Whole Transcriptome Analysis of Botrychium ternatum and Screening of Plant Hormone Signal Transduction-Related Genes (Postprint)

Authors: Zhang Linsu, Han Zhongyao, Wang Chuanming, Deng Xiankuo

Date: 2020-01-12T00:00:00+00:00

Abstract

Botrychium is a commonly used medicinal plant in the family Botrychiaceae and genus Botrychium, whose growth and development are representative. To obtain its transcriptome and other biological information, next-generation sequencing and analysis were performed. This study used fresh whole plants of Botrychium as material for whole-transcriptome sequencing on the Illumina HiSeq 2500 platform. Clean reads were assembled to obtain unigenes, which were then subjected to bioinformatic analysis against the non-redundant protein database (Nr), Nucleotide Sequence Database (Nt), Gene Ontology (GO), Clusters of Eukaryotic Orthologous Groups (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), SwissProt, and Interpro. The results showed that a total of 6.67 Gb of clean reads were obtained, yielding 58,646 unigenes with an average length of 1,023 bp. The overall annotation rate of unigenes across the aforementioned databases was 69.25%. Specifically, in the GO database, 20,762 genes were annotated to 52 functional groups within three ontologies: biological process, cellular component, and molecular function. COG annotated 20,633 genes and classified them into 25 functional categories. In the KEGG database, 29,377 genes were annotated and could be divided into 5 major categories and 19 subcategories of metabolic pathways. Based on this analysis, 41 gene families related to eight categories of plant hormone signal transduction were identified. Through alignment, a total of 43,102 coding sequences (CDS) were obtained, with an average length of 749 bp and an N50 of 1,137. Sixty transcription factor (TF) families comprising 1,502 transcription factor genes were identified. A total of 17,195 single nucleotide polymorphism (SNP) sites were discovered, including 11,122 transitions and 6,073 transversions. Additionally, 8,245 simple sequence repeats (SSRs) were found, with dinucleotide repeats and trinucleotide repeats being the most abundant. These results provide comprehensive transcriptome information for Botrychium at both functional and structural levels, as well as

potential genes involved in plant hormone signal transduction, thereby furnishing fundamental molecular biological data for further in-depth research on the growth and development, genetics, and variety identification of Botrychium.

Full Text

Preamble

Global Transcriptome Analysis of *Botrychium ternatum* and Screening of Plant Hormone Signal Transduction-Related Genes

Zhang Linsu, Han Zhongyao, Wang Chuanming, Deng Xiankuo
Department of Pharmacy, Qiannan Medical College for Nationalities, Duyun 558000, Guizhou, China

---

Abstract

Botrychium ternatum is a commonly used medicinal plant in the Ophioglossaceae family whose growth and development exhibit representative characteristics of ferns. To obtain its biological information including transcriptome data, we performed next-generation sequencing and analysis. Using fresh whole plants of *B. ternatum* as material, we conducted global transcriptome sequencing on the Illumina HiSeq 2500 platform. Clean reads were assembled into unigenes, which were then subjected to bioinformatic analysis across seven databases: the non-redundant protein/nucleotide database (Nr), Nucleotide Sequence Database (Nt), Gene Ontology (GO), Clusters of Eukaryotic Orthologous Groups (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), SwissProt, and InterPro. The results yielded 6.67 Gb of clean sequences, which assembled into 58,646 unigenes with an average length of 1,023 bp. The overall annotation rate was 69.25% across all databases. Specifically, 20,762 genes were annotated in GO across three ontologies (biological process, cellular component, and molecular function) covering 52 functional groups. COG annotated 20,633 genes and classified them into 25 functional clusters. KEGG annotated 29,377 genes that could be divided into 5 major categories and 19 sub-pathways, from which we screened 41 gene families related to eight classes of plant hormone signal transduction. BLAST analysis identified 43,102 coding sequences (CDS) with an average length of 749 bp and N50 of 1,137. We screened 60 transcription factor (TF) families comprising 1,502 TF genes. Additionally, we discovered 17,195 single-nucleotide polymorphism (SNP) sites, including 11,122 transitions and 6,073 transversions, and 8,245 simple sequence repeats (SSRs), with dinucleotide and trinucleotide repeats being the most abundant. These results provide comprehensive transcriptome information for *B. ternatum* from both functional and structural perspectives, including potential genes involved in plant hormone signal transduction, offering foundational molecular data for further research on its growth, development, genetics, and variety identification.

Keywords: *Botrychium ternatum*, transcriptome, plant hormones, signal transduction, gene screening

Introduction

Plant hormones are small signaling molecules that play crucial roles in plant growth and development, functioning through plant hormone signal transduction systems. Internal or external stimuli induce expression of a series of plant hormone genes, which act on corresponding hormone receptors or components to ultimately manifest different traits (Su et al., 2008). Common plant hormones include auxin, cytokinin, gibberellin, abscisic acid, ethylene, brassinosteroid, jasmonic acid, and salicylic acid. In these signal transduction systems, some receptors or key components interact or exhibit crosstalk, creating synergistic or antagonistic effects that network the signaling pathways (Ohri et al., 2015). For example, light signals can regulate root development through crosstalk with auxin signaling pathways (Kumari & Panigrahi, 2019), while phytochrome-interacting factors (PIFs) can respond to gibberellin, brassinosteroid, jasmonic acid, auxin indole-3-acetic acid (IAA), abscisic acid, and ethylene signaling pathways, linking these hormone signaling networks through this "hub" molecule (Ren et al., 2016). Plant hormones also promote flowering through epigenetic regulation; gibberellin, jasmonic acid, abscisic acid, and auxin play important roles in DNA methylation and histone post-translational modification-mediated chromatin compaction, thereby affecting flowering (Campos-Rivero et al., 2017). Additionally, plants have evolved complex hormone signaling networks to protect themselves against soil pathogen attacks (Berens et al., 2017). Thus, plant hormone signal transduction systems are vital for plant growth, development, defense, and environmental adaptation.

Botrychium ternatum, also known as "Yi Duo Yun," "Xiao Chun Hua," "She Bu Jian," and other local names, belongs to the Ophioglossaceae family and Ophioglossum genus. It is an annual herbaceous medicinal plant that primarily reproduces via spores, and its growth and development are representative of ferns. Widely used in traditional Chinese medicine, particularly in Guizhou and Fujian provinces, it has heat-clearing, detoxifying, cough-relieving, and hemostatic effects, and is mainly used to treat infantile convulsions due to high fever, lung-heat cough, hemoptysis, whooping cough, snakebites, conjunctivitis, and corneal opacity (Qi, 2012; Zhao et al., 2008; Ruan, 2002). Current research on *B. ternatum* has focused primarily on chemical constituents, clinical and pharmacological effects, and classification and distribution surveys, with limited molecular biology information that restricts deeper investigation.

The transcriptome refers to the complete set of transcripts in a cell under specific physiological conditions, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNA. With the development and popularization of sequencing technology, transcriptome sequenc-

ing (RNA-seq) has become an important method for studying genes and their regulation at the molecular level. In this study, we obtained the global transcriptome of *B. ternatum* through high-throughput sequencing and analyzed it using bioinformatic methods to obtain overall annotation information, screen for potential genes involved in plant hormone signal transduction, and identify single-nucleotide polymorphisms (SNPs) and short sequence repeat polymorphisms (SSRs). This provides valuable molecular data for further research on the growth, development, and variety identification of *B. ternatum*.

1.1 Materials

Fresh, mature whole plants of *B. ternatum* (including roots, stems, leaves, and spores) were collected in July 2016 from the Doupeng Mountain area of Duyun City, Qiannan Prefecture, Guizhou Province (altitude ~1,500 m, 107°20 - 107°27 E, 26°12 -26°16 N). The plants were identified by Associate Professor Wang Chuanming of Qiannan Medical College for Nationalities. Samples were immediately rinsed with clean water, dried with absorbent paper, placed in a dry ice box, and transported back for RNA extraction.

1.2 cDNA Library Preparation and Sequencing

Whole plants were ground into powder in liquid nitrogen. Total RNA was extracted using an RNA extraction kit (Aidlab, Beijing) with DNA digestion. mRNA was enriched using magnetic beads with Oligo(dT), and quality was assessed via agarose electrophoresis and NanoDrop micro-volume nucleic acid detection. After passing quality control, cDNA was synthesized using a kit, followed by purification, end-repair, 3 -end adenylation, adapter ligation, fragment size selection, and PCR amplification to construct the cDNA library. The qualified library was then sequenced on the Illumina HiSeq platform.

1.3 De Novo Assembly

Raw reads were filtered to remove low-quality sequences, adapter contamination, and reads with excessive unknown bases (N), yielding clean reads. Trinity software (v2.0.6) (Grabherr et al., 2011) was used for de novo assembly of clean reads, followed by clustering and redundancy removal using TGICL software (v2.0.6) (Pertea et al., 2003) to obtain unigenes for subsequent analysis.

1.4 Unigene Functional Annotation and Analysis

To understand unigene functions, we annotated them across seven functional databases using bioinformatic software: BLAST (v2.2.23) for NT, NR, COG, KEGG, and SwissProt annotations; Blast2GO (v2.5.0) (Conesa et al., 2005) with NR annotation results for GO annotation; and InterProScan5 (Quevillon et al., 2005) for InterPro annotation. Plant hormone signal transduction-related

genes were identified by classifying annotated genes according to the KEGG signaling pathway map04075.

1.5 Transcriptome Structure Analysis

CDS Prediction: Based on functional annotation results, the best-aligned segment from NR, SwissProt, KEGG, and COG databases was selected as the CDS for each unigene following database priority order. Unigenes without annotation were modeled using ESTScan (v3.0.2) (Iseli et al., 1999) for CDS prediction.

TF Coding Capacity Prediction: Open reading frames (ORFs) were detected using getorf (EMBOSS:6.5.7.0) (Rice et al., 2000). ORFs were aligned to transcription factor protein domains from PlantTFDB using hmmsearch (v3.0) (Mistry et al., 2013), and unigenes were identified for TF coding capacity based on TF family characteristics described in PlantTFDB (Jin et al., 2017).

SSR and SNP Detection: SSRs were detected in unigenes using MISA (v1.0) (Thiel et al., 2003). Clean reads were aligned to unigenes using HISAT (v0.1.6-beta) (Kim et al., 2015), and SNPs were detected using GATK (v3.4-0) (McKenna et al., 2010).

2.1 Sequencing and Assembly Results

The Illumina HiSeq platform generated 55.52 Mb of raw reads, which after filtering yielded 44.45 Mb of clean reads, achieving a clean read rate of 80.6%. The sequencing depth was considered "deep" (>15 Mb). The total clean bases reached 6.67 Gb, assembling into 58,646 unigenes with an average length of 1,023 bp. Both N50 and N70 exceeded 1,000 bp (Table 1). All unigenes were longer than 300 bp, with the majority (25.5%) ranging from 300-400 bp, and 39% exceeding 1,000 bp (Figure 1 [Figure 1: see original paper]), indicating good sequencing continuity and assembly quality.

2.2 Unigene Functional Annotation

Unigenes were annotated across seven functional databases (NR, NT, GO, COG, KEGG, SwissProt, and InterPro). The results are summarized in Table 2 . NR (NCBI protein database) provided the most annotations (65.4%), with an overall annotation rate of 69.25% across all databases. Based on NR annotation results, we analyzed species distribution (Figure 2 [Figure 2: see original paper]). Fern species *Physcomitrella patens* and *Selaginella moellendorffii* accounted for 24% of annotations, consistent with *B. ternatum* being a fern. Additionally, the reference species *Picea sitchensis* showed high annotation frequency (14.21%), likely due to its well-annotated genome (Ralph et al., 2008). The Venn diagram of NR, COG, KEGG, SwissProt, and InterPro annotations (Figure 3 [Figure

3: see original paper]) revealed 12,522 unigenes (21.4%) shared across all five databases, indicating high annotation reliability for these genes.

2.3 GO Annotation Results

GO annotation assigned 20,762 *B. ternatum* genes or gene products to three main categories: molecular function, cellular component, and biological process. The functional distribution is shown in Figure 4 [Figure 4: see original paper]. In biological process, the top three terms were metabolic process, cellular process, and single-organism process. In cellular component, "cell" was most abundant while "nucleotide" was least common. In molecular function, catalytic activity and binding were most numerous, followed by transporter activity.

2.4 COG Functional Annotation

COG database comparison annotated 20,633 *B. ternatum* unigenes, with results shown in Figure 5 [Figure 5: see original paper]. The largest cluster was "general function prediction only" (4,559 genes). Eight clusters contained 1,000-2,000 genes, including essential biological activities such as translation, ribosomal structure and biogenesis, and transcription. Notably, 995 genes were identified as "function unknown."

2.5 KEGG Pathway Analysis and Plant Hormone Signal Transduction Gene Screening

A total of 29,377 genes were mapped to six major categories and 21 sub-pathways (Figure 6 [Figure 6: see original paper]). Metabolism pathways contained the most genes (17,698; 60%), while human disease-related genes were least abundant (141 genes, as expected for a plant species). Environmental adaptation-related genes in the organismal systems category numbered 1,266. Based on KEGG signaling pathway map04075, we classified annotated genes to identify candidate genes related to plant hormone signal transduction (Table 3).

2.6 Transcriptome Structure

CDS: BLAST analysis identified 38,212 CDS, and ESTScan predicted an additional 4,890 CDS, totaling 43,102 CDS with an average length of 749 bp and N50 of 1,137.

TF: We screened 60 transcription factor gene families comprising 1,502 TF genes. Families with over 100 members included C3H, MYB, MYB-related, and bHLH. Other abundant families included AP2-EREBP, WRKY, and GRAS.

SNP: A total of 17,195 SNP sites were discovered, including 11,122 transitions (5,452 A-G and 5,670 C-T) and 6,073 transversions (1,444 A-C, 1,729 A-T, 1,418 C-G, and 1,482 G-T).

SSR: Dinucleotide repeats were most abundant (3,666), followed by trinucleotide repeats (3,439), then mononucleotide (563), hexanucleotide (260), tetranucleotide (169), and pentanucleotide repeats (148).

Discussion

B. ternatum is a non-model medicinal plant commonly used in traditional medicine, with flavonoids and polysaccharides as its main active components. RNA-seq yielded 6.67 Gb of clean bases, with 各项指标 indicating good sequencing depth and assembly quality. Annotation across seven databases revealed NR had the highest annotation rate (65.4%). Since NR contains both verified and predicted proteins with large data volume, its high annotation rate should be interpreted in combination with other databases. The Venn diagram analysis of NR, COG, KEGG, SwissProt, and InterPro annotations (Figure 3 [Figure 3: see original paper]) showed 12,522 unigenes shared across all five databases, suggesting high annotation reliability for this subset.

GO, COG, and KEGG annotations are essential for gene annotation, describing genes from perspectives of ontology, clustering, and pathways, respectively. COG analysis identified 995 unigenes of unknown function (4.8%). *B. ternatum* belongs to Eusporangiopsida class and Ophioglossales order, representing numerous fern species that are widely distributed and ancient, with most reproducing via spores. Ferns occupy a transitional position between lower and higher plants, exhibiting unique developmental characteristics (Christenhusz & Chase, 2014; Zhang et al., 2016). Comparing the *B. ternatum* transcriptome with other ferns such as *Huperzia serrata* (Yang et al., 2017) revealed similar overall annotation rates (55-60%). In both plants, metabolism pathways contained the most KEGG-annotated genes, and environmental adaptation genes represented a significant proportion, suggesting ferns mobilize numerous genes for environmental adaptation, enabling their survival as an ancient plant lineage. Therefore, studying the growth and development of ferns like *B. ternatum* is meaningful, particularly the screened but functionally unknown genes warranting further investigation.

Plant hormones are crucial factors affecting plant growth and development. KEGG analysis identified genes in all eight known plant hormone signal transduction pathways. Most gene families contained few members, facilitating subsequent cloning, analysis, and functional characterization. However, some families

were large, such as TF and DELLA in gibberellin signaling, BRI1 in brassinosteroid signaling, B-ARR in cytokinin signaling, and PP2C in abscisic acid signaling, each containing dozens or even hundreds of members. Further work should identify key genes and narrow the research scope through differential expression analysis across tissues, organs, or treatments—a limitation of the current study. Additionally, although both are ferns, *B. ternatum* and *H. serrata* exhibit vastly different growth characteristics: *B. ternatum* grows rapidly and reproduces annually via spores, while *H. serrata* grows slowly with spore germination requiring several years (Guo et al., 2009). Comparing hormone signal transduction gene numbers (our lab data deposited at http://bigd.big.ac.cn/gsa, accession PRJCA001325) revealed significant differences in some gene families with more than two-fold variation. *B. ternatum* contained more genes in gibberellin signaling TF (PIF4 and PIF3), brassinosteroid signaling BAK1, BZR1/2, and TCH4 families, while *H. serrata* had more genes in cytokinin signaling CRE1 and A-ARR, brassinosteroid signaling BSK, and abscisic acid signaling NPR1 and TGA families. This suggests these genes may have different growth regulatory roles, potentially contributing to the distinct growth patterns of these two ferns.

Besides endogenous hormones, external environmental factors such as light, temperature, stimuli, and soil affect plant growth and development. Many studies aim to identify key nodes of these influencing factors (e.g., light-temperature interactions, pests, and microbes). Multi-omics and systems biology approaches are valuable strategies for studying such complex regulatory networks (Meena et al., 2017; Myburg et al., 2019; Choi, 2019).

Acknowledgments

We thank the Beijing Genomics Institute (BGI) for technical support and Mr. Guo Shichuan for assistance with plant sample collection.

References

Berens ML, Berry HM, Mine A, et al., 2017. Evolution of hormone signaling networks in plant defense[J]. *Annu Rev Phytopathol*, 55: 401-425.

Campos-Rivero G, Osorio-Montalvo P, Sanchez-Borges R, et al., 2017. Plant hormone signaling in flowering: An epigenetic point of view[J]. *J Plant Physiol*, 214: 16-27.

Choi HK, 2019. Translational genomics and multi-omics integrated approaches as a useful strategy for crop breeding[J]. *Genes Genom*, 41(2): 133-146.

Christenhusz MJ, Chase MW, 2014. Trends and concepts in fern classification[J]. *Ann Bot*, 113(4): 571-594.

Conesa A, Götz S, García-Gómez JM, et al., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research[J]. *Bioinformatics*, 21(18): 3674-3676.

Grabherr MG, Haas BJ, Yassour M, et al., 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data[J]. *Nat Biotechnol*, 29(7): 644-652.

Guo B, Xu L, Wei Y, et al., 2009. Research progress of *Huperzia serrata*[J]. *Zhongguo Zhong Yao Za Zhi*, (16): 2018-2022.

Iseli C, Jongeneel CV, Bucher P, 1999. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences[J]. *Proc Int Conf Intell Syst Mol Biol*, 99: 138-148.

Jin JP, Tian F, Yang DC, et al., 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants[J]. *Nucleic Acids Res*, 45(D1): D1040-D1045.

Kim D, Langmead B, Salzberg SL, 2015. HISAT: a fast spliced aligner with low memory requirements[J]. *Nat Methods*, 12(4): 357-360.

Kumari S, Panigrahi KCS, 2019. Light and auxin signaling cross-talk programme root development in plants[J]. *J Biosci*, 44(1): 26.

McKenna A, Hanna M, Banks E, et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data[J]. *Genome Res*, 20(9): 1297-1303.

Meena KK, Sorty AM, Bitla UM, et al., 2017. Abiotic stress responses and microbe-mediated mitigation in plants: The omics strategies[J]. *Front Plant Sci*, 8: 172.

Mistry J, Finn RD, Eddy SR, et al., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions[J]. *Nucleic Acids Res*, 41(12): e121.

Myburg AA, Hussey SG, Wang JP, 2019. Systems and synthetic biology of forest trees: A bioengineering paradigm for woody biomass feedstocks[J]. *Front Plant Sci*, 10: 1145.

Ohri P, Bhardwaj R, Bali S, et al., 2015. The common molecular players in plant hormone crosstalk and signaling[J]. *Curr Protein Pept Sci*, 16(5): 369-388.

Pertea G, Huang X, Liang F, et al., 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets[J]. *Bioinformatics*, 19(5): 651-652.

Qi JH, 2012. A summary of recent studies on *Botrychium* Sw[J]. *J Xi' an Univ Arts Sci (Nat Sci Ed)*, 15(2): 48-50.

Quevillon E, Silventoinen V, Pillai S, et al., 2005. InterProScan: protein domains identifier[J]. *Nucleic Acids Res*, 33: W116-W120.

Ralph SG, Chun HJ, Kolosova N, et al., 2008. A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*)[J]. *BMC Genomics*, 9: 484.

Ren XY, Wu MQ, Chen JM, et al., 2016. The molecular mechanisms of phytochrome interacting factors (PIFs) in phytohormone signaling transduction[J]. *J Plant Physiol*, 52(10): 1466-1473.

Rice P, Longden I, Bleasby A, 2000. EMBOSS: the European Molecular Biology Open Software Suite[J]. *Trends Genet*, 16(6): 276-277.

Ruan JS, 2002. Research progress of *Sceptridium ternatum* and its effective ingredients[J]. *J Chin Pharm Univ*, 33: 328-329.

Su Q, An D, Wang K, 2008. Phytohormone receptors and induced genes in plants[J]. *Plant Physiol Mol Biol*, 44(6): 1202-1208.

Thiel T, Michalek W, Varshney RK, et al., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.)[J]. *Theor Appl Genet*, 106(3): 411-422.

Yang M, You W, Wu S, et al., 2017. Global transcriptome analysis of *Huperzia serrata* and identification of critical genes involved in the biosynthesis of huperzine A[J]. *BMC Genomics*, 18: 245.

Zhang KM, Shen Y, Liu Y, et al., 2016. Research progress on development and physio-ecology of fern gametophytes[J]. *Guihaia*, 36(4): 419-424.

Zhao JH, Zhao NW, Wang PS, et al., 2008. Study on the species and distribution of *Adiantum* and *Botrychium* medicinal plants from Tujia medicine of Guizhou Province origin[J]. *J Med Pharm Chin Minor*, 5: 44-46.

Note: Figure translations are in progress. See original paper for figures.