Angelica dahurica Whole Genome Sequencing Analysis and BGLU Gene Family Analysis Post-print

Authors: Wang Yalan, Zhou Luojing, Zhang Lingyu, Zhang Jing, Bian Jinhui, Gao Jihai

Date: 2023-08-25T00:00:00+00:00

Abstract

Angelica dahurica is a commonly used species with both medicinal and edible properties, serving as both a clinically employed traditional Chinese medicine and a spice, with extensive applications. To obtain whole-genome sequence information of Angelica dahurica, this study employed leaf DNA of Hangbai Zhi (Angelica dahurica var. formosana) as material, utilized Nanopore sequencing technology to construct a whole-genome database of Hangbai Zhi, and applied bioinformatics methods to conduct assembly, functional annotation, and evolutionary analysis of the acquired nucleotide sequences. The results demonstrated: (1) Following filtration of raw sequencing data, 662 Gb of third-generation data was obtained, with a Read N50 of approximately 32 932 bp; assembly yielded a genome size of 5.6 Gb for Hangbai Zhi, with a Contig N50 of approximately 806 638 bp. (2) Through comparison of assembled sequences against functional databases including KOG, GO, and KEGG, 66.47% of genes received functional annotation. KOG functional annotation results indicated that protein functions of Hangbai Zhi were primarily concentrated in general function prediction, post-translational modification, protein turnover, chaperones, and signal transduction mechanisms; GO functional classification revealed that genes of Hangbai Zhi were enriched in biological processes and cellular components; KEGG pathway annotation demonstrated that genes participating in metabolic pathways constituted the predominant category. (3) Hangbai Zhi contained 45 genes in the BGLU family. This study represents the first comprehensive analysis of the Hangbai Zhi whole genome utilizing third-generation sequencing technology, establishing a foundation for systems biology research on Hangbai Zhi and facilitating further in-depth development and utilization of this species. Simultaneously, it provides a preliminary analysis of BGLU family genes in Hangbai Zhi, offering an important theoretical basis for subsequent investigations into the functions of BGLU in the growth and development of Hangbai Zhi.

Full Text

Complete Genome Sequencing and BGLU Gene Family Analysis of Angelica dahurica

WANG Yalan, ZHOU Luojing, ZHANG Lingyu, ZHANG Jing, BIAN Jinhui, GAO Jihai*

Key Laboratory of Distinctive Chinese Medicine Resources in Southwest China, Chengdu University of Traditional Chinese Medicine, Chengdu 611137, China

Abstract

Angelica dahurica is a common plant with both medicinal and edible value, widely used in clinical practice as a traditional Chinese medicine and also as a spice. To obtain the whole genome sequence information of A. dahurica, this study used leaf DNA from A. dahurica var. formosana as material, constructed a whole-genome database using Nanopore sequencing technology, and performed assembly, functional annotation, and evolutionary analysis of the obtained nucleotide sequences through bioinformatic methods. The results showed: (1) After filtering the raw sequencing data, 662 Gb of third-generation data were obtained with a Read N50 of approximately 32,932 bp. The assembled genome size of A. dahurica var. formosana was 5.6 Gb with a Contig N50 of approximately 806,638 bp. (2) Through comparison with functional databases including KOG, GO, and KEGG, 66.47% of the assembled sequences received functional annotation. KOG functional annotation results indicated that the protein functions of A. dahurica var. formosana were mainly concentrated in general function prediction, posttranslational modification, protein turnover, chaperones, and signal transduction mechanisms. GO functional classification showed that genes were concentrated in biological processes and cellular components. KEGG pathway annotation revealed that genes involved in metabolic pathways were predominant. (3) The study identified 45 BGLU family genes in A. dahurica var. formosana. This study represents the first analysis of the whole genome of A. dahurica var. formosana using third-generation sequencing technology, laying a foundation for systematic biological research on this species and facilitating its further development and utilization. Meanwhile, the preliminary analysis of the BGLU family genes provides an important theoretical basis for subsequent studies on the function of BGLU in the growth and development of A. dahurica var. formosana.

Keywords: Angelica dahurica var. formosana, genome, third-generation sequencing technology, BGLU gene family, medicinal plant

Supported by: Major Increase and Decrease of Expenditures of the Central Government (2060302); National Administration of Traditional Chinese Medicine (ZYYCXTD-D-202209); Science and Technology Program of Science and Technology Department of Sichuan Province (2020YFN0152, 22CXTD0009); Sichuan Provincial Administration of Traditional Chinese

Medicine (2022C001); Talent Promotion of Chengdu University of Traditional Chinese Medicine Project (QNXZ2018017, QNXZ2019001)

First author: WANG Yalan (1998-), Master's degree candidate, research direction: analysis and application of active components in traditional Chinese medicine, (E-mail) wangyalan@stu.cdutcm.edu.cn.

Corresponding author: GAO Jihai, Ph.D., Associate Professor, mainly engaged in molecular pharmacognosy research, (E-mail) gaojihai@cdutcm.edu.cn.

Introduction

Angelica dahurica (Apiaceae) is the dried root of Angelica dahurica or A. dahurica var. formosana, mainly produced in Sichuan, Hangzhou, and other regions, mostly cultivated. As a common medicinal and edible herb, it is clinically used for various pain symptoms such as headache from colds, supraorbital bone pain, toothache, and sore swellings from boils [National Pharmacopoeia Commission, 2020], and also used as a spice in daily life. Due to its aromatic properties, it is widely applied in cosmetics and personal care products [Yu et al., 2014]. A. dahurica contains various active components, including coumarins, volatile oils, polysaccharides, and alkaloids [Li et al., 2014; Zhao et al., 2022]. Modern research indicates that its main active ingredients are coumarins and volatile oils, which exhibit multiple pharmacological effects such as antipyretic and analgesic, anti-inflammatory, antimicrobial, antitumor, blood pressure lowering, and hepatoprotective activities [Ji et al., 2020; Wang et al., 2020].

The application prospects of A. dahurica are extensive, but recent research has mostly focused on chemical composition, cultivation techniques, and pharmacological effects, with few studies on its genetic information. Currently, only transcriptome sequencing analyses [Wu et al., 2020] and studies on specific gene families such as CONSTANS-like [Jiang et al., 2021], NAC [Huang et al., 2021], and MYB-related [Yao et al., 2022] have been reported, all based on transcriptome data. The lack of genomic data for A. dahurica prevents acquisition of complete genetic information and hinders further research, making whole-genome sequencing essential.

Coumarins are both medicinal and aromatic components in A. dahurica. These compounds are widely present in various plants from families such as Apiaceae, Rutaceae, and Moraceae [Venugopala et al., 2013]. Recent research has extensively investigated coumarin biosynthesis pathways, with clear elucidation of some key enzymes and their functions [Duan et al., 2022]. Among these, β-glucosidase (BGLU) plays important regulatory roles not only in coumarin biosynthesis but also in various physiological processes including plant hormone signal activation [Sun et al., 2014] and secondary metabolism [Sampedro et al., 2017]. Studies have shown that the β-glucosidase family plays crucial reg-

ulatory roles in coumarin synthesis in Melilotus [Wu, 2021]. In maize, BGLU can hydrolyze β-glycosidic bonds between carbohydrate moieties and coumarin core structures to produce coumarin aglycone forms. Aspergillus niger-derived β-glucosidase can specifically hydrolyze scopolin in crude extracts of Erycibe and increase its content by 47% [Yu et al., 2023]. Three β-glucosidases isolated from Arabidopsis thaliana can specifically hydrolyze scopolin. Scopolin is hydrolyzed into scopoletin under the action of β-glucosidase, and scopoletin belongs to coumarin components that also exist in A. dahurica. Our research group hypothesizes that BGLU genes also play key roles in coumarin component synthesis in A. dahurica.

As no high-quality genome studies on A. dahurica have been reported and few studies have analyzed coumarin synthesis pathways in this species, this study performed second- and third-generation genome sequencing of A. dahurica var. formosana. Through assembly, annotation, and other analyses of the sequencing data, we obtained a high-quality genome, performed functional annotation and gene family clustering analyses, and mined key BGLU genes in the coumarin synthesis pathway. Using online software for basic characteristic analysis of BGLU sequences extracted from the genome, this study aims to address: (1) the general profile of the A. dahurica var. formosana genome; (2) which biological processes and metabolic pathways the gene functions are mainly concentrated in; and (3) the basic characteristics of the BGLU gene family. This study will provide data and molecular foundations for subsequent research on A. dahurica and preliminary basis for in-depth studies on the function of the BGLU gene family in coumarin synthesis pathways.

Materials and Methods

1.1 Materials and DNA Extraction

A. dahurica var. formosana plants were obtained from the Medicinal Botanical Garden of Chengdu University of Traditional Chinese Medicine and identified by Associate Professor Gao Jihai, an expert from the National Chinese Herbal Medicine Germplasm Resources Bank, as A. dahurica var. formosana (Apiaceae). Fresh, young, and pest-free leaves were collected, washed with distilled water, cleaned three times with 75% ethanol, dried, and stored at -80°C. DNA was extracted from A. dahurica var. formosana leaves using the CTAB method as described by Sha [2018]. The extracted DNA was assessed for concentration via agarose gel electrophoresis and Qubit Fluorometer, and for purity and integrity using Nanodrop.

1.2 Library Construction and Sequencing

(1) MGISEQ-200 sequencing: After qualified genomic DNA extraction, the DNA was randomly fragmented by enzymatic digestion. End repair, A-tailing, sequencing adapter ligation, purification, and PCR amplification were

performed to construct a DNA library with an insert size of 150 bp. The constructed library was subjected to paired-end sequencing on the MGISEQ-200 platform.

(2) Nanopore sequencing: Qualified DNA was enriched and purified using magnetic beads. After damage repair, end repair, and A-tailing, the purified DNA was further purified. The product was ligated with sequencing-related adapters and purified to obtain the final library for sequencing. The constructed DNA library was accurately quantified using Qubit. A certain amount of DNA library was mixed with sequencing reagents and loaded into the flow cell for single-molecule sequencing on the GridION sequencer to obtain raw data.

1.3 Quality Control of Genome Sequencing Data

Raw second-generation sequencing data contains adapter sequences, low-quality bases, and undetermined bases (represented by N) that can significantly interfere with subsequent bioinformatic analyses. These interfering elements were filtered using FastQC v0.11.9 and Trimmomatic v0.39 software to obtain clean reads for downstream analysis. For third-generation Nanopore sequencing data, NanoPlot v1.20.0 was used for quality assessment, followed by filtering of low-quality and short reads using NanoFilt v2.8.0.

1.4 Genome Size and Heterozygosity Assessment

Using reads data obtained from MGISEQ-200 sequencing, Jellyfish v1.1.10 was employed for survey analysis to estimate genome size, heterozygosity rate, and repetitive sequence proportion, thereby determining genome complexity. K-mer analysis was used for this assessment.

1.5 Genome Assembly and Evaluation

To obtain highly accurate third-generation assembly results, Canu v2.1.1 [Koren et al., 2017] was first used to correct clean data. The corrected data were then assembled, and the assembly results were polished using Racon v1.0.0 [Senol et al., 2019]. Pilon v1.22 was subsequently used to correct the assembly using second-generation data. Finally, BUSCO v5.1.2 [Simão et al., 2015] was used to assess the completeness of the assembled genome.

1.6 Sequence Prediction

First, based on structural prediction and ab initio principles, LTR Finder v1.05 [Xu et al., 2007], RepeatScout v1.0.6, and PILER-DF v2.4 software were used to construct a repetitive sequence database. PASTEClassifier v2.0 was used to classify the constructed repetitive sequence library. The library was then merged with the Repbase database (https://www.girinst.org/repbase/) as the final repetitive sequence database for A. dahurica var. formosana. Finally,

RepeatMasker v4.1.2 was used to predict repetitive sequences in A. dahurica var. formosana based on the constructed database.

Gene prediction was performed based on both ab initio and homologous species prediction principles. First, Genscan v1.0, Augustus v3.3.1, GlimmerHMM v3.0.4, GeneID v1.4, and SNAP v8.0 were used for ab initio prediction. GeMoMa v1.3.1 was then used for homologous species-based prediction. Finally, EvidenceModeler v1.1.0 was used to integrate and correct the prediction results from the above methods. For non-coding RNA prediction, including microRNA, rRNA, and tRNA with known functions, rRNA and microRNA were predicted based on the Rfam database [Finn et al., 2006] and miRBase using Infernal v1.1.3. tRNA was identified using tRNAscan-SE v2.0.7.

1.7 Functional Gene Annotation

Predicted gene sequences were compared against functional databases including NR (Non-Redundant Protein Database), KOG (EuKaryotic Orthologous Groups), KEGG (Kyoto Encyclopedia of Genes and Genomes), and TrEMBL using BLAST v2.2.31 with a threshold e-value $<$ 1e-5 to obtain gene functional annotations. Based on NR database comparison results, Blast2GO v5.2.5 was used for GO database functional annotation.

1.8 Gene Family Clustering and Phylogenetic Analysis

To identify gene families, protein sequences of A. dahurica var. formosana and its congeneric species were compared. Protein sequences of celery (Apium graveolens) [Song et al., 2021] and carrot (Daucus carota subsp. sativus) [Iorizzo et al., 2016] were downloaded from the NCBI database, while coriander (Coriandrum sativum) [Song et al., 2020] protein sequences were obtained from CGDB (http://cgdb.bio2db.com). OrthoMCL v2.0 [Li et al., 2003] was used to cluster all-vs-all blastp results of protein sequences from all species. Single-copy protein sequences extracted from OrthoMCL clustering results were aligned using Muscle v3.8.31 [Edgar, 2004], and a phylogenetic tree was constructed using RAxML v8.2.12 [Guindon & Gascuel, 2003] with the maximum likelihood method (ML TREE).

1.9 Mining of BGLU Gene Family Members in A. dahurica var. formosana

Using the SMART database, typical domain sequences of Arabidopsis thaliana BGLU gene family were obtained and used for tBLASTN search (P=0.001) against the A. dahurica var. formosana genome database. All BGLU gene family members in A. dahurica var. formosana were identified through the Pfam database.

1.10 Analysis of Physicochemical Properties, Subcellular Localization, Protein Secondary Structure, and Conserved Domains of BGLU Family Genes

The ProtParam tool (https://web.expasy.org/protparam/) [Wilkins et al., 1999] was used for physicochemical property analysis of BGLU family proteins. Subcellular localization was comprehensively analyzed using Plant-mPLoc (http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/) and PSORT (https://wolfpsort.hgc.jp/). Secondary structure was analyzed using SOMPA (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_{automat}.pl?page=npsa_{sopma}.html). Conserved domains were analyzed using MEME (https://meme-suite.org/meme/tools/meme).

1.11 Phylogenetic Analysis of BGLU Family

Clustal W v2.0 [Larkin et al., 2007] in MEGA software was used to align BGLU family protein sequences of A. dahurica var. formosana and Arabidopsis thaliana. A phylogenetic tree was constructed from the alignment results using the neighbor-joining method.

Results

2.1 Genome Sequencing

Whole-genome sequencing of A. dahurica var. formosana leaves was performed using sequencing platforms. After initial filtering of raw reads quality values to remove low-quality and short reads, 150 Gb of second-generation raw data and 662 Gb of third-generation raw data were obtained. In the third-generation data, the Read N50 was 32,932 bp, the longest read length was 422,833 bp, and the average length was 27,750 bp, meeting the requirements for subsequent assembly. Survey analysis estimated the genome size of A. dahurica var. formosana to be approximately 5.2 Gb.

2.2 Genome Assembly and Evaluation

Using Canu software for error correction and assembly of A. dahurica var. formosana, the genome size was approximately 5.6 Gb with a Contig N50 of 806,638 bp. The longest contig was 21,677,961 bp, and the GC content was 35.73%. The assembled genome was evaluated using BUSCO v5.1.2 software. A total of 1,580 complete BUSCO genes were identified in the assembled genome, including 1,272 complete single-copy genes, 18 fragmented BUSCO genes, and 16 genes not found in the Embryophyta_{odb10} database. The BUSCO assessment indicated a genome completeness of 97.9%, demonstrating that the assembly was relatively complete.

2.3 Gene Prediction Results

Repetitive sequence prediction using RepeatMasker v4.1.2 software revealed that the A. dahurica var. formosana genome contained 5.4 Gb of repetitive sequences, accounting for 91.36% of the genome. Among these, long interspersed nuclear elements (LINEs) numbered 21,726 (0.41%), short interspersed nuclear elements (SINEs) numbered 0, long terminal repeats (LTRs) numbered 3,550,524 (69.07%) including 1,083,004 copia (30.01%) and 989,985 gypsy (24.56%) elements, rolling-circles numbered 2,893 (0.03%), and simple sequence repeats (SSRs) numbered 7,710 (0.03%).

Among the 67,004 predicted genes, 34,119 (93.1%) received support from homologous identification in other species or RNA-seq data. A total of 2,749 noncoding RNAs (ncRNAs) were identified, including 20 ribosomal RNAs (rRNAs), 781 transfer RNAs (tRNAs), 97 small RNAs (microRNAs), and 15,505 small nuclear RNAs (snRNAs).

2.4 Gene Function Annotation and Analysis

KOG functional annotation (Figure 1 [Figure 1: see original paper]) revealed that 29,788 genes in the A. dahurica var. formosana genome received annotation, accounting for 44.46% of the total predicted genes. The results showed that protein functions were mainly concentrated in secondary metabolite biosynthesis, transport, and metabolism (10.8%), followed by signal transduction mechanisms (10.1%), transcription (6.7%), carbohydrate transport and metabolism (3.7%), and general function prediction (22.8%). Differential expression of these genes can provide data support for future in-depth research on A. dahurica var. formosana.

GO annotation of A. dahurica var. formosana (Figure 2 [Figure 2: see original paper]) indicated that 44,540 genes had GO annotation functions, accounting for 66.47% of the total predicted genes. Genes with functions distributed in reproduction, cellular processes, stress response, cell, and cell part were predominant, with genes involved in reproduction being the most abundant.

KEGG pathway annotation (Figure 3 [Figure 3: see original paper]) annotated 15,263 genes in A. dahurica var. formosana, accounting for 22.78% of the total predicted genes. The annotation results indicated that genes involved in "metabolism" were predominant, with major metabolic pathways including microbial metabolism in diverse environments, carbon metabolism, and amino acid biosynthesis.

2.5 Gene Family Clustering and Phylogenetic Analysis

Comparison of protein sequences between A. dahurica var. formosana and its congeneric species coriander, celery, and carrot identified 23,151 gene families among the 67,004 protein sequences in the A. dahurica var. formosana genome. Among these, 4,004 gene families containing 18,151 genes were specific to A.

dahurica var. formosana, while 1,030 gene families were shared among the four plants (Figure 4 [Figure 4: see original paper]).

To further investigate the phylogenetic relationships of A. dahurica var. formosana, 96 single-copy protein sequences were selected for comparative analysis. A phylogenetic tree was constructed with A. dahurica var. formosana and seven species with known genome information, including Arabidopsis thaliana, Zea mays, Amborella trichopoda, and Apiaceae species coriander, celery, carrot, and Angelica sinensis (Figure 5 [Figure 5: see original paper]). The results showed that A. dahurica var. formosana clustered with coriander, indicating a close phylogenetic relationship between the two species.

2.6 Physicochemical Properties and Subcellular Localization Analysis of BGLU Family Genes in A. dahurica var. formosana

A total of 45 BGLU family genes were identified in the whole genome of A. dahurica var. formosana and named AdBGLU01~AdBGLU45. Physicochemical property analysis was performed using ProtParam Tool, and subcellular localization was predicted using Plant-mPLoc and WoLF PSORT (Table 1). The results showed that the number of amino acids encoded by the 45 BGLU genes ranged from 51 to 930, with the longest containing 930 amino acid residues (AdBGLU32) and the shortest containing 51 amino acid residues (AdBGLU30). The instability index ranged from 11.18 to 61.86, with 38 genes having instability coefficients less than 40, suggesting they encode stable proteins, while the remaining 7 were unstable proteins. The aliphatic index ranged from 56.76 to 113.25, indicating good thermal stability of the proteins. The grand average of hydropathicity ranged from -0.643 to 0.35, with 7 being positive and 38 negative, suggesting the proteins are mainly hydrophilic. The isoelectric point ranged from 4.24 to 10.35, indicating the amino acids are mostly weakly acidic or weakly basic. Subcellular localization predictions placed AdBGLU family members in the nucleus, cytoplasm, chloroplast, and vacuole. The significant differences in physicochemical properties and diverse subcellular localizations among AdBGLU gene family members suggest that this gene family has diverse functions and participates in different physiological processes in vivo.

2.7 Secondary Structure and Conserved Domain Analysis of BGLU Family Proteins in A. dahurica var. formosana

Online analysis of the secondary structure of A. dahurica var. formosana BGLU family proteins (Table 2) showed that α-helices and random coils accounted for the largest proportions, with α-helices being most abundant in 27 members and random coils in 18 members. Random coils are unstable coding regions in proteins, suggesting that more random coils may indicate more diverse functions of family members [Yao et al., 2022].

Conserved domain analysis (Figure 6 [Figure 6: see original paper]) showed that Motif 8 was the shortest, containing 29 amino acid residues; Motif 6 was slightly

longer with 35 residues; Motifs 2, 3, and 7 were longer with 41 residues; and Motifs 1, 4, and 5 were the longest, each containing 50 residues. The conserved motif structure revealed that Motif 5 had relatively high conservation. Different genes contained different numbers of conserved domains, with Motif 1 showing the highest frequency among all motifs, suggesting it may be a characteristic motif.

A phylogenetic tree constructed based on protein sequences of A. dahurica var. formosana and Arabidopsis thaliana (Figure 7 [Figure 7: see original paper]) divided AdBGLU genes into six subfamilies (A-F). Both AdBGLU and AtBGLU genes were present in subfamilies B-F, indicating conserved gene functions in these subfamilies [Zhang et al., 2022]. Subfamily A contained 3 AtBGLU genes but no AdBGLU genes; subfamily B contained 1 AdBGLU and 4 AtBGLU genes; subfamily C contained 13 AdBGLU and 14 AtBGLU genes; subfamily D contained 5 AdBGLU and 8 AtBGLU genes; subfamily E contained 14 AdBGLU and 17 AtBGLU genes; and subfamily F contained 12 AdBGLU and 2 AtBGLU genes. The similar gene numbers of A. dahurica var. formosana and Arabidopsis thaliana in subfamily C suggest that homologous genes in this subfamily may play similar roles in both species [Liu, 2020]. In contrast, significant differences in gene numbers in other subfamilies may indicate the presence of key genes regulating coumarin synthesis in A. dahurica var. formosana, which requires further verification.

Discussion and Conclusion

Studies have shown that genome size is positively correlated with ploidy level and corresponding chromosome number in species [Mank & Avise, 2006]. Research on 282 Poaceae species revealed that as chromosome ploidy increased from diploid to octoploid, genome size increased significantly, showing a highly significant positive correlation with ploidy and chromosome number [Li et al., 2012]. This study obtained a genome of approximately 5.6 Gb for A. dahurica var. formosana. Other Apiaceae species with completed genome sequencing include Centella asiatica (~430 Mb), celery (~3.33 Gb), Angelica sinensis (~2.37 Gb) [Han et al., 2022], Oenanthe javanica (~1.28 Gb), Bupleurum chinense (~621.42 Mb), carrot (~421.5 Mb), wild carrot (~371.6 Mb), and coriander (~2,130.29 Mb). Among these, A. dahurica var. formosana, celery, A. sinensis, and coriander have chromosome numbers of 2n=22, while C. asiatica, carrot, and wild carrot have 2n=18, and B. chinense has 2n=12. Except for B. chinense, the results support a positive correlation between chromosome number and genome size. The plant height of A. dahurica var. formosana and celery can reach 1.5 m, while other plants do not exceed 1 m, suggesting a preliminary positive correlation between genome size and plant height in Apiaceae species [Shao et al., 2021], providing a reference for future genome studies of related species.

Coumarins are natural compounds with important medicinal value, classified into four categories: simple coumarins, furanocoumarins, pyranocoumarins, and other coumarins [Wang et al., 2022]. In plants, coumarins are synthesized through the phenylpropanoid metabolic pathway, with many studies revealing key genes involved in this biosynthetic pathway. For example, the PAL gene extracted from Photorhabdus luminescens can convert L-phenylalanine to cinnamic acid and L-tyrosine to p-coumaric acid [Zhang et al., 2021]. Studies on sunflower identified three C4H genes that catalyze the conversion of cinnamic acid to p-coumaric acid, and functional exploration of C4H genes from Peucedanum praeruptorum and P. decursivum showed the same catalytic function [Wang et al., 2020]. Research on Melilotus albus also found that the MaBGLU1 gene plays a key role in converting scopolin to scopoletin [Wu et al., 2022]. In the related species Angelica sinensis, PT genes were found to potentially play key roles in furanocoumarin formation. PAL and C4H are relatively upstream genes in the coumarin biosynthesis pathway and have been extensively studied, while downstream BGLU genes have been less investigated, particularly in A. dahurica. Studies have shown that BGLU is involved in various aspects of plant physiology, especially responses to biotic and abiotic stresses, by activating plant hormones and defense compounds. For example, five GhBGLU genes in Gossypium hirsutum may positively regulate resistance to Verticillium wilt, AtBGLU10 in Arabidopsis thaliana can catalyze the production of free ABA, AtBGLU21-23 regulate scopolin hydrolysis in roots, and AtBGLU42 participates in inducing resistance to cellular diseases. The genome of A. dahurica var. formosana obtained in this study provides a valuable foundation for mining genes related to coumarin component synthesis.

Currently, 48 BGLU family genes have been identified in Arabidopsis thaliana, 26 in maize [Gómez-Anduro et al., 2011], 40 in rice [Opassiri et al., 2006], 42 in soybean [Ke et al., 2019], 53 in upland cotton [Zhang et al., 2022], and 51 in Medicago [Yang et al., 2021]. This study identified 45 BGLU family genes in A. dahurica var. formosana and analyzed their physicochemical properties and secondary structures. The subcellular localization was mainly in the cytoplasm, chloroplast, and vacuole, consistent with the localization of β-glucosidase in maize [Kristoffersen et al., 2000]. The significant differences in physicochemical properties, secondary structures, and subcellular localizations among AdBGLU gene family members indicate a complex structure and diverse functions, with different genes having distinct functional roles and participating in various metabolic processes. A. dahurica var. formosana contains various coumarin compounds such as imperatorin, isoimperatorin, byakangelicin, and bergapten, with complex biosynthetic pathways that may be related to the diverse functions of AdBGLU genes. The preliminary analysis of AdBGLU provides an important foundation for further revealing and utilizing key genes in coumarin biosynthesis pathways in A. dahurica var. formosana.

Data Availability

Raw sequencing data have been uploaded to the China National GeneBank DataBase (CNGBdb, https://db.cngb.org/) under project number CNP0003549.

References

DUAN Z, WU F, YAN Q, et al., 2022. Research progress on plant coumarin biosynthesis pathway and the genes encoding the key enzymes[J]. Acta Pratacult Sin, 31(1): 217-228. [Duan Z, Wu F, Yan Q, et al., 2022. Research progress on plant coumarin biosynthesis pathway and the genes encoding the key enzymes[J]. Acta Prataculturae Sinica, 31(1): 217-228.]

EDGAR RC, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput[J]. Nucl Acids Res, 32(5): 1792-1797.

FINN RD, MISTRY J, SCHUSTER-BÖCKLER B, et al., 2006. Pfam: clans, web tools and services[J]. Nucl Acid Res, 34(Database issue): D247-D251.

GÓMEZ-ANDURO G, CENICEROS-OJEDA EA, CASADOS-VÁZQUEZ LE, et al., 2011. Genome-wide analysis of the beta-glucosidase gene family in maize (Zea mays L. var B73)[J]. Plant Mol Biol, 77(1-2): 159-183.

GUINDON S, GASCUEL O, 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood[J]. Syst Biol, 52(5): 696-704.

HAN X, LI C, SUN S, et al., 2022. The chromosome-level genome of female ginseng (Angelica coumarin into molecular mechanisms and evolution of insights sinensis) provides biosynthesis[J]. Plant J, 112(5): 1224-1237.

HUANG WJ, XU X, CHEN JS, et al., 2021. Bioinformatics analysis and expression pattern of NAC transcription factor family of Angelica dahurica var. formosana from Sichuan province[J]. Chin J Chin Mat Med, 46(7): 1769-1782. [Huang WJ, Xu X, Chen JS, et al., 2021. Bioinformatics analysis and expression pattern of NAC transcription factor family of Angelica dahurica var. formosana from Sichuan province[J]. Chinese Journal of Chinese Materia Medica, 46(7): 1769-1782.]

IORIZZO M, ELLISON S, SENALIK D, et al., 2016. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution[J]. Nat Genet, 48(6): 657-666.

JI Q, MA YH, ZHANG Y, 2020. Research progress on chemical constituents and pharmacological effects of Angelicae dahuricae radix[J]. Food Drug, 22(6): 509-514. [Ji Q, Ma YH, Zhang Y, 2020. Research progress on chemical constituents and pharmacological effects of Angelicae dahuricae radix[J]. Food and Drug, 22(6): 509-514.]

JIANG YJ, JIANG YM, YAO F, et al., 2021. Bioinformatics analysis on the CONSTANS-like protein family in Angelica dahurica var. formosana[J]. Mol Plant Breed, 19(12): 3923-3931. [Jiang YJ, Jiang YM, Yao F, et al., 2021. Bioinformatics analysis on the CONSTANS-like protein family in Angelica dahurica var. formosana[J]. Molecular Plant Breeding, 19(12): 3923-3931.]

KE DX, LIU YH, ZHANG JJ, et al., 2019. Genome-wide identification and expression analysis of BGLU family genes in Soybean[J]. J Xinyang Norm Univ(Nat Sci Ed), 32(3): 372-378. [Ke DX, Liu YH, Zhang JJ, et al., 2019. Genome-wide identification and expression analysis of BGLU family genes in Soybean[J]. Journal of Xinyang Normal University (Natural Science Edition), 32(3): 372-378.]

KOREN S, WALENZ BP, BERLIN K, et al., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation[J]. Genome Res, 27(5): 722-736.

KRISTOFFERSEN P, BRZOBOHATY B, HÖHFELD I, et al., 2000. Developmental regulation of the maize Zm-p60.1 gene encoding a beta-glucosidase located to plastids[J]. Planta, 210(3): 407-415.

LARKIN MA, BLACKSHIELDS G, BROWN NP, et al., 2007. Clustal W and Clustal X version 2.0[J]. Bioinformatics, 23(21): 2947-2948.

LI B, ZHANG X, WANG J, et al., 2014. Simultaneous characterisation of fifty coumarins from the roots of Angelica dahurica by off-line two-dimensional high-performance liquid chromatography coupled with electrospray ionisation tandem mass spectrometry[J]. Phytochem Analysis, 25(3): 229-240.

LI GS, CAO B, BAI CK, 2012. Correlation analysis between genome size and seed characteristics in poaceae plants[J]. Bull Bot Res, 32(6): 701-706. [Li GS, Cao B, Bai CK, 2012. Correlation analysis between genome size and seed characteristics in poaceae plants[J]. Bulletin of Botanical Research, 32(6): 701-706.]

LI L, STOECKERT CJ Jr, ROOS DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes[J]. Genome Res, 13(9): 2178-2189.

LIU YX, 2020. Identification and expression analysis of WRKY gene family in Solanum lycopersicum[D]. Shenyang: Shenyang Agricultural University: 1-79. [Liu YX, 2020. Identification and expression analysis of WRKY gene family in Solanum lycopersicum[D]. Shenyang: Shenyang Agricultural University: 1-79.]

LIU Y, 2019. Studies on bacteriostatic mechanism of Angelica dahurica and excavation of key genes of coumarin biosynthesis[D]. Chengdu: Sichuan Agricultural University: 1-69. [Liu Y, 2019. Studies on bacteriostatic mechanism of Angelica dahurica and excavation of key genes of coumarin biosynthesis[D]. Chengdu: Sichuan Agricultural University: 1-69.]

MANK JE, AVISE JC, 2006. Cladogenetic correlates of genomic expansions in the recent evolution of actinopterygiian fishes[J]. Proceed Royal Soc B Biol Sci,

273(1582):33-38.

NATIONAL PHARMACOPOEIA COMMISSION, 2020. Pharmacopoeia of People's Republic of China: 1[M]. Beijing: China Medical Science Press: 109-110. [National Pharmacopoeia Commission, 2020. Pharmacopoeia of People's Republic of China: Volume 1[M]. Beijing: China Medical Science Press: 109-110.]

OPASSIRI R, POMTHONG B, ONKOKSOONG T, et al., 2006. Analysis of rice glycosyl hydrolase family 1 and expression of Os4bglu12 beta-glucosidase[J]. BMC Plant Biol, 6: 33.

SAMPEDRO J, VALDIVIA ER, FRAGA P, et al., 2017. Soluble and membrane-bound β- glucosidases are involved in trimming the xyloglucan backbone[J]. Plant Physiol, 173(2): 1017-1029.

SENOL CALI D, KIM JS, GHOSE S, et al., 2019. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions[J]. Brief Bioinform, 20(4): 1542-1559.

SHA LP, 2018. Examples of CTAB method, SDS method and salting-out method for crude extraction of plant DNA[J]. Teach Middle Sch Biol, 21: 65-67. [Sha LP, 2018. Examples of CTAB method, SDS method and salting-out method for crude extraction of plant DNA[J]. Teaching of Middle School Biology, 21: 65-67.]

SHAO C, LI YQ, LUO A, et al., 2021. Relationship between functional traits and genome size variation of angiosperms with different life forms[J]. Biodivers Sci, 29(5): 575-585. [Shao C, Li YQ, Luo A, et al., 2021. Relationship between functional traits and genome size variation of angiosperms with different life forms[J]. Biodiversity Science, 29(5): 575-585.]

SIMÃO FA, WATERHOUSE RM, IOANNIDIS P, et al., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs[J]. Bioinformatics, 31(19): 3210-3212.

SONG X, WANG J, LI N, et al., 2020. Deciphering the high-quality genome sequence of coriander that causes controversial feelings[J]. Plant Biotechnol J, 18(6): 1444-1456.

SONG X, SUN P, YUAN J, et al., 2021. The celery genome sequence reveals sequential paleo- polyploidizations, karyotype evolution and resistance gene reduction in apiales[J]. Plant Biotechnol J, 19(4): 731-744.

SUN HH, XUE YM, LIN YF, 2014. Enhanced catalytic efficiency in quercetin-4'-glucoside hydrolysis of Thermotoga maritima β-glucosidase A by site-directed mutagenesis[J]. J Agric Food Chem, 62(28): 6763-6770.

VENUGOPALA KN, RASHMI V, ODHAV B, 2013. Review on natural coumarin lead compounds for their pharmacological activity[J]. Biomed Res Int, 2013: 963248.

WANG R, LIU J, YANG DY, et al., 2020. Research progress in chemical constituents and pharmacological action of Angelica dahurica[J]. Inf Trad Chin Med, 37(2): 123-128. [Wang R, Liu J, Yang DY, et al., 2020. Research progress in chemical constituents and pharmacological action of Angelica dahurica[J]. Information on Traditional Chinese Medicine, 37(2): 123-128.]

WANG RX, SONG J, SUN B, et al,. 2022. Research progress of function and biosynthesis of coumarins[J]. Chin Biotechnol, 42(12): 79-90. [Wang RX, Song J, Sun B, et al., 2022. Research progress of function and biosynthesis of coumarins[J]. China Biotechnology, 42(12): 79-90.]

WANG Z, JIAN X, ZHAO Y, et al.,2020. Functional characterization of cinnamate 4-hydroxylase from Helianthus annuus Linn using a fusion protein method[J]. Gene, 758: 144950.

WILKINS MR, GASTEIGER E, BAIROCH A, et al., 1999. Protein identification and analysis tools in the ExPASy server[J]. Meth Mol B, 112: 531-552.

WU F, DUAN Z, XU P, et al., 2022. Genome and systems biology of Melilotus albus provides insights into coumarins biosynthesis[J]. Plant Biotechnol J, 20(3): 592-609.

WU F, 2021. Study on whole genome sequencing and functional genes of key traits in Cleistogenes songorica and Melilotus albus[D]. Lanzhou: Lanzhou University: 1-185. [Wu F, 2021. Study on whole genome sequencing and functional genes of key traits in Cleistogenes songorica and Melilotus albus[D]. Lanzhou: Lanzhou University: 1-185.]

WU P, GUO JX, WANG XY, et al., 2020. High-throughput transcriptome sequencing of roots of Angelica dahurica and data analyses[J]. Mol Plant Breed, 2020, 18(10): 3207-3216. [Wu P, Guo JX, Wang XY, et al., 2020. High-throughput transcriptome sequencing of roots of Angelica dahurica and data analyses[J]. Molecular Plant Breeding, 18(10): 3207-3216.]

XU Z, WANG H, 2007. LTR_{FINDER}: an efficient tool for the prediction of full-length LTR retrotransposons[J]. Nucl Acid Res, 35(Web Server issue): W265-W268.

YANG J, MA L, JIANG W, et al., 2021. Comprehensive identification and characterization of abiotic stress and hormone responsive glycosyl hydrolase family 1 genes in Medicago truncatula[J]. Plant Physiol Biochem, 158: 21-33.

YAO F, JIANG MY, YANG YS, et al., 2022. Bioinformatics and expression analysis on MYB- related family in Angelicae dahuricae var. formosana[J]. Chin J Chin Mat Med, 47(7): 1831- 1846. [Yao F, Jiang MY, Yang YS, et al., 2022. Bioinformatics and expression analysis on MYB-related family in Angelicae dahuricae var. formosana[J]. Chinese Journal of Chinese Materia Medica, 47(7): 1831-1846.]

YU KP, PENG C, LIN YL, et al., 2023. Expression of β-glucosidase An-bgl3 from Aspergillus niger for conversion of scopoline[J]. Chin J Biotechnol, 39(3):

1232-1246. [Yu KP, Peng C, Lin YL, et al., 2023. Expression of β-glucosidase An-bgl3 from Aspergillus niger for conversion of scopoline[J]. Chinese Journal of Biotechnology, 39(3): 1232-1246.]

YU J, ZHU YH, 2014. Summary of the application of Angelica dahurica in ancient prescription[J]. Heilongjiang Med J, 27(1): 156-158. [Yu J, Zhu YH, 2014. Summary of the application of Angelica dahurica in ancient prescription[J]. Heilongjiang Medical Journal, 27(1): 156-158.]

ZHANG F, REN J, ZHAN J, 2021. Identification and characterization of an efficient phenylalanine ammonia-lyase from Photorhabdus luminescens[J]. Appl Biochem Biotechnol, 193(4): 1099-1115.

ZHANG M, WANG ZC, LIU ZW, et al., 2022-09-17. Genome-wide identification and analysis of BGLU genes family in Gossypium hirsutum[J/OL]. J Agric Sci Technol: 1-12. [Zhang M, Wang ZC, Liu ZW, et al., 2022-09-17. Genome-wide identification and analysis of BGLU genes family in Gossypium hirsutum[J/OL]. Journal of Agricultural Science and Technology: 1-12.]

ZHAO H, FENG YL, WANG M, et al., 2022. The Angelica dahurica: a review of traditional uses, phytochemistry and pharmacology[J]. Front Pharmacol, 13: 896637.

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