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

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

Date: 2023-08-04T00:00:00+00:00

Abstract

Angelica dahurica is a commonly used species exhibiting medicine-food homology, serving both as a clinically used 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 Hangbai Zhi (Angelica dahurica var. formosana) as material, employed Nanopore sequencing technology to construct a whole-genome database for Hangbai Zhi, and applied bioinformatic methods to assemble, functionally annotate, and perform evolutionary analysis on the obtained nucleotide sequences. The results demonstrated: (1) After filtering the raw sequencing data, 662 Gb of third-generation sequencing data was obtained, with a Read N50 of approximately 32,932 bp; assembly yielded a Hangbai Zhi genome size of 5.6 Gb, with a Contig N50 of approximately 806,638 bp. (2) The assembled sequences were compared against functional databases including KOG, GO, and KEGG, resulting in functional annotation of 66.47% of genes. KOG functional annotation results indicated that Hangbai Zhi protein functions were primarily concentrated in general function prediction, post-translational modification, protein turnover, chaperones, and signal transduction mechanisms; GO functional classification revealed that Hangbai Zhi genes were concentrated in biological processes and cellular components; KEGG pathway annotation demonstrated that genes involved in metabolic pathways constituted the majority. (3) Hangbai Zhi genes were concentrated in 45 BGLU family genes. This study represents the first comprehensive analysis of the Hangbai Zhi whole genome using 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. Additionally, 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 Hangbai Zhi growth and development.

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 widely used medicinal and edible plant, serving both as a common clinical traditional Chinese medicine and as a spice. To obtain the complete 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 on 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) By comparing the assembled sequences with functional databases including KOG, GO, and KEGG, 66.47% of the genes were functionally annotated. KOG annotation revealed that protein functions were primarily concentrated in general functional prediction, posttranslational modification, protein turnover, chaperones, and signal transduction mechanisms. GO functional classification indicated that genes were mainly involved in biological processes and cellular components. KEGG pathway annotation showed that genes participating in metabolic pathways dominated. (3) The genome contained 45 BGLU family genes. This study represents the first resolution of the *A. dahurica* var. *formosana* genome using third-generation sequencing technology, laying a foundation for systematic biological research on *A. dahurica* and facilitating its further development and utilization. Simultaneously, the preliminary analysis of the BGLU gene family provides an important theoretical basis for future investigations into BGLU functions in the growth, development, and coumarin biosynthesis of *A. dahurica* var. *formosana*.

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

Introduction

Angelica dahurica (Bai Zhi) is the dried root of *Angelica dahurica* or *A. dahurica* var. *formosana* (Apiaceae), primarily cultivated in Sichuan and Hangzhou re-

gions. As a common medicinal and edible herb, it is clinically used to treat various pain symptoms including headache from common cold, supraorbital bone pain, toothache, and sore swellings [?, ?]. In daily life, it serves as a spice and, due to its aromatic properties, is widely applied in cosmetics and personal care products [?, ?, ?]. *A. dahurica* contains multiple active components such as coumarins, volatile oils, polysaccharides, and alkaloids [?, ?, ?, ?, ?, ?, ?, ?, ?], with modern research indicating that coumarins and volatile oils are the main active ingredients responsible for various pharmacological effects including antipyretic and analgesic, anti-inflammatory, antimicrobial, antitumor, blood pressure-lowering, and hepatoprotective activities [?, ?, ?, ?, ?, ?, ?, ?].

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 [?, ?, ?, ?, ?] and analyses of specific gene families including CONSTANS-like [?, ?, ?, ?, ?], NAC [?, ?, ?, ?, ?], and MYB-related [?, ?, ?, ?, ?] have been reported, all based on transcriptome data. The lack of genomic data for *A. dahurica* has hindered acquisition of complete genetic information and prevented further in-depth research, making whole-genome sequencing essential.

Coumarin compounds, which serve as both medicinal and aromatic components in *A. dahurica*, are widely distributed in various plant families including Apiaceae, Rutaceae, and Moraceae [?, ?, ?, ?]. Recent research has elucidated the coumarin biosynthetic pathway and key enzymes involved [?, ?, ?, ?, ?], including β-glucosidase (BGLU). The BGLU gene family plays crucial regulatory roles not only in coumarin biosynthesis but also in diverse physiological processes such as plant hormone signal activation [?, ?, ?, ?] and secondary metabolism [?, ?, ?, ?, ?]. Studies have demonstrated that BGLU genes regulate coumarin synthesis in *Melilotus albus* [?, ?], catalyze hydrolysis of β-glycosidic bonds between carbohydrate moieties and coumarin core structures in maize, and specifically hydrolyze scopolin in *Aspergillus niger* extracts, increasing its content by 47% [?, ?, ?, ?, ?]. Three BGLU enzymes isolated from *Arabidopsis thaliana* can specifically hydrolyze scopolin into scopoletin, a coumarin compound also present in *A. dahurica*. Our research group hypothesizes that BGLU genes play key roles in coumarin biosynthesis in *A. dahurica*.

Given the absence of high-quality genome studies on *A. dahurica* and limited analysis of its coumarin biosynthetic pathway, this study performed second- and third-generation genome sequencing of *A. dahurica* var. *formosana* to obtain a high-quality genome through assembly and annotation, followed by functional annotation and gene family clustering analysis. We specifically mined for key BGLU genes in the coumarin biosynthetic pathway and analyzed their basic characteristics using online tools to address: (1) the general features of the *A. dahurica* var. *formosana* genome; (2) the biological processes and metabolic pathways where gene functions are concentrated; and (3) the fundamental char-

acteristics 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 biosynthesis.

Materials and Methods

1.1 Plant Material 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 Resource 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 leaf tissues using the CTAB method as described by SHA LP (2018). Extracted DNA quality was assessed via agarose gel electrophoresis, concentration was measured using a Qubit Fluorometer, and purity and integrity were evaluated with Nanodrop.

1.2 Library Construction and Sequencing

MGISEQ-200 Sequencing: Qualified genomic DNA was randomly fragmented by enzymatic digestion. Libraries with 150 bp insert sizes were constructed through end repair, A-tailing, adapter ligation, purification, and PCR amplification, followed by paired-end sequencing on the MGISEQ-200 platform.

Nanopore Sequencing: Qualified DNA was enriched and purified using magnetic beads, then subjected to damage repair and end repair with A-tailing. After purification, sequencing adapters were ligated and the final library was quantified precisely using Qubit. A measured amount of DNA library was mixed with sequencing reagents and loaded into a flow cell for single-molecule sequencing on the GridION instrument to obtain raw data.

1.3 Quality Control of Genome Sequencing Data

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

1.4 Genome Size and Heterozygosity Assessment

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

quence proportion, thereby evaluating genome complexity. K-mer analysis was performed to assess these parameters.

1.5 Genome Assembly and Evaluation

To achieve high-accuracy third-generation assembly, Canu v2.1.1 [?, ?, ?, ?, ?] was first used to correct errors in clean data. The corrected data were then assembled, and the assembly was polished using Racon v1.0.0 [?, ?, ?, ?, ?]. Pilon v1.22 was subsequently applied for correction using second-generation data. Finally, BUSCO v5.1.2 [?, ?, ?, ?, ?] was used to assess genome completeness.

1.6 Gene and Repeat Sequence Prediction

Repeat sequences were predicted using a combination of structure-based and ab initio approaches. LTR Finder v1.05 [?, ?, ?], RepeatScout v1.0.6, and PILER-DF v2.4 were used to construct a repeat 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 repeat sequence database for *A. dahurica* var. *formosana*. RepeatMasker v4.1.2 was employed for repeat sequence prediction based on this database.

Gene prediction was performed using both ab initio and homology-based approaches. Ab initio prediction utilized Genscan v1.0, Augustus v3.3.1, GlimmerHMM v3.0.4, GeneID v1.4, and SNAP v8.0.0. Homology-based prediction was conducted using GeMoMa v1.3.1. All predictions were integrated and refined using EvidenceModeler v1.1.0. Non-coding RNAs, including microRNA, rRNA, and tRNA, were predicted using Infernal v1.1.3 based on Rfam [?, ?, ?, ?, ?] and miRBase databases, while tRNAscan-SE v2.0.7 was used for tRNA identification.

1.7 Functional Gene Annotation

Predicted gene sequences were aligned 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 alignment 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*) [?, ?, ?, ?, ?], carrot (*Daucus carota* subsp. *sativus*) [?, ?, ?, ?, ?], and coriander (*Coriandrum sativum*) [?, ?, ?, ?, ?] were downloaded from NCBI and CGDB (http://cgdb.bio2db.com). OrthoMCL v2.0 [?, ?, ?, ?] was used for all-vs-all blastp clustering analysis. Single-copy protein sequences

extracted from OrthoMCL results were aligned using Muscle v3.8.31 [?, ?], and a maximum likelihood (ML) phylogenetic tree was constructed using RAxML v8.2.12 [?, ?, ?].

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

The typical domain sequences of *Arabidopsis thaliana* BGLU genes were obtained from the SMART database and used to search the *A. dahurica* var. *formosana* genome database via tBLASTN (P=0.001). All BGLU gene family members in *A. dahurica* were identified using the Pfam database.

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

Physicochemical properties of BGLU family proteins were analyzed using the ProtParam tool (https://web.expasy.org/protparam/) [?, ?, ?, ?, ?]. Subcellular localization was predicted using Plant-mPLoc (http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/) and WoLF 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 identified using MEME (https://meme-suite.org/meme/tools/meme).

1.11 Phylogenetic Analysis of BGLU Family Genes

BGLU family protein sequences from *A. dahurica* var. *formosana* and *A. thaliana* were aligned using Clustal W v2.0 [?, ?, ?, ?, ?] in MEGA software, and a neighbor-joining phylogenetic tree was constructed.

Results

2.1 Genome Sequencing

Whole-genome sequencing of *A. dahurica* var. *formosana* leaves yielded 150 Gb of raw second-generation data and 662 Gb of raw third-generation data after initial quality filtering to remove low-quality and short reads. For the third-generation data, the Read N50 was 32,932 bp, the longest read measured 422,833 bp, and the average read length was 27,750 bp, meeting the requirements for subsequent assembly. Survey analysis estimated the genome size of *A. dahurica* var. *formosana* at approximately 5.2 Gb.

2.2 Genome Assembly and Evaluation

Assembly using Canu software produced a genome of approximately 5.6 Gb with a Contig N50 of 806,638 bp and a longest contig of 21,677,961 bp. 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 missing genes from the Embryophyta_{odb10} database, yielding a BUSCO completeness score of 97.9% and indicating a relatively complete assembly.

2.3 Gene Prediction Results

Repeat sequence prediction using RepeatMasker v4.1.2 revealed that the *A. dahurica* var. *formosana* genome contained 5.4 Gb of repetitive sequences, accounting for 91.36% of the genome. This 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 elements (30.01%), 989,985 gypsy elements (24.56%), and 2,893 rolling-circles (0.03%), plus 7,710 simple sequence repeats (SSRs, 0.03%).

Among the 67,004 predicted genes, 34,119 (93.1%) received support from homologous identification 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] assigned functions to 29,788 genes (44.46% of total predicted genes). Protein functions were predominantly 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 functional prediction (22.8%). These differentially expressed genes provide valuable data for future research on *A. dahurica* var. *formosana*.

GO annotation [Figure 2: see original paper] assigned functional categories to 44,540 genes (66.47% of total predicted genes), with predominant representation in reproduction, cellular processes, stress response, cell, and cell part categories, with reproduction-related genes showing the highest proportion.

KEGG pathway annotation [Figure 3: see original paper] assigned pathways to 15,263 genes (22.78% of total predicted genes), with metabolic pathways being most represented. Major metabolic pathways included microbial metabolism in diverse environments, carbon metabolism, and amino acid biosynthesis.

2.5 Gene Family Clustering and Phylogenetic Analysis

Comparison of *A. dahurica* var. *formosana* protein sequences with those of coriander, celery, and carrot 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 species [Figure 4: see original paper].

To investigate phylogenetic relationships, 96 single-copy protein sequences were compared across seven species with known genome information: *Arabidopsis*

thaliana, *Zea mays*, *Amborella trichopoda*, and four Apiaceae species (coriander, celery, carrot, and *Angelica sinensis*). The resulting 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 of BGLU Family Genes

A total of 45 BGLU family genes were identified in the *A. dahurica* var. *formosana* genome and designated AdBGLU01 through AdBGLU45. Physicochemical property analysis using ProtParam Tool and subcellular localization prediction using Plant-mPLoc and WoLF PSORT revealed that AdBGLU genes encoded proteins ranging from 51 to 930 amino acids in length (AdBGLU32 being longest at 930 residues, AdBGLU30 shortest at 51 residues). 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 (7 positive, 38 negative) suggested predominantly hydrophilic proteins. Isoelectric points ranged from 4.24 to 10.35, indicating mostly weakly acidic or basic amino acids. Subcellular localization predictions placed AdBGLU family members in the nucleus, cytoplasm, chloroplast, and vacuole. The diverse physicochemical properties and varied subcellular localizations suggest functional diversity among AdBGLU family members, with involvement in different physiological processes.

2.7 Secondary Structure and Conserved Domain Analysis of BGLU Family Proteins

Secondary structure analysis showed that α-helices and random coils constituted the largest proportions in BGLU family proteins, with α-helices being most prevalent in 27 members and random coils in 18 members. Random coils represent unstable coding regions in proteins, suggesting that more random coils may indicate greater functional diversity among family members [**?**, **?**, **?**, **?**, **?**].

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

Phylogenetic analysis based on protein sequences from *A. dahurica* var. *formosana* and *A. thaliana* [Figure 7: see original paper] divided AdBGLU genes into six subfamilies (A–F). AdBGLU and AtBGLU genes coexisted in subfamilies B–F, indicating conserved functions in these subfamilies [**?**, **?**, **?**, **?**, **?**]. Subfamily A contained 3 AtBGLU genes but no AdBGLU genes. Subfamily distributions were: B (1 AdBGLU, 4 AtBGLU), C (13 AdBGLU, 14 AtBGLU), D

(5 AdBGLU, 8 AtBGLU), E (14 AdBGLU, 17 AtBGLU), and F (12 AdBGLU, 2 AtBGLU). The similar gene numbers in subfamily C suggest that homologous genes may perform similar functions in both species [?, ?], while the substantial numerical differences in other subfamilies may indicate key genes regulating coumarin synthesis in *A. dahurica* var. *formosana*, requiring further validation.

Discussion and Conclusion

Previous studies have demonstrated a positive correlation between genome size and ploidy level/chromosome number [?, ?, ?]. Research on 282 Poaceae species revealed that genome size increased significantly with ploidy from diploid to octoploid, showing extremely significant positive correlations with both ploidy and chromosome number [?, ?, ?, ?]. The *A. dahurica* var. *formosana* genome obtained in this study is approximately 5.6 Gb. Among sequenced Apiaceae species, genome sizes vary considerably: *Centella asiatica* (~430 Mb), celery (~3.33 Gb), *Angelica sinensis* (~2.37 Gb) [?, ?, ?, ?, ?], *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). Chromosome numbers are 2n=22 for *A. dahurica*, celery, *A. sinensis*, and coriander; 2n=18 for *C. asiatica*, carrot, and wild carrot; and 2n=12 for *B. chinense*. Except for *B. chinense*, the data support a positive correlation between chromosome number and genome size. Notably, *A. dahurica* and celery (both 2n=22) can reach 1.5 m in height, while other species do not exceed 1 m, suggesting a potential positive correlation between genome size and plant height in Apiaceae [?, ?, ?, ?, ?], providing a reference for future genomic studies of related species.

Coumarins are valuable medicinal compounds classified into simple coumarins, furanocoumarins, pyranocoumarins, and other coumarins [?, ?, ?, ?, ?]. In plants, coumarins are synthesized via the phenylpropanoid pathway, with many key genes already identified. For example, PAL genes from *Photorhabdus luminescens* convert L-phenylalanine to cinnamic acid and L-tyrosine to p-coumaric acid [?, ?, ?, ?]. 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* [?, ?, ?, ?, ?]. Research on *Melilotus albus* demonstrated that MaBGLU1 is crucial for converting scopolin to scopoletin [?, ?, ?, ?, ?], while studies on *Angelica sinensis* suggested PT genes may be key determinants for furanocoumarin formation. Although upstream genes such as PAL and C4H have been well studied, downstream BGLU genes remain poorly characterized, particularly in *A. dahurica*. BGLU genes are associated with 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* catalyzes free ABA production, AtBGLU21-23 regulate scopolin hydrolysis in roots, and AtBGLU42 participates in disease resistance induction. The high-quality genome obtained in this study provides a valuable foundation for mining genes related

to coumarin biosynthesis in *A. dahurica*.

The BGLU gene family has been characterized in several species: 48 members in *Arabidopsis*, 26 in maize [**?**, **?**, **?**, **?**, **?**], 40 in rice [**?**, **?**, **?**, **?**, **?**], 42 in soybean [**?**, **?**, **?**, **?**, **?**], 53 in cotton [**?**, **?**, **?**, **?**, **?**], and 51 in alfalfa [**?**, **?**, **?**, **?**, **?**]. This study identified 45 BGLU family genes in *A. dahurica* var. *formosana* and analyzed their physicochemical properties and secondary structures. Subcellular localization predictions predominantly placed them in the cytoplasm, chloroplast, and vacuole, consistent with β-glucosidase localization in maize [**?**, **?**, **?**, **?**, **?**]. The substantial variation in physicochemical properties, secondary structures, and subcellular localizations among AdBGLU family members suggests a complex family structure with diverse functions and differential functional division of labor among genes, participating in various metabolic processes. *A. dahurica* contains multiple coumarin compounds including imperatorin, isoimperatorin, byakangelicin, and bergapten, with complex biosynthetic pathways that may be related to the functional diversity of AdBGLU genes. This preliminary analysis of AdBGLU genes provides an essential foundation for further elucidating and utilizing key genes in coumarin biosynthesis pathways in *A. dahurica*.

Data Availability

Raw sequencing data have been deposited in 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., 2022-09-17. Genome-wide identification and analysis of BGLU genes family in *Gossypium hirsutum*[J/OL]. 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.