frontiers in Plant Science

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/fpls.2020.563237

GelFAP: Gene Functional Analysis Platform for *Gastrodia elata*

Jiaotong Yang[1]†, Qiaoqiao Xiao[1]†, Jiao Xu[1], Lingling Da[2], Lanping Guo[3], Luqi Huang[3], Yue Liu[4], Wenying Xu[2], Zhen Su[2], Shiping Yang[2], Qi Pan[1], Weike Jiang[1] and Tao Zhou[1]**

[1] Source Institute for Chinese and Ethnic Materia Medica, Guizhou University of Traditional Chinese Medicine, Guiyang, China, [2] College of Biological Sciences, China Agricultural University, Beijing, China, [3] National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China, [4] College of Horticulture, Qingdao Agricultural University, Qingdao, China

Gastrodia elata, also named Tianma, is a valuable traditional Chinese herbal medicine. It has numerous important pharmacological roles such as in sedation and lowering blood pressure and as anticonvulsant and anti-aging, and it also has effects on the immune and cardiovascular systems. The whole genome sequencing of *G. elata* has been completed in recent years, which provides a strong support for the construction of the *G. elata* gene functional analysis platform. Therefore, in our research, we collected and processed 39 transcriptome data of *G. elata* and constructed the *G. elata* gene co-expression networks, then we identified functional modules by the weighted correlation network analysis (WGCNA) package. Furthermore, gene families of *G. elata* were identified by tools including HMMER, iTAK, PfamScan, and InParanoid. Finally, we constructed a gene functional analysis platform for *G. elata*[1]. In our platform, we introduced functional analysis tools such as BLAST, gene set enrichment analysis (GSEA), and *cis*-elements (motif) enrichment analysis tool. In addition, we analyzed the co-expression relationship of genes which might participate in the biosynthesis of gastrodin and predicted 19 mannose-binding lectin antifungal proteins of *G. elata*. We also introduced the usage of the *G. elata* gene function analysis platform (GelFAP) by analyzing *CYP51G1* and *GFAP4* genes. Our platform GelFAP may help researchers to explore the gene function of *G. elata* and make novel discoveries about key genes involved in the biological processes of gastrodin.

Keywords: *Gastrodia elata*, co-expression network, functional module, gene functional analysis platform, functional enrichment analysis

OPEN ACCESS

Edited by:
Maud Fagny,
UMR 7206 Eco-Anthropologie et Ethnobiologie (EAE), France

Reviewed by:
Jinpeng Wang,
Institute of Botany, Chinese Academy of Sciences, China
Won Kyong Cho,
Seoul National University, South Korea
Etienne Delannoy,
UMR 9213 Institut des Sciences des Plantes de Paris Saclay (IPS2), France

***Correspondence:**
Tao Zhou
taozhou88@163.com
Jiaotong Yang
y_jiaotong@163.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Plant Science

Received: *18 May 2020*
Accepted: *16 September 2020*
Published: *22 October 2020*

Citation:
Yang J, Xiao Q, Xu J, Da L, Guo L, Huang L, Liu Y, Xu W, Su Z, Yang S, Pan Q, Jiang W and Zhou T (2020) GelFAP: Gene Functional Analysis Platform for Gastrodia elata. Front. Plant Sci. 11:563237. doi: 10.3389/fpls.2020.563237

INTRODUCTION

Gastrodia elata, a kind of perennial herb of Orchidaceae, is one of the traditional Chinese herbal medicines. The growth cycle of *G. elata* is generally about 3 years, including the development stages of the seed, protocorm, juvenile tuber, immature tuber, mature tuber, and scape (Yuan et al., 2018). *G. elata* is a typical heterotrophic plant, which has a symbiotic relationship with at least two fungi during its life cycle. One is *Mycena* that offers nutrition for the seed germination of *G. elata*, and the other is *Armillaria mellea* that offers nutrition and energy for the vegetative propagation corms of *G. elata* development into tubers (Xu, 1981, 1989). The mannose-binding

[1] http://www.gzybioinformatics.cn/Gel

lectin antifungal proteins of *G. elata* (GAFPs) play important roles in its growth during *G. elata* and *A. mellea*, establishing a stable symbiotic association (Yuan et al., 2018). *G. elata* has important functions such as in sedation and lowering blood pressure and as anticonvulsant and anti-aging, and it also has effects on the immune and cardiovascular systems. Its pharmacological action makes it widely used in clinical settings (Shan et al., 2016). As an important medicinal plant, *G. elata* has many active chemical ingredients, such as gastrodins, 4-hydroxybenzyl alcohols, vanillyl alcohols, vanillins, polysaccharides, sterols, and organic acids (Shan et al., 2016). Among them, gastrodin is one of the important components for its beneficial effects. Gastrodin biosynthesis pathway from toluene to 4-hydroxytoluene can be catalyzed by monooxygenase of cytochrome P450 (CYP450) (Carmona et al., 2009), and then CYP450 further catalyzes the oxidation of 4-hydroxytoluene to p-hydroxybenzyl alcohol; finally, glycogenase is synthesized through glycosyltransferase (UGT) (Tsai et al., 2016). Therefore, exploring the function of genes that can catalyze the synthesis of gastrodin from the CYP450 and UGT gene family will help to explore the molecular mechanism of gastrodin biosynthesis.

The development of high-throughput sequencing technology has greatly enriched the research methods in the field of life sciences, and it not only improves the efficiency of scientific research but also promotes the development of basic research. In the past decade, whole genome sequencing had been completed in typical model plants and crops, and many species even owned their gene function analysis platforms, which were established by the integration of multiple omics data. Reiser et al. (2017) had established the Arabidopsis Information Resource (TAIR) platform, which covered detailed functional annotation information of each gene and various auxiliary analysis tools, thereby greatly improving research efficiency in scientific fields. Tian et al. (2018) had also built a gene function analysis platform MCENet, which contained a large number of *Zea mays* gene co-expression networks constructed by transcriptomic data and gene function analysis tools, so as to study gene function and synergy between different genes. Recently, Wang et al. (2020) analyzed the genomics data of 13 species in 9 genera of Malvaceae, such as genome-wide association analysis site (GWAS) information and single nucleotide mutation site (SNP) information, as well as a total of 374 sets of transcriptomic and proteomic data, and established a functional genomic hub for Malvaceae plants, which provided a powerful online analysis tool for scientists to carry out mallow family gene function analysis. Therefore, it is necessary to develop a gene function analysis platform for *G. elata* by integrating various annotations, which may contribute to deeper gene function analysis and mining.

The whole genome sequencing of *G. elata* was completed in 2018 (Yuan et al., 2018), making a certain accumulation in transcriptome data of *G. elata*. We collected the transcriptome data of 39 samples, and of these samples, 27 were from the Sequence Read Archive (SRA) in the National Center for Biotechnology Information (NCBI) and 12 were generated by our group. In order to use these data adequately and effectively, we constructed the co-expression network of *G. elata* and identified its functional modules to predict gene function. Furthermore, we constructed a *G. elata* gene function analysis platform (GelFAP) with analysis tools, such as BLAST, GSEA, and *cis*-element enrichment analysis tools, which will help to further explore the novel functions of genes in *G. elata*.

MATERIALS AND METHODS

RNA-Seq Data Processing

The quality control of *G. elata* transcriptome data was performed by FastQC software (version 0.11.2). After removing the unqualified transcriptome data samples, we used TopHat (version 2.1.0) (Trapnell et al., 2009) to map the clean reads to the reference genome and calculated the fragments per kilobase of exon model per million reads mapped (FPKM) values by the Cufflinks software (version 2.2.1) (Trapnell et al., 2010).

Co-expression Network Construction

Here, the Pearson correlation coefficient (PCC) algorithm was used to construct the gene co-expression networks of *G. elata*. We firstly calculated the correlation between different genes according to the expression values of genes in all 37 samples. Genes with high correlation had similar expression patterns in different samples, which could be considered as gene pairs with co-expression relationship. Then, we calculated the network density and the scale-free topology fitting index R^2 based on the PCC changes and selected the appropriate PCC to construct the gene co-expression network based on the maximizing scale-free topology fitting index R^2 and relative small network density. Correlation can be evaluated by PCC, and the formula is as follows:

$$\mathrm{PCC}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

PCC_{xy} is the Pearson correlation coefficient between gene x and gene y, n represents the total number of samples, x_i represents the FPKM values of gene x in the i sample, y_i represents the FPKM value of gene y in sample i, $\overline{x}$ represents the average value of gene x in n samples, and $\overline{y}$ is the average value of gene y in n samples.

Gene Set Enrichment Analysis

Gene set enrichment analysis was used as a method for annotating gene sets by calculating the degree of overlap between a specific gene set and various clearly defined gene sets and then defining an enriched gene set by the hypergeometric test, Fisher's exact test, or χ^2 test. Multiple test correction methods for GSEA, including Yekutieli, Bonferroni, Hochberg, Hochberg, Hommel, and Holm, could be used to reduce the false positive rate of GSEA analysis. These methods could perform enrichment analysis on gene ontology (GO) annotations, Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations, and Pfam domain of specific gene sets (Yi et al., 2013). The

hypergeometric test was set as a default method for users to perform gene set enrichment analysis. The formula is as follows:

$$P = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}}$$

N represents the number of genes in *G. elata*, K represents the number of genes in an annotated gene set a, n represents the number of genes submitted by the user, and k represents the overlapped number of genes submitted by the user and the same genes in gene set a.

Enrichment Analysis of *Cis*-Elements (Motifs)

For the genes which needed to be analyzed, we used the following steps to calculate the Z score and P value of each motif. Firstly, we scanned the promoter region (1k, 2k, or 3k from annotated genes based on the gene structure "gff" file) of each gene that was submitted by the user and obtained the number of matches for each motif. Secondly, we selected genes to form a gene list from *G. elata* genome for 1,000 times randomly, and the number of genes was equal to the number of users who have submitted. Thirdly, we scanned the 3-kb promoter region of each gene list and calculated the average number of each motif. Finally, we calculated the Z score and P value of each motif based on the following formula. If the P value was less than 0.05, it meant that the motif was significantly enriched.

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$P \text{ value} = 1 - P \text{ norm}\left(\overline{X}, \mu, \frac{\sigma}{\sqrt{n}}\right)$$

Module Identification and Annotation

We used the weighted gene correlation network analysis (WGCNA) package (Langfelder and Horvath, 2008) of R language to identify the functional modules. The process mainly included four steps. Firstly, we defined the gene co-expression correlated matrix, which weighted the correlation between genes, and determined the software threshold β based on the maximizing scale-free topology fitting index (R^2). Secondly, the blockwiseModules function was used to construct a scale-free network, and then module partition analysis was executed to identify functional modules. Thirdly, modules were defined by the dynamic tree cutting algorithm. Lastly, modules with high similarity were merged to get the final modules. Through this package, we identified the functional modules of *G. elata* co-expression network and further annotated their functions *via* gene set enrichment analysis.

Orthologous Protein Prediction and Protein–Protein Interaction Network Construction

InParanoid (Sonnhammer and Ostlund, 2015) was a software developed by Perl script for constructing orthologous groups, and its normal operation could not do without the BLAST software. We used InParanoid software (Sonnhammer and Ostlund, 2015) to predict orthologous relationship between rice/maize and *G. elata* with a cutoff over 60% bootstrap. We then mapped the protein–protein interaction (PPI) network of maize and rice to *G. elata* to construct *G. elata* PPI networks.

Gene Family Classification

We used the localized iTAK software to predict the transcription factors and transcription regulators of *G. elata* with default parameters, and the operation command was "iTAK.pl+protein_sequence." We downloaded the hidden Markov model file of the conserved domain of ubiquitin proteases from the Ubiquitin and Ubiquitin-like Conjugation Database (UUCD) (Gao et al., 2013) and used the HMMER software to predict the ubiquitin proteases of *G. elata*. The *e*-value parameter used in this calculation process was derived from the threshold recommended by the UUCD (Gao et al., 2013). In order to predict EAR motif-containing proteins and CYP450 proteins, we first collected 20,542 EAR motif-containing proteins and 19,221 CYP450 protein sequences from the PlantEAR (Yang et al., 2018) and CYP450 databases (Nelson, 2009), respectively. Then, we predicted the orthologous relationship between collected proteins and *G. elata* proteins by InParanoid (bootstrap >60%) and further defined the EAR motif-containing proteins and CYP450 proteins based on the orthologous relationship.

Search and Visualization Platform Construction

GelFAP was constructed based on CentOS Linux, Apache server, MySQL database, and PHP language. The software used for network visualization in the platform was a JavaScript package Cytoscape.js with open resources (Franz et al., 2016).

PLATFORM CONTENTS

Data Resources and Functional Annotation

Gastrodia elata genomic data, including 3,779 scaffold sequences, gene location files, gene sequences, 18,969 transcript sequences, and 18,969 protein sequences, was derived from the National Genomics Data Center (NGDC) (Accession number: GWHAAEX00000000) of China produced by the National Resource Center for Chinese Materia Medica of China Academy of Chinese Medical (Yuan et al., 2018). The gene functions of *G. elata* were annotated by comparing nucleic acids or protein sequences with various functional annotation databases, including nr, KOG, TAIR (Reiser et al., 2017), COG, Swiss-Prot, and TrEMBL (**Figure 1A**). In addition, 27 transcriptome data samples were obtained from the SRA in NCBI (Accession

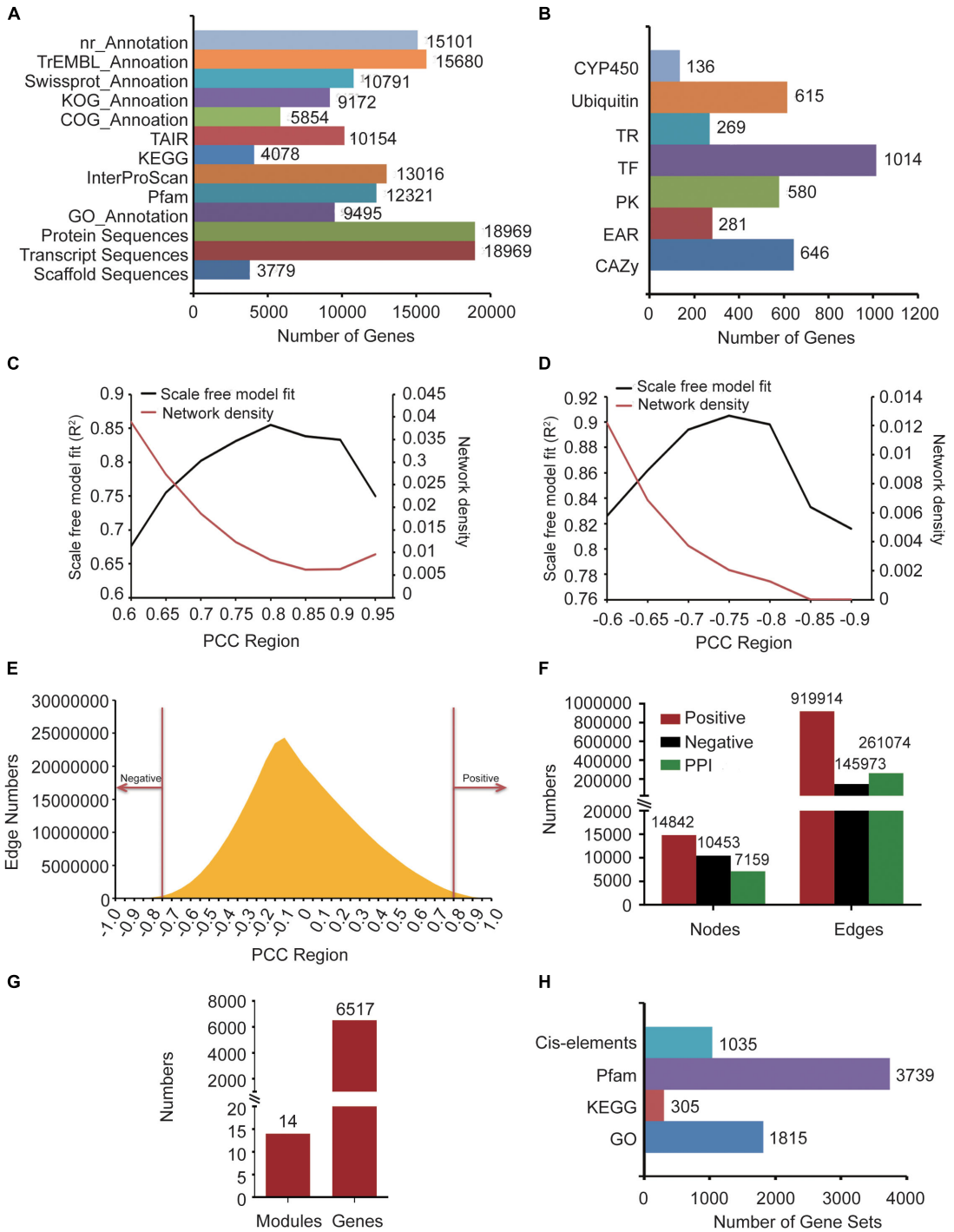

FIGURE 1 | The information about *G. elata* gene function analysis platform. **(A)** Gene function annotation information. **(B)** Gene family classification information. **(C)** Network density and scale-free model fitting (R^2) of the positive co-expression network based on changing Pearson correlation coefficient (PCC) cutoffs. **(D)** Network density and scale-free model fitting (R^2) of the negative co-expression network based on changing PCC cutoffs. **(E)** Distribution diagram of the relationship between PCC and the number of edges. **(F)** Statistics of nodes and edges in the positive co-expression network, negative co-expression network, and PPI network. **(G)** Predicted gene functional modules and involved genes. **(H)** The background gene sets of the GSEA and motif enrichment analysis tools.

number: SRP064423, SRP108465 and SRP118053) and 12 samples were produced by our group. We used the InterProScan (Jones et al., 2014) software to obtain GO terms of 9,495 genes and InterProScan domain annotations of 13,016 genes. The GO annotations were obtained from Gene Ontology Consortium (Gene Ontology Consortium, 2015). Pfam domain annotation

information of 12,321 genes was predicted by the local PfamScan tool (El-Gebali et al., 2019). KEGG orthology annotation information of 4,078 genes was predicted by GhostKOALA (Kanehisa et al., 2016), which was supported by the KEGG website. Finally, the orthologous relationship between *G. elata* and *Arabidopsis thaliana* was analyzed by the InParanoid tool, and *Arabidopsis thaliana* annotation information of 10,154 genes in *G. elata* was obtained (**Figure 1A**).

Gene Family Identification

Pfam is a protein family database, which contained multiple sequence alignment results and hidden Markov model (HMM) profiles of conserved regions from many gene families (El-Gebali et al., 2019). HMMER is a homolog searching tool based on HMM profiles (Potter et al., 2018). The gene families could be identified by combining Pfam with HMMER. Several platforms could also be used to identify gene families; for example, the analysis tools provided by the iTAK website were used for the identification of transcriptional regulators and protein kinases (Zheng et al., 2016), and HMM profiles offered by the UUCD database were used to identify members of the ubiquitin protease family (Gao et al., 2013). In addition, gene families could also be predicted by the orthologous relationship between different species.

To identify the CYP450 gene family numbers, 20,657 CYP450 protein sequences were downloaded from the CYP450 website (Nelson, 2009). Then, we constructed a library according to the downloaded CYP450 protein sequences and aligned the *G. elata* protein sequences with this library. From the results, we obtained 1,455 protein sequences whose *e*-value was less than 1e-5. Among them, 136 protein sequences with the CYP450 domain (PF00067.21) were identified as candidate members of the CYP450 family by HMMER. We used the iTAK software to identify the transcription factors, transcription regulators, and protein kinases of *G. elata* and obtained 1,014 transcription factors, 269 transcription regulators, and 580 protein kinases. We also used UUCD's HMM profile to predict the ubiquitin proteases of G. elata, and 615 ubiquitin proteases were identified. To identify the carbohydrate-active enzymes (CAZy), we downloaded the genes of A. thaliana CAZy gene family from the CAZy database (Lombard et al., 2014), matched the CAZy gene family to *G. elata* according to their orthologous relationship, and predicted 646 CAZy genes of *G. elata* (**Figure 1B**). We also collected the EAR motif-containing proteins of 71 plants from the PlantEAR platform (Yang et al., 2018) and identified 281 EAR motif-containing proteins in *G. elata* according to their orthologous relationship (**Figure 1B**).

Network Construction and Functional Module Identification

Co-expression Network

After removing the non-compliant transcriptome data samples by FastQC tools, we obtained 39 *G. elata* transcriptome data samples, including RNA-seq samples of SRP108465, SRP064423, SRP279888, and SRP118053 in SRA (**Supplementary Table S1**). The reads of RNA-seq samples were mapped to the *G. elata* genome and detailed alignment information was obtained by TopHat (**Supplementary Table S1**). In addition, the FPKM expression values of genes in each sample were obtained by computation using the Cufflinks software. Then, we calculated the PCC value between every two genes in different samples by WGCNA package of R language. Biological networks are usually scale-free networks and the network density is relatively low. Based on this principle, we analyzed PCC value over 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9 and 0.95 to evaluate the scale-free model fitting index R^2 and network density of the positive co-expression network. PCC $>$ 0.8 had the largest scale-free model fitting index (R^2) and the network density was relatively low (**Figure 1C**). We also chose the PCC threshold of the negative co-expression network based on the same method (**Figure 1D**). Finally, we chose PCC $>$ 0.8 and PCC $<$ −0.75 to determine the positive co-expression network and the negative co-expression network, respectively, (**Figure 1E**). We obtained a positive co-expression network with 14,842 nodes and 919,914 edges and a negative co-expression network with 10,453 nodes and 145,973 edges (**Figure 1F**).

Protein–Protein Interaction Network

The PPI network of maize and rice had been constructed in recent years (Zhu et al., 2016; Liu et al., 2017). So, we constructed the *G. elata* PPI network by predicting the orthologous relationship between maize and *G. elata* and mapped the maize PPI network to *G. elata*. By the same method, we also mapped the rice PPI network to *G. elata*. Finally, we obtained a PPI network with 7,159 nodes and 261,074 edges (**Figure 1F**).

Functional Module Identification

The co-expression network we constructed covered 14,842 genes, so we used the WGCNA to divide these genes into modules. WGCNA is a method used to construct a gene co-expression network based on gene expression profiles. By evaluating the relationship between soft threshold and scale-free model fitting index, we chose 7 as the soft threshold (**Supplementary Figure S1A**). Similarly, the relationship between soft threshold and mean connectivity showed that a soft threshold of 7 had a lower mean connectivity (**Supplementary Figure S1B**). Finally, we merged the modules after performing the dynamic tree cutting algorithm and then further identified gene functional modules based on the similarity between modules (**Supplementary Figure S1C**). We obtained 14 functional modules with 6,517 genes (**Figure 1G**).

Functional Enrichment Analysis Tools

We annotated *G. elata* genes by gene sets of 1,815 GO annotations, 305 KEGG orthology and 3,739 Pfam (**Figure 1H**). Then, we constructed the GSEA online tool by the algorithm described in the "Materials and Methods" section.

Motifs are short and conserved sequences of the gene promoter region. It could be recognized by various transcription factors and participated in the regulation of gene expression. We also collected 1,035 motifs from the PlantEAR (Yang et al., 2018) and ccNET platforms (**Figure 1H**; You et al., 2017). Using the motif analysis algorithm in the "Materials and Methods" section,

we constructed an online motif enrichment analysis tool, which could perform motif analysis for the gene of *G. elata*.

The Structure of GelFAP

Based on the constructed gene co-expression networks, gene family classification, and functional analysis tools, the *G. elata* gene function analysis platform was constructed. The platform contained six main sections, namely Home, Browse, Gene family, Tools, KEGG, and Download and Help (**Figure 2**). Among them, there were network search and module search secondary menu functions under the network. The Tools section contained four secondary menus – Search, BLAST analysis, GSEA analysis, and cis-element analysis. The Gene family section contained CYP450, transcription factors, protein kinases, ubiquitin proteases, carbohydrate-active enzyme families, and EAR motif-containing proteins. The Pathway section contained pathways predicted by GhostKOALA (Kanehisa et al., 2016). In addition, the platform also provided the Download and Help page to assistant users to obtain data sources and help. The construction of the platform may contribute to the functional analysis of *G. elata* genes.

APPLICATION

Analysis of Putative Gastrodin Biosynthesis-Related Genes

The gastrodin biosynthesis may be regulated by CYP450, UGT, PAL, C4H, 4-HBS, and ADH family genes (Bai et al., 2016; Tsai et al., 2016). As shown in **Supplementary Figure S2**, many genes in this pathway had an obvious co-expression relationship. The PAL gene had a co-expression relationship with the C4H, CYP450, and ADH genes, and the CYP450 gene also had a co-expression relationship with UTG and ADH. Therefore, there may be an important synergistic relationship between them and they further participated in the regulation of gastrodin biosynthesis (**Supplementary Figure S2**).

A previous study had indicated that the CYP51G1 gene may be involved in the biosynthesis of gastrodin (Tsai et al., 2016), so we guessed that the function of this gene may be regulated by transcription factors that targeted on its upstream. We used the motif enrichment analysis tool to predict the transcription factors that might target on the CYP51G1 gene promoter region and found that multiple transcription factors were significantly enriched, including DRE1, MADS, and HD-zip transcription factors (**Supplementary Figure S3**). Therefore, these transcription factors may be the most probable genes that participated in the biosynthesis of gastrodin by regulating the CYP51G1 gene.

We selected the top 300 genes co-expressed with Arabidopsis CYP51G1 from the ATTED-II database (Obayashi et al., 2018) and compared them with the top 300 co-expressed genes of *G. elata* CYP51G1 (**Supplementary Figure S4**). The results demonstrated that there were 19 pairs of orthologous relationship. It had been reported that many genes of Arabidopsis had different functions (**Supplementary Figure S4**). For example, CPI1 (AT5G50375) was related to plant defense response (Cao et al., 2020), mMDH1 (AT1G53240) may be related to plant response to low temperature (Nakaminami et al., 2014), PGD1 (AT1G64190) could regulate the growth of Arabidopsis (Lim et al., 2009), TBL35 (AT5G01620) was related to xylan acetylation and growth (Yuan et al., 2016), and ARA12 (AT5G67360) was

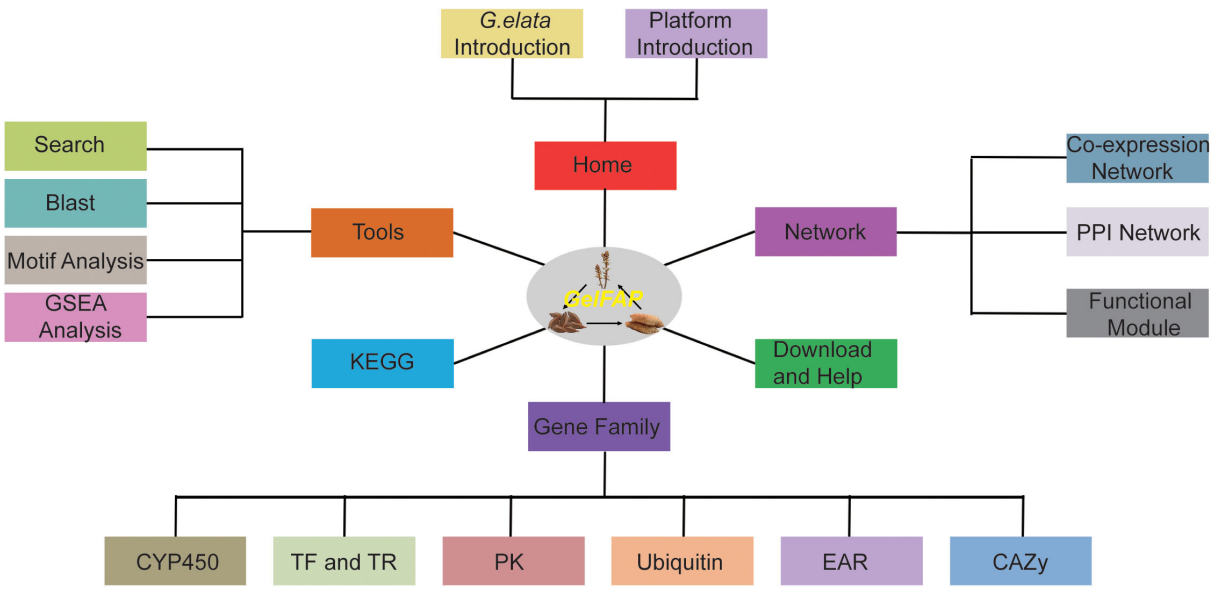

FIGURE 2 | The framework structure of *G. elata* gene function analysis platform. GelFAP included six main sections. Home section was the introduction of *G. elata* and the platform. The network section contained the co-expression network, PPI network, and function modules. The gene family section contained CYP450 family genes, transcription factors, transcription regulators, protein kinases, ubiquitin proteasomes, CAZy genes, and EAR motif-containing proteins. Tools section included Search, BLAST, motif analysis, and GSEA toolkit. KEGG, Download and Help became section separately.

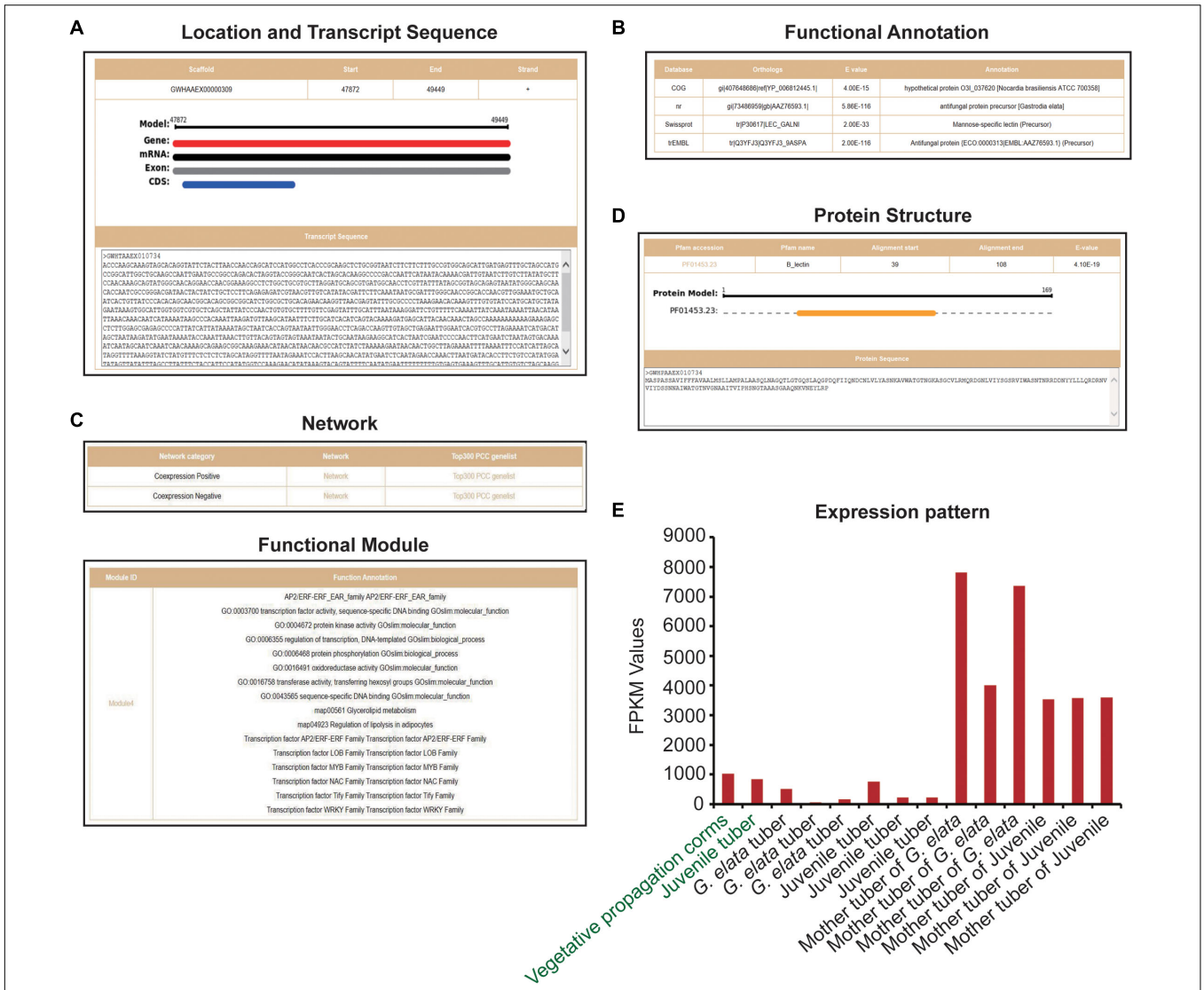

FIGURE 3 | Gene details of *GAFP4*. **(A)** Location and transcript sequences. **(B)** Functional annotations. **(C)** Network and functional modules. **(D)** Protein structure. **(E)** Expression pattern of *GAFP4* gene; green represents transcriptome samples of SRP064423, and black represents transcriptome samples produced by us.

related to release mucilage of seed coat (Rautengarten et al., 2008). Therefore, these reported genes in Arabidopsis may help to predict the function of *G. elata* CYP51G1 gene.

GAFP Identification and Functional Analysis

We obtained 12 *G. elata* mannose-binding lectin antifungal protein (GAFP) sequences from previous researches (Wang et al., 2016; Wang et al., 2019) and GenBank. By comparing these sequences with *G. elata* protein sequences, we obtained 23 protein sequences (*e*-value < 1e-3) and further identified them by the protein domain B_lection (PF01453.23). Finally, 19 *G. elata* proteins were identified as mannose-binding lectin antifungal proteins (**Supplementary Table S2**). The heterologous expression of *G. elata* GAFP4 gene (GWHGAAEX010734) in Arabidopsis thaliana could increase the resistance against *Botrytis cinerea*, and the heterologous expression of GAFP4 in cotton could also increase the resistance against Verticillium wilt (Wang et al., 2016; Wang et al., 2019). Here, we took *G. elata* antifungal protein GAFP4 as an example to analyze its functions by GelFAP. We searched the gene details and obtained the structure information and transcript sequences (**Figure 3A**), annotation information (**Figure 3B**), networks and functional modules (**Figure 3C**), protein structure and sequences (**Figure 3D**), and expression values (**Figure 3E**). We found that this gene had only one exon and CDS, and gene length was 1,557bp (**Figure 3A**). In addition, the functional annotation information indicated that this gene was annotated as an antifungal protein in the nr and TrEMBL databases (**Figure 3B**). Protein structure and sequence information suggested that this protein had a B_lectin domain. Related researches showed that the protein with B_lectin domains

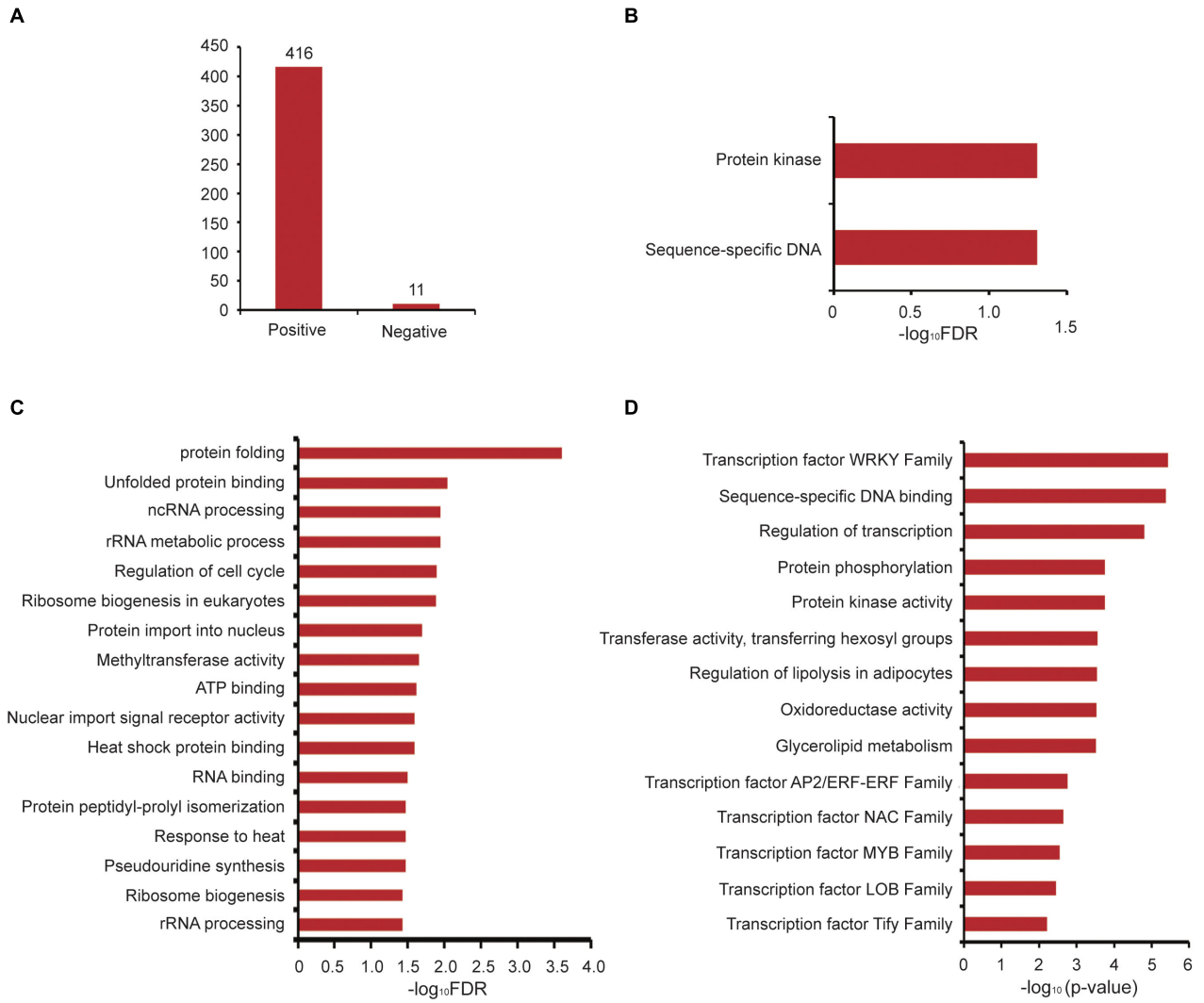

FIGURE 4 | Co-expression and functional module analysis of *GAFP4*. **(A)** The number of positive and negative co-expressed genes of *GAFP4* gene. **(B)** GSEA of *GAFP4* positive co-expressed genes. **(C)** GSEA of *GAFP4* negative co-expressed genes. **(D)** GSEA annotation for the module that contained *GAFP4*.

had antibacterial and antiviral functions (Cox et al., 2006; Sun et al., 2016; Wang et al., 2016; Wang et al., 2019; Yin et al., 2019; **Figure 3D**). Therefore, this domain may be an important structure for GAFP4 to perform its function.

Next, we analyzed GAFP4 gene function by the co-expression network, and the search results showed that GAFP4 had a positive co-expression relationship with 416 genes and a negative co-expression relationship with 11 genes (**Figure 4A** and **Supplementary Table S3**). GSEA of GAFP4 positive co-expressed genes revealed that this gene might have functions of protein kinase activity and sequence-specific DNA binding (Fisher's exact test, FDR $<$ 0.05) (**Figure 4B**). Therefore, GAFP4 may be co-expressed with several transcription factors (TFs) to perform its DNA-binding function. GSEA of GAFP4 negative co-expressed genes revealed its possible function in heat response, protein folding, methyltransferase, regulation of cell cycle, and so on (Fisher's exact test, FDR $<$ 0.05) (**Figure 4C**). In addition, we obtained a function module that contained GAFP4 (**Supplementary Table S4**). GSEA of this module showed significant enriched transcription factor family members, including ERF, MYB, and WRKY families. Moreover, protein kinase activity, transferase activity, oxidoreductase activity, and glycerolipid metabolism were also enriched in the module (Fisher's exact test, P value $<$ 0.05) (**Figure 4D**). When plants were infected by bacteria or viruses, the plant transcription factor families ERF (Wang et al., 2018; Zhu et al., 2019), MYB (Ibraheem et al., 2015; Shan et al., 2016), WRKY (Chen et al., 2013; Peng et al., 2016; Wang et al., 2017; Gao et al., 2018; Liu et al., 2018; Li et al., 2020), and protein kinase (Kim and Hwang, 2011; Shen et al., 2012) showed response functions. Therefore, GAFP4 may be co-expressed with many antibacterial TFs or form functional modules with TFs to further play its role in antibacterial defense response. Therefore, by analysis of GAFP4 gene in GelFAP, we found that it might have antibacterial effect functions. At

present, its antibacterial function has been verified in cotton and Arabidopsis (Wang et al., 2016; Wang et al., 2019), and many other functions still need to be explored in the future.

DISCUSSION

Gastrodia elata is a valuable traditional Chinese herbal medicine and has numerous important pharmacological roles. The whole genome sequencing of *G. elata* has been completed in recent years and its transcriptome data also has a certain accumulation (Tsai et al., 2016; Yuan et al., 2018). In this study, we firstly used the genome and transcriptomes of *G. elata* to construct *G. elata* gene co-expression networks and functional modules and provided related gene function analysis and annotation tools, including the BLAST search tool, GSEA tool, and motif enrichment analysis tool. The gene co-expression networks were of great significance for exploring gene functions, such as comparing networks between orthologous gene pairs in model specie and *G. elata*, which could provide more information for gene function researches. Similarly, gene function enrichment analysis tools also played important roles in *G. elata* gene functional researches. For example, gene enrichment analysis tools could analyze possible downstream functions of differentially expressed genes in the transcriptome. Finally, the gene families such as CYP450, transcription factors, protein kinases, ubiquitin proteases, and carbohydrate-active enzymes were classified and predicted, and the results were integrated into the *G. elata* gene functional analysis platform. Therefore, our platform can provide more data sources and analysis methods for researchers to study the gene function of *G. elata*, which may improve the efficiency of the research for *G. elata* genes.

Gastrodia elata established a symbiotic relationship with Armillaria during the growth process, and it was reported that GAFPs played important roles in establishing this relationship, but which kind of GAFPs were not mentioned. We identified 19 *G. elata* GAFPs based on the sequence information provided by the platform, and provided candidate genes for the follow-up research on the establishment of symbiotic relationship. Furthermore, we took the *GAFP4* gene as an example to introduce the application method of the platform. The analysis results indicated that *GAFP4* might be involved in various regulatory processes including antibacterial, and it had also been reported to have the function of antibacterial (Wang et al., 2016; Wang et al., 2019). Therefore, the platform we built has a certain feasibility and practicality.

The *G. elata* gene function analysis platform is established by us for the first time. Users can submit their interesting genes to the platform and then obtain information of various existing and processed annotations. However, there is still much room for improvement in the accumulation of omics data. In the future, we will continue to update and maintain the *G. elata* gene function analysis platform, such as collecting and integrating more transcriptome, proteome, metabolome data, etc. We expect that this platform will contribute to the study of molecular mechanisms in the process of gastrodin biosynthesis, and further help to solve the problems about variety and quality improvement of *G. elata*.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP064423, https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP108465, https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP118053, and https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP279888.

AUTHOR CONTRIBUTIONS

JY designed this study. JY and QX constructed the platform and completed the draft. LD, LG, and JX produced and processed the transcriptome. ZS, WX, and YL participated in the construction of this platform. SY and QP participated in the revision of the manuscript. TZ, WJ, and LH directed this work and provided financial support.

FUNDING

This work was supported by the ability establishment of sustainable use for valuable Chinese Medicine Resources (Grant No. 2060302), the High-level Innovative Talents of Guizhou Province of China (Qian Ke He Platform and Talent [2018]5638), Guizhou Education Department Innovation Group Major Research Projects (Qian Jiao He KY Zi [2018]022), Ph.D. Startup Foundation of Guizhou University of Traditional Chinese Medicine [2019]141 and [2020]32, the Science and Technology Project in Guizhou Province of China (Qian Ke He Platform and Talent [2019]5611), and National Natural Science Foundation of China [81960694].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2020.563237/full#supplementary-material

REFERENCES

Bai, Y., Yin, H., Bi, H., Zhuang, Y., Liu, T., and Ma, Y. (2016). De novo biosynthesis of Gastrodin in *Escherichia coli*. *Metab. Eng.* 35, 138–147. doi: 10.1016/j.ymben.2016.01.002

Cao, Y., He, Q., Qi, Z., Zhang, Y., Lu, L., Xue, J., et al. (2020). Dynamics and endocytosis of Flot1 in *Arabidopsis* require CPI1 function. *Int. J. Mol. Sci.* 21:1552. doi: 10.3390/ijms21051552

Carmona, M., Zamarro, M. T., Blazquez, B., Durante-Rodriguez, G., Juarez, J. F., Valderrama, J. A., et al. (2009). Anaerobic catabolism of aromatic

compounds: a genetic and genomic view. *Microbiol. Mol. Biol. Rev.* 73, 71–133. doi: 10.1128/MMBR.00021-08

Chen, L., Zhang, L., Li, D., Wang, F., and Yu, D. (2013). WRKY8 transcription factor functions in the TMV-cg defense response by mediating both abscisic acid and ethylene signaling in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 110, E1963–E1971. doi: 10.1073/pnas.1221347110

Cox, K. D., Layne, D. R., Scorza, R., and Schnabel, G. (2006). Gastrodia antifungal protein from the orchid Gastrodia elata confers disease resistance to root pathogens in transgenic tobacco. *Planta* 224, 1373–1383. doi: 10.1007/s00425-006-0322-0

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., and Bader, G. D. (2016). Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 32, 309–311. doi: 10.1093/bioinformatics/btv557

Gao, J., Bi, W., Li, H., Wu, J., Yu, X., Liu, D., et al. (2018). WRKY transcription factors associated with NPR1-mediated acquired resistance in barley are potential resources to improve wheat resistance to *Puccinia triticina*. *Front. Plant Sci.* 9:1486. doi: 10.3389/fpls.2018.01486

Gao, T. S., Liu, Z. X., Wang, Y. B., Cheng, H., Yang, Q., Guo, A. Y., et al. (2013). UUCD: a family-based database of ubiquitin and ubiquitin-like conjugation. *Nucleic Acids Res.* 41, D445–D451. doi: 10.1093/nar/gks1103

Gene Ontology Consortium (2015). Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056. doi: 10.1093/nar/gku1179

Ibraheem, F., Gaffoor, I., Tan, Q., Shyu, C. R., and Chopra, S. (2015). A sorghum MYB transcription factor induces 3-deoxyanthocyanidins and enhances resistance against leaf blights in maize. *Molecules* 20, 2388–2404. doi: 10.3390/molecules20022388

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W. Z., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* 428, 726–731. doi: 10.1016/j.jmb.2015.11.006

Kim, D. S., and Hwang, B. K. (2011). The pepper receptor-like cytoplasmic protein kinase CaPIK1 is involved in plant signaling of defense and cell-death responses. *Plant J.* 66, 642–655. doi: 10.1111/j.1365-313X.2011.04525.x

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Li, H., Wu, J., Shang, X., Geng, M., Gao, J., Zhao, S., et al. (2020). WRKY transcription factors shared by BTH-induced resistance and NPR1-mediated acquired resistance improve broad-spectrum disease resistance in wheat. *Mol. Plant Microbe Interact.* 33, 433–443. doi: 10.1094/MPMI-09-19-0257-R

Lim, H., Cho, M. H., Jeon, J. S., Bhoo, S. H., Kwon, Y. K., and Hahn, T. R. (2009). Altered expression of pyrophosphate: fructose-6-phosphate 1-phosphotransferase affects the growth of transgenic *Arabidopsis* plants. *Mol. Cells* 27, 641–649. doi: 10.1007/s10059-009-0085-0

Liu, Q., Li, X., Yan, S., Yu, T., Yang, J., Dong, J., et al. (2018). OsWRKY67 positively regulates blast and bacteria blight resistance by direct activation of PR genes in rice. *BMC Plant Biol.* 18:257. doi: 10.1186/s12870-018-1479-y

Liu, S., Liu, Y., Zhao, J., Cai, S., Qian, H., Zuo, K., et al. (2017). A computational interactome for prioritizing genes associated with complex agronomic traits in rice (Oryza sativa). *Plant J.* 90, 177–188. doi: 10.1111/tpj.13475

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

Nakaminami, K., Matsui, A., Nakagami, H., Minami, A., Nomura, Y., Tanaka, M., et al. (2014). Analysis of differential expression patterns of mRNA and protein during cold-acclimation and de-acclimation in *Arabidopsis*. *Mol. Cell Proteomics* 13, 3602–3611. doi: 10.1074/mcp.M114.039081

Nelson, D. R. (2009). The cytochrome p450 homepage. *Hum. Genomics* 4, 59–65. doi: 10.1186/1479-7364-4-1-59

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: a plant Coexpression database Based on Investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.* 59:440. doi: 10.1093/pcp/pcx209

Peng, X., Wang, H., Jang, J. C., Xiao, T., He, H., Jiang, D., et al. (2016). OsWRKY80-OsWRKY4 module as a positive regulatory circuit in rice resistance against *Rhizoctonia solani*. *Rice (NY)* 9:63. doi: 10.1186/s12284-016-0137-y

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448

Rautengarten, C., Usadel, B., Neumetzler, L., Hartmann, J., Bussis, D., and Altmann, T. (2008). A subtilisin-like serine protease essential for mucilage release from *Arabidopsis* seed coats. *Plant J.* 54, 466–480. doi: 10.1111/j.1365-313X.2008.03437.x

Reiser, L., Subramaniam, S., Li, D., and Huala, E. (2017). Using the *Arabidopsis* information resource (TAIR) to find information about *Arabidopsis* genes. *Curr. Protoc. Bioinformatics* 60, 1.11.1–1.11.45. doi: 10.1002/cpbi.36

Shan, T., Rong, W., Xu, H., Du, L., Liu, X., and Zhang, Z. (2016). The wheat R2R3-MYB transcription factor TaRIM1 participates in resistance response against the pathogen Rhizoctonia cerealis infection through regulating defense genes. *Sci. Rep.* 6:28777. doi: 10.1038/srep28777

Shen, Q., Bao, M., and Zhou, X. (2012). A plant kinase plays roles in defense response against geminivirus by phosphorylation of a viral pathogenesis protein. *Plant Signal. Behav.* 7, 888–892. doi: 10.4161/psb.20646

Sonnhammer, E. L., and Ostlund, G. (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* 43, D234–D239. doi: 10.1093/nar/gku1203

Sun, Y. Y., Liu, L., Li, J., and Sun, L. (2016). Three novel B-type mannose-specific lectins of *Cynoglossus semilaevis* possess varied antibacterial activities against Gram-negative and Gram-positive bacteria. *Dev. Comp. Immunol.* 55, 194–202. doi: 10.1016/j.dci.2015.10.003

Tian, T., You, Q., Yan, H., Xu, W., and Su, Z. (2018). MCENet: a database for maize conditional co-expression network and network characterization collaborated with multi-dimensional omics levels. *J. Genet. Genomics* 45, 351–360. doi: 10.1016/j.jgg.2018.05.007

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Tsai, C. C., Wu, K. M., Chiang, T. Y., Huang, C. Y., Chou, C. H., Li, S. J., et al. (2016). Comparative transcriptome analysis of *Gastrodia elata* (Orchidaceae) in response to fungus symbiosis to identify gastrodin biosynthesis-related genes. *BMC Genomics* 17:212. doi: 10.1186/s12864-016-2508-6

Wang, D., Fan, W., Guo, X., Wu, K., Zhou, S., Chen, Z., et al. (2020). MaGenDB: a functional genomics hub for Malvaceae plants. *Nucleic Acids Res.* 48, D1076–D1084. doi: 10.1093/nar/gkz953

Wang, J., Tao, F., Tian, W., Guo, Z., Chen, X., Xu, X., et al. (2017). The wheat WRKY transcription factors TaWRKY49 and TaWRKY62 confer differential high-temperature seedling-plant resistance to *Puccinia striiformis* f. sp. tritici. *PLoS One* 12:e0181963. doi: 10.1371/journal.pone.0181963

Wang, M., Zhu, Y., Han, R., Yin, W., Guo, C., Li, Z., et al. (2018). Expression of *Vitis amurensis* VaERF20 in *Arabidopsis thaliana* improves resistance to *Botrytis cinerea* and *Pseudomonas* syringae pv. Tomato DC3000. *Int. J. Mol. Sci.* 19:696. doi: 10.3390/ijms19030696

Wang, Y., Liang, C., Wu, S., Jian, G., Zhang, X., Zhang, H., et al. (2019). Vascular-specific expression of Gastrodia antifungal protein gene significantly enhanced cotton *Verticillium wilt* resistance. *Plant Biotechnol. J.* 18, 1498–1500. doi: 10.1111/pbi.13308

Wang, Y., Liang, C., Wu, S., Zhang, X., Tang, J., Jian, G., et al. (2016). Significant improvement of cotton *Verticillium wilt* resistance by manipulating the expression of gastrodia antifungal proteins. *Mol. Plant* 9, 1436–1439. doi: 10.1016/j.molp.2016.06.013

Xu, J. T. (1981). [A brief report on the nutrition sources of seed germination of *Gastrodia elata* (author's transl)]. *Zhong Yao Tong Bao* 6:2.

Xu, J. T. (1989). [Studies on the life cycle of Gastrodia elata]. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* 11, 237–241.

Yang, J., Liu, Y., Yan, H., Tian, T., You, Q., Zhang, L., et al. (2018). PlantEAR: functional analysis platform for plant EAR motif-containing proteins. *Front. Genet.* 9:590. doi: 10.3389/fgene.2018.00590

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic Acids Res.* 41, W98–W103. doi: 10.1093/nar/gkt281

Yin, X., Mu, L., Li, Y., Wu, L., Yang, Y., Bian, X., et al. (2019). Identification and characterization of a B-type mannose-binding lectin from *Nile tilapia* (*Oreochromis nilot*icus) in response to bacterial infection. *Fish Shellfish Immunol.* 84, 91–99. doi: 10.1016/j.fsi.2018.09.072

You, Q., Xu, W., Zhang, K., Zhang, L., Yi, X., Yao, D., et al. (2017). ccNET: database of co-expression networks with functional modules for diploid and polyploid Gossypium. *Nucleic Acids Res.* 45, 5625–5626. doi: 10.1093/nar/gkw1342

Yuan, Y., Jin, X., Liu, J., Zhao, X., Zhou, J., Wang, X., et al. (2018). The *Gastrodia elata* genome provides insights into plant adaptation to heterotrophy. *Nat. Commun.* 9:1615. doi: 10.1038/s41467-018-03423-5

Yuan, Y., Teng, Q., Zhong, R., and Ye, Z. H. (2016). Roles of *Arabidopsis* TBL34 and TBL35 in xylan acetylation and plant growth. *Plant Sci.* 243, 120–130. doi: 10.1016/j.plantsci.2015.12.007

Zheng, Y., Jiao, C., Sun, H. H., Rosli, H. G., Pombo, M. A., Zhang, P. F., et al. (2016). iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* 9, 1667–1670. doi: 10.1016/j.molp.2016.09.014

Zhu, G., Wu, A., Xu, X. J., Xiao, P. P., Lu, L., Liu, J., et al. (2016). PPIM: a protein-protein interaction database for maize. *Plant Physiol.* 170, 618–626. doi: 10.1104/pp.15.01821

Zhu, Y., Li, Y., Zhang, S., Zhang, X., Yao, J., Luo, Q., et al. (2019). Genome-wide identification and expression analysis reveal the potential function of ethylene responsive factor gene family in response to *Botrytis cinerea* infection and ovule development in grapes (*Vitis vinifera* L.). *Plant Biol. (Stuttg)* 21, 571–584. doi: 10.1111/plb.12943

Conflict of Interest: The authors 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.