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

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

Date: 2023-08-24T00:00:00+00:00

Abstract

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

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 medicinal and edible homologous species widely used both as a clinical traditional Chinese medicine and as a spice. To obtain 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 revealed that protein functions in *A. dahurica* var. *formosana* were mainly concentrated in general functional prediction, posttranslational modification, protein turnover, chaperones, and signal transduction mechanisms. GO functional classification indicated that genes were concentrated in biological processes and cellular components. KEGG pathway annotation showed that genes involved in metabolic pathways were predominant. (3) The study identified 45 BGLU family genes in *A. dahurica* var. *formosana*. This research represents the first analysis of the whole genome of *A. dahurica* var. *formosana* using third-generation sequencing technology, laying a foundation for systematic biological studies of this species and facilitating its further development and utilization. The preliminary analysis of the BGLU gene family also provides an important theoretical basis for future studies on the function of BGLU genes 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

---

Angelica dahurica (Apiaceae) is the dried root of *Angelica dahurica* or *A. dahurica* var. *formosana*, mainly cultivated in Sichuan, Hangzhou, and other regions of China. As a common medicinal and edible herb, it is clinically used for various pain symptoms including headache from colds, supraorbital bone pain, toothache, and sore swelling from ulcers (National Pharmacopoeia Commission, 2020), and also serves 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 multiple active components such as 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 various pharmacological effects including antipyretic and analgesic, anti-inflammatory, antimicrobial, antitumor, blood pressure-lowering, and hepatoprotective activities (Ji et al., 2020; Wang et al., 2020).

Despite its broad application prospects, recent research on *A. dahurica* has primarily focused on chemical composition, cultivation techniques, and pharmacological effects, with limited investigation into its genetic information. Currently, only transcriptome sequencing studies have been reported (Wu et al., 2020), and research on gene families such as CONSTANS-like (Jiang et al., 2021), NAC (Huang et al., 2021), and MYB-related (Yao et al., 2022), as well as mining of key genes for coumarin synthesis (Liu, 2019), have relied on transcriptome data. The lack of genomic data for *A. dahurica* prevents acquisition of complete genetic information and hinders further in-depth research, making whole-genome sequencing essential.

Coumarin compounds are both medicinal and aromatic components in *A. dahurica* and are widely distributed in various plants including 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 an important regulatory role 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 BGLU family plays a crucial regulatory role in coumarin synthesis in sweet clover (Wu, 2021). In maize, BGLU can catalyze the hydrolysis of 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 obtusifolia*, increasing its content by 47% (Yu et al., 2023). Three β-glucosidases isolated from *Arabidopsis thaliana* can specifically hydrolyze scopolin into scopoletin, a coumarin component also present in *A. dahurica*. Our research group hypothesizes that BGLU genes also play a key role in coumarin component synthesis in *A. dahurica*.

As no high-quality genome studies of *A. dahurica* have been reported and analysis of its coumarin synthesis pathway remains limited, this study performed second- and third-generation genome sequencing of *A. dahurica* var. *formosana* to obtain a high-quality genome through assembly and annotation. Functional annotation and gene family clustering analyses were conducted, followed by mining of key BGLU genes in the coumarin synthesis pathway. Basic characteristic analysis of BGLU sequences extracted from the genome was performed using online software to address: (1) the general profile of the *A. dahurica* var.

formosana genome; (2) the biological processes and metabolic pathways where gene functions are concentrated; and (3) the basic characteristics of the BGLU gene family. This study provides both data and molecular foundations for subsequent research on *A. dahurica* and establishes a basis for further investigation of BGLU gene family functions in coumarin synthesis pathways.

1.1 Materials and DNA Extraction

Angelica 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 Medicine Germplasm Resources Bank. 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 quality verification, the extracted genomic DNA was randomly fragmented by enzymatic digestion. Libraries with 150 bp insert fragments were constructed through end repair, A-tailing, adapter ligation, purification, and PCR amplification. The constructed libraries were 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 products were purified again and ligated with sequencing adapters. The final library was precisely quantified using Qubit. A measured amount of DNA library was mixed with sequencing reagents, loaded into a flow cell, and subjected to 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 as N) that can significantly interfere with subsequent bioinformatic analysis. These contaminants 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

Jellyfish v1.1.10 was employed for survey analysis using reads obtained from MGISEQ-200 sequencing to estimate genome size, heterozygosity rate, and

repetitive sequence proportion, thereby evaluating genome complexity. K-mer analysis was performed for gene 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 was polished using Racon v1.0.0 (Senol et al., 2019). Pilon v1.22 was subsequently applied for correction 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

Repetitive sequences were predicted using a combination of structure-based and ab initio approaches. 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, which was classified using PASTEClassifier v2.0. This database was then merged with Repbase (https://www.girinst.org/repbase/) to create the final repetitive sequence database for *A. dahurica* var. *formosana*. RepeatMasker v4.1.2 was employed for repetitive sequence prediction based on this database.

Gene prediction was performed using both ab initio and homology-based approaches. Genscan v1.0, Augustus v3.3.1, GlimmerHMM v3.0.4, GeneID v1.4, and SNAP v8.0.0 were used for ab initio prediction, while GeMoMa v1.3.1 performed homology-based prediction. EvidenceModeler v1.1.0 integrated and corrected predictions from all methods. Non-coding RNAs, including microRNA, rRNA, and tRNA, were predicted using Infernal v1.1.3 based on Rfam (Finn et al., 2006) and miRBase databases for rRNA and microRNA, and tRNAscan-SE v2.0.7 for tRNA identification.

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 an e-value threshold of <1e-5. GO functional annotation was performed using Blast2GO v5.2.5 based on NR database comparison results.

1.8 Gene Family Clustering and Phylogenetic Analysis

To identify gene families, protein sequences of *A. dahurica* var. *formosana* were compared with those of related Apiaceae species. Protein sequences of celery (*Apium graveolens*) (Song et al., 2021) and carrot (*Daucus carota* subsp. *sativus*) (Iorizzo et al., 2016) were downloaded from NCBI, and coriander (*Coriandrum sativum*) (Song et al., 2020) from CGDB (http://cgdb.bio2db.com). OrthoMCL

v2.0 (Li et al., 2003) performed clustering analysis based on all-vs-all blastp similarity relationships. Single-copy protein sequences extracted from OrthoMCL results were aligned using Muscle v3.8.31 (Edgar, 2004), and a maximum likelihood (ML) phylogenetic tree was constructed using RAxML v8.2.12 (Guindon & Gascuel, 2003).

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

The SMART database was used to obtain typical domain sequences of the *Arabidopsis thaliana* BGLU gene family. tBLASTN (P=0.001) searches were performed against the *A. dahurica* var. *formosana* genome database, and all BGLU gene family members 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 predicted using PlantmPLoc (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), and 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 from *A. dahurica* var. *formosana* and *Arabidopsis thaliana*. A phylogenetic tree was constructed from the alignment using the neighbor-joining method.

2.1 Genome Sequencing

Whole-genome sequencing of *A. dahurica* var. *formosana* leaves was performed using sequencing platforms. After initial filtering of raw reads to remove low-quality and short fragments, 150 Gb of second-generation raw data and 662 Gb of third-generation raw data were obtained. For the third-generation data, the Read N50 was 32,932 bp, the longest read was 422,833 bp, and the average length was 27,750 bp, meeting the quality 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, 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%. BUSCO v5.1.2 assessment identified 1,580 complete BUSCO genes 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 completeness score of 97.9% indicated a relatively complete assembly.

2.3 Gene Prediction Results

RepeatMasker v4.1.2 prediction identified 5.4 Gb of repetitive sequences, accounting for 91.36% of the genome. These included 21,726 long interspersed nuclear elements (LINEs, 0.41%), 0 short interspersed nuclear elements (SINEs), 3,550,524 long terminal repeats (LTRs, 69.07%) comprising 1,083,004 copia (30.01%), 989,985 gypsy (24.56%), and 2,893 rolling-circles (0.03%) elements, plus 7,710 simple sequence repeats (SSRs, 0.03%).

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

2.4 Gene Function Annotation and Analysis

KOG functional annotation [Figure 1: see original paper] revealed that 29,788 genes (44.46% of total predicted genes) received annotation. 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%). These differentially expressed genes provide data support for future in-depth studies of *A. dahurica* var. *formosana*.

GO annotation [Figure 2: see original paper] showed that 44,540 genes (66.47% of total predicted genes) had GO functional annotations. Genes involved in reproduction, cellular processes, stress response, cell, and cell part categories were predominant, with reproduction-related genes being the most abundant.

KEGG pathway annotation [Figure 3: see original paper] assigned pathway annotations to 15,263 genes (22.78% of total predicted genes). The 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 from *A. dahurica* var. *formosana* with those of related Apiaceae species identified 23,151 gene families among 67,004 protein

sequences. Of these, 4,004 gene families containing 18,151 genes were specific to *A. dahurica* var. *formosana*, while 1,030 gene families were shared among all four plant species [Figure 4: see original paper].

To further investigate the phylogenetic relationships, 96 single-copy protein sequences were compared among seven species with known genome information: *Arabidopsis thaliana*, *Zea mays*, *Amborella trichopoda*, and four Apiaceae species including coriander, celery, carrot, and *Angelica sinensis*. The phylogenetic tree [Figure 5: see original paper] showed that *A. dahurica* var. *formosana* clustered with coriander, indicating a close phylogenetic relationship.

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 *A. dahurica* var. *formosana* genome and designated AdBGLU01–AdBGLU45. Physicochemical property analysis using ProtParam Tool and subcellular localization prediction using Plant-mPLoc and WoLF PSORT revealed that the encoded amino acid numbers ranged from 51 (AdBGLU30) to 930 (AdBGLU32). Instability indices ranged from 11.18 to 61.86, with 38 proteins predicted to be stable (instability coefficient <40) and 7 unstable. Aliphatic indices of 56.76–113.25 indicated good thermal stability. Grand average hydropathicity values of -0.643 to 0.35 showed that 38 proteins were hydrophilic (negative values) and 7 were hydrophobic (positive values). Isoelectric points ranged from 4.24 to 10.35, indicating mostly weakly acidic or basic amino acids. Subcellular localization predicted AdBGLU family members in the nucleus, cytoplasm, chloroplast, and vacuole. The significant variation in physicochemical properties and diverse subcellular localizations among AdBGLU gene family members suggest functional diversity and involvement in different physiological processes.

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

Secondary structure analysis showed that α-helices and random coils were the most abundant structures in the BGLU family, with α-helices predominating in 27 members and random coils in 18 members. As random coils represent unstable coding regions in proteins, more random coils may indicate greater functional diversity of family members (Yao et al., 2022).

Conserved domain analysis [Figure 6: see original paper] revealed that Motif 8 was the shortest with 29 amino acid residues, Motif 6 contained 35 residues, Motifs 2, 3, and 7 contained 41 residues, and Motifs 1, 4, and 5 were the longest with 50 residues each. Motif 5 showed high conservation. Different genes contained varying numbers of conserved domains, with Motif 1 appearing most frequently, suggesting it may be a characteristic motif.

Phylogenetic tree construction based on protein sequences from *A. dahurica* var. *formosana* and *Arabidopsis thaliana* [Figure 7: see original paper] divided

AdBGLU genes into six subfamilies (A–F). AdBGLU and AtBGLU genes co-existed in subfamilies B–F, indicating conserved 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. Subfamily F contained 12 AdBGLU and 2 AtBGLU genes. The similar gene numbers in subfamily C between *A. dahurica* var. *formosana* and *Arabidopsis* suggest that homologous genes in this subfamily may perform similar functions (Liu, 2020). The large differences in other subfamilies may indicate the presence of key genes regulating coumarin synthesis in *A. dahurica* var. *formosana*, though this requires further verification.

3 Discussion and Conclusion

Studies have shown that genome size correlates positively with ploidy level and chromosome number (Mank & Avise, 2006). Research on 282 Poaceae species demonstrated that genome size increased significantly with ploidy from diploid to octoploid, showing a highly significant positive correlation with both ploidy and chromosome number (Li et al., 2012). The *A. dahurica* var. *formosana* genome obtained in this study is approximately 5.6 Gb. Other sequenced Apiaceae species 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). *A. dahurica*, celery, *A. sinensis*, and coriander have 2n=22 chromosomes, while *C. asiatica* and carrot species have 2n=18, and *B. chinense* has 2n=12. Except for *B. chinense*, the data support a positive correlation between chromosome number and genome size. Additionally, *A. dahurica* and celery can grow up to 1.5 m, while other species do not exceed 1 m, suggesting a potential positive correlation between genome size and plant height in Apiaceae (Shao et al., 2021), providing a reference for future genomic studies of related species.

Coumarins are natural compounds with important medicinal value, classified into simple coumarins, furanocoumarins, pyranocoumarins, and other coumarins (Wang et al., 2022). In plants, coumarins are synthesized through the phenylpropanoid metabolic pathway, with many key genes already identified. For example, PAL genes from *Photorhabdus luminescens* can convert L-phenylalanine to cinnamic acid and L-tyrosine to p-coumaric acid (Zhang et al., 2021). Studies in sunflower identified three C4H genes that catalyze cinnamic acid to p-coumaric acid, with similar functions confirmed in *Peucedanum praeruptorum* and *P. decursivum* (Wang et al., 2020). Research on white sweet clover revealed that MaBGLU1 is crucial for converting scopolin to scopoletin (Wu et al., 2022), and studies on *A. sinensis* suggested that PT genes may play a key role in furanocoumarin formation. While upstream genes like PAL and C4H have been extensively studied, downstream BGLU genes have received less

attention, particularly in *A. dahurica*. BGLU genes are involved in multiple aspects of plant physiology, especially responses to biotic and abiotic stresses, through activation of plant hormones and defense compounds. For instance, five GhBGLU genes in cotton may positively regulate resistance to verticillium wilt, AtBGLU10 in *Arabidopsis* can catalyze free ABA production, AtBGLU21-23 regulate scopolin hydrolysis in roots, and AtBGLU42 participates in disease resistance induction. The *A. dahurica* var. *formosana* genome obtained in this study provides a valuable foundation for mining genes related to coumarin biosynthesis.

Previous studies have identified 48 BGLU family genes in *Arabidopsis*, 26 in maize (Gómez-Anduro et al., 2011), 40 in rice (Opassiri et al., 2006), 42 in soybean (Ke et al., 2019), 53 in cotton (Zhang et al., 2022), and 51 in alfalfa (Yang et al., 2021). This study identified 45 BGLU family genes in *A. dahurica* var. *formosana* and analyzed their physicochemical properties and secondary structures. Subcellular localization was primarily predicted in the cytoplasm, chloroplast, and vacuole, consistent with β-glucosidase localization in maize (Kristoffersen et al., 2000). The substantial variation in physicochemical properties, secondary structures, and subcellular localizations among AdBGLU gene family members indicates a complex structure and diverse functions, with different genes participating in various metabolic processes. *A. dahurica* contains multiple coumarin compounds including imperatorin, isoimperatorin, byakangelicin, and bergapten, with complex biosynthetic pathways likely related to the functional diversity of AdBGLU genes. This preliminary analysis of AdBGLU genes is important for understanding coumarin biosynthesis in *A. dahurica* and provides a foundation for further functional characterization of key genes in coumarin synthesis pathways.

4 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.

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 sinensis*) provides insights into coumarin biosynthesis and evolution[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.

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.

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.

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.

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 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 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.

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.

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.

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 RX, SONG J, SUN B, et al., 2022. Research progress of function and biosynthesis of coumarins[J]. Chin Biotechnol, 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 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.

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.

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 J, ZHU YH, 2014. Summary of the application of *Angelica dahurica* in ancient prescription[J]. Heilongjiang Med J, 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. Genome-wide identification and analysis of BGLU genes family in *Gossypium hirsutum*[J]. J Agric Sci Technol: 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.