Transcriptome-wide identification and characterization of the MYB gene family in sickle seagrass (Thalassia hemprichii)

Shen Jie
Hainan academy of ocean and fisheries sciences
Cai Zefu
Hainan academy of ocean and fisheries sciences
Chen Shiquan
Hainan academy of ocean and fisheries science
Wang Daoru
Hainan academy of ocean and fisheries sciences
Zhongjie Wu (blandman@126.com)
Hainan academy of ocean and fisheries science

Research article

Keywords: Sickle seagrass, Thalassia hemprichii, Transcriptome-wide analysis, MYB transcription factors, Functional homology analysis

DOI: https://doi.org/10.21203/rs.3.rs-26345/v1

Abstract

Background: Sickle seagrass (Thalassia hemprichii) is one of the most important marine plants living in the tropical climate, mainly distributed in Southeast Asian waters. It is an important food source for marine herbivores and plays important roles in nitrogen fixation, water purification and maintaining the balance of marine ecology. In recent years, the area of aquatic plants has declined rapidly, affecting the ecological balance. However, the molecular mechanism of aquatic plants has been poorly studied.

Results: In this study, all transcriptional information of T. hemprichii was obtained by using high-throughput sequencing technology, and 32,097 unigenes were identified by annotation. In addition, 119 MYB transcription factors were screened, and 61 genes with complete ORF were sequenced. Furthermore, 17 clays were identified according to the information of Arabidopsis .

Conclusions: This study provides useful information for enriching the genetic information of T. hemprichii , and further exploring the molecular mechanisms of the evolution, development, and physiological functions of Sickle seagrass.

Background

Seagrasses are marine flowering plants that form habitats in coastal areas. The seagrass can stabilize sediment, reduce water flow rate, and provide habitat for other aquatic animals and plants. Coastal seagrass habitat provides a wide variety of ecosystem services for marine life. Also, it is also a source of food for large herbivores, nursery for most fish and invertebrates, and plays an important role in the nutrition, nitrogen fixation and physical and chemical properties of marine organisms[1–4]. In recent years, seagrass meadows have been lost, leading to marine ecological damage. Scientists have carried out extensive research on genetics, artificial cultivation, and ecological restoration of different species of seagrass population worldwide[5–7].

Sickle seagrass, *Thalassia hemprichii* is a common seaweed species worldwide, especially biogeographic sub-region of the Indian Ocean stretching on a latitudinal scale from Somalia to the east coast of South Africa[1, 8]. *T. Hemprichii* can be propagated sexually through seeds or asexually through rhizomes. The range of each component of reproduction has an important impact on local population statistics, diffusion, biogeography and genetic diversity[1]. *T. hemprichii* seeds seem to sink under the water for 24 hours and may float ten to hundreds of kilometers for about a month on the water. Studies have shown that adult plants of *T. hemprichii* can float for several months and remain alive, and are likely to settle in new areas[9]. Their mature stems and fruits are mainly distributed by passive drifting, and how to improve seagrass management in the western Indian Ocean (WIO) area to maintain high adaptability to the environmental changes. The healthy seaweed population is of great significance.

The MYB family is one of the largest transcription factor (TF) families, and it is named after a highly conservative sequence (MYB DNA-binding domain) located in the N-terminus of these proteins[10]. MYB TF regulates the physiological and biochemical processes, such as the development and growth of various plants, and the synthesis of flavonol by participating in many physiological and biochemical processes such as petal morphogenesis and cell etc[11]. MYB TF plays a central role in the transcriptional regulation of plant secondary cell wall (SCW) deposition. SCW deposition and lignification have very important values for plant growth and development[12]. MYB family genes have been found in *Arabidopsis*, corn, soybean, maize,pineapple, bamboo and many other plants.[10, 13–17].

T. hemprichii is an important seaweed, and *T. hemprichii* meadow has significance in marine ecology. The transcriptome data of *T. hemprichii* were studied, and the similarity and classification of *T. hemprichii* MYB were compared with *Arabidopsis* MYB. The complete transcriptome information of *T. hemprichii* was obtained for the first time. Also, the genes of the MYB gene family in *T. hemprichii* were analyzed.

Result

Transcriptome data assembly and acquisition

In general, the high-quality sequence (6.47 GB) comes from the transcriptome sequencing of *T. hemprichii*. The average error rate of the sequences was 0.01%, and more than 87% of clean reads (Figure 1A). Using Benchmarking Universal Single-Copy Orthologs (BUSCO) v3, a single copy homologous database, and the integrity of the transcriptome assembly was demonstrated by comparing with the conserved genes (Figure 1B). The assembled sequence data for these raw reads were deposited in the Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/Traces/sra) (Accession no. SAMN13255716 and SAMN13255717). The sequencing data were assembled into 51,619 transcripts with the sequencing lengths ranging from 200 to 19,192 bases (Figure 1C) (mean length = 1045 bases, N50 = 1798 bases, and GC content = 45.73%). And the longest Unigene is 5.40 Mb (53,976,312 bases). All the data show that the production capacity of the product is high enough and the assembly quality is very good, which can be used for subsequent analysis.

Gene annotation and functional classification

Seven databases (KEGG), Gene Oncology (GO), NR, NT, SwissProt(SWISS-PROT), Protein families (Pfam), and EuKaryotic Orthologous Groups (KOG) were used to annotate the obtained genes. In total, 32,097 Unigenes (62.18% of the total Unigenes) were annotated in the databases in this study (Table 1). Transcoder software was used to identify the candidate coding regions (CDs) in Unigene. First, the longest open reading frame was obtained. Then, through Swissport database and blast in hmmscan, the homologous sequence of Pfam protein database was found and its CDs was predicted.. 24,870 CDS were predicted, with a total length of 26,466,750, N50 of 1,407 bases, length range of 297 to 10,956 bases and GC content of 49.09% (Figure 1D). According to the results, 11,804 genes were annotated in all databases. Most of the unigenes get annotation information in the NR database (30,772 and 59.52%) (Figure 2). Regarding to GO classification, the highest number of annotations was in the molecular function (MF) category, for which catalytic activity (6317), binding (6269) and transporter activity (676) were the top 3 GO terms; the second-largest amount of annotations was in the cellular component (CC) category, the top 3 GO terms were membrane part (3948), cell (3746) and organelle part (1512); and the third-largest amount of annotations was in the biological process (BP) category, the top 3 GO terms were cellular process (3613), biological regulation (1404) and localization (879) (Figure 3A). According to the classification of KEGG, the most annotated genes are metabolic related pathways, and the most annotated pathway is metabolic pathway (15,152). In this category, the Global and overview maps pathway enriched the most genes (5,855) (Figure 3B). Furthermore, Classification of Unigene according to KOG. The top 3 classes were general function prediction(5,970), signal transduction mechanisms(3,334), and function unknown(2,071) (Figure 3C).

Functional homology analysis of whole transcriptome

In the results of gene function annotation, NR database had 59.52% of the annotation proportion. The proportion of different species on the NR annotation was calculated, and the species distribution map was drawn (Figure 4). The NR annotation of top 5 species were *Elaeis guineensis* (5,272 and 17.16%), *Phoenix dactylifera* (3,436 and 11.19%), *Ananas comosus* (1,896 and 6.17%), *Nelumbo nucifera* (1,756 and 5.72%) and *Zostera marina* (1,672 and 5.44%). *T. hemprichii* and *E. guineensis* had more closely related genes and were closer to the terrestrial plants.

Identification and distribution of MYB TFs encoding genes in transcriptome

The genes were predicted, which were capable of encoding transcription factors (TFs). Then,we use getorf to detect the ORF of Unigene, and then useing hmmsearch to compare ORF with transcription factor protein domain (the data comes from TF), and then recognize the ability of UniGene according to the characteristics of transcription factor family described by TFDB. The TFs family was classified and counted (Additional files 1: Figure S1). There were 119 MYB genes family in all TF genes (Additional files 2: Table S1).

GO classification was used; the results showed that these MYB family genes were divided into 10 subclasses. The most annotated genes was in the molecular function (MF) category; binding (69), and transcription regulator activity (13). In the cellular component (CC) category, the 3 GO terms were cell (42), organelle part (2), and membrane part (1). In the biological process (BP) category, the 5 GO terms were cellular process (39), biological regulation (39), response to a stimulus (6), cellular component organization or biogenesis (2), and multicellular organismal process (2) (Figure 5A). According to the KEGG classification, All MYB family genes were involved in 9 pathways, among which these pathway was Environmental adaptation (24 genes), Global and overview maps (9 genes), Biosynthesis of other secondary metabolites (4 genes), Transport and catabolism (4 genes), Lipid metabolism (3 genes), Translation (1 gene), Transcription (1 gene), Carbohydrate metabolism (1 gene), and Energy metabolism (1 gene) (Figure 5B). Also, 24 of these genes were enriched in the Circuit rhythm pathway.

conserved DNA-binding domain analysis and of Protein characteristics MYB TFs.

Through comparative analysis, a set of 119 MYB Unigenes containing MYB DNA-binding domains was identified in *T. hemprichii* (Additional files 3: TableS 2), which included 67 full-length CDS and 52 fragment sequences. MYB Unigenes (ThMYB1–ThMYB119) were identified, respectively. The full-length sequence of MYB ranges from 588 to 3393. The amino acid sequence length of the protein is 148-1131 amino acids, the calculated molecular weight of ThMYB is 15.87-127.50 kDa, and the calculated pi is 4.34-110.01. The majority of ThMYB protein is about 400 amino acids, and its molecular weight is about 50 kDa. Among the 119 ThMYBs, ThMYB042 was the longest protein (1,131 amino acids), while the shortest protein was ThMYB038 (148 amino acids).

In order to study and identify the characteristics of homologous domain of ThMYB protein, 119 amino acid sequences of ThMYBs were used for multiple sequence alignment. According to the domain classification, ThMYB protein contains type 11 domain, which shows that ThMYBs has similar domain with other species (Figure 6).

Putative functions of ThMYBs in *T. hemprichii*

MYB TFs in *A. thaliana* contains 27 clades, and the function of each clades was annotated [14, 18-21]. The conglomerated homologous proteins usually have similar functions, which indicates that ThMYB has similar functions with AtMYB in the same clades. Therefore, through the conclusion and discussion compared with AtMYBs, the function of ThMYBs is predicted and summarized (Figure 7; Additional files 4: Table S3). Constructing NJ unrooted phylogenetic tree with 09 ThMYBs and 110 AtMYBs (Figure 7). The results showed that MYB TFs of the two species gathered in 36 clades (C1-C36), and 17 of which were found in both species.

Discussion

In this study, high-throughput sequencing technology was used to obtain the complete transcriptome information of *T. hemprichii*, and 32,097 unigenes were identified by annotation. In addition, we showed that there were more than 52 proteins with one domain, 43 proteins with two domains, and ThMYB095 protein contained three domains. Interestingly, it was found that 7 proteins did not have typical MYB domains, but they were still annotated as MYB genes. Enable ThMYB042 contained a Myb_Cef domain. ThMYB004, ThMYB011, and ThMYB086 contained a SWIRM domain. ThMYB087 and ThMYB019 contained a ZZ domain, ThMYB053 and ThMYB022 contained a DnaJ domain, ThMYB098 contained a DMAP1 domain, respectively.

And we found that 60 ThMYBs were clustered into 17 clades of known functional annotation, and 48 ThMYBs were clustered into 9 clades of unknown functional annotation. According to the annotation results, ThMYBs can be divided into six functional classes.There are six clades in class I (C1, C12, C14, C16, C22 and C32), which regulate the biosynthesis and deposition of lignin, cellulose and hemicellulose, and are responsible for the formation of secondary cell wall (SCW). Class II includes five clades (C2, C6, C13, C28 and C34), which can regulate the ABA pathway to participate in the response of biotic and abiotic stresses. The third category includes five clades (C4, C20, C25, C29, C31), which play an important role in morphogenesis and organogenesis of root, anther, embryogenesis, epidermal cell, vegetative cell and stomatal cell development, etc. The fourth group includes two clades (C7, C17), which are mainly involved in the regulation of secondary metabolism. Class IV consists of two clades (C4, C31), which are mainly involved in the regulation of secondary metabolism.

Seagrasses are important marine ecological plants that provide sustainable and reliable guarantees for the health status of marine ecology. Seagrasses play important roles in cleaning water, producing oxygen, fixing nitrogen, providing food for aquatic animals and maintaining seabed stability. In recent years, the reduction of aquatic plants has seriously affected the ecological stability of the marine ecosystem. The studies have carried out extensive research in various fields, such as physiology, biochemistry, genetic breeding, and geographic pedigree to keep seabed stability by the artificial restoration of seagrass. *T. hemprichii* is a common water-borne grass. It grows fast and can be spread by ocean currents. It is rich in nutrients and is a food of many herbivores.

Conclusions

At present, there are a few studies on the molecular biology and genomics of *T. hemprichii*. Here, we obtained the full transcriptome information of *T. hemprichii* through assembly prediction. A total of 32,097 unigenes were also obtained. This provides useful genetic information for the further study of the evolution, development, molecular characteristics and molecular breeding of *T. hemprichii*. Besides, 119 MYB transcription factors were obtained by screening the whole transcript. Therefore, it provides a molecular basis for a better understanding of *T. hemprichii*.

The similar number of genes of *T. hemprichii* and *E. guineensis*, *P. dactylifera*, and *A. comosus* have been found, which indicates that *T. hemprichii* is closer to terrestrial plants. In the evolutionary process, *T. hemprichii* might migrate from land to water, which is worthy of being studied in the future.

Methods

Collection of plants

T. hemprichii was collected at three separate occasions (10 days before the start of an experimental run) from March 2019 at the Lingshui Xincun port seagrass special protected area, Hainan, China. Xincun port (21.97km^2) is about 4km long from north to south, and the port door is from Xin cun jiao (18°24'42"N and 109°57'58"E) to Shitoucun Shazui (18°24'34"N and 109°57'42"E). The selected grassland is located in the shallow water area (0.5 - 1 m deep) on the coast of Xincun Lingshui, about 5 km northwest of Xincun Lingshui.. An iron shovel was used to dig up the ground, and the lower part of T. hemprichii together to ensure the integrity of the roots, rhizomes, erect stems, and leaves, and clean the shore with seawater. The seagrasses were packed in tissues wetted with seawater, placed inside plastic bags and transported within 48 h to the the Molecular Biology Laboratory in Hainan Academy of Ocean and Fisheries Sciences. We identify the collected samples by morphological identification in order to ensure the accuracy of the samples. The whole plant packed in the sealed pocket and brought to the laboratory. After cleaning, the liquid nitrogen treatment was carried out for 15 min and stored at - 80 °C. We identify the collected samples by morphological identification. According to the website of World Register of Marine Species, Petermanns Geogr named the *Thalassia hemprichii* in 1871 (Aphia ID: 208931), and it was described as follows: Rhizomes terete, with persistent leaf sheaths. Leaves curved, 6–12(–40) cm×4–8 mm. Peduncle of male inflorescence 2–3 cm, female inflorescence without peduncle; spathe linear. Male flower on a pedicel 2–3 cm; perianth segments elliptic, petaloid; anthers oblong; female flower with ovary of 6 carpels; stigmatic branches 1–1.5 cm. Fruit greenish, 2–2.5×1.8–3.2 cm.

After identification by associate fellow Shiquan Chen and proofreading by Associate Senior Engineer Zefu Cai, we thought that the sample was Thalassia hemprichii.. The "Plant Sample Collection Information Record Form" is attached.

RNA extraction, library construction and high-throughput sequencing

RNA of each sample was extracted using the TRIzol extraction method and Purification of RNA products using RNeasy® MinElute® cleanup kit (Qiagen®). RiboZeroTM Magnetic (Plant leaf) kit (Epicentre®) was used to remove rRNA from 4 µg of each sample. The removal of rRNA was confirmed using a Agilent 2100 BioAnalyser system (Agilent Technologies).

After extracting the total RNA from each sample, Oligo DT magnetic beads were used to enrich mRNA and fragment buffer was used to split mRNA into short fragments. Using mRNA as template, the first cDNA strand was synthesized with six base random primers, and buffer was added. which was end-repaired after purifying by Qiaquick PCR purification kit and eluting with EB buffer, plus floya and ligation of the sequencing link. The agarose gel electrophoresis was done for fragment selection. Finally, The constructed library was sequenced with bgi-500（BGI.

Transcriptome assembly, gene annotation and ontology

Short read assembly software, SOAPdenovo, which was used to perform transcriptome assembly from scratch[22]. SOAPdenovo first concatenated reads with a certain length of overlapping fragments. These N-free assembly fragments obtained from the overlapping reads are called contig. Then, the reads back to contig were compared. Different contigs from the same transcript and the distance between these contigs were determined. The contigs were linked together. The paired-end reads were used to make a hole in the Scaffold (N, unknown intermediate sequence), and the sequence with the least N was identified, and there was no longer extension at both ends, which is called Unigene. Unigene assembled from different samples, which was further sequenced and de-duplicated by sequence clustering software to obtain the longest non-redundant Unigene. Finally, the Unigene sequences were aligned with no redundant (Nr) protein, Swiss-Prot, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Clusters of Orthologous Group (COG) databases (E-value <0.00001), and the best-aligned protein was determined for the sequence orientation of Unigene. If there was a contradiction between different libraries, then the sequence direction of Unigene was determined according to the priority of Nr, Swiss-Prot, KEGG, and COG. ESTScan software was used to predict the coding region and determine the direction of the Unigene, which is incomparable with the above four libraries (Iseli et al., 1999).

Phylogenetic analysis and ThMYB protein domain prediction

To examine the evolutionary relationships among MYBs in *T. hemprichii*119 sequences were used. Also, the MYB sequences of *Arabidopsis* were downloaded from The *Arabidopsis* Information Resource (TAIR) (http://www.arabidopsis.org). We constructed a neighbor-joining (NJ) phylogenetic tree of ThMYB and AtMYB amino acid sequences by using the method of ClustalW in MEGA software (version X) with the following parameters: pairwise deletion, bootstrap analysis with 1,000 replicates and Poisson correction,. Additionally, the protein domain of ThMYBs was predicted according to the bio sequence analysis using profile hidden Markov models (HMMER) (https://www.ebi.ac.uk/Tools/hmmer).

Abbreviations

TF
Transcription factor;
T. hemprichii
Thalassia hemprichii;
TAIR
The Arabidopsis Information Resource;
NJ
Neighbor-Joining Phylogenetic Tree;
BUSCO
Benchmarking Universal Single-Copy Orthologs;
SRA
Sequence Read Archive;
KEGG
Kyoto Encyclopedia of Genes and Genomes;
GO
Gene Oncology ;
NR
Non-Redundant Protein Sequence Database ;
NT
Nucleotide Sequence Database;
SWISS-PROT
SwissProt Database;
Pfam
Protein families;
KOG
EuKaryotic Orthologous Groups;
CDS
the candidate coding regions .

Declarations

Ethics approval and consent to participate

The plant material used in our study has been generated by ourselves by the methods described.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available in the supplemental information files or from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by Hainan Provincial Natural Science Foundation of China(319QN251); The Central Government Guided Local Science and Technology Development Project in 2019 (ZY2019HN03); National Key Research and Development Project of China (2017YFC0506104) and Department budget projects of Hainan provincial in 2019. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors ' Contributions

SJ and WZJ conceived the ideas and designed the experiments; SJ collected and analyzed the data; CZF and CSQ Collecting samples; SJ and WDR drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank MogoEdit and Shanghai Keran Biological Technology Co., Ltd. , for editing the English text of a draft of this manuscript.

Author information

Hainan Academy of Ocean and Fisheries Sciences,Haikou,China.

Key laboratory of Utilization and Conservation for Tropical Marine Bioresources(Hainan Tropical Ocean University),Ministry of Education, Sanya, China.

References

1. Jahnke M, Gullstrom M, Larsson J, Asplund ME, Mgeleka S, Silas MO, Hoamby A, Mahafina J, Nordlund LM. Population genetic structure and connectivity of the seagrass Thalassia hemprichii in the Western Indian Ocean is influenced by predominant ocean currents. Ecol Evol. 2019;9(16):8953–64.
2. Lee CL, Huang YH, Chen CH, Lin HJ. Remote underwater video reveals grazing preferences and drift export in multispecies seagrass beds. J Exp Mar Biol Ecol. 2016;476:1–7.
3. Nordlund LM, Koch EW, Barbier EB, Creed JC. **Seagrass Ecosystem Services and Their Variability across Genera and Geographical Regions**. *Plos One* 2016, 11(10).
4. van de Koppel J, van der Heide T, Altieri AH, Eriksson BK, Bouma TJ, Olff H, Silliman BR. Long-distance interactions regulate the structure and resilience of coastal ecosystems. Ann Rev Mar Sci. 2015;7:139–58.
5. Jahnke M, Jonsson PR, Moksnes PO, Loo LO, Jacobi MN, Olsen JL. Seascape genetics and biophysical connectivity modelling support conservation of the seagrass Zostera marina in the Skagerrak-Kattegat region of the eastern North Sea. Evol Appl. 2018;11(5):645–61.
6. Sinclair EA, Anthony JM, Greer D, Ruiz-Montoya L, Evans SM, Krauss SL, Kendrick GA. Genetic signatures of Bassian glacial refugia and contemporary connectivity in a marine foundation species. J Biogeogr. 2016;43(11):2209–22.
7. Wainwright BJ, Arlyza IS, Karl SA. Population genetic subdivision of seagrasses, Syringodium isoetifolium and Thalassia hemprichii, in the Indonesian Archipelago. Bot Mar. 2018;61(3):235–45.
8. Obura D. The diversity and biogeography of Western Indian Ocean reef-building corals. Plos One. 2012;7(9):e45013.
9. Wu KY, Chen CNN, Soong K. **Long Distance Dispersal Potential of Two Seagrasses Thalassia hemprichii and Halophila ovalis**. *Plos One* 2016, 11(6).
10. Yang K, Li Y, Wang S, Xu X, Sun H, Zhao H, Li X, Gao Z. Genome-wide identification and expression analysis of the MYB transcription factor in moso bamboo (Phyllostachys edulis). PeerJ. 2019;6:e6242.
11. Baumann K, Perez-Rodriguez M, Bradley D, Venail J, Bailey P, Jin HL, Koes R, Roberts K, Martin C. Control of cell and petal morphogenesis by R2R3 MYB transcription factors. Development. 2007;134(9):1691–701.

12. Oh S, Park S, Han KH. Transcriptional regulation of secondary growth in Arabidopsis thaliana. J Exp Bot. 2003;54(393):2709–22.
13. Du H, Wang YB, Xie Y, Liang Z, Jiang SJ, Zhang SS, Huang YB, Tang YX. Genome-Wide Identification and Evolutionary and Expression Analyses of MYB-Related Genes in Land Plants. DNA Res. 2013;20(5):437–48.
14. Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L. MYB transcription factors in Arabidopsis. Trends Plant Sci. 2010;15(10):573–81.
15. Liu CY, Xie T, Chen CJ, Luan AP, Long JM, Li CH, Ding YQ, He YH. **Genome-wide organization and expression profiling of the R2R3-MYB transcription factor family in pineapple (Ananas comosus)**. *Bmc Genomics* 2017, 18.
16. Salih H, Gong WF, He SP, Sun GF, Sun JL, Du XM. **Genome-wide characterization and expression analysis of MYB transcription factors in Gossypium hirsutum**. *Bmc Genet* 2016, 17.
17. Wilkins O, Nahal H, Foong J, Provart NJ, Campbell MM. Expansion and Diversification of the Populus R2R3-MYB Family of Transcription Factors. Plant Physiol. 2009;149(2):981–93.
18. Zhong RQ, Lee CH, Zhou JL, McCarthy RL, Ye ZH. A Battery of Transcription Factors Involved in the Regulation of Secondary Cell Wall Biosynthesis in Arabidopsis. Plant Cell. 2008;20(10):2763–82.
19. Li Z, Peng R, Tian Y, Han H, Xu J, Yao Q. Genome-Wide Identification and Analysis of the MYB Transcription Factor Superfamily in Solanum lycopersicum. Plant Cell Physiol. 2016;57(8):1657–77.
20. Li S, Wang W, Gao J, Yin K, Wang R, Wang C, Petersen M, Mundy J, Qiu JL. MYB75 Phosphorylation by MPK4 Is Required for Light-Induced Anthocyanin Accumulation in Arabidopsis. Plant Cell. 2016;28(11):2866–83.
21. Zhou J, Lee C, Zhong R, Ye ZH. MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis. Plant Cell. 2009;21(1):248–66.
22. Li RQ, Zhu HM, Ruan J, Qian WB, Fang XD, Shi ZB, Li YR, Li ST, Shan G, Kristiansen K, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010;20(2):265–72.

Tables

Table 1. Different database annotation results

Total	NR	NT	Swissport	KEGG	KOG	Pfam	GO	Inter section	Over all
51,619	30,722	17,450	22,393	24,594	24,629	21,409	12,927	6,166	32,097
100%	59.52%	33.81%	43.38%	47.65%	47.71%	41.48%	25.04%	11.95%	62.18%

Additional Information

Additional files1: Figure S1. Classification and statistics of different TF family genes.

Additional files2: Table S1. MYB Unigene in *Thalassia hemprichii*.

Additional files3: Table S2. Nomenclature and protein information of MYBs in *Thalassia hemprichii*.

Additional files4: Table S3. The functions of ThMYBs were predicted and summarized by conclusion and discussion compared with those of AtMYBs

Figures

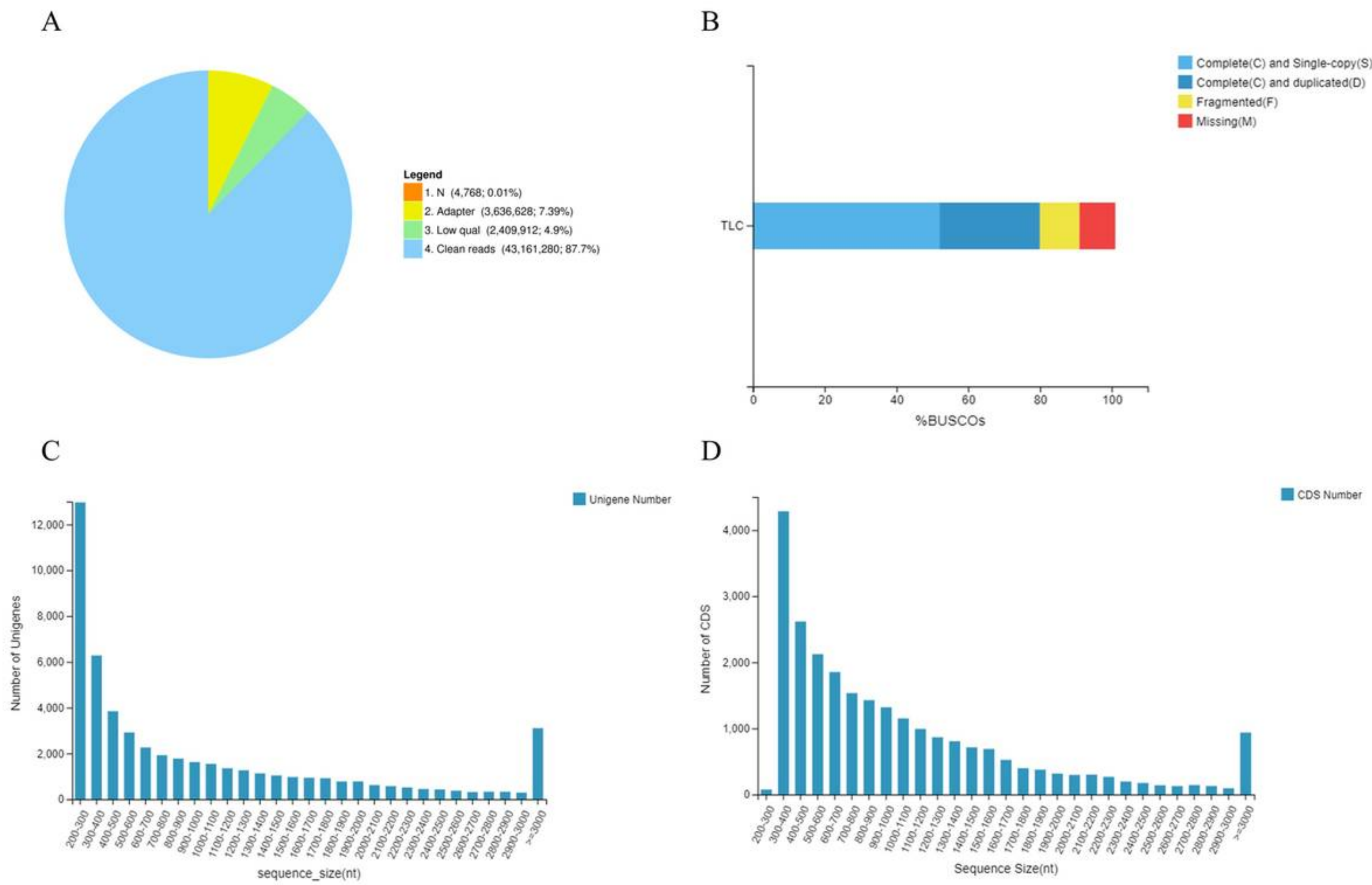

Figure 1

Unigene assembly and prediction. A. The sequences after insert recognition (Reads Of insert, ROI) processing can be divided into four categories; B. Correspondence before and after Isoform clustering. The quality of assembled transcripts was assessed using a single copy ortholog database BUSCO Gene comparison, to some extent, illustrates the integrity of transcriptome assembly; C. Classification statistics of UNIGENE gene with different length intervals; D. Classification statistics of different length intervals of CDS sequences.

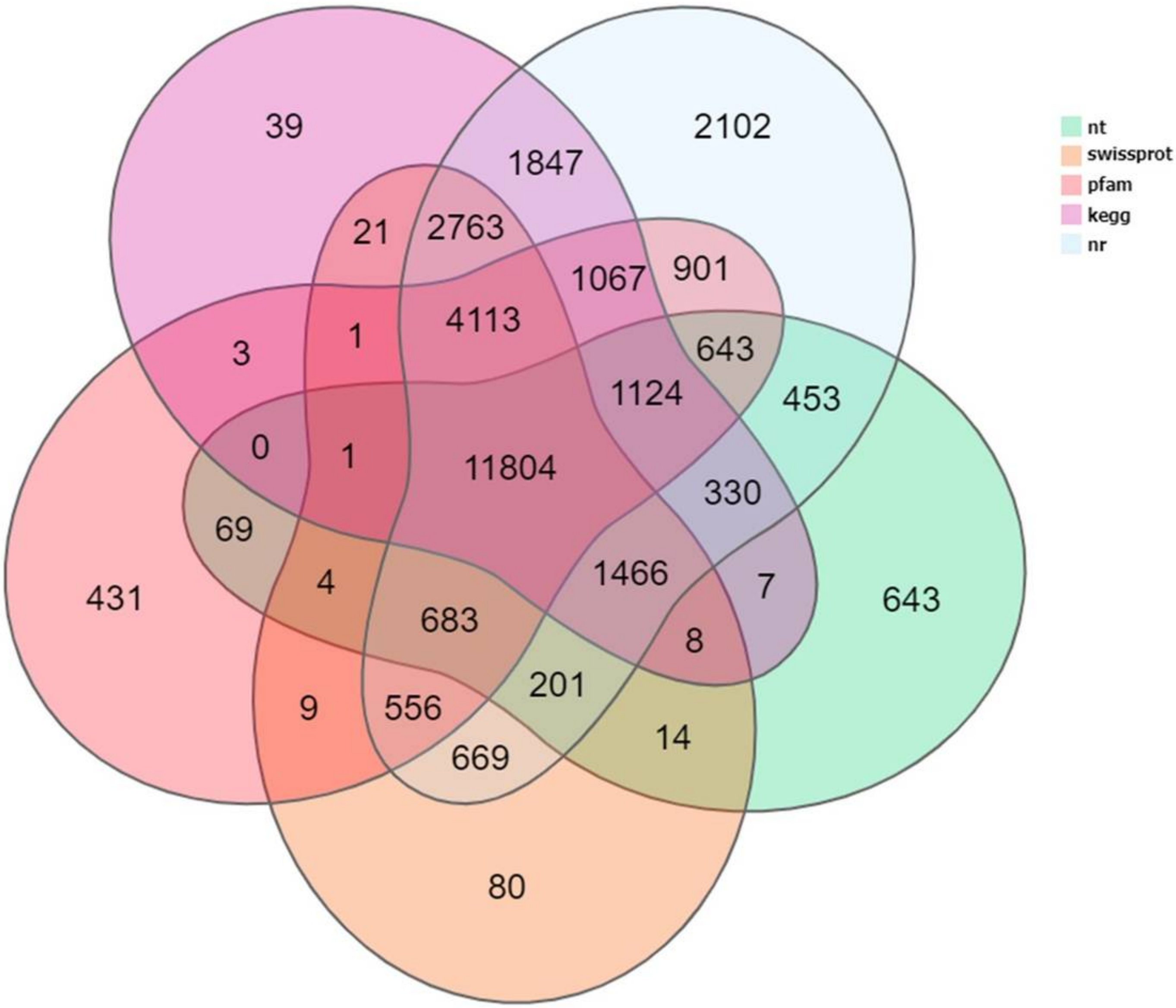

Figure 2

Venn diagrams comparing annotation results from different databases. Use Venn diagrams to show annotated results for NR, NT, KEGG, Pfam, and SwissProt.

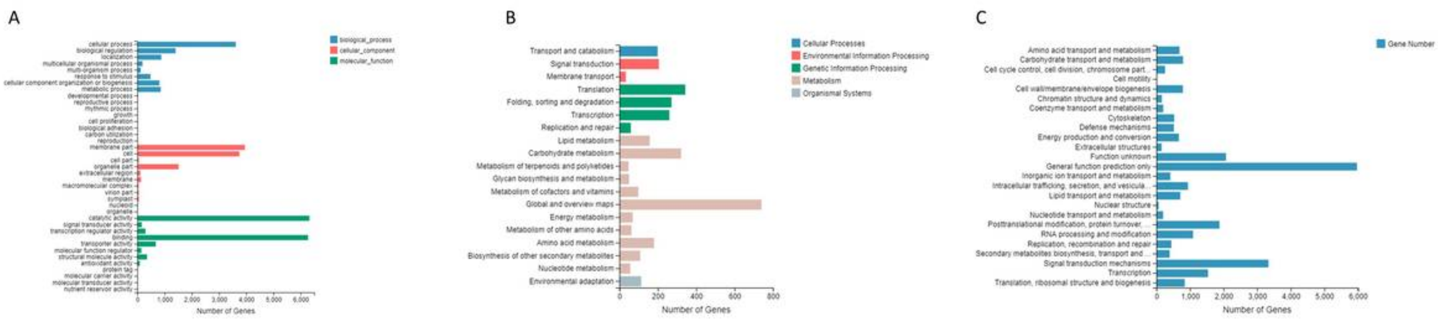

Figure 3

GO, KEGG and KOG classification. After the de-redundancy is complete, we will perform an annotation on the seven major functions of the Uniq Isoform obtained by the de-redundancy. A. Blast2GO software was used to annotate the Isoform results of all comparisons to the NR database to the GO database, and statistically annotate to the three aspects of GO: Biological Process, Cellular Component, and Molecular Function. Classification diagram; B. Compare all Isoforms to the KEGG database, and count the Isoforms on the level1 and level2 levels of the KEGG database, and draw a statistical diagram of the KEGG function distribution; C. Annotate Isoform to the KOG database, and compare the statistics of Isoform in the 25 functional groups of the KOG database.

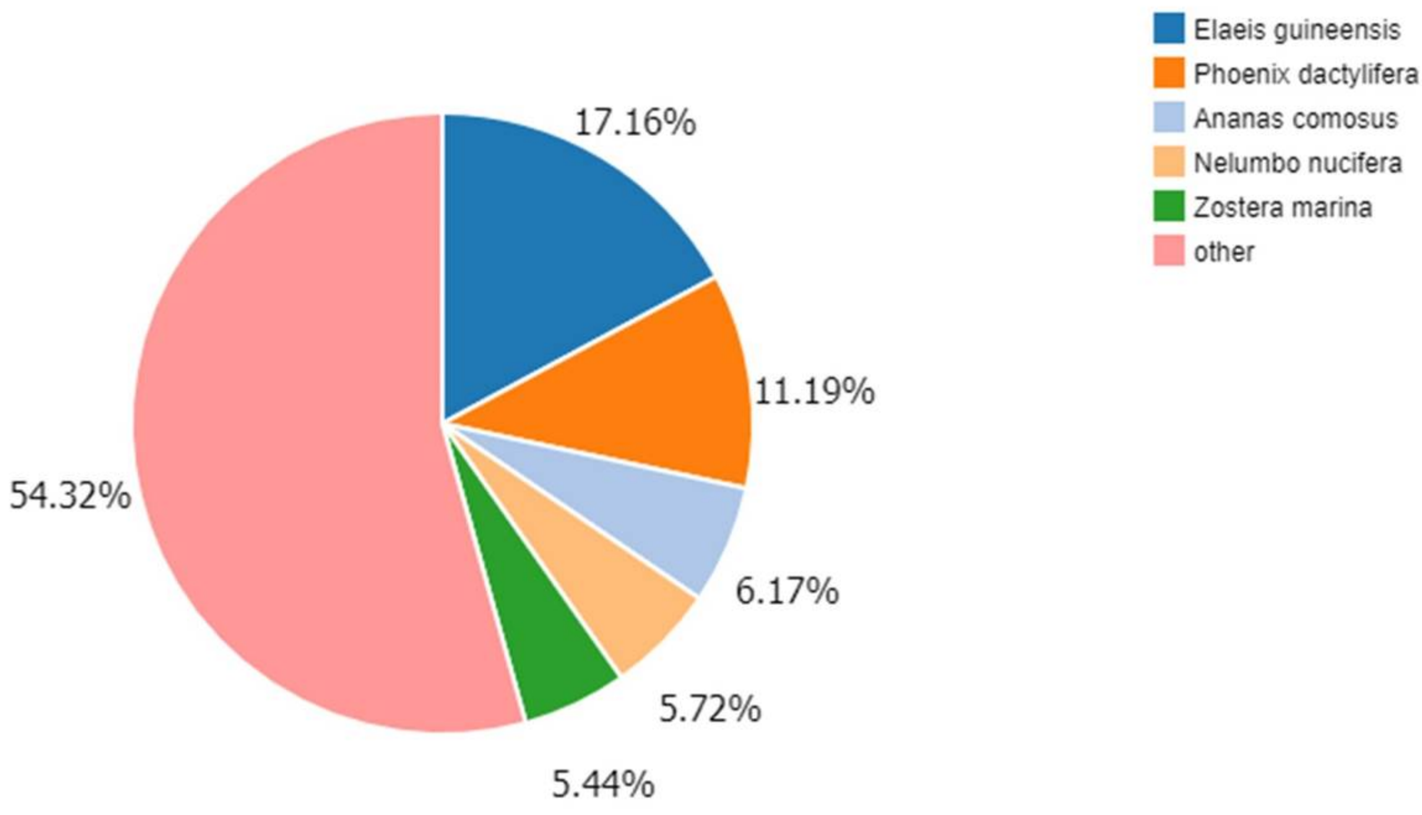

Figure 4

According to the results of NR, the proportion of different species on the annotation is calculated. According to the annotation results of the NR database, the proportion of different species on the isoform annotation is counted, and the species distribution map is drawn

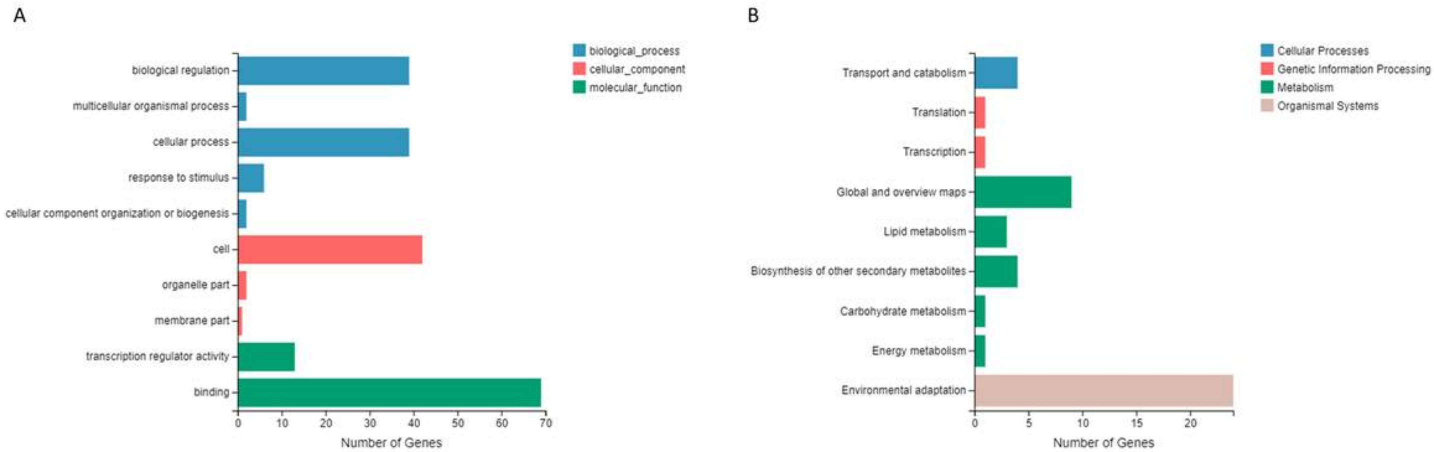

Figure 5

MYB gene family GO and KEGG classification. A. The GO classification of T. hemprichii (Pvalue <= 0.001; FDR <= 0.01), Gene Ontology is divided into three functional categories: molecular function, cellular component and biological process. Functional classification based on differential gene test results. There are sub-categories for each level under each major category; B. The KEGG in the differentially expressed genes of T. hemprichii (Pvalue <= 0.001; FDR <= 0.01). Is classified according to the results of KEGG annotations and official classification. We classify the MYB genes into biological pathways and divide the genes involved in the KEGG metabolic pathway into 10 branches. Statistics are further classified under each branch.

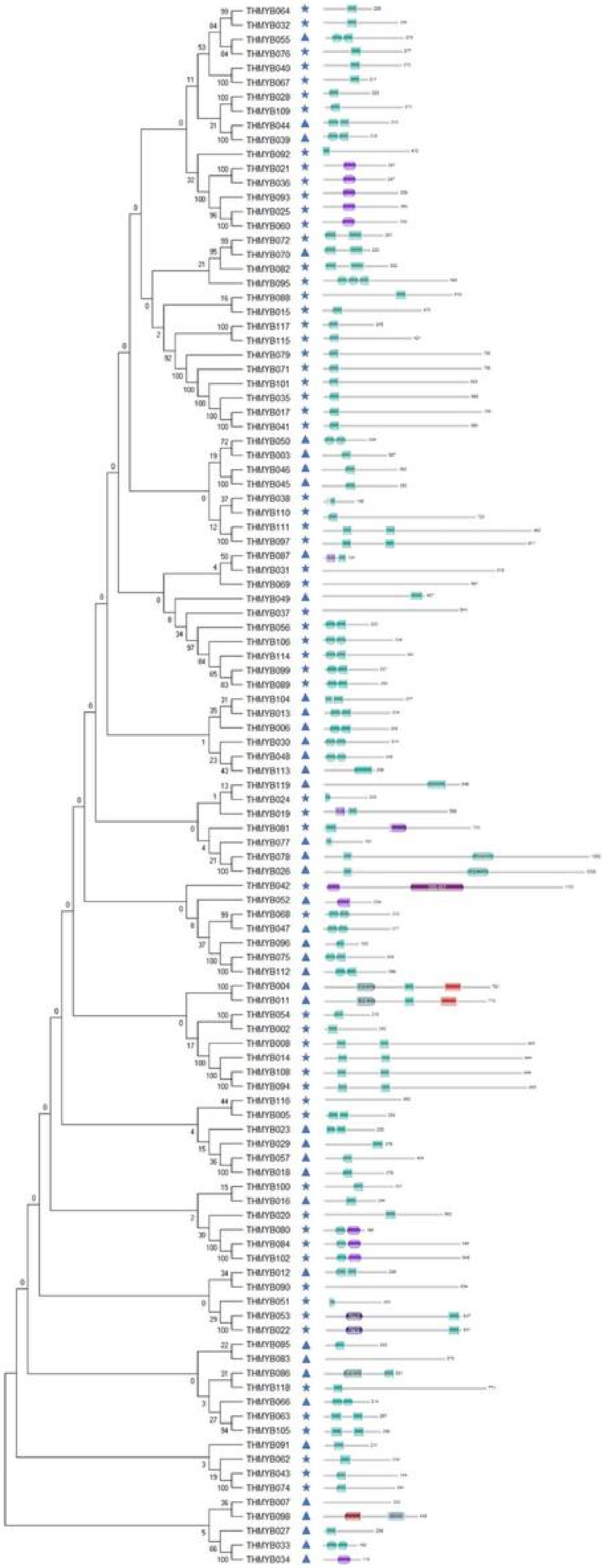
THMYB064
THMYB032
THMYB055
THMYB076
THMYB040
THMYB067
THMYB028
THMYB109
THMYB044
THMYB039
THMYB092
THMYB021
THMYB036
THMYB093
THMYB025
THMYB060
THMYB072
THMYB070
THMYB082
THMYB095
THMYB088
THMYB015
THMYB117
THMYB115
THMYB079
THMYB071
THMYB101
THMYB035
THMYB017
THMYB041
THMYB050
THMYB003
THMYB046
THMYB045
THMYB038
THMYB110
THMYB111
THMYB097
THMYB087
THMYB031
THMYB069
THMYB049
THMYB037
THMYB056
THMYB106
THMYB114
THMYB099
THMYB089
THMYB104
THMYB013
THMYB006
THMYB030
THMYB048
THMYB113
THMYB119
THMYB024
THMYB019
THMYB081
THMYB077
THMYB078
THMYB026
THMYB042
THMYB052
THMYB068
THMYB047
THMYB096
THMYB075
THMYB112
THMYB004
THMYB011
THMYB054
THMYB002
THMYB008
THMYB014
THMYB108
THMYB094
THMYB116
THMYB005
THMYB023
THMYB029
THMYB057
THMYB018
THMYB100
THMYB016
THMYB020
THMYB080
THMYB084
THMYB102
THMYB012
THMYB090
THMYB051
THMYB053
THMYB022
THMYB085
THMYB083
THMYB086
THMYB118
THMYB066
THMYB063
THMYB105
THMYB091
THMYB062
THMYB043
THMYB074
THMYB007
THMYB098
THMYB027
THMYB033
THMYB034

Figure 6

Phylogenetic relationships and protein domain predict. The amino acid sequences of 119 ThMYBs were aligned by the Clustal W program in MEGA, and the phylogenetic tree Was constructed by the NJ method with 1,000 bootstrap replicates. Bootstrap values >50 were indicated on the nodes. Different subgroups were marked with alternating tones of a gray background to make subgroups identification easier.

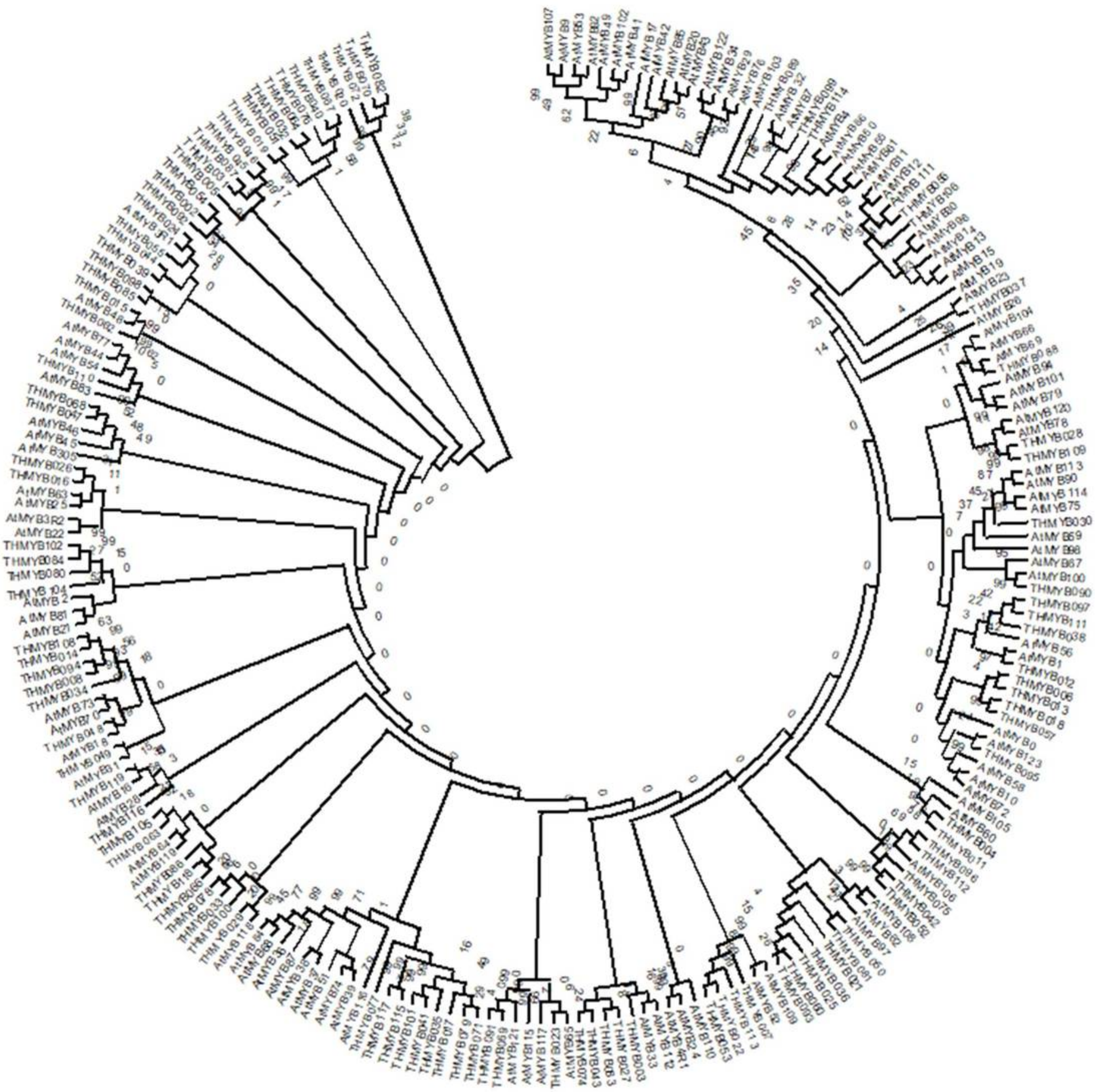

Figure 7

Putative functions of the MYB proteins in T. hemprichii based on the phylogenetic tree. Along with MYBs from Arabidopsis. The circular unrooted tree was generated by NJ method with 1,000 bootstrap replicates.

Supplementary Files

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

- Additionalfiles4.xlsx
- Additionalfiles3.xlsx
- FigureS1.JPG
- Additionalfiles2.xlsx