Genome-wide identification and characterization of AP2 gene family in *Artemisia annua* L.: Different tissues expression profiles

Yu Liang (✉ liangyu94@126.com)

Dali University <https://orcid.org/0000-0003-4084-1421>

Ying Meng

Shandong University of Traditional Chinese Medicine

Baosheng Liao

China Academy of Chinese Medical Sciences Institute of Chinese Materia Medica

Wenrui He

Chengdu University of Traditional Chinese Medicine

Xiaofeng Shen

China Academy of Chinese Medical Sciences Institute of Chinese Materia Medica

Xiaohui Qiu

Second Clinical Medical College of Guangzhou University of Chinese Medicine: Guangdong Provincial Hospital of Traditional Chinese Medicine

Zhihai Huang

Second Clinical Hospital of Guangzhou University of Chinese Medicine: Guangdong Provincial Hospital of Traditional Chinese Medicine

Baozhong Duan

Dali University

Jiang Xu

China Academy of Chinese Medical Sciences Institute of Chinese Materia Medica

Research Article

Keywords: *Artemisia annua*¹, AP2 gene family², Genome-wide³, co-expression analysis⁴, Regulation of artemisinin synthesis⁵

Posted Date: March 3rd, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1389320/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Artemisia annua produces a mass of artemisinin, an effective drug widely used in malaria treatment. The AP2 transcription factors family which is one of the largest families in higher plants plays significant roles in growth and development regulating plant secondary metabolites and tolerances to environmental stresses. However, the available data on the contributions of the AP2 gene family in the whole genome of *Artemisia annua* is limited. In this study, 195 AP2 genes were identified and compared with AP2 genes from *Arabidopsis thaliana*. These AP2 genes in *A. annua* were classified into five subfamilies: DREB (85), ERF (84), AP2 (21), RAV (3), as well as two Soloists. Chromosome locations and synteny alignment of AP2 genes indicated that whole-genome duplication (WGD) gene retention and tandem repeat events contributed to the expansion of the AP2 family in *A. annua*. The expression profiles of root, stem, leaf, and flower of *A. annua* two strains revealed that ADS genes were co-expressed with AP2 genes. In addition, AaAP2_096 and AaAP2_177 were suggested to have potential in regulating artemisinin content in *Artemisia annua*. AaAP2_096 and AaAP2_177 might be helpful to breeding *A. annua*. In conclusion, our results provide a foundation for the further mechanism and function study of the AP2 gene family in *A. annua*, and promote their potential application in the biosynthesis of artemisinin.

1 Introduction

Artemisia annua Linn. Sp, with abundant artemisinin, is a significant variety of the *Artemisia* genus of the Compositae family. Artemisinin is currently the finest drug for treating cerebral malaria caused by *Plasmodium falciparum* [1]. However, the artemisinin content in *A. annua* is low, accounting for only 0.01% – 0.8% in dry weight, which greatly limits the commercialization of artemisinin [2]. It is crucial to find the key regulators in *A. annua* to increase the yield of artemisinin. It is reported that the AP2 family relates to the biosynthesis of artemisinin by regulating the expression of ADS genes [3]. AaERF1, AaERF2, and AaORA could actively regulate the artemisinin content of *Artemisia annulata* through gene overexpression and RNA interference (RNAi) experiments [3, 4]. The AP2 transcription factor proteins, specifically expressed in the trichomes of *A. annua*, were likely to play an important role in the regulation of the artemisinin biosynthesis pathway and plant growth and development.

Several AP2 family regulatory functions have been thoroughly researched. Plants' development and stress responses are influenced by many members of the AP2 family [5, 6]. The gene expression about organ differentiation, stress signaling pathways, response to biotic and abiotic stress, and metabolite synthesis was closely regulated by the AP2 family in plants [7]. A similar study of burley tobacco and legume using transcription and metabolic histochemistry showed that AP2 TFs genes were the dominant regulatory genes for leaf and root to resist drought [8]. The TtAP2/ERF genes could cope with drought and salinity stresses and might contribute to increase the stress tolerance of durum wheat [9]. In addition, the family AP2 genes control the identity of flower organs, flowering time (TOE1), as well as ontogeny of embryonic cells in plants [5]. The AP2 family plays crucial roles in seed germinating, flower differentiation, and leaf shape [10]. According to the advice, several proteins have been isolated and identified in studies over the past few years. NTERF189, a member of the AP2 family in *Nicotiana*

tabacum, was a key regulator of nicotine biosynthesis [11]. *GAME9/JRE4* gene shared by tomato and potato was the main regulatory factor of steroidal glycoalkaloid biosynthesis [12]. Using molecular biology experiments such as gene cloning, CRISPR-Cas9, and ChIP-seq, researchers had demonstrated that AP2 proteins played a leading role in the transcriptional regulation of a variety of biological processes in plants. The stimulation of biotic and abiotic factors and the development of plants were intimately associated with these biological processes [13].

The AP2 domain defines the APETALA2/ethylene factor family by the shared functional sequence of transcription factors. The AP2 TFs can be identified by the presence of the AP2 domain which consists of about 54 amino acids and is involved in DNA binding. The AP2 family can be further subdivided into 5 subfamilies, referred to as AP2, ERF, DREB, RAV, and Soloist, respectively. All members of the AP2 family shared a specific single 54-amino-acid domain, and most of the AP2 subfamily proteins had two same domains. The RAV subfamily proteins generally contained an AP2 and an associated B3 domain.

Many species have carried out gene family analysis due to a large number of published genomes. For example, the AP2 family was identified 145, 132, 200, and 163 genes, severally, in the *Arabidopsis* [14], grapevine [15], poplar [16], and rice [17] genomes, respectively. However, the regulatory mechanism of AP2 family transcription factors involved in artemisinin biosynthesis was still not fully understood. To better understand the diversity and specificity of the AP2 superfamily, unlock the secrets about the functional and structural evolution of its members as a whole, and their potential contributions and correlations with artemisinin biosynthesis, we performed comprehensive analyses of the AP2 family with *A. annua* in this study.

2 Results

2.1 Identification of AP2 Genes in *A. annua*.

An entire 195 AP2 family genes, named as *AaAP2_001* to *AaAP2_195*, were identified in *A. annua* (Supplementary Table 1). The predicted AaAP2s proteins ranged from 59 (AaAP2_112) to 612 (AaAP2_045) amino acids with an average of 270, and their calculated MW varies from 6.943 to 68.612 kDa with an average of 29.868 kDa. The lowest theoretical PI value was 4.36 for AaAP2_149, whereas the highest PI value was 11.54 for AaAP2_168. PFAM analysis showed that most AaAP2 proteins contained a single AP2 domain, and 19 AaAP2 proteins were possessing additional domains, such as two AP2 domains, Myb DNA-bind 3, and B3 domain. Among them, 175 AP2 genes were predicted to be located in the nucleus, 1 in the endomembrane system, and 19 in the chloroplast (Supplementary Table 2).

2.2 Phylogenetic Classification of the AP2 Gene Family of *A. annua*.

An unrooted phylogenetic tree of *A. annua* was structured by the Neighbor-Joining program. According to the phylogenetic tree and the *Arabidopsis* subfamily, the AP2 family genes of *A. annua* could be separated into five subfamilies (Fig. 1). In this study, 195 AP2 genes were identified by MUSCLE software

and compared with AP2 genes from Arabidopsis. The 195 AP2 genes in *A. annua* were classified into five subfamilies: DREB (85), ERF (84), AP2 (21), RAV (3), as well as two soloists (Supplementary Table 3). A comparison of the AP2 gene family of *A. annua* clearly showed that DREB and ERF are the prominent subfamilies that have more gene members than other subfamilies. Most members of the DREB and ERF subfamily had a single AP2 domain, while only the *AaAP2_039* which was part of the ERF subgroup had 3 AP2 domains. All genes with two AP2 domains belong to the AP2 subfamily, and not all members of this subfamily have two AP2 domains. And all three members of the RAV subfamily have two domains in tandem with AP2 and B3. It is interesting that the *AaAP2_173* subordinated to the ERF subgroup also has two concatenated domains, AP2 and Myb DNA-bind. Generally, the AP2 proteins with high sequence similarity in the same subfamily might develop similar functions. In contrast, the RAV and soloist subfamilies have a greater genetic distance than other groups. Their branches formed two small distinct clades in the phylogenetic tree, which suggested that RAV and soloist might have evolutionary trajectories independent of the other genes.

Additionally, 3 *AaAP2*s (*AaAP2_160*, *AaAP2_046*, and *AaAP2_021*) with high similarity to Arabidopsis *TAR1* (*TRICHOME AND ARTEMISININ REGULATOR 1*) gene and *A. annua* *TAR1* gene were clustered in group DREB. *TAR1* gene played pivotal roles in the regulates of glandular secretory trichomes morphological features and its artemisinin metabolism. *TAR1* expression was mainly concentrated in young leaves and buds and was expressed in the nucleus generally. *TAR1* in *A. annua* regulated the biosynthesis of artemisinin by influencing the expression of key enzyme genes, such as *ADS* and *CYP71AV1* genes [18]. *AaAP2_160*, *AaAP2_046* and *AaAP2_021* can be grouped with the *TAR1* gene in genetic distance proximity. These results suggest that *A. annua* members in group DREB may have similar functions with the *TAR1* gene.

2.3 Conserved Motifs and Gene Structure Analysis of the AP2 Gene Family of *A. annua*.

Only one conserved motif has been identified in *A. annua* *AaAP2* protein sequences (Supplementary Fig. 1). As shown in Supplementary Fig. 1, all genes contained the single same conserved motif, a conserved sequence annotated as an AP2 domain. We found that it was unique in subfamily AP2 to have two same motifs (2 AP2 domains). We found that the conserved gene structure of DREB and ERF subfamilies were almost indistinguishable from each other only based on motifs and exon-intron features (Fig. 2). The members of DREB and ERF subfamilies typically lacked introns or have only one intron, as opposed to the AP2 subfamily, which had multiple introns from 5 to 10. *AaAP2* proteins within the same subfamilies had a similar structure, sequence, and motif, suggesting that AP2 protein structure is more conserved within a specific sub. We could well distinguish between the different subgroups except for DREB and ERF as shown in Fig. 2.

The intron-exon structure of the AP2 gene is preserved as a trace of the evolutionary process, which shows the evolutionary relationship of the AP2 gene family. Genes within different subgroups usually had inconsistent genetic structures. But ERF and DREB subfamilies were very similar which means they might

be more closely related. The structural characteristics of the AP2 subfamily were disparate to ERF in that AP2 subfamily members possessed 1–10 exons (Fig. 2). For example, most members contained 1 exon and 0 introns except *AaAP2_065* which contained 10 exons, and 9 introns which subordinated to the AP2 subgroup. In general, specific domains and roughly the same exon-intron structures were shared by the same AP2 subfamily members, which powerfully verified the authenticity of the subfamily by comparison with the phylogenetic analysis.

2.4 Chromosome Location and Synteny Analysis of the AP2 Gene Family in *A. annua*

To explore whether the AP2 gene family is distributed regularly, the chromosomal locations of all AP2 genes were mapped. The location of genes on chromosomes is shown in Fig. 3. According to karyotype analysis, the chromosome number was set to 9 and the longest chromosome was labeled number 1, while chromosome 9 is the shortest. *AaAP2_169* to *AaAP2_195* cannot be located on any of the 9 chromosomes of *A. annua*, and the remaining 166 AP2 genes were located on 9 chromosomes. We found that 25 genes were mapped onto chr1, which was the longest chromosome; chr7 contained the largest number of AP2 genes (26) in *A. annua*, whereas chr9 had the least numbers with 4 genes (Supplementary Table 1). The AP2 gene family members were not clustered and were instead dispersed on diverse chromosomes. In particular, many genes are distributed on one side of chromosomes 3, 5, and 7. Therefore, the longest chromosome H1 did not have the most genes located, which showed there were no positive associations between chromosome length and the number of AP2 genes.

The synteny analysis of the AP2 gene family in the *A. annua* haplotype genome is shown in Fig. 3. Some tandem duplication events within 20k were observed in the *A. annua* genome, including *AaAP2_037*, *AaAP2_038*, *AaAP2_118*, *AaAP2_119*, *AaAP2_133*, and *AaAP2_134*. The aggregation and enhancement of function caused by gene tandem duplication were common in plants. These results suggested that part of the *AaAP2* genes might be generated by post-WGD gene retention and tandem repeat events, which might be the main driving force of the evolution of *AaAP2*.

2.5 Expression Pattern of AP2 Genes in Different Tissues of *A. annua*.

To further analyze the expression pattern of *A. annua*, we explored the different tissues of *A. annua* between two strains which were 'HAN1' and 'LQ-9'. About half of the *AaAP2* genes had no expression in the 5 tissues of *A. annua*, which showed dark blue in the heatmap of Fig. 4. However, the expression levels of some *AaAP2* genes showed significant differences in the root, stem, leaf, flower, and trichome. The correlation heat map between samples (SupplementaryFigure2) shows that stems, leaves, and flowers of 'HAN1' and 'LQ-9' have a strong correlation, while roots ('HAN1') and trichomes of 'Huhao' (Hu_Trichome) have a weak correlation. For example, *AaAP2_136* showed high expression levels in the 'HAN1' root and relatively mean expression levels in other organizations, but extremely low expression levels in the 'LQ-9' root. *AaAP2_085* and *AaAP2_177* were abundantly expressed in glandular secretory

trichomes. This indicated that it might be involved in tissue development. *AaAP2_085* and *AaAP2_177* showed only high expression levels in trichomes, indicating that they might be associated with trichomes development. In addition, some genes showed lower expression levels in all organs than any other gene such as *AaAP2_077*, *AaAP2_110*, *AaAP2_163*, and so on. *AaAP2_033* and *AaAP2_144* genes showed relatively high expression levels in all tissues. Further, *AaAP2_026* and *AaAP2_056* genes showed high expression levels in the root of 'HAN1'. *AaAP2_033* and *AaAP2_144* genes showed high expression levels in the leaf of 'HAN1'. Transcriptome results from 5 different tissues demonstrated that members of the AP2 gene family might be involved in the morphological development of *A. annua* different tissues because of their extremely significant tissue specificity. The results shown in Fig. 4 were consistent with the transcription abundance data in Supplementary Table 4. According to the heatmap expression profile of *A. annua*, the expression of *AaAP2_107* was only high in the trichomes, the expression of *AaAP2_160* was relatively high in all tissues, but the expression of *AaAP2_151* was almost not in all tissues (Fig. 4). These three genes belong to the DREB subfamily, and they are likely to be closely associated with glandular secretory trichomes development.

2.6 Co-expression analysis of *A. annua* AP2 genes

Based on the WGCNA analysis using *A. annua* RNA-seq datasets showed in Supplementary Table 5. And 10,594 genes were clustered into 29 modules with different colors (Fig. 5. a). We explored the correlation between modules and tissues by graphing a heatmap that showed in Fig. 5. b. After calculating the correlation coefficient, two modules were strongly associated with trichomes, indicating it might be highly contributed to glandular secretory trichomes initiation and formation (Fig. 5. b). The MEyellow module exhibited extremely exclusive correlations with trichome traits ($r = 0.99$, $p\text{-value} = 9e-26$). The 809-gene module MEyellow and MEtan(231) are closely related to trichomes, and they are likely to be involved in the development of glandular secretory trichomes and the synthesis of specific secondary metabolites. These modules could be analyzed further. There are 10 AaAP2 family members in MEyellow module: *AaAP2_017*, *AaAP2_020*, *AaAP2_050*, *AaAP2_054*, *AaAP2_076*, *AaAP2_085*, *AaAP2_096*, *AaAP2_173*, *AaAP2_176*, *AaAP2_177*. These genes are highly expressed in glandular secretory trichomes. The Gene Ontology (GO) analysis showed that these genes from various modules had diverse functions in various facets of plant development. In addition, we also identified magenta, darkred, turquoise, and orange modules significantly and substantially correlated with leaf development of *A. annua*. It suggested genes involved in these modules most probably participated in leaf differentiation and elongation during plant developmental stages.

2.7 Expression patterns of AaAP2 genes in different tissues

The genes of the AaAP2 family regulate both biological and abiotic rejoinder and the synthesis of secondary metabolites, meanwhile, there are countless AaAP2 family associates in *A. annua*. Therefore, we obtained three AaAP2 genes, *AaAP2_020*, *AaAP2_096*, and *AaAP2_177*, by combining multiple aspects. Firstly, these genes were located in the MEyellow module in the weighted gene co-expression network, which was closely related to the development of glandular secretory trichomes as well as the

synthesis of artemisinin. We analyzed the transcriptomic expressions of *ADS*, *CYP71AV1*, *DBR2*, *ADH1*, *ALDH1*, and other key genes in the artemisinin synthesis pathway [19]. In addition, we conducted correlation analysis with the AaAP2 family, to find the AaAP2 gene with high correlation with these key genes. Combining the above two results, we come across *AaAP2_020*, *AaAP2_096*, and *AaAP2_0177* that may regulate artemisinin synthesis and proved these genes by qRT-PCR. The results of the qRT-PCR experiment were shown in Fig. 6. And the data in Fig. 6 were stored in Supplementary Table 7. In both HAN1 and LQ-9, the gene expression level of leaf (HAN1_Y and LQ.9_Y) tissues was relatively high, and the expression level of flower (LQ.9_H) was also high. On the contrary, the expression level of each gene was low or not expressed in stems (HAN1_J and LQ.9_J) and roots (HAN1_G and LQ.9_G) in equally strain. Since glandular secretory trichomes were mainly distributed on leaves and flowers, with little presence on roots and old stems, this result fitted our prediction. *AaAP2_096* and *AaAP2_177* showed comparable expression patterns. They were all accompanied by the expression of *ADS*, a key enzyme gene for artemisinin synthesis. Moreover, these genes were highly expressed in artemisinin-containing tissues.

3 Methods

3.1 Plant materials and genome sequences acquisition

In this study, two different strains of *Artemisia annua*. 'HAN1' and 'LQ-9' were planted in the experimental field of the Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences with seeds provided by the research group. In this study, 195 AaAP2s identified and analyzed were derived from the LQ-9 genome. The genome sequences and gene annotation of 'HAN1' and 'LQ-9' downloaded from GPGD (Global Pharmacopoeia Genome Database) (<http://www.gpggenome.com/>) were used in this study [20]. The *A. annua* genome used in this study was a high-quality chromosomal genome sequenced, assembled, and annotated by our research group. The *A. annua* genome data has been uploaded to GPGD.

3.2 Identification and classification of AP2 genes in *Artemisia annua*

PfamScan (<http://pfam.xfam.org/>) was used to search for candidate members of AP2 genes by searching annotated proteins in the PFAM database (34.0) with evaluate $< 1e-5$ of *A.annua* genomes [21]. Genes with hits to PFAM ID PF00847 (SWIRM/SWIRM-associated) were considered as candidate AP2 genes. Then, AP2 gene models and evidence tracks were integrated and corrected into Apollo for manual inspection [22]. Reference annotations, shreds of evidence including next-generation sequencing, third-generation sequencing data, and gene predictions were integrated into Apollo [23]. Additional evidence tacks based on the alignments between genome sequences and corresponding transcriptome with HISAT2 [24] were added to evaluate the quality of the gene models. All candidate AP2 domains were calibrated manually by comparing the sequence of *A. annua* with BLASTP and BLASTN [25]. Once again, the artificially corrected AP2 family genes were compared globally in the *A. annua* genome by BLASTN

and TBLASTN [25]. These comparison parameters were 50% query length coverage 50% and sequence identity 80% to avoid lost false-negative identification. Finally, all the AP2 genes were reconfirmed by PfamScan to exclude false-positive results. Based on PfamScan results of AP2 domains number and the sequence similarity with *Arabidopsis thaliana*, AP2 genes were divided into 5 subfamilies.

3.3 Sequence information, gene structure, and phylogenetic analysis of AP2 genes

The online tools Compute pI/Mw (<https://web.expasy.org/protparam/>) to calculate the theoretical isoelectric point (pI) and molecular weight (MW). Each AP2 protein subcellular localization was predicted using the online BUSCA tool (<http://busca.biocomp.unibo.it/>). Using MEME Suite Web Server Version 5.1.0 (<https://meme-suite.org/>) to predict the conserved motifs in 195 AaAP2 genes. The Gene Structure Display Server (GSDS 2.0) (Gene Structure Display Server) was applied to analyze the gene structure of the *A. annua* AP2 gene family by genomic sequence and CDS sequence, including the length and number of introns-exons. The chromosomal locations about AP2 genes were computed using the online MapGene2Chrom web v2 tool (http://mg2c.iask.in/mg2c_v2.0/)[26]. To analyze the phylogenetic relationship of AP2 TFs of *A. annua*, multiple sequence alignments were carried out according to domain amino acid sequences using the MUSCLE software. A multiple sequence alignment of *A. annua* AP2 genes was performed and a neighbor-joining (NJ) phylogenetic tree was constructed using MEGA 6 with 1000 bootstrap replicates [27] and painted by Interactive Tree Of Life (iTOL) (<https://itol.embl.de/itol.cgi>) [28]. TBtools was employed to synthesize the NJ phylogenetic tree, motif analysis, and gene structure data of *A. annua* to plot the distribution of the genes [29]. Ultimately, visualized the phylogenetic tree, 5 conserved motifs, and exon-intron structure of 23 genes from different subgroups in *A. annua* by Tltools. In addition, a colossal NJ tree containing 164 additional *Arabidopsis thaliana* AP2 protein sequences and AaAP2s was built using the same method to Separate subgroups. The homology of AP2 genes between *A. annua* and *Arabidopsis* was detected by BLAST. The protein sequences of *Arabidopsis thaliana* were downloaded from EnsemblPlants (<http://plants.ensembl.org/index.html>). The synteny analysis was constructed by using the protein sequence and location information of AaAP2 family genes. Among the set of *A. annua* haplotype genomes, there were 183 pairs of correspondence with identity greater than 75% by BLASTP comparison. We selected the genes that were localized on the chromosomes and they were visualized using the Synteny Visualization tool in Tltools software [29].

3.4 Expression analysis of *Artemisia annua* AP2 genes by RNA sequencing data

Based on the RNA-seq data for *A. annua*, expression data for five different tissues (root, stem, leaf, flower, and trichome) were obtained from GPGD (<http://www.gpggenome.com/>) and shown in Supplementary Table 5. Transcriptome data of root, stem, leaf, and flower were sequenced by our team, and trichome data (Hu_Trichome) were downloaded [30]. Expression analysis was performed based on fragments transcript per million (TPM) values calculating by HISAT2 and StringTie. Hierarchical clustering and construction of heatmaps were performed using R packages “pheatmap” based on Log fold change [31].

We finally used Log (TPM + 1) to process the expression quantity data of different strains and different tissues to draw the expression heatmap.

3.5 Co-expression analysis of *A. annua* AP2 genes

We used the same expression data as RNA-seq to construct gene co-expression analysis. Gene co-expression analysis could help us to find the relationship between each gene and different tissue-specific functions, focusing on the specific sites of artemisinin synthesis - glandular secretory trichomes. To construct the co-expression network, transcripts with the sum of TPM count > 10 counts and filtered according to the variance. Finally, 10,594 genes with the variance > 0.25 were selected from 56,033 genes to use and all the outliers were removed. A co-expression network of *A. annua* genes was built by the Weighted Gene Co-expression Network Analysis (WGCNA) package in R [32]. In the WGCNA analysis, we removed the samples with sizable differences to make the outlier value less than 50000, confirmed RsquaredCut = 0.9, and selected Soft Threshold (power) = 9 according to the system recommendation to construct the co-expression matrix. The calculated co-expression matrix was visualized for modules, and module-trait relationships were finally obtained by associating the modules with traits.

3.6 Expression analysis by qRT-PCR

Total RNA was extracted from *A. annua* plants using RN38 EASYspin Plus Plant RNA Rapid Extraction Kit (Aidlab, Beijing, China). The plant samples were from the remaining parts of frozen transcriptomic sequencing samples, including the roots, stems, and leaves of HAN1 and the roots, stems, leaves, and flowers of LQ-9 (HAN1 flower samples had been used up and no remaining). The extracted RNA was quality-checked by NanoDrop Spectrophotometer 2000 (Thermo Fisher Scientific, Waltham, USA) and 2% agarose gel electrophoresis. According to the instructions, total RNA was transformed into cDNA using the FastKing cDNA Synthesis Kit KP118 (Tiangen, Beijing, China). Finally, the target genes were mixed with ChamQ Universal SYBR qPCR Master Mix (Vazyme, Nanjing, China) and performed real-time quantitative PCR (qRT-PCR) by SLAN-96S automatic fluorescent quantitative PCR (Hongshi, Shanghai, China). QRT-PCR system was 20μl with 1μL template cDNA added. The reaction parameters of qRT-PCR were pre-denaturation at 95°C for 30 seconds; There were 40 cycles in the extension stage with 95°C 10 seconds and 60°C 30 seconds; Finally, the dissociation curve was detected. All primers in the experiment were enrolled in Supplementary Table S6. In qRT-PCR, *A. annua* *β-actin* gene was used as an internal reference gene and LQ-9 root (LG) was used as the control group [33]. The relative gene expression of qRT-PCR was calculated by $2^{-\Delta\Delta CT}$ method [34].

4 Discussion

Many organics' primary developmental processes are controlled by transcription factors (TFs), obligatory roles in diverse biological processes. Plenty of vital cellular processes, such as morphogenesis, metabolism, and resistance to environmental stresses are affected by gene expression regulation. Jofuku first described the AP2 family conservative structural domains in the APETALA2 [35]. The AP2 TFs family, one of the greatest transcription factor families in plants, is diffusely found in eukaryotes. It plays a

pivotal role in regulating abiotic stress responses, flower development [35, 36], Synthesis of secondary metabolites [37], and so on. We confirm that a large number of plant species have been reported about the AP2 gene family but not in *A. annua*.

According to previous studies, the AP2 gene family might rise when an enzyme migrates from bacteria or viruses to plants [38]. It was hypothesized that the transfer of HNH-AP2 endonuclease from bacterial or viral levels to plants may lead to the origin of the AP2/ERF TFs family through the process of transposition and homing [39]. Whole-genome duplication (WGD), tandem repeat, and segmental duplication lead to the expansion of the HNH-AP2 family domains and the number of members during plant evolution.

In this study, as a transcription factors family, the protein of the AP2 genes is relatively small, with an average of only 54 amino acids if there is only one AP2 domain (Supplementary Table 1). According to the structural characteristics of AP2 genes and the similarity of *Arabidopsis thaliana*, *A. annua* was divided into five subgroups [9]. These subfamilies are DREB (85), ERF (84), AP2 (21), RAV (3), and Soloist (2) (Fig. 1). It suggested that the AP2 gene family was conserved in evolutionary history [14, 40, 41]. Among these subgroups, AP2 and RAV subfamilies were of significant particularity. All AaAP2 genes with two AP2 domains were subordinate to the AP2 subfamily. All RAV subgroup members had the B3 domain except for the AP2 domain which was shared by the AP2 family. These results from Fig. 2 about characteristics of subfamily structure are consistent with previous studies [14, 42]. A special case of all genes was *AaAP2_039*, belonging to the ERF subfamily, which had three closely spaced continuous AP2 domains. We believe that *AaAP2_039* is the only gene with three AP2 domains in *A. annua*. This is probably a consequence of the unique and crucial function of *AaAP2_039*. We found that the conserved gene structure of DREB and ERF subfamilies were almost indistinguishable from each other only based on motifs and exon-intron features (Fig. 2). But the AP2 subfamily had a profound difference from the DREB and ERF. The members of DREB and ERF subfamilies typically lacked introns or have only one intron, as opposed to the AP2 subfamily, which had multiple introns up to a maximum of 10. What DREB and ERF subfamilies genes need to respond quickly to environmental choice led to the absence of most introns during plants' growth and development. The introns, belonging to DREB and ERF subfamilies, might have been lost along with functional differentiation during the evolution of higher plants [43]. In this report, the 9 chromosomes of *A. annua* were numbered chr1 to chr9 according to their length, and chr1 was the longest chromosome. The high-quality chromosome level genome of *A. annua* was assembled by our research group, and the chromosome number in this study was consistent with it. The *AaAP2_001* to *AaAP2_168* could be located on the 9 different chromosomes, except for the remaining 29 AP2 genes, which could not be located on any chromosomes. Some genes were concentrated in the chromosomes, such as chr1, chr3, chr6, and chr7 (Fig. 3). Each line in the haplotype genome synteny analysis of AP2 genes indicated that the pair of genes are highly homologous, which they were most likely from the same ancestral genes. In a pair of genes linked by a magenta line that the two genes were very close in proximity possibly due to tandem duplication, and if the two genes were on different and distant chromosomes, they might be caused by tandem repeat and WGD-retention events (Fig. 4). Whole-

genome duplications and tandem repeats usually give rise to multiple copies of genes in a gene family [44, 45].

As a TF of the AP2 family, *TAR1* could bind to the GCC-box promoter elements and increase the content of effective substances by stimulating the expression of genes such as *ADS*. For example, *AaAP2_160*, *AaAP2_107*, and *AaAP2_151*, which are similar to *TAR1*, can bind to cis-elements in the *ADS* and *CYP71AV1* genes promoter to participate in the Artemisinin synthesis pathway [18]. *AaAP2_107* was specifically expressed in trichomes (Fig. 5). Results inferred tandem repeat and WGD-retention events played momentous roles in the formation of the AP2 gene family. For example, *AaAP2_076* and *AaAP2_075* were the results of tandem repeats, a primeval replication event. The expression levels of *AaAP2_136* and *AaAP2_056* were very high and specific in the roots. However, roots were not the main site of artemisinin production in *A. annua*. These two genes might be related to tissue formation and organ differentiation, and had a low correlation with artemisinin synthesis (SupplementaryFigure2). In the co-expression analysis, the genes of different modules showed strong tissue specificity. The Results implied that their function might be related to the division of labor among different tissues (Fig. 6. b). We found that not only the *ADS* and *CYP71AV1* that can regulate amorpho-4,11-diene and artemisinic alcohol, but also the *DBR2* promoting the accumulation of dihydroartemisinic alcohol are concentrated in the MEyellow module. We also gained 10 AaAP2 family members from MEyellow module. MEyellow module showed a significant correlation with glandular secretory trichomes, a factory for artemisinin production and storage (Fig. 6. b). And the genes in this module are likely to be closely related to the whole artemisinin pathway, which has great potential for exploration. In Fig. 7. e, one particular gene is *AaAP2_177*. The Identity between *AaAP2_177* and *AaORA* was 98.9%. In the previous study, *AaORA* is a trichome-specific transcription factor that regulated the expression of *ADS*, *CYP71AV1*, and *DBR2* genes also increased the production of artemisinin [3]. *AaORA* was expressed in the young tissues of *A. annua*, such as bud, young leaf, and flower bud. QRT-PCR test in this study also verified this conclusion, *AaAP2_177* in HAN1 and LQ-9 lines were highly expressed in leaf and flower, and almost no expression in stem and root (Fig. 7. e). There were more glandular secretory trichomes on the surfaces of leaves and flowers, and the content of artemisinin was also higher, resulting in higher expression of regulatory genes. The correlation between RNA-seq data of *AaAP2_020*, *AaAP2_096*, and *ADS* gene expression was greater than 0.8. However, in the qRT-PCR experiment, the results showed that *AaAP2_096* and *AaAP2_177* had a stronger correlation with *ADS* expression. Although the overall qRT-PCR data were considered to be consistent with RNA-seq data. We compared the qRT-PCR expression results with the transcriptome expression data of the heatmap. Considering all tissues with qRT-PCR results, the Spearman correlation was 0.703 (P-value = 0.003), indicating a significant correlation between the two results, and the expression trend of each tissue was consistent. Screening and expression of AaAP2s members, highly correlated with the *ADS* gene, will facilitate our understanding of the signaling network in regulating sesquiterpene biosynthesis in *A. annua*. Plants' development and secondary metabolite synthesis were regulated by some members of the AaAP2 TFs family. Therefore, the further describe the molecular characteristics of the AaAP2 family in *A. annua*, which is of great significance for understanding the regulation of the artemisinin biosynthetic pathway in this study.

Artemisinin was an effective ingredient in treating malaria. However, the utilization of *A. annua* was limited due to its low yield. Since artemisinin content was closely related to the changes of glandular secretory trichomes [46], the development of trichomes of *A. annua* deserved further study. Following, we will continue to verify the direct relationship between AaAP2 TFs family members (*AaAP2_096* and *AaAP2_177*) and key genes such as *ADS*. It was of great potential in plants secondary metabolism regulatory to work the downstream of transcription factors.

5 Conclusion

In this study, 195 AP2 family members of *Artemisia annua* were identified based on genome-wide analysis. The 195 AaAP2 genes were grouped into 5 subfamilies, AP2, DREB, ERF, RAV, and Soloist. We implemented the Sequence characteristics and gene structure analysis as well as chromosomal location, synteny, expression pattern analysis, and co-expression analysis of 195 AaAP2 genes undergoing different tissues. Genome distribution and chromosomal localization indicate that the expansion of AaAP2 might be contributed by tandem repeat and whole-genome duplication. The study found that the expression profiles of *AaAP2_096*, *AaAP2_177*, and the key genes of artemisinin synthesis *ADS* were similar in different tissues of *A. annua*, showing enormous potential in regulating artemisinin synthesis. Our findings provided abundant data support for functional evolution and global molecular characterization of AP2 gene in the *A. annua* different tissues.

Declarations

Acknowledgments

We are grateful to Shuai Guo, Xuejiao Liao, and Qianwen Liu for their assistance and discussion.

• Funding

This work was supported by the National Natural Science Foundation of China and Karst Science Research Center of Guizhou Province (U1812403-1), Innovation Team Project for Traditional Chinese Medicine Resources & Ethnic Medicine of Dali University (ZKLX2019318), and the Discipline construction of Materia medica genomics by Fundamental Research Funds for the Central public welfare research institutes (L2019016).

• Conflict of Interest

We declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

• Author Contributions

XJ conceived and designed the experiments. LY, MY, LBS, HWR, SXF, QXH contributed to reagents, materials, and analysis tools. XJ, LBS, DBZ, and HZH guided the whole manuscript. LY analyzed data and

wrote the article. LY and MY conducted experiments. All authors read and approved the final manuscript.

References

1. Weathers PJ, Elkholy S, Wobbe KK (2006) Artemisinin: The biosynthetic pathway and its regulation in *Artemisia annua*, a terpenoid-rich species. *Vitro Cell Dev Biology - Plant* 42:309
2. Czechowski T, Larson TR, Catania TM, Harvey D, Wei C, Essome M, Brown GD, Graham IA (2018) : Detailed Phytochemical Analysis of High- and Low Artemisinin-Producing Chemotypes of *Artemisia annua*. *Frontiers in Plant Science*.9.
3. Lu X, Zhang L, Zhang F, Jiang W, Shen Q, Zhang L, Lv Z, Wang G, Tang K (2013) AaORA, a trichome-specific AP2/ERF transcription factor of *Artemisia annua*, is a positive regulator in the artemisinin biosynthetic pathway and in disease resistance to *Botrytis cinerea*. *New Phytol* 198:1191
4. Yu Z, Li J, Yang C, Hu W, Wang L, Chen X (2012) The Jasmonate-Responsive AP2/ERF Transcription Factors AaERF1 and AaERF2 Positively Regulate Artemisinin Biosynthesis in *Artemisia annua* L. *Mol Plant* 5:353
5. Licausi F, Ohme-Takagi M, Perata P (2013) APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. *New Phytol* 199:639
6. Phukan UJ, Jeena GS, Tripathi V, Shukla RK (2017) : Regulation of Apetala2/Ethylene Response Factors in Plants. *Frontiers in Plant Science*.8.
7. Huang PY, Catinot J, Zimmerli L (2016) Ethylene response factors in *Arabidopsis* immunity. *J Exp Bot* 67:1231
8. Rabara RC, Tripathi P, Reese RN, Rushton DL, Alexander D, Timko MP, Shen QJ, Rushton PJ (2015) : Tobacco drought stress responses reveal new targets for Solanaceae crop improvement. *BMC Genomics*.16.
9. Faraji S, Filiz E, Kazemitabar SK, Vannozzi A, Palumbo F, Barcaccia G, Heidari P (2020) The AP2/ERF Gene Family in *Triticum durum*: Genome-Wide Identification and Expression Analysis under Drought and Salinity Stresses. *Genes* 11:1464
10. Horstman A, Willemsen V, Boutilier K, Heidstra R (2014) AINTEGUMENTA-LIKE proteins: hubs in a plethora of networks. *Trends Plant Sci* 19:146
11. Shoji T, Kajikawa M, Hashimoto T (2010) Clustered Transcription Factor Genes Regulate Nicotine Biosynthesis in Tobacco. *The Plant Cell* 22:3390
12. Nakayasu M, Shioya N, Shikata M, Thagun C, Abdelkareem A, Okabe Y, Ariizumi T, Arimura G, Mizutani M, Ezura H, Hashimoto T, Shoji T (2018) JRE4 is a master transcriptional regulator of defense-related steroidal glycoalkaloids in tomato. *Plant J* 94:975
13. Rashid M, Guangyuan H, Guangxiao Y, Hussain J, Xu Y (2012) AP2/ERF Transcription Factor in Rice: Genome-Wide Canvas and Syntenic Relationships between Monocots and Eudicots. *Evol Bioinform Online* 8:321

14. Sakuma Y, Liu Q, Dubouzet JG, Abe H, Shinozaki K, Yamaguchi-Shinozaki K (2002) DNA-Binding Specificity of the ERF/AP2 Domain of Arabidopsis DREBs, Transcription Factors Involved in Dehydration- and Cold-Inducible Gene Expression. *Biochem Biophys Res Commun* 290:998
15. Public C (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463
16. Zhuang J, Cai B, Peng R, Zhu B, Jin X, Xue Y, Gao F, Fu X, Tian Y, Zhao W, Qiao Y, Zhang Z, Xiong A, Yao Q (2008) Genome-wide analysis of the AP2/ERF gene family in *Populus trichocarpa*. *Biochem Biophys Res Commun* 371:468
17. Sharoni AM, Nuruzzaman M, Satoh K, Shimizu T, Kondoh H, Sasaya T, Choi I, Omura T, Kikuchi S (2011) Gene Structures, Classification and Expression Models of the AP2/EREBP Transcription Factor Family in Rice. *Plant and Cell Physiology* 52:344
18. Tan H, Xiao L, Gao S, Li Q, Chen J, Xiao Y, Ji Q, Chen R, Chen W, Zhang L (2015) TRICHOME AND ARTEMISININ REGULATOR 1 Is Required for Trichome Development and Artemisinin Biosynthesis in *Artemisia annua*. *Mol Plant* 8:1396
19. Yang K, Monafared RS, Wang H, Lundgren A, Brodelius PE (2015) The activity of the artemisinic aldehyde $\Delta 11(13)$ reductase promoter is important for artemisinin yield in different chemotypes of *Artemisia annua* L. *Plant Mol Biology* 88:325
20. Liao B, Hu H, Xiao S, Zhou G, Sun W, Chu Y, Meng X, Wei J, Zhang H, Xu J, Chen S (2021) : Global Pharmacopoeia Genome Database is an integrated and mineable genomic database for traditional medicines derived from eight international pharmacopoeias. *Science China Life Sciences*.
21. Dunn N, Unni D, Diesh C, Munoz-Torres M, Harris NL, Yao E, Rasche H, Holmes IH, Elisk CG, Lewis SE (2019) : Apollo: Democratizing genome annotation. *PLOS Comput. Biol.*
22. Wang L, Huang Q, Zhang L, Wang Q, Liang L, Liao B (2020) : Genome-Wide Characterization and Comparative Analysis of MYB Transcription Factors in *Ganoderma* Species. *G3 (Bethesda, Md.)*.10. 2653
23. Campbell MS, Holt C, Moore B, Yandell M (2014) : Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics*.48.
24. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protocols* 11:1650
25. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinf* 10:421
26. Zhang S, Jiang Z, Chen J, Han Z, Chi J, Li X, Yu J, Xing C, Song M, Wu J, Liu F, Zhang X, Zhang J, Zhang J (2021) : The cellulose synthase (CesA) gene family in four *Gossypium* species: phylogenetics, sequence variation and gene expression in relation to fiber quality in Upland cotton. *Molecular genetics and genomics: MGG*. 296. 355
27. Riaz MW, Lu J, Shah L, Yang L, Chen C, Mei XD, Xue L, Manzoor MA, Abdullah M, Rehman S, Si H, Ma C (2021) : Expansion and Molecular Characterization of AP2/ERF Gene Family in Wheat (*Triticum aestivum* L.). *Frontiers in Genetics*.12.

28. Letunic I, Bork P (2019) Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 47:W256
29. Chen C, Xia R, Chen H, He Y (2018) : TBtools - an integrative toolkit developed for interactive analyses of big biological data. *BioRxiv*.289660
30. Shen Q, Zhang L, Liao Z, Wang S, Yan T, Shi P, Liu M, Fu X, Pan Q, Wang Y, Lv Z, Lu X, Zhang F, Jiang W, Ma Y, Chen M, Hao X, Li L, Tang Y, Lv G, Zhou Y, Sun X, Brodelius PE, Rose J, Tang K (2018) The Genome of *Artemisia annua* Provides Insight into the Evolution of Asteraceae Family and Artemisinin Biosynthesis. *Mol Plant* 11:776
31. Wang L, Lu C, Fan M, Liao B (2021) : *Coriopsis trogii* hydrophobin genes favor a clustering distribution and are widely involved in mycelial growth and primordia formation. *Gene*.802.
32. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf* 9:559
33. Hu Z, Han Z, Song N, Chai L, Yao Y, Peng H, Ni Z, Sun Q (2013) Epigenetic modification contributes to the expression divergence of three TaEXPA1 homoeologs in hexaploid wheat (*Triticum aestivum*). *The New phytologist* 197:1344
34. Livak KJ, Schmittgen TD (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2 – $\Delta\Delta CT$ Method. *Methods* 25:402
35. Jofuku KD, den Boer BG, Van Montagu M, Okamoto JK (1994) Control of *Arabidopsis* flower and seed development by the homeotic gene APETALA2. *Plant Cell* 6:1211
36. Kuluev B, Avalbaev A, Nurgaleeva E, Knyazev A, Nikonorov Y, Chemeris A (2015) Role of AINTEGUMENTA-like gene NtANTL in the regulation of tobacco organ growth. *J Plant Physiol* 189:11
37. Buseman CM, Tamura P, Sparks AA, Baughman EJ, Maatta S, Zhao J, Roth MR, Esch SW, Shah J, Williams TD, Welti R (2006) Wounding Stimulates the Accumulation of Glycerolipids Containing Oxophytodienoic Acid and Dinor-Oxophytodienoic Acid in *Arabidopsis* Leaves. *Plant Physiol* 142:28
38. Tang Y, Qin S, Guo Y, Chen Y, Wu P, Chen Y, Li M, Jiang H, Wu G (2016) : Genome-Wide Analysis of the AP2/ERF Gene Family in Physic Nut and Overexpression of the JcERF011 Gene in Rice Increased Its Sensitivity to Salinity Stress. *PLOS ONE*. 11. e150879
39. Magnani E, Sjolander K, Hake S (2004) From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. *Plant Cell* 16:2265
40. Nakano T, Suzuki K, Fujimura T, Shinshi H (2006) Genome-Wide Analysis of the ERF Gene Family in *Arabidopsis* and Rice. *Plant physiology (Bethesda)*. 140:411
41. Cao D, Lin Z, Huang L, Damaris RN, Yang P (2021) Genome-wide analysis of AP2/ERF superfamily in lotus (*Nelumbo nucifera*) and the association between NnADAP and rhizome morphology. *BMC Genomics* 22:171
42. Zhou M, Tang Y, Wu Y (2012) Genome-Wide Analysis of AP2/ERF Transcription Factor Family in *Zea Mays*. *Curr Bioinf* 7:324

43. Yamada Y, Nishida S, Shitan N, Sato F (2020) Genome-wide identification of AP2/ERF transcription factor-encoding genes in California poppy (*Eschscholzia californica*) and their expression profiles in response to methyl jasmonate. *Sci Rep* 10:18066
44. Licausi F, Giorgi FM, Zenoni S, Osti F, Pezzotti M, Perata P (2010) Genomic and transcriptomic analysis of the AP2/ERF superfamily in *Vitis vinifera*. *BMC Genomics* 11:719
45. Yadav CB, Muthamilarasan M, Pandey G, Prasad M (2015) Identification, Characterization and Expression Profiling of Dicer-Like, Argonaute and RNA-Dependent RNA Polymerase Gene Families in Foxtail Millet. *Plant Mol Biology Report* 33:43
46. Arsenault PR, Vail D, Wobbe KK, Erickson K, Weathers PJ (2010) Reproductive development modulates gene expression and metabolite levels with possible feedback inhibition of artemisinin in *Artemisia annua*. *Plant Physiol* 154:958

Figures

Figure 1

Neighbor-Joining phylogenetic tree of the AP2 gene family in *Artemisia annua*. AaAP2s were divided into five subfamilies: AP2, ERF, DREB, RAV, and soloist.

Figure 2

Motif composition and gene structure of 23 *Artemisia annua* AP2 proteins. The five subfamilies had different motif and gene structures.

Figure 3

Haplotype genome synteny analysis of *AP2* genes in *Artemisia annua*. The central string connected the pairs of syntenic genes. There were 183 pairs of genes with more than 75% homology. The gray line connected all the syntenic pairs, and the dark red lines showed duplicated by tandem duplication. The gray rings on the outside represented the different locations of the 9 chromosomes and *AP2* genes.

Figure 4

Heat map of all AP2 genes expression in *Artemisia annua* different tissues. RNA-seq expression heatmap was finally processed and plotted by the Log (TPM+1) of R package 'Pheatmap' for different lines and tissues. Different colors represented expression level; Red represented high expression and blue represented low expression. HAN1_G, HAN1_J, HAN1_Y, and HAN1_H represented different tissues, roots, stems, leaves, and flowers of HAN1, respectively. LQ.9_G, LQ.9_J, LQ.9_Y, and LQ.9_H represented different

tissues, roots, stems, leaves, and flowers of LQ-9, respectively. Hu_Trichome represented the glandular secretory trichomes hairs of ‘HuHao’.

Figure 5

Weighted Gene Co-expression Network (WGCNA) analysis results in *Artemisia annua*. **a** Module identification by weighted gene co-expression network analysis in *Artemisia annua*. The same color represented the same module. **b** Module-trait relationships of *Artemisia annua*. Modules with similar functions of different gene compositions were associated with different tissues, and red represented a high correlation between this module and this tissue. For example, MEyellow is strongly related to Trichome, representing genes in MEyellow that were intimately linked to Trichome function.

Figure 6

The relative expression level of *AaAP2* genes in different tissues. The internal reference gene was β -ACTIN, and the control group was LQ-9 root. Error bars indicate the standard error (SE) between replicates. **a** The relative expression of “LQ-9” ACTIN, ADS, AaAP2_020, AaAP2_096, and AaAP2_177. **b** The relative expression of “HAN1” ACTIN, ADS, AaAP2_020, AaAP2_096, and AaAP2_177.

Supplementary Files

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

- [SupplementaryMaterials.zip](#)