Postprint: Genome Sequencing and Analysis of Magnolia officinalis Using PacBio Third-Generation Sequencing Technology

Authors: Yin Yanpeng, Ding Qiaojiao, Luo Jiawei, Lin Xinna, Zhang Min, Peng Cheng, Gao Jihai

Date: 2020-05-28T00:00:00+00:00

Abstract

Magnolia officinalis is a renowned traditional medicinal plant belonging to the Magnoliaceae family and Magnolia genus, widely cultivated in China. Its bark, root bark, branch bark, leaves, flowers, and fruits can all be used for medicinal purposes or as food. To obtain the whole-genome sequence information of Magnolia officinalis, this study utilized leaf DNA as material and employed PacBio Sequel third-generation sequencing technology to construct a whole-genome database, followed by bioinformatics-based assembly, functional annotation, and evolutionary analysis of the obtained nucleotide sequences. The results demonstrated that after filtering the raw sequencing data, 140.91 Gb of third-generation data was acquired, with a read N50 of approximately 13,784 bp. Assembly yielded a Magnolia officinalis genome size of 1.68 Gb, with a contig N50 of approximately 222,069 bp and single-copy gene completeness of 78.05%. The assembled sequences were compared against functional databases including NR, KOG, and KEGG, resulting in functional annotation for 98.40% of the genes. KOG functional annotation revealed that protein functions in Magnolia officinalis were primarily concentrated in general function prediction, post-translational modification, protein turnover, chaperones, and signal transduction mechanisms. GO functional classification indicated that Magnolia officinalis genes were concentrated in cellular components and biological processes. KEGG analysis revealed that genes involved in metabolic pathways predominated. Comparative analysis with the genomes of grape, Arabidopsis, rice, poplar, ginkgo, Amborella, tea plant, and Cinnamomum kanehirae revealed that among 23,424 genes in Magnolia officinalis, 20,801 genes could be classified into 12,129 families, including 515 gene families specific to Magnolia officinalis. Magnolia officinalis exhibited a close phylogenetic relationship with Cinnamomum kanehirae (Lauraceae), with divergence time estimated at approximately 122.5 million years ago (mya). This study represents the first whole-genome

analysis of Magnolia officinalis using third-generation sequencing technology, which will facilitate its further in-depth development and utilization, and also establish a foundation for whole-genome studies of other medicinal plants.

Full Text

Genomic Sequencing Analysis of *Magnolia officinalis* Based on PacBio Third-Generation Sequencing Technology

YIN Yanpeng, DING Qiaojiao, LUO Jiawei, LIN Xinna, ZHANG Min, PENG Cheng, GAO Jihai*

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

Abstract: *Magnolia officinalis* is a renowned traditional medicinal plant belonging to the family Magnoliaceae and genus *Magnolia*, widely cultivated throughout China. Its bark, root bark, branch bark, leaves, flowers, and fruits all have medicinal or edible applications. To obtain the complete genome sequence information of *M. officinalis*, we constructed a whole-genome database using PacBio Sequel third-generation sequencing technology with leaf DNA as starting material, followed by bioinformatic assembly, functional annotation, and evolutionary analysis. The results showed that after filtering raw sequencing data, we obtained 140.91 Gb of third-generation data with a read N50 of approximately 13,784 bp. The assembled *M. officinalis* genome size was 1.68 Gb with a contig N50 of approximately 222,069 bp, and single-copy gene completeness reached 78.05%. Comparative analysis against functional databases including NR, KOG, and KEGGG revealed that 98.40% of genes received functional annotation. KOG annotation indicated that *M. officinalis* proteins function primarily in general function prediction, posttranslational modification, protein turnover, chaperone activities, and signal transduction mechanisms. GO functional classification showed gene enrichment in cellular components and biological processes. KEGG analysis revealed that genes involved in metabolic pathways predominated. Comparative genomic analysis with *Vitis vinifera*, *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Ginkgo biloba*, *Amborella trichopoda*, *Camellia sinensis*, and *Cinnamomum kanehirae* demonstrated that 20,801 of 23,424 *M. officinalis* genes could be classified into 12,129 families, including 515 gene families unique to *M. officinalis*. Phylogenetic analysis showed that *M. officinalis* (Magnoliaceae) is closely related to *C. kanehirae* (Lauraceae), with a divergence time of approximately 122.5 million years ago (mya). This study represents the first whole-genome analysis of *M. officinalis* using third-generation sequencing technology, providing valuable resources for further development and utilization of this species and establishing a foundation for whole-genome studies of other medicinal plants.

Keywords: *Magnolia officinalis*; genome; third-generation sequencing technology; gene annotation

Introduction

The completion of the Human Genome Project and the continuous development and maturation of genome sequencing technologies, particularly the advancement of third-generation sequencing to single-molecule real-time sequencing, have accelerated whole-genome research in plants. Genome size, defined as the total number of DNA base pairs in a haploid genome, forms the foundation of species genomics research. Magnoliaceae occupies a relatively primitive position in plant evolution and taxonomy. In recent years, chloroplast genome sequencing of *Magnolia* species has been extensively studied both domestically and internationally. For instance, Li et al. (2012; Li et al., 2013) established a standard sequencing pipeline for the chloroplast genome of *M. officinalis* using the 454 FLX second-generation high-throughput sequencing platform to distinguish *M. officinalis* from related species, and obtained the complete chloroplast genome sequence of *Magnolia grandiflora*, providing valuable information for elite cultivar breeding, chloroplast genetic engineering, molecular marker development, and phylogenetic analysis. Cui et al. (2019) sequenced the complete chloroplast genome of *M. sieboldii*, a congeneric species of *M. officinalis*, identifying 111 unique genes including 78 protein-coding genes, 29 tRNA genes, and 4 rRNA genes.

Magnolia officinalis, a species of Magnoliaceae, is mainly produced in eastern Sichuan and western Hubei provinces, with wild populations protected as a second-class national conservation species in China (Xue et al., 2019). Its bark, branch bark, root bark, and buds are all used medicinally and widely applied in clinical practice. Additionally, its large and beautiful flowers are listed as health food products, while its seeds can be pressed for oil with benefits for improving vision and replenishing qi. As a genuine medicinal material, its main active components are phenolic compounds represented by magnolol and honokiol, which have demonstrated significant antibacterial, anti-inflammatory, anti-tumor, and antiviral pharmacological effects (Wang et al., 2005). Cha et al. (2015) investigated the biosynthetic pathway of terpenoid compounds in *M. officinalis* through transcriptomics, revealing the regulatory mechanisms of the mevalonate (MVA) pathway in synthesizing terpenoid secondary metabolites. Shi et al. (2018) further explored the phenylpropanoid and terpenoid synthesis pathways in *M. officinalis* secondary metabolism, obtaining information on related enzymes and genes in these metabolic pathways.

M. officinalis has a long natural growth cycle and relatively low yield, yet market demand remains high, leading to extensive artificial cultivation and rich germplasm resources (Zhang et al., 2013). However, current research on *M. officinalis* lacks understanding of its genetic information, evolutionary history, and the molecular basis of trait formation, resulting in unclear regulatory mechanisms for the synthesis of core secondary metabolites such as magnolol and hon-

okiol. This knowledge gap hinders effective molecular-assisted breeding and the identification of genes related to growth, development, disease resistance, and stress tolerance, leading to low resource utilization and insufficient development of *M. officinalis*. Therefore, based on the scarcity of genomic information for *M. officinalis*, this study conducted whole-genome sequencing to enrich genetic and evolutionary research resources and establish a foundation for exploring elite cultivar breeding, biosynthetic pathways and regulatory mechanisms of active components, and comprehensive utilization of medicinal plants.

1.1 Sample Collection and DNA Extraction

M. officinalis plants were selected from the Medicinal Botanical Garden of Chengdu University of Traditional Chinese Medicine. Fresh, young, disease-free leaves were harvested, washed with distilled water, then cleaned three times with 75% ethanol, dried, and stored at -80°C for future use. DNA was extracted from *M. officinalis* leaves using the CTAB method (Sha, 2018). The procedure involved: (1) grinding samples in liquid nitrogen and aliquoting into centrifuge tubes; (2) adding cetyltrimethylammonium bromide (CTAB) solution, incubating at 65°C for 1 hour, centrifuging at 10,800 rpm for 10 minutes, and collecting the supernatant; (3) adding an equal volume of chloroform:isoamyl alcohol (24:1), mixing thoroughly, centrifuging at 4°C and 10,800 rpm for 10 minutes, collecting the supernatant, and repeating this step twice; (4) adding isopropanol and sodium acetate solution to the supernatant, centrifuging, and discarding the supernatant; (5) adding 75% ethanol, centrifuging, and discarding the supernatant; (6) air-drying and dissolving in TE buffer, then storing at 4°C for future use.

1.2 Library Construction and Sequencing

DNA samples were first sheared using g-TUBE fragmentation tubes. Fragmented DNA (5 g) was then processed using the SMRTbell Template Prep Kit for damage repair, end repair, and adapter ligation. The adapter-ligated products were size-selected using the BluePippin Size-Selection System and purified with AMpure PB magnetic beads. The purified products underwent secondary damage repair using the SMRTbell Damage Repair Kit, followed by another round of purification with AMpure PB magnetic beads. The final library (secondary damage repair product) was assessed for concentration (Qubit) and size (Agilent 2100) to obtain the sequencing library. Single-molecule sequencing was performed on the PacBio Sequel third-generation sequencing platform. Raw data were evaluated and filtered to obtain high-quality data for genome assembly and quality assessment.

1.3 Genome Assembly and Evaluation

Low-quality and short reads were filtered from raw PacBio sequencing data. The Canu software (Koren et al., 2017) was used for initial assembly of filtered data,

followed by LACHESIS (Belton et al., 2012) for grouping, ordering, and orienting the assembled sequences. Each scaffold was broken into 50 kb fragments, and Hi-C (high-throughput chromosome conformation capture) technology (Marbout & Koszul, 2015) was applied for reassembly. Positions that could not be restored to the original assembly were identified as candidate error regions, and locations with low Hi-C coverage depth within these regions were identified as error points, thereby completing error correction of the initial assembly to improve genome quality. BUSCO v2.0 (Simao et al., 2015) was used to assess assembly completeness by comparing against 1,440 conserved core genes in the Embryophyta_odb9 database, and interaction heatmaps were generated to evaluate Hi-C assembly results. LACHESIS parameters were: (1) CLUSTER_MIN_RE_SITES=52; (2) CLUSTER_MAX_LINK_DENSITY=2; (3) CLUSTER_NONINFORMATIVE_RATIO=2; (4) ORDER_MIN_N_RES_IN_TRUN=46; (5) ORDER_MIN_N_RES_IN_SHRED=42.

1.4 Gene Prediction

Repeat sequences were predicted using LTR_FINDER v1.05 (Zhao & Wang, 2007), RepeatScout v1.0.5 (Price et al., 2005), and PILER-DF v2.4 (Edgar & Myers, 2005) based on structural prediction and *ab initio* principles to construct a repeat sequence database. The constructed repeat library was classified using PASTEClassifier (Wicker et al., 2007) and merged with the Repbase database (https://www.girinst.org/repbase/) as the final repeat sequence database for *M. officinalis*. RepeatMasker v4.0.6 (Tarailo & Chen, 2009) was then used to predict repeat sequences in *M. officinalis* based on this database.

Gene prediction was performed using both *ab initio* and homolog-based approaches. *Ab initio* prediction employed Genscan (Burge & Karlin, 1997), Augustus v2.4 (Stanke & Waack, 2003), GlimmerHMM v3.0.4 (Majoros et al., 2004), GeneID v1.4 (Blanco et al., 2007), and SNAP (version 2006-07-28) (Blanco et al., 2007). Homolog-based prediction used GeMoMa v1.3.1 (Jens et al., 2016). EVM v1.1.1 integrated results from all methods. Non-coding RNA prediction, including microRNA, rRNA, and tRNA, was performed using Infernal 1.1 (Nawrocki & Eddy, 2013) against Rfam (Griffithsjones et al., 2005) and miRBase (Griffithsjones et al., 2006) databases for rRNA and microRNA, and tRNAscan-SE v1.3.1 (Lowe & Eddy, 1997) for tRNA identification.

1.5 Functional Gene Annotation

Predicted gene sequences were compared against functional databases including NR (Non-Redundant Protein Database) (Aron et al., 2011), KOG (EuKaryotic Orthologous Groups) (Tatusov et al., 2001), KEGG (Kyoto Encyclopedia of Genes and Genomes) (Minoru & Susumu, 2000), and TrEMBL (Boeckmann et al., 2003) using BLAST v2.2.31 (Altschul et al., 1990) with an e-value threshold of <1e-5. GO (Dimmer et al., 2012) functional annotation was performed using Blast2GO (Conesa et al., 2005) based on NR database alignment results.

1.6 Comparative Genomics Analysis

Protein sequences from *M. officinalis* and eight other species [*Vitis vinifera*, *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, *Ginkgo biloba*, *Amborella trichopoda*, *Camellia sinensis*, and *Cinnamomum kanehirae*] were aligned (NCBI database https://www.ncbi.nlm.nih.gov/). Sequence and structure comparisons of known genes were performed to analyze interspecies evolution and classification of species-specific genes. OrthoMCL (Li, 2003) (parameters: Pep_length: 10, Stop_coden: 20, PercentMatchCutoff: 50, EvalueExponentCutoff: -5, Mcl: 1.5 #1.2~4.0) was used for family classification to identify gene families unique to *M. officinalis*. Single-copy protein sequences extracted from OrthoMCL clustering results were aligned using Muscle (http://www.ebi.ac.uk/Tools/msa/muscle/), and a phylogenetic tree was constructed using PHYML (Stéphane et al., 2010) (parameters: -gapRatio 0.5, -badRatio 0.25, -model HKY85, -bootstrap 1000) via maximum likelihood (ML) to study evolutionary relationships. Divergence times were estimated using mcmctree (http://abacus.gene.ucl.ac.uk/software/paml.html) after querying fossil times between species from Timetree (http://www.timetree.org/). MCScanX (Wang et al., 2012) was used for synteny analysis within *M. officinalis* (parameters: -s 10, -b 1, other parameters default) and with the closely related species *C. kanehirae* (parameters: -s 10, -b 2, other parameters default) to count syntenic gene numbers and blocks.

2.1 Genome Sequencing

Whole-genome sequencing of *M. officinalis* leaves was performed using the third-generation sequencing platform. Raw reads were filtered for quality, removing low-quality and short fragments. The filtered dataset comprised 140.91 Gb of third-generation raw data with a read N50 of 13,784 bp, maximum read length of 128,492 bp, and average length of 8,654 bp, meeting quality requirements for subsequent assembly.

2.2 Genome Assembly and Evaluation

Initial assembly using Canu software yielded results shown in Table 1 . After Hi-C error correction and assembly, the genome size was approximately 1.68 Gb with a contig N50 of 222,069 bp and longest contig of 2,700,203 bp. GC content was 40.65%. Hi-C assembly anchored 1.67 Gb (99.66%) of genome sequences to 19 chromosomes, representing 11,470 sequences (99.20%). Among anchored sequences, 1.53 Gb (91.21%) with determined order and orientation comprised 8,689 sequences (75.75% of anchored sequences).

BUSCO assessment identified 1,340 complete BUSCO genes (93.05% completeness), including 1,124 single-copy genes, 176 duplicated genes, 61 fragmented genes, and 93 missing genes from the Embryophyta_odb9 database. Hi-C assembly heatmap analysis (Figure 1 [Figure 1: see original paper]) clearly distinguished 19 chromosome groups, with diagonal interaction signals stronger

than off-diagonal positions, indicating high interaction intensity between adjacent sequences (diagonal) and weak signals between non-adjacent sequences (off-diagonal), confirming high-quality genome assembly.

2.3 Gene Prediction Results

RepeatMasker v4.0.6 predicted 1.37 Gb of repetitive sequences (81.60% of the genome). This included 450,863 long interspersed nuclear elements (LINEs, 8.47%), 18,530 short interspersed nuclear elements (SINEs, 0.2%), 997,318 long terminal repeats (LTRs, 44.04%), 145,539 terminal inverted repeats (TIRs, 4.5%), and 10,506 simple sequence repeats (SSRs, 0.47%). Gene prediction identified 23,424 protein-coding genes and 1,096 non-protein-coding genes, including 72 microRNA genes, 575 tRNA genes, and 449 rRNA genes (Table 3).

2.4 Functional Annotation and Analysis

KOG functional annotation (Figure 2 [Figure 2: see original paper]) assigned functions to 13,845 genes (59.11% of predicted genes). Protein functions were concentrated in "posttranslational modification, protein turnover, chaperones" (O, 10%), "signal transduction mechanisms" (T, 9%), "carbohydrate transport and metabolism" (G, 5%), and "transcription" (K, 5%). "General function prediction only" (R) accounted for 22%. These differentially expressed genes provide data support for future investigations into *M. officinalis* environmental response mechanisms during evolution.

GO annotation (Figure 3 [Figure 3: see original paper]) assigned functions to 13,438 genes (57.37% of predicted genes). Genes involved in "cell,""binding,""catalytic activity," "cellular process," and "metabolic process" were predominant. Across all categories, cellular components comprised 33%, molecular functions 21%, and biological processes 45%, indicating initial enrichment of *M. officinalis* genes in metabolic processes within biological processes.

KEGG pathway annotation (Figure 4 [Figure 4: see original paper]) assigned pathways to 8,253 genes (35.23% of predicted genes), distributed as 5.40% "cellular processes," 4.50% "environmental information processing," 29.85% "genetic information processing," 55.09% "metabolism," and 5.16% "organismal systems." Metabolic pathway genes were predominant, with starch and sucrose metabolism (ko00500), amino acid biosynthesis (ko01230), and carbon metabolism (ko01200) as major pathways.

2.5 Comparative Genomics Analysis

Protein sequence comparison between *M. officinalis* and eight other species revealed that 20,801 of 23,424 predicted genes could be classified into 12,129 families, including 515 families unique to *M. officinalis* (Figure 5 [Figure 5: see

original paper], Table 4). Phylogenetic analysis using single-copy protein sequences (Figure 6 [Figure 6: see original paper]) confirmed that *M. officinalis* clusters with *C. kanehirae*, indicating close phylogenetic relationship. Divergence time analysis (Figure 7 [Figure 7: see original paper]) estimated their separation at approximately 122.5 mya. Synteny analysis (Figure 8 [Figure 8: see original paper]) comparing *M. officinalis* and *C. kanehirae* genomes revealed limited collinear fragments, suggesting substantial genomic differences between the two species.

3 Discussion

Advances in genome sequencing and bioinformatics technologies, coupled with reduced sequencing costs and improved analytical methods, have greatly facilitated whole-genome studies of non-model medicinal plants like *M. officinalis*. Common methods for determining genome size include flow cytometry (Lin et al., 2019), second-generation high-throughput sequencing (Li et al., 2012), and third-generation single-molecule sequencing (Liu et al., 2015). Our third-generation sequencing yielded a *M. officinalis* genome size of approximately 1.68 Gb, consistent with the 1.59 Gb reported for *M. officinalis* subsp. *biloba* (2n=2x=38) using flow cytometry (Ye et al., 2015). Genome size correlates positively with ploidy level and chromosome number (Ye et al., 2015). Chromosome analysis of *M. officinalis* callus tissue revealed 2n=38 chromosomes (Wang et al., 2005), and the aforementioned subspecies *biloba* is also diploid with 38 chromosomes, confirming that our genome size aligns with its ploidy and chromosome number.

Genome functional annotation is crucial for functional gene analysis. Our GO annotation revealed gene enrichment in "metabolic process" within biological processes, consistent with KEGG results showing metabolism pathway genes as predominant, particularly starch and sucrose metabolism, amino acid biosynthesis, and carbon metabolism. Transcriptome analysis of different *M. officinalis* tissues using Illumina high-throughput sequencing also identified carbohydrate metabolism, amino acid metabolism, and energy metabolism as major pathways (Yang et al., 2019), corresponding to our genomic annotations. Integrating genomic and transcriptomic analyses will facilitate functional gene discovery and analysis in *M. officinalis*.

Most previous studies on medicinal plants like *M. officinalis* focused on chloroplast genomes due to the challenges of nuclear genome sequencing: large size, complex structure, polyploidy, and abundant repetitive sequences (Chen et al., 2014). However, validation against previous studies confirms that our *M. officinalis* whole-genome sequence is of high quality and represents the first nuclear genome sequence for a *Magnolia* species, providing a reference for studying origin and evolution of Magnoliaceae. Completion of the *M. officinalis* whole-genome sequence represents an important step toward molecular-assisted breeding of medicinal plants. Integrating genomics, proteomics, and germplasm information through bioinformatic analysis enables screening of optimal genotypes

and breeding strategies (Ma & Mo, 2017), offering a novel approach for clinically important medicinal plants. The whole-genome sequence also provides data support for functional genomics studies (Wang et al., 2009), enabling identification of key enzymes in secondary metabolite synthesis, metabolic pathway elucidation, and screening of superior trait loci related to growth, development, disease resistance, and stress tolerance through transcriptomics and metabolomics—an effective strategy for addressing insufficient resource development. This whole-genome sequencing study enhances molecular-level understanding of *M. officinalis*, provides a reference for other medicinal plants, and establishes a molecular biology foundation for modernizing traditional Chinese medicine resources.

References

Altschul SF, Gish W, Miller W, et al., 1990. Lipman DJ: Basic local alignment search tool[J]. *J Mol Biol*, 215: 403-410.

Aron MB, Shennan L, Anderson JB, et al., 2010. CDD: A conserved domain database for the functional annotation of proteins[J]. *Nucl Acid Res*. 39(Suppl_1): D225-D229.

Belton JM, McCord RP, Gibcus JH, et al., 2012. Hi–C: A comprehensive technique to capture the conformation of genomes[J]. *Methods*, 58(3): 268-276.

Blanco E, Genis P, Roderic G, 2007. Using geneid to identify genes[J]. *Current Protocols*, 18(1).

Boeckmann, Brigitte, et al., Apweiler R, et al., 2003. The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003[J]. *Nucl Acid Res*, 31(1): 365-370.

Burge C, Karlin S, 1997. Prediction of complete gene structure in human genomic DNA[J]. *J Mol Biol*, 268(1): 78-94.

Chen Y, Liu YS, Zeng JG, 2014. Progress in plant genome sequencing [J]. *Life Science Res*, 18(1): 66-74.

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

Cui Y, Li C, Zhang Y, et al., 2019. The complete chloroplast genome of Siebold' s magnolia: *Magnolia sieboldii* (Magnoliaceae), a highly ornamental species with attractive aromatic flowers[J]. *Conserv Genet Resour*, 11(3): 299-301.

Dimmer, Emily C, et al., 2012. Eberhardt R: The UniProt-GO annotation database in 2011[J]. *Nucl Acid Res*, 40: D565-D570.

Edgar RC, Myers EW, 2005. PILER: identification and classification of genomic repeats[J]. *Bioinformatics*, 21(Suppl. 1): i152-i158.

Griffithsjones S, Grocock RJ, Dongen SV, et al., 2006. miRBase: microRNA sequences, targets and gene nomenclature[J]. *Nucl Acid Res*, 34(Suppl. 1): 140-4.

Griffithsjones S, Moxon S, Marshall M, et al., 2005. Rfam: Annotating Non-Coding RNAs in Complete Genomes[J]. *Nucl Acid Res*, 33(Database issue): D121-4.

Jens K, Michael W, Erickson JL, et al., 2016. Using intron position conservation for homology-based gene prediction[J]. *Nucl Acid Res*, 44(9): e89-e89.

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.

Li XW, Gao HH, Wang YT, et al., 2012. High throughput sequencing and structural analysis of the whole chloroplast genome of Evergreen magnolia [J]. *Chinese Science: Life Science*, 42 (12): 947-956.

Li XW, Gao H, Wang Y, et al., 2013. Complete chloroplast genome sequence of *Magnolia grandiflora* and comparative analysis with related species[J]. *Sci Chin Life Sci*, 56(2): 189-198.

Li XW, Hu ZG, Lin XH, et al., 2012. Whole chloroplast genome sequencing of *Magnolia Officinalis* based on 454 FLX high throughput technology and its application[J]. *Acta Pharm Sin*, 47(1): 124-130.

Li L, Stoeckert CJ, Roos DS, 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes[J]. *Genome Res*, 13(9): 2178-2189.

Lin H, Han XW, Lan SR, et al., 2019. Determination of genome size of two orchids based on flow cytometry [J]. *J For Environ*, 39(6): 616-620.

Liu YH, Wang L, Yu L, 2015. Principle and application of single molecule real-time sequencing [J]. *Genetics*, 37(3): 259-268.

Lowe TM, Eddy SR, 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence[J]. *Nucl Acid Res*, 25(5): 955-964.

Ma XJ, Mo CM, 2017. Prospects for molecular breeding of medicinal plants [J]. *Chin J Trad Chin Med*, 42(11): 2021-2031.

Majoros WH, Pertea M, Salzberg SL, 2004. TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders[J]. *Bioinformatics*, 20(16): 2878-2879.

Marbouty M, Koszul R, 2015. Metagenome analysis exploiting high-throughput chromosome conformation capture (3C) data[J]. *Trends Genet*, 31(12): 673-682.

Minoru K, Susumu G, 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes[J]. *Nucl Acid Res*, 28(1): 27-30.

Nawrocki EP, Eddy SR, 2013. Infernal 1.1: 100-fold faster RNA homology searches[J]. *Bioinformatics*, 29(22): 2933-2935.

Price AL, Jones NC, Pevzner PA, 2005. *De novo* identification of repeat families in large genomes[J]. *Bioinformatics*, 21(Suppl 1): i351-i358.

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

Shi XD, Gu YX, Dai J, et al., 2018. Gene mining and analysis of *Magnolia officinalis* secondary metabolite pathway based on transcriptome[J]. *Lishizhen Med Mat Med Res*, 29(1): 247-250.

Simao 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.

Stanke M, Waack S, 2003. Gene prediction with a hidden Markov model and a new intron submodel[J]. *Bioinformatics*, 19(Suppl 2): ii215-ii225.

Stephane G, Dufayard JF, Lefort V, et al., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0[J]. *Syst Biol*, 59(3).

Tatusov RL, Natale DA, Garkavtsev IV, et al., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes[J]. *Nucl Acid Res*, 29(1): 22-28.

Tarailo GM, Chen N, 2009. Using RepeatMasker to identify repetitive elements in genomic sequences[J]. *Current Protocols*, 25(1): 4.10. 1-4.10. 14.

Wang LQ, Jiang RG, Chen HF, 2005. Research progress on pharmacological effects of magnolol and honokiol[J]. *Chin Trad Herb Drugs*, (10): 155-158.

Wang YB, Liu Z, Zhao AH, et al., 2009. Application of functional genomics in the study of secondary metabolites of medicinal plants [J]. *Chin J Trad Chin Med*, 34(1): 6-10.

Wang Y, Tang H, Debarry JD, et al., 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity[J]. *Nucl Acid Res*, 40(7): e49-e49.

Wang YH, Xu WJ, Ma DW, et al., 2005. Chromosome production and karyotype analysis of *Magnolia officinalis* [J]. *J Sichuan Norm Univ (Nat Sci Ed)*, 28(2): 242-244.

Wicker T, Sabot F, Hua V, et al., 2007. A unified classification system for eukaryotic transposable elements[J]. *Nat Rev Genet*, 8(12): 973-982.

Xue ZZ, Zhang RX, Yang B, 2019. Research progress of *Magnolia officinalis* authenticity[J]. *Chin J Chin Mat Med*, 44(17): 3601-3607.

Yang X, Yang ZL, Tan M, et al., 2019. Analysis of transcriptome characteristics of *Magnolia officinalis* and development of EST-SSR markers [J]. *J Nucl Agric*, 33 (7): 1318-1329.

Ye LJ, Zhang ZR, Sun ZX, et al., 2015. Determination of nuclear DNA content (2C value) in the main genera of Magnoliaceae [J]. *J Plant Classif Resour*, 37(5): 605-610.

Zha LP, Yuan Y, Huang LQ, et al., 2015. Identification and bioinformatics analysis of *Magnolia officinalis* MVA related genes[J]. *Chin J Chin Mat Med*, 40(11): 2077-2083.

Zhang LF, Huang SJ, Jiang JL, et al., 2013. Study on the current situation and resource development of *Magnolia officinalis* forest [J]. *Fujian For*, (2): 28-30.

Zhao X, Wang H, 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons[J]. *Nucl Acid Res*, 35(Suppl. 2): W265-W268.

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