data) and CCnet website (http://structuralbiology.cau.edu.cn/gossypium), respectively. The heatmap charts were drawn according to gene expression values (FPKM).

Quantitative RT-PCR (qRT-PCR) for GhAHL genes

The upland cotton (TM-1) seeds were germinated on a wet germinated disc for 3 days at 28 °C, and then transferred to a liquid culture medium. Total RNA was extracted from the seedlings. The true leaves, root and stem were collected and were immediately frozen in liquid nitrogen for RNA extraction. Blossom in full bloom, and then take the first 3 days, 1 day, 0, 1 days after flowering, flowering after 3 days, 5 days after flowering ovule and flower liquid nitrogen treatment - 80 ℃ after preservation; Total RNA was extracted from the seedlings. cDNA was synthesized by using an EASYspin Plus Plant RNA Kit (Aidlab) with gDNA Eraser (Takara). The qRT-PCR reactions were conducted using a SYBR Green I Master mixture

(Roche, Basel, Switzerland) according to the manufacturer's protocol on a Light Cycler 480II system (Roche, Switzerland). Cotton ACTIN14 (GenBank accession number: AY305733) was used as an internal control in the PCR assays. The primers designed for qRT-PCR were showed in Additional file 11. The qRT-PCR was completed with three biological replicates, each comprising three technical replicates. The PCR conditions were as follows: 95 °C for 30 s; 40 cycles of 95 °C for 5 s, 60°C for 1 min, and 72 °C for 10 s; 50 °C for 30 s. The relative gene expression levels were calculated based on the $2^{-\Delta\Delta CT}$ method [31].

Declarations

Conflicting of Interests

The authors declare no conflict of interest.

Authors' contributions

Z.-Y.S. planned and designed the research. Z.-L.J. wrote the manuscript. C. W, Y.-J.B, L.Y., L.-Q.L., P.-J.W., F.-S.T performed the experiments. S.J. supervised the research. Z.-L.J and L.-Y.J. contributed equally. All authors read and approved the final manuscript.

Acknowledgements

The study was supported in part by National Science Foundation in China (31871680), the project of state key lab of cotton biology (CB2017C12).

Auther details:

[1]Institute of Cotton Research of CAAS, Anyang, Henan, China, 455000

[2] Research Base, Anyang Institute of Technology, State Key Laboratory of Cotton Biology, Anyang, Henan, China, 455000

[3] Shihezi University, Shihezi, Xinjiang, China, 832003

References

1. Zhao J, Favero DS, Qiu J, Roalson EH, Neff MM (2014) Insights into the evolution and diversification of the AT-hook Motif Nuclear Localized gene family in land plants. BMC Plant Biol 14:1-19
2. Aravind L, Landsman D (1998) AT-hook motifs identified in a wide variety of DNA-binding proteins. Nucleic Acids Res 26:4413
3. Huth JR, Bewley CA, Nissen MS, Evans JN, Reeves R, Gronenborn AM, Clore GM (1997) The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. Nat Struct Biol 4:657-665

4. Zhao J, Favero DS, Peng H, Neff MM (2013) Arabidopsis thaliana AHL family modulates hypocotyl growth redundantly by interacting with each other via the PPC/DUF296 domain. Proc Natl Acad Sci U S A 110:E4688-4697
5. Favero DS, Jacques CN, Iwase A, Le KN, Zhao J, Sugimoto K, Neff MM (2016) SUPPRESSOR OF PHYTOCHROME B4-#3 Represses Genes Associated with Auxin Signaling to Modulate Hypocotyl Growth. Plant Physiol 171:2701-2716
6. Street IH, Shah PK, Smith AM, Avery N, Neff MM (2008) The AT-hook-containing proteins SOB3/AHL29 and ESC/AHL27 are negative modulators of hypocotyl growth in Arabidopsis. Plant J 54:1-14
7. Xiao C, Chen F, Yu X, Lin C, Fu YF (2009) Over-expression of an AT-hook gene, AHL22, delays flowering and inhibits the elongation of the hypocotyl in Arabidopsis thaliana. Plant Mol Biol 71:39-50
8. Gallavotti A, Malcomber S, Gaines C, Stanfield S, Whipple C, Kellogg E, Schmidt RJ (2011) BARREN STALK FASTIGIATE1 Is an AT-Hook Protein Required for the Formation of Maize Ears. Plant Cell 23:1756-1771
9. Jia QS, Zhu J, Xu XF, Lou Y, Zhang ZL, Zhang ZP, Yang ZN (2015) Arabidopsis AT-hook protein TEK positively regulates the expression of arabinogalactan proteins for Nexine formation. Molecular Plant 8:251-260
10. Jin Y, Luo Q, Tong H, Wang A, Cheng Z, Tang J, Li D, Zhao X, Li X, Wan J et al. (2011) An AT-hook gene is required for palea formation and floral organ number control in rice. Dev Biol 359:277-288
11. Liu X, Zhao B, Zheng HJ, Hu Y, Lu G, Yang CQ, Chen JD, Chen JJ, Chen DY, Zhang L et al. (2015) Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. Sci Rep 5:14139
12. Xu Y, Wang Y, Stroud H, Gu X, Sun B, Gan E-S, Ng K-H, Jacobsen SE, He Y, Ito T (2013) A matrix protein silences transposons and repeats through interaction with retinoblastoma-associated proteins. Nucleus 23:345-350
13. Ng KH, Ito T (2010) Shedding light on the role of AT-hook/PPC domain protein in Arabidopsis thaliana. Plant Signaling Behavior 5:200-201
14. Ng KH, Yu H, Ito T (2009) AGAMOUS controls GIANT KILLER, a multifunctional chromatin modifier in reproductive organ patterning and differentiation. PLoS Biol 7:e1000251
15. Yun J, Kim YS, Jung JH, Seo PJ, Park CM (2012) The AT-hook motif-containing protein AHL22 regulates flowering initiation by modifying FLOWERING LOCUS T chromatin in Arabidopsis. J Biol Chem 287:15307-15316
16. Kim SY, Kim YC, Seong ES, Lee YH, Park JM, Choi D (2010) The chili pepper CaATL1: an AT-hook motif-containing transcription factor implicated in defence responses against pathogens. Mol Plant Pathol 8:761-771
17. Lu H, Zou Y, Feng N (2010) Overexpression of AHL20 negatively regulates defenses in Arabidopsis. J Integr Plant Biol 52:801-808

18. Matsushita A, Furumoto T, Ishida S, Takahashi Y (2007) AGF1, an AT-Hook Protein, Is Necessary for the Negative Feedback of AtGA3ox1 Encoding GA 3-Oxidase. Plant Physiol 143:1152-1162
19. Zhou L, Liu Z, Liu Y, Kong D, Li T, Yu S, Mei H, Xu X, Liu H, Chen L (2016) A novel gene OsAHL1 improves both drought avoidance and drought tolerance in rice. Sci Rep 6:30264
20. Delaney SK, Orford SJ, Martinharris M, Timmis JN (2007) The Fiber Specificity of the Cotton FSltp4 Gene Promoter is Regulated by an AT-Rich Promoter Region and the AT-Hook Transcription Factor GhAT1. Plant Cell Physiology 48:1426
21. Wang K, Wang Z, Li F, Ye W, Wang J, Song G, Yue Z, Cong L, Shang H, Zhu S et al. (2012) The draft genome of a diploid cotton Gossypium raimondii. Nat Genet 44:1098-1103
22. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, Llewellyn D, Showmaker KC, Shu S, Udall J (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres.
23. Li F, Fan G, Wang K, Sun F, Yuan Y, Song G, Li Q, Ma Z, Lu C, Zou C et al. (2014) Genome sequence of the cultivated cotton Gossypium arboreum. Nat Genet 46:567-572
24. Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J et al. (2015) Genome sequence of cultivated Upland cotton (Gossypium hirsutum TM-1) provides insights into genome evolution. Nat Biotechnol 33:524-530
25. Zhu T, Liang C, Meng Z, Sun G, Meng Z, Guo S, Zhang R (2017) CottonFGD: an integrated functional genomics database for cotton. BMC Plant Biol 17:101
26. Wang, Y.; Tang, H.; DeBarry, J.D.; Tan, X.; Li, J.; Wang, X.; Lee, T.-H.; Jin, H.; Marler, B.; Guo, H. MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res.2012, 40, e49
27. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947-2948
28. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870-1874
29. Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13, 2178–2189. doi: 10.1101/gr.1224503
30. Dapeng Wang, Yubin Zhang, Zhang Zhang, Jiang Zhu, and Jun Yu(2010). KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. Genomics Proteomics Bioinfo. 8: 77-80
31. Schmittgen, T.D.; Livak, K.J. Analyzing Real-time PCR data by the comparative c(t) method. *Nat. Protoc.* 2008, *3*, 1101–1108

Table

Table 1. Information of the *AHL* genes in *G. raimondii*

Gene name	Sequence ID	Gene (bp)	CDS (bp)	Protein(aa)	Intron	MW[a] (kDa)	pI[b]	Charge
GrAHL1-1	Gorai.003G167100.1	2846	984	327	4	33.734	9.912	8.5
GrAHL1-2	Gorai.004G203700.1	4309	1026	341	4	35.305	9.994	8.5
GrAHL1-3	Gorai.007G021700.1	3420	984	327	4	33.742	10.038	9.5
GrAHL3	Gorai.008G283600.1	3507	1008	335	4	35.308	8.45	6.5
GrAHL5-1	Gorai.004G211500.1	3767	1026	341	4	35.485	10.638	18
GrAHL5-2	Gorai.008G246700.1	4352	1023	340	4	35.256	10.307	16
GrAHL7-1	Gorai.004G161300.1	3148	1011	336	4	34.954	9.402	6
GrAHL7-2	Gorai.007G091400.1	4226	996	331	4	34.149	9.686	5.5
GrAHL9-1	Gorai.004G158000.1	3982	1023	340	4	35.007	9.669	10
GrAHL9-2	Gorai.008G122100.1	2887	990	329	4	34.067	10.453	14
GrAHL9-3	Gorai.007G098600.1	4773	1023	340	4	34.978	10.486	14
GrAHL10	Gorai.002G112700.1	6764	1095	364	4	37.03	10.27	14
GrAHL13-1	Gorai.004G186000.1	4606	1191	396	4	41.381	10.057	16
GrAHL13-2	Gorai.008G227100.1	3931	1176	391	4	40.457	9.13	8
GrAHL14-1	Gorai.007G280000.1	6475	1035	344	5	35.976	9.59	10.5
GrAHL14-2	Gorai.007G345200.1	4396	1038	345	5	36.075	9.86	12
GrAHL14-3	Gorai.013G186600.1	3669	1035	344	5	36.192	9.356	11.5
GrAHL-X1	Gorai.002G160000.1	5307	1101	366	4	38.034	9.785	9
GrAHL-X2	Gorai.006G158700.1	2517	1053	350	4	36.311	8.883	8.5
GrAHL-X3	Gorai.009G408800.1	3520	993	330	4	33.683	8.456	4
GrAHL-X4	Gorai.001G119100.1	2223	1095	364	4	38.243	7.851	4.5
GrAHL-X5	Gorai.012G024700.1	8394	633	210	4	22.749	5.77	-1.5
GrAHL15	Gorai.011G267800.1	3780	933	310	0	32.515	5.898	-4
GrAHL16-1	Gorai.006G007800.1	771	771	256	0	27.271	9.111	7.5
GrAHL16-2	Gorai.007G070000.1	1215	759	252	0	26.816	8.113	5
GrAHL17-1	Gorai.001G133900.1	1721	894	297	0	30.26	9.235	10.5
GrAHL17-2	Gorai.005G096700.1	921	921	306	0	31.725	7.472	5.5
GrAHL17-3	Gorai.006G120100.1	684	684	227	0	24.17	7.234	3
GrAHL17-4	Gorai.006G124100.1	684	684	227	0	24.232	6.944	2
GrAHL17-5	Gorai.009G075100.1	1020	882	293	0	30.273	8.108	8.5
GrAHL17-6	Gorai.009G230300.1	7250	957	318	0	33.38	8.214	11.5
GrAHL17-7	Gorai.010G035300.1	1448	864	287	0	29.728	7.054	3
GrAHL17-8	Gorai.013G253800.1	1304	987	328	0	33.21	10.682	10
GrAHL20-1	Gorai.005G048000.1	1403	888	295	0	30.487	5.349	-3.5
GrAHL20-2	Gorai.006G247900.1	1008	852	283	0	28.921	5.62	-3
GrAHL20-3	Gorai.007G280400.1	909	909	302	0	30.368	5.954	-2
GrAHL22-1	Gorai.001G173500.1	870	870	289	0	30.852	7.738	4
GrAHL22-2	Gorai.004G160700.1	1004	927	308	0	32.41	6.97	2
GrAHL22-3	Gorai.007G091800.1	1768	903	300	0	31.472	6.512	0
GrAHL23-1	Gorai.003G181200.1	804	804	267	0	27.75	6.856	2
GrAHL23-4	Gorai.008G226900.1	1477	801	266	0	27.957	6.739	1.5
GrAHL23-2	Gorai.004G185900.1	960	864	287	0	29.853	6.794	1.5
GrAHL23-3	Gorai.006G216300.1	828	828	275	0	28.779	6.221	-2
GrAHL24-1	Gorai.003G167700.1	1374	930	309	0	33.106	6.798	2.5
GrAHL24-2	Gorai.006G211500.1	900	900	299	0	31.775	6.704	1.5
GrAHL24-3	Gorai.008G240700.1	1751	924	307	0	32.791	6.584	0.5
GrAHL25-1	Gorai.005G215400.1	1776	846	281	0	28.239	9.008	4.5
GrAHL25-2	Gorai.012G138000.1	1800	846	281	0	28.419	8.982	5

[a] Molecular weight of the amino acid sequence, [b] Isoelectric point.

Additional File Legend

Additional file 1 - Information of *AHLs* in G. arboretum. a Molecular weight of the amino acid sequence, b Isoelectric point.

Additional file 2 - Information of *AHLs* in G. hirsutum. a Molecular weight of the amino acid sequence, b Isoelectric point.

Additional file 3 - The orthologous relationship and type of AHL proteins in *V. vinifera, T. cacao, A.thaliana*and *G. raimondii.* The forms in pink indicated the Type-I *AHL* genes, those in yellow indicated the Type-III *AHL* genes and those in blue indicated the Type-II *AHL* genes. The lines repented the loss of orthologous gene.

Additional file 4 - Phylogenetic relationship of AHL proteins in cotton. *AHL* proteins from *G. raimondii*, *G. arboreum* and *G. hirsutum* are marked with blue rhombus, green squares, and red rhombus squares, respectively.

Additional file 5 - The variations of gene structures and motifs of *AHL20-2* from *G. raimondii*, *G. arboreum* and *G. hirsutum.*Gene structure and conserved motifs were predicted from the GSDS and MEME website. The length of proteins and DNA sequence was estimated using the scale at the bottom. The motifs were displayed in different colored boxes with Arabic numerals, black line indicated the non-conserved amino acid or intron. Gray boxes indicate untranslated 5- and 3-regions, blue boxes indicate exons. The sequences of motifs were listed in additional file 6.

Additional file 6 - The sequences of motifs predicted by MEME (http://meme-suite.org/tools/meme)

Additional file 7 - The expression profiles of *GrAHLs*. The heatmap was generated on the basis of RNA-seq data from the website (http://structuralbiology.cau.edu.cn/gossypium), the color scale was shown at the right. Higher expression levels were shown in red, and lower in blue. DPA represented the day of ovule after anthesis.

Additional file 8 - The circos map of *AHL*s in *G.raimondii, G.arbretum* and *G.hirsutum.*The collinearity of *AHL* genes between *G.raimondii and D-subgenome in G.hirsutum* were showed in orange lines, that between *G.arbretum* and the A-subgenome in *G.hirsutum in blue lines. AHL* genes located in *Scalffolds* were showed in red lines, and the locations of scalffolds were putatived

Additional file 9 - Ka, Ks and Ka/Ks ratio between orthologous genes pairs from *G.raimondii and D-subgenome in G.hirsutum*

Additional file 10 - Ka, Ks and Ka/Ks ratio between orthologous gene pairs from *G.arbretum and A-subgenome of G.hirsutum*

Additional file 11 - The primers designed for qRT-PCR

Figures

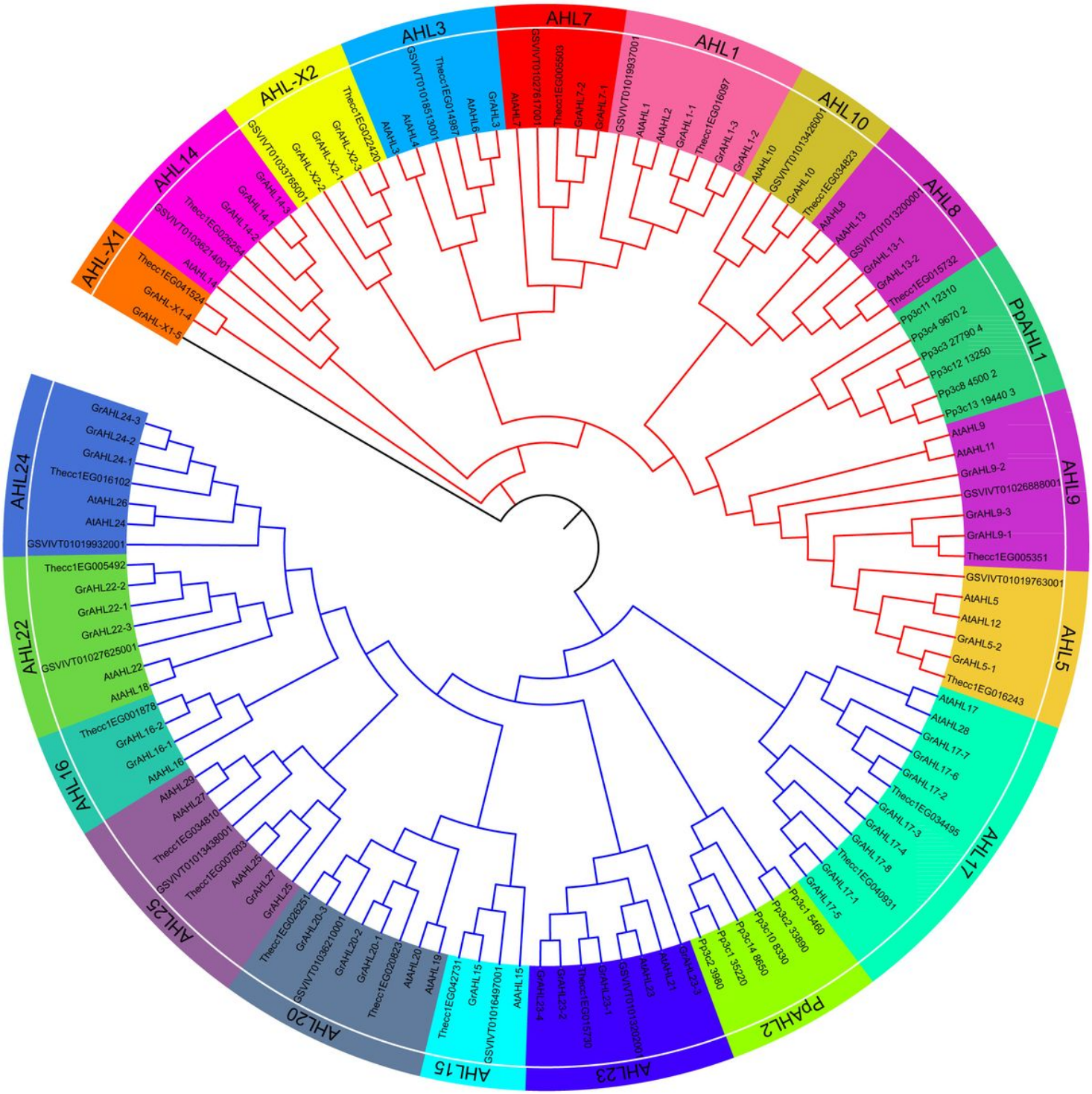

Figure 1

Phylogenetic relationships among AHL proteins. The maximum-likelihood (ML) phylogenetic tree was constructed by MEGA7.0 using 1000 bootstrap replicates for the AHL proteins from V. vinifera (Vv), A. thaliana (At), T. cacao (Tc), G. raimondii (Gr) and P. patens (Pp). Clade-A indicated in blue branch lines and Clade-B in red branch lines. The black line showed the pesudogene.

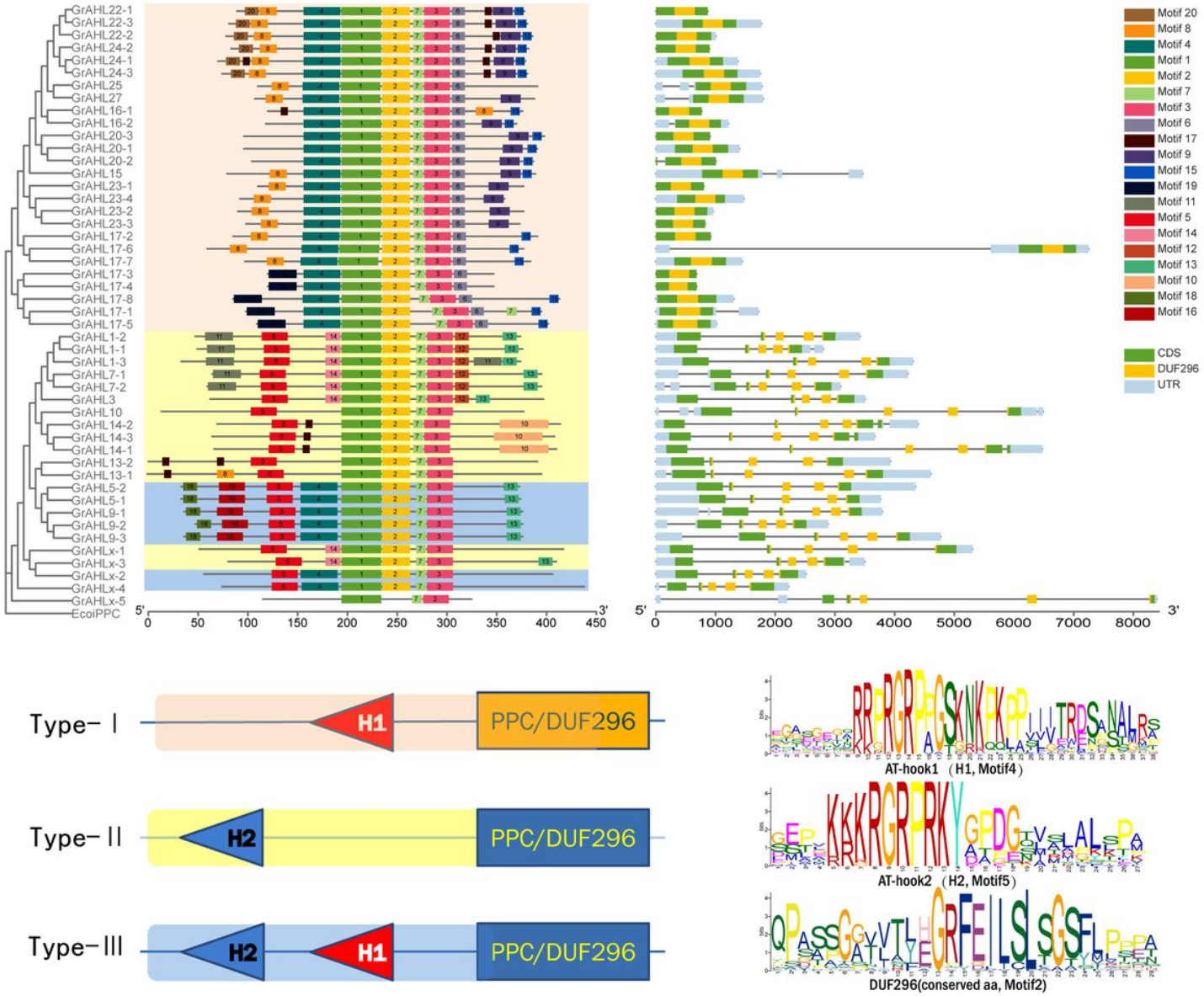

Figure 2

The conserved motifs, Exon–intron structures of GrAHLs. The maximum-likelihood phylogenetic tree was constructed by MEGA7.0 using 1000 bootstrap replicates. Conserved motifs and gene structure were predicted from MEME and GSDS website (http://meme-suite.org/tools/meme, http://gsds.cbi.pku.edu.cn/chinese.php). The length of proteins and DNA sequence was estimated using the scale at the bottom. The motifs were displayed in different colored boxes with various number, black line indicated the non-conserved amino acid or introns. Light blue boxes indicate untranslated 5- and 3-regions; Green boxes indicate exons. The PPC domains were highlighted by yellow boxes. The topology of three types of AHL proteins in cotton based on the AT-hook motifs and PPC domain (motif 2). H1 represented the AT-hook1 containing the conserved peptide in motif4, H2 represented the AT-hook2 containing the conserved peptide in motif5.

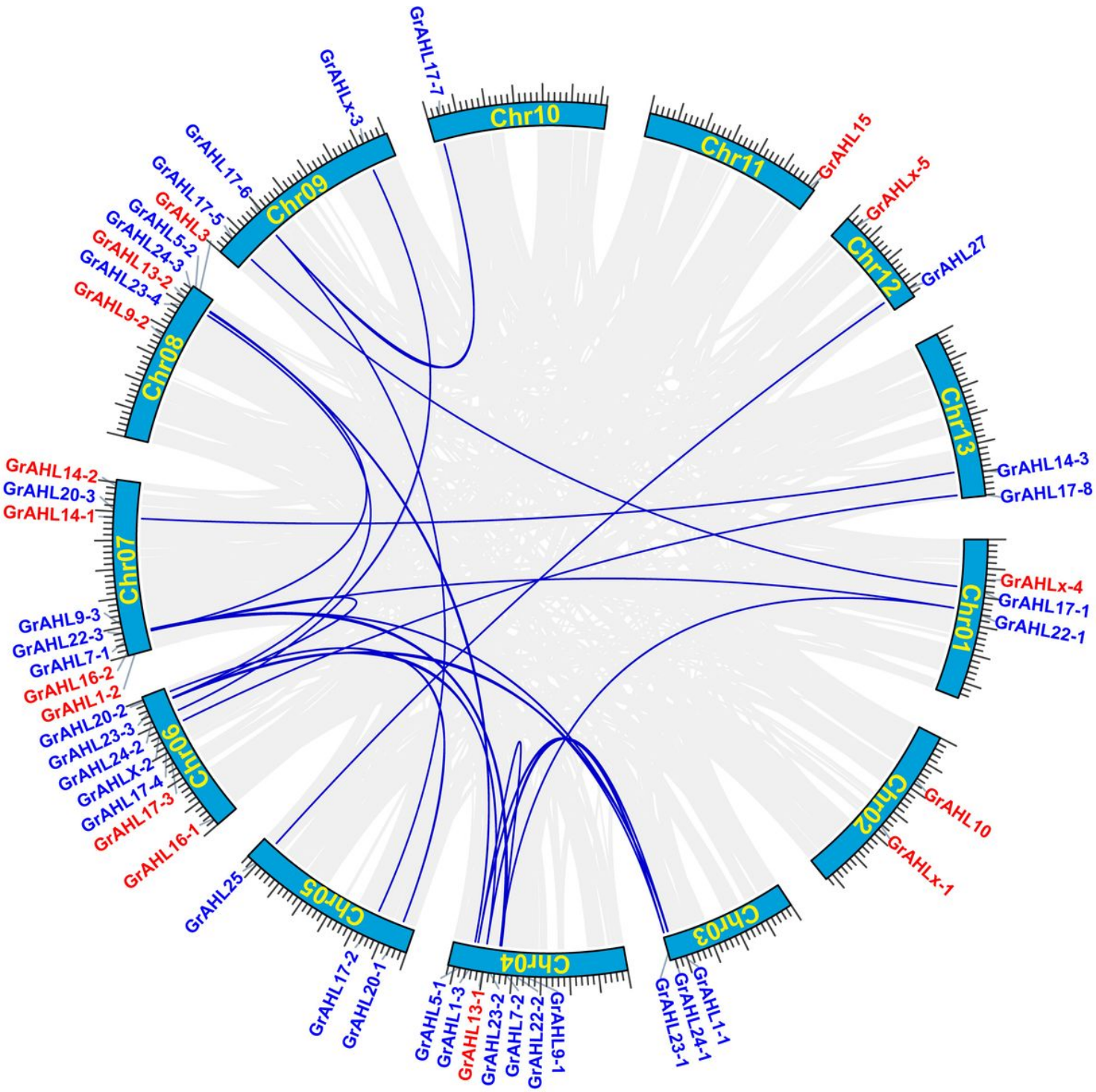

Figure 3

The synteny and collinearity analysis of AHL genes among grape, cacao, and cotton. The positions of AHL genes were depicted in chromosome of V. vinifera (Vv, orange band), T. cacao (Tc, blue band) and G. raimondii (Gr, red band). The Arabic numerals in bars represented different chromosome respectively. The picture was drawn with Circos.

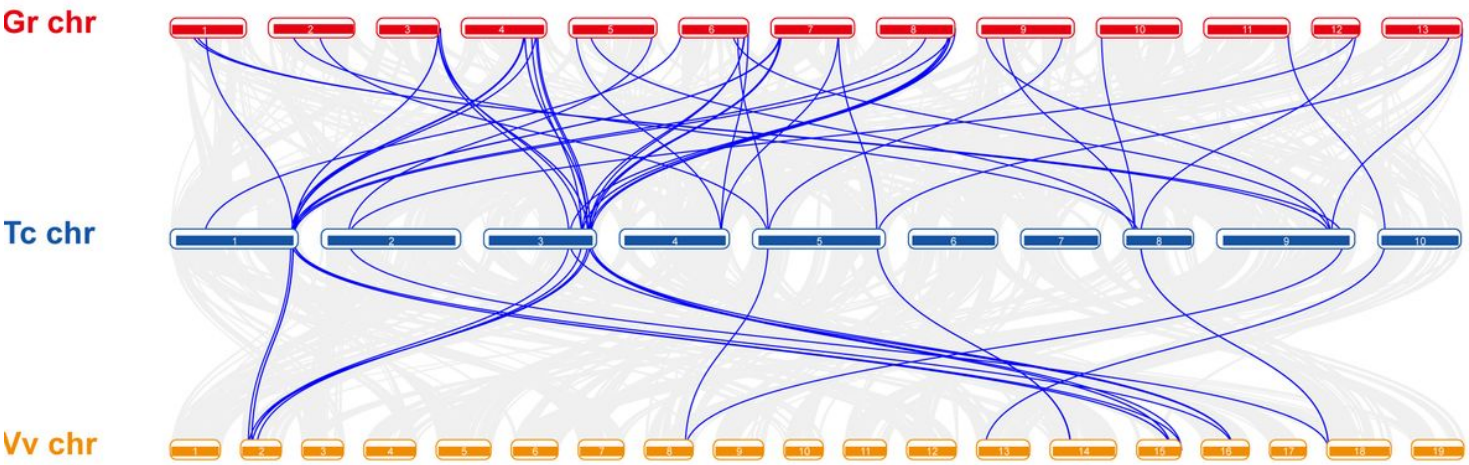

Figure 4

Distribution and gene duplications of GrAHL genes.The scale on the circle is in Megabases. Each colored bar represents a chromosome as indicated. Gene IDs are labeled on the basis of their positions on the chromosomes. AHL name in red indicated the singleton, AHL name in blue indicated the synteny or collinearity among chromosomes. The whole genome duplication (WGD) or segmental duplication was linked by blue lines, gray lines in the background indicated the collinear blocks among different chromosomes.

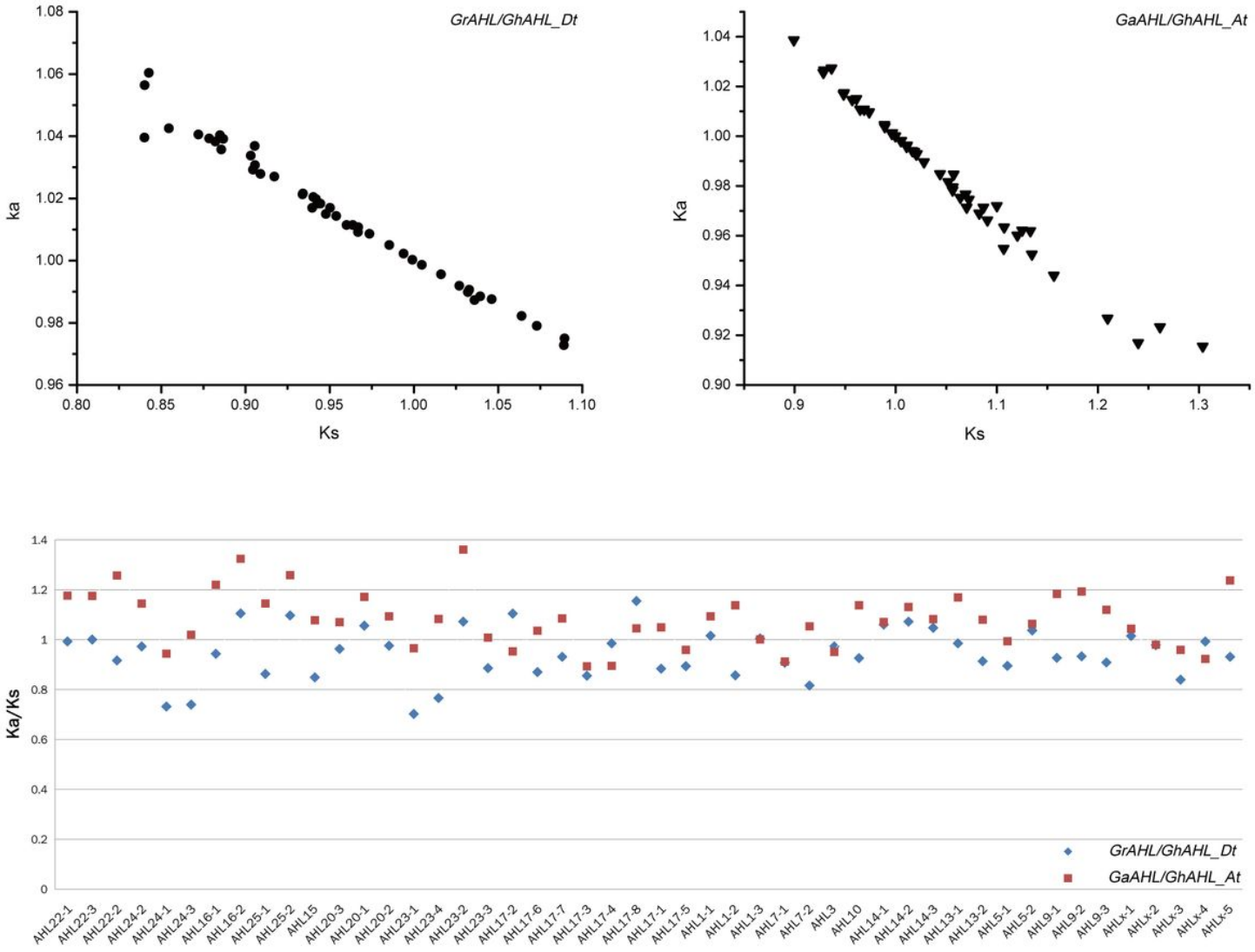

Figure 5

The distribution of non-synonymous (Ka) and synonymous (Ks) nucleotide substitution values of and Ka/Ks ratio of orthologous pairs between GrAHL, GaAHL and GhAHL.

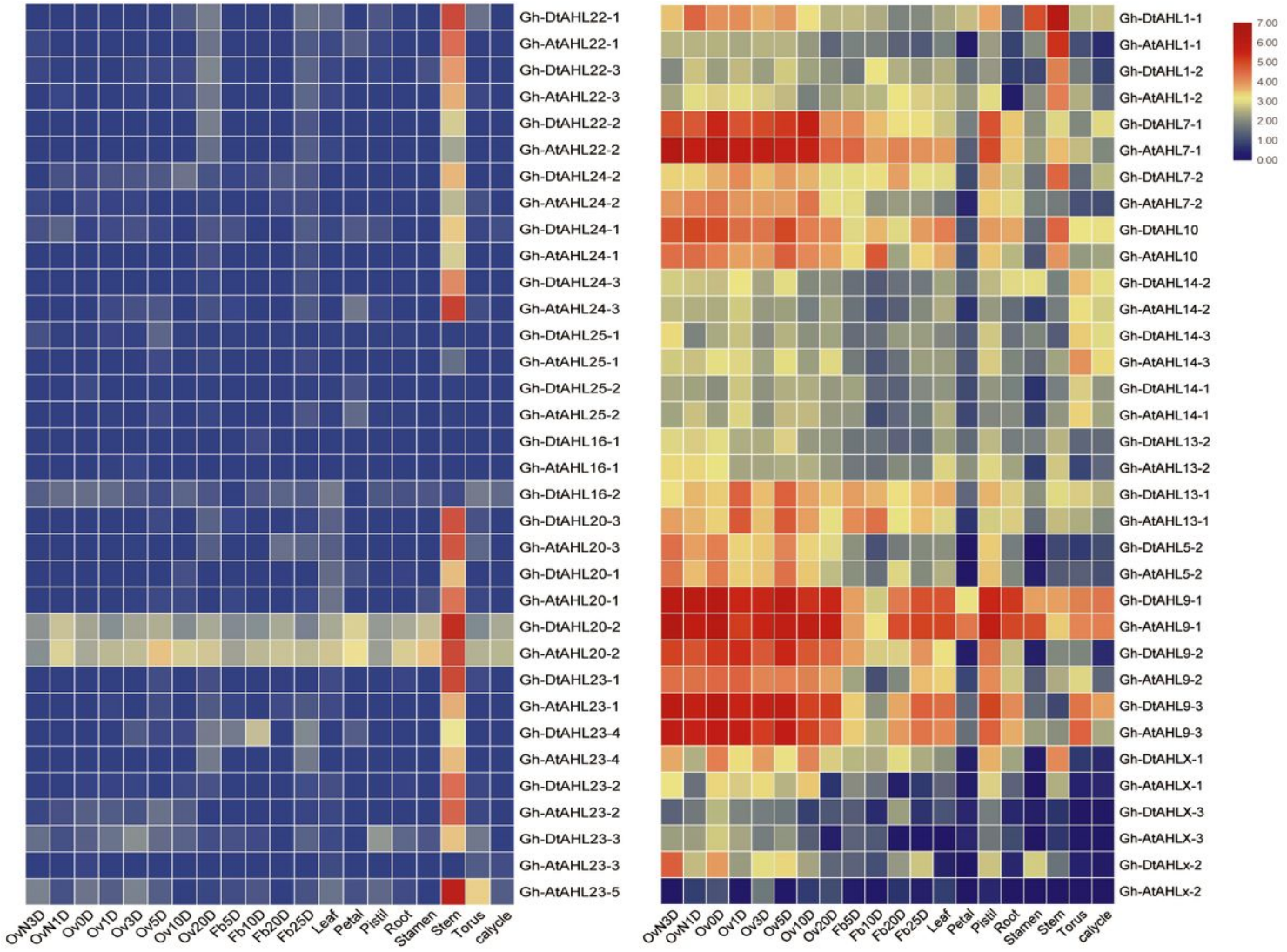

Figure 6

The expression profiles of GhAHL genes. The heatmap was generated on the basis of RNA-seq data from the website (http://www.cottonfgd.com), the color scale was shown at the right of the figure. Higher expression levels were shown in red, and lower in blue. OvN3D, represented the ovule in 3rd days before anthesis, Ov0d= the ovule in 0 day of anthesis, fb5d =the fiber in 5th day after anthesis.

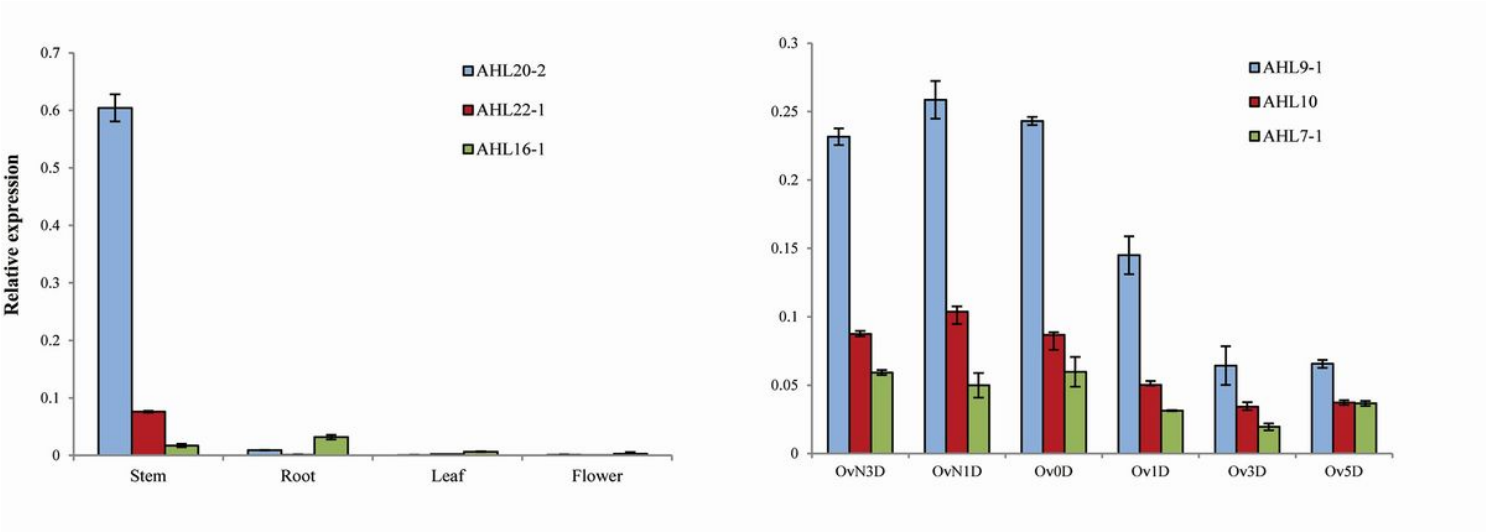

Figure 7

The expression patterns of six AHL genes in G.hirsutum. qRT-PCR was conducted to analyze the relative expression of six AHL genes in stem, root, leaf, flower and ovule(-3,-1,0,1,3,5 PDA).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.docx
- supplement2.docx
- supplement2.docx
- supplement3.docx
- supplement4.docx
- supplement6.docx
- supplement7.tif
- supplement8.tif
- supplement9.tif
- supplement10.tif
- supplement11.tif