Genome Survey and SSR Molecular Marker Analysis of *Macaranga indica* and *Macaranga denticulata* Postprint

Authors: Li Jiangying, Lu Tianquan, Yang Junbo, Tian Bo

Date: 2020-08-02T00:00:00+00:00

Abstract

Macaranga indica and Macaranga denticulata are plants of the genus Macaranga in the family Euphorbiaceae. Plants of this genus have multiple medicinal values and are widely used in the treatment of numerous diseases in folk medicine. The nervonic acid contained in the seeds of these two plant species has also attracted considerable attention from researchers. To determine suitable whole-genome sequencing research strategies for Macaranga indica and Macaranga denticulata, this study employed second-generation high-throughput sequencing technology combined with bioinformatics methods to determine for the first time genomic information such as genome size, heterozygosity rate, and repeat rate of these two species, and preliminarily analyzed the

Full Text

Genome Survey and SSR Molecular Marker Analysis of *Macaranga indica* and *M. denticulata*

LI Jiangying[1],[3], LU Tianquan[1], YANG Junbo[2], TIAN Bo[1]*

[1]Key Laboratory of Tropical Plant Resources and Sustainable Use, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming 650223, China
[2]The Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650204, China
[3]University of Chinese Academy of Sciences, Beijing 101408, China

Abstract

Macaranga indica and *M. denticulata* are species of the genus *Macaranga* in the Euphorbiaceae family. Plants of this genus possess various medicinal properties

and are widely used in folk medicine for treating numerous diseases. The nervonic acid present in the seeds of these two species has also attracted significant research attention. To determine an appropriate whole-genome sequencing strategy for these species, this study employed second-generation high-throughput sequencing technology combined with bioinformatics methods to characterize, for the first time, the genome size, heterozygosity rate, repeat content, and other genomic features of *M. indica* and *M. denticulata*. The SSR sequence characteristics of both species were also preliminarily analyzed. The results showed that: (1) the genome sizes of *M. indica* and *M. denticulata* were 986.84 Mb and 946.23 Mb, respectively; (2) the heterozygosity rates were 0.75% and 0.65%, respectively, while repeat sequence proportions were 73.02% and 71.5%; (3) SSR feature analysis of the two species identified 4,499,185 SSRs in *M. indica* and 4,969,098 SSRs in *M. denticulata*. These findings provide theoretical guidance for SSR molecular marker screening and development, as well as for deep whole-genome sequencing of *M. indica* and *M. denticulata*.

Keywords: *Macaranga indica*, *Macaranga denticulata*, nervonic acid, genome survey, SSR

Macaranga indica and *M. denticulata* belong to the genus *Macaranga* (tribe Acalypheae, family Euphorbiaceae). Both are tall trees with peltate leaves widely distributed in valleys, secondary forests, and evergreen broad-leaved forests in southwestern China. According to *Zhonghua Bencao* (Chinese Materia Medica), the roots and bark of *M. denticulata*—the primary medicinal parts—possess jaundice-reducing and heat-clearing properties and can be used to treat gastric pain and hepatitis (Huang et al., 2015). Previous studies have reported the isolation of various compounds from *M. indica*, including ellagic acid and prenylated flavonoids, which exhibit multiple biological activities such as antioxidant and anti-inflammatory effects, suggesting potential for industrial extraction (Yang et al., 2015). Analysis of fatty acid composition in the seeds of both species revealed the presence of nervonic acid, a very-long-chain monounsaturated fatty acid. This finding confirms earlier work by Wang et al. (2006) who identified *M. adenantha* (now revised in the new *Flora of China* as conspecific with *M. indica*) as a woody plant with high nervonic acid content in seed oil, making it an ideal resource for nervonic acid product development. Nervonic acid is a core natural component of brain fibers and nerve cells, essential for brain nerve biosynthesis. It plays important biological roles in promoting brain development, improving memory, and delaying brain aging (Li et al., 2019). Nervonic acid intake can prevent and treat various neurological diseases including Alzheimer' s disease, post-stroke sequelae, brain atrophy, cerebral palsy, memory loss, and insomnia (Tian et al., 2015). Consequently, nervonic acid development has attracted considerable attention from experts worldwide, and utilizing nervonic acid-rich plants has become the primary approach to meet growing demand.

Current research on *M. indica* and *M. denticulata* has primarily focused on pharmacological activities of chemical constituents from common medicinal parts

and seed fatty acid composition. However, no studies have reported genomic information for these species, which significantly hinders efficient utilization of wild resources and breeding of new varieties. As woody plants with unknown genome sizes, various factors have slowed molecular biology research on *Macaranga* species. Therefore, before conducting deep whole-genome sequencing, low-coverage genome surveys are necessary to understand genomic composition characteristics and patterns (Li et al., 2019). Analyzing genetic information in DNA is a massive undertaking, and the primary challenge is overcoming technical difficulties (Albach et al., 2007). The rapid development of plant whole-genome research has benefited from advancing next-generation sequencing technologies (Shi et al., 2012). With maturing sequencing technologies and decreasing costs, genome sequencing has been widely applied to species with scientific, economic, and ornamental value. Genome sequencing helps elucidate regulatory mechanisms of life phenomena, population evolution, growth, and development. Current methods for determining genome size include flow cytometry, Feulgen spectrophotometry, pulsed-field gel electrophoresis, and rapidly developing high-throughput sequencing technology (Wu et al., 2014). In the Euphorbiaceae family, genomic information has been reported for several plants including physic nut (*Jatropha curcas*), castor bean (*Ricinus communis*), cassava (*Manihot esculenta*), and rubber tree (*Hevea brasiliensis*) (Chan et al., 2010; Shusei et al., 2011; Simon et al., 2012; Zou & Yang, 2019), providing valuable references for studying *Macaranga* genomes.

This study employed Illumina second-generation high-throughput sequencing technology to conduct the first genome survey of *M. indica* and *M. denticulata*. Bioinformatics methods were used to estimate repeat content, heterozygosity, and SSR characteristics based on the genome survey data. The objectives were to provide a basis for developing whole-genome sequencing and assembly strategies, support further research and utilization of *Macaranga* species, facilitate genetic improvement, and offer references for germplasm resource conservation and genetic diversity studies using SSR markers.

1.1 Materials

The experimental materials—wild *M. indica* and *M. denticulata* plants with normal flowering and fruiting—were collected in July 2019 from the roadside in Mengsong Village, Menglong Town, Jinghong City, Xishuangbanna Dai Autonomous Prefecture. Samples were brought to the laboratory, snap-frozen in liquid nitrogen, and stored at -80°C for subsequent use.

1.2 Genomic DNA Extraction, Detection, and Sequencing

Genomic DNA was extracted from young leaves of *M. indica* and *M. denticulata* using the CTAB method. Sample concentration was measured using a UV spectrophotometer, and integrity was assessed via agarose gel electrophoresis. Extracted DNA samples were sent for library construction and sequencing. Based on genome sizes of other woody plants and C-value ranges for Euphorbiaceae

species, a genome size of approximately 1 Gb was used to evaluate sequencing coverage for *M. indica* and *M. denticulata*.

1.3 Library Construction and Data Statistics

Whole-genome shotgun (WGS) libraries with insert sizes of 350 bp and 500 bp were constructed for both species and sequenced on the Illumina HiSeq™ 2000 platform using paired-end sequencing. Raw reads were generated and processed for image recognition, decontamination, and adapter removal. Statistical results included read count, data yield, sequencing error rate, Q20 content, Q30 content, and GC content.

1.4 Genome Size Prediction and Heterozygosity Estimation

Genome size and heterozygosity of *M. indica* and *M. denticulata* were estimated using K-mer analysis (K=17) of the sequencing data. K-mer distribution plots were used to assess repetitive sequence content; high repeat proportions produce right-skewed tails in the distribution. K-mer depth follows a Poisson distribution, and the expected K-mer depth obtained from the curve was used to estimate genome size (Zhou et al., 2019). Additionally, the presence of a small peak on either side of the main peak in the K-mer distribution curve indicates high heterozygosity, while its absence suggests low heterozygosity.

1.5 Sample Contamination Assessment

Contamination assessment is critical in genome research to ensure sequence integrity, data validity, and reliable results. Contaminated data cannot provide accurate information. For quality control, 10,000 high-quality reads (5,000 read1 and 5,000 read2) were randomly selected from the filtered data and compared against the NCBI nucleotide database (NT) using BLAST. Homologous matches indicated no exogenous contamination, while matches to distantly related species suggested potential contamination (Yan, 2018).

1.6 SSR Analysis

The microsatellite identification tool MISA (http://pgrc.Ipk-gatersleben.de/misa/) was used to search for SSR loci in all sequences with the following parameters: mono-10, di-6, tri-5, tetra-5, penta-5, hexa-6. The maximum distance between two different SSRs in compound sequences was set to 100 bp (Zhang et al., 2019).

2.1 DNA Extraction from Materials

Genomic DNA was extracted from young leaves of *M. indica* and *M. denticulata* using the CTAB method. Electrophoresis showed good quality DNA for both species [Figure 1: see original paper]. The DNA concentration was 15.42 ng ·

L^{-1} for *M. indica* and 10.46 ng· L^{-1} for *M. denticulata*, suitable for subsequent analysis.

2.2 Sequencing Data Output Statistics

High-throughput paired-end sequencing of both species was performed on the Illumina platform. After stringent filtering of raw data, high-quality clean data were obtained. The output statistics for four libraries of *M. indica* and *M. denticulata* are shown in , including read count, data yield, sequencing error rate, Q20 content, Q30 content, and GC content. After removing low-quality data, 53.56 Gb and 68.07 Gb of data were obtained for *M. indica* and *M. denticulata*, respectively. Both species showed normal base quality with Q20 and Q30 values exceeding 90% and sequencing error rates of 0.04%. GC content was 33.89% for *M. indica* and 33% for *M. denticulata*, indicating good raw sequencing quality suitable for subsequent analysis.

2.3 K-mer Analysis and Genome Size Estimation

K-mer analysis (K=17) was performed on 53.56 Gb and 68.07 Gb of data from *M. indica* and *M. denticulata*, respectively. The 17-mer distributions are shown in [Figure 2: see original paper], with the x-axis representing total K-mer occurrences and the y-axis representing K-mer frequency (Tang et al., 2015). Heterozygous peaks appeared before the main peak in both species, indicating certain heterozygosity levels. Both species showed severe rightward tailing in their 17-mer curves, indicating high repeat sequence proportions. As shown in , sequencing depths were 40× and 54× for *M. indica* and *M. denticulata*, respectively. Total K-mer counts were 39,725,851,195 for *M. indica* and 51,594,983,117 for *M. denticulata*. Genome size was estimated using the formula G = K-num/K-depth, where K-depth is the expected sequencing depth and K-num is the total K-mer count (Liu, 2019, http://blog.sciencenet.cn/u/lyao222lll). The estimated genome size was 993.15 Mb (revised to 986.84 Mb) for *M. indica* and 955.46 Mb (revised to 946.23 Mb) for *M. denticulata*. Heterozygosity rates were 0.75% and 0.65%, while repeat rates were 73.02% and 71.5% for *M. indica* and *M. denticulata*, respectively. These results indicate that both species have highly repetitive, moderately heterozygous genomes.

2.4 Sample Contamination Assessment–Nucleotide Comparison Results

From each 350 bp and 500 bp library of *M. indica* and *M. denticulata*, 10,000 filtered high-quality single-end reads (5,000 read1 and 5,000 read2) were randomly selected for BLAST comparison against the NT database. The top six species by match ratio are shown in . Castor bean (*Ricinus communis*) showed the highest match ratios in all libraries: 1.61% and 1.9% in the 350 bp and 500 bp libraries of *M. indica*, and 1.78% and 1.65% in the corresponding libraries of *M. denticulata*. Taxonomically, castor bean belongs to Euphorbiaceae and is a close relative of both species. All other matched species were plants, with no

high-ratio matches to animals or microorganisms, confirming that the samples were contamination-free and suitable for subsequent genome survey analysis.

2.5 SSR Analysis of *M. indica* and *M. denticulata* Genomes

MISA was used to search for SSRs in all preliminarily assembled sequences of both species. As shown in , 4,499,185 SSRs were identified in *M. indica*, with 445,117 sequences containing more than one SSR and 492,341 SSRs present in compound formation. In *M. denticulata*, 4,969,098 SSRs were found, with 458,726 sequences containing multiple SSRs and 507,887 SSRs in compound formation. SSR nucleotide types were categorized as mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide repeats. In *M. indica*, these numbered 2,800,292 (62.24%), 1,199,707 (26.66%), 432,509 (9.61%), 48,890 (1.09%), 10,498 (0.23%), and 7,289 (0.16%), respectively. In *M. denticulata*, the corresponding numbers were 3,037,613 (61.13%), 1,321,752 (26.60%), 522,801 (10.52%), 63,973 (1.29%), 11,254 (0.23%), and 11,705 (0.24%). Further classification of SSR repeat motifs by sequence composition was performed, with partial results shown in .

Genome size is fundamental to comparative and evolutionary genomics, while heterozygosity and repeat content are key determinants of assembly quality. Evaluating these parameters helps identify appropriate assembly strategies (Bi et al., 2019). The 17-mer analysis estimated genome sizes of approximately 987 Mb and 946 Mb for *M. indica* and *M. denticulata*, respectively. These sizes are larger than those of cassava (770 Mb) (Simon et al., 2012), castor bean (350 Mb) (Shusei et al., 2011), and physic nut (410 Mb) (Chan et al., 2010), but slightly smaller than rubber tree (1.1 Gb) (Zou & Yang, 2019). This variation likely reflects inter-generic differences within Euphorbiaceae, as the compared species belong to different genera. The completed genome size estimation provides a reference for understanding genome size variation patterns in *Macaranga*.

Assessing heterozygosity facilitates selection of appropriate assembly methods. Genomes can be classified as moderately heterozygous ($0.5\% \leq$ heterozygosity $< 0.8\%$), highly heterozygous (heterozygosity $\geq 0.8\%$), or highly repetitive (repeat content $\geq 50\%$) (Wang et al., 2018). The heterozygosity rates of 0.75% and 0.65%, and repeat rates of 73.03% and 71.6% for *M. indica* and *M. denticulata*, respectively, indicate moderately heterozygous, highly repetitive genomes. Both species are dioecious, which may contribute to their relatively high heterozygosity. Consequently, WGS-based analysis poses certain risks and challenges. We recommend a combined strategy using second-generation (Illumina) and third-generation (PacBio) sequencing technologies, supplemented with Hi-C technology for chromosome-level assembly, to obtain high-quality genome maps for both species.

SSR markers offer advantages including ease of use, high polymorphism, low cost, and abundance. Based on the genome survey data, SSR analysis revealed an average density of one SSR per 2,251 bp in *M. indica* and one per 2,348 bp in *M. denticulata*, with rich repeat types. Both species showed significant

base preferences: A/T content exceeded C/G content in mononucleotide repeats, while AT/AT was the most abundant dinucleotide motif and CG/CG the least abundant. This may result from methylation-induced conversion of C residues to T, creating large differences between the two nucleotide repeats (Zhou et al., 2017). Studies suggest that higher proportions of low-level repeat units indicate higher evolutionary levels, while higher proportions of high-level repeat units suggest shorter evolutionary time or lower mutation frequency (Yu et al., 2019). Therefore, large-scale development and screening of SSR markers based on genome survey sequencing will provide references for future research on genetic map construction, genetic diversity analysis, and QTL mapping.

References

Albach DC, Li HQ, Zhao N, et al., 2007. Molecular systematics and phytochemistry of *Rehmannia* (Scrophulariaceae)[J]. Biochem Syst Ecol, 35(5): 293-300.

Bi QX, Zhao Y, Cui YF, et al., 2019. Genome survey sequencing and genetic background characterization of yellow horn based on next-generation sequencing[J]. Mol Biol Rep, 46(4): 3729-3738.

Chan AP, Crabtree J, Zhao Q, et al., 2010. Draft genome sequence of the oilseed species *Ricinus communis*[J]. Nat Biotechnol, 28(9): 951-956.

Huang JY, Lu WJ, Tan X, et al., 2015. Chemical constituents from *Macaranga denticulata* root[J]. Chin Med Mat, 38(8): 1671-1673.

Li GQ, Song LX, Jin CQ, et al., 2019. Genome survey and SSR analysis of *Apocynum venetum*[J]. Biosci Rep, 39(6): BSR20190146. doi: https://doi.org/10.1042/BSR20190146

Li Q, Chen J, Yu XZ, et al., 2019. A mini review of nervonic acid: Source, production, and biological functions[J]. Food Chem, 125286. doi: https://doi.org/10.1016/j.foodchem.2019.125286

Shi JS, Wang ZJ, Chen JH, 2012. Progress on whole genome sequencing in woody plants[J]. Hereditas, 34(2): 145-156.

Shusei S, Hideki H, Sachiko I, et al., 2011. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L.[J]. DNA Res, 18(1): 65-76.

Simon P, Pradeep R, Brian D, et al., 2012. The cassava genome: Current progress, future directions[J]. Trop Plant Biol, 5(1): 88-94.

Tang Q, Ma XJ, Mo CM, et al., 2015. Genome survey analysis in *Siraitia grosvenorii*[J]. Guihaia, 35(6): 786-791.

Tian DY, Wang SA, Wang LH, et al., 2015. The biosynthesis and metabolic engineering of very long-chain monounsaturated fatty acid[J]. Biotechnol Bull, 31(12): 42-49.

Wang XY, Fan JS, Wang SQ, 2006. Development situation and outlook of nervonic acid plants in China[J]. Chin Oils Fats, 3: 69-71.

Wang X, Zhou JY, Sun HG, et al., 2018. Genomic survey sequencing and estimation of genome size of *Ammopiptanthus mongolicus*[J]. J Plant Genet Resourc, 19(1): 143-149.

Wu YF, Xiao FM, Xu HN, et al., 2014. Genome survey in *Cinnamomum camphora* L. presl[J]. J Plant Genet Resourc, 15(1): 149-152.

Yang DS, Peng WB, Yang YP, et al., 2015. Cytotoxic prenylated flavonoids from *Macaranga indica*[J]. Fitoterapia, 103: 187-191.

Yu FL, Huang M, Zhang YB, et al., 2019. Genome survey and characteristic analysis of SSR in *Callicarpa nudiflora*[J]. Chin J Chin Mat Med, 44(18): 3974-3978.

Zhang JX, Tu MW, Xue S, et al., 2019. Genome survey and analysis of SSR molecular markers on traditional Chinese medicine *Nauclea officinalis*[J]. Mol Plant Breed, 17(23): 7829-7833.

Zhou JY, Wang X, Gao F, et al., 2017. Genome survey and SSR analysis of *Ammopiptanthus mongolicus*[J]. Genom Appl Biol, 36(10): 4334-4338.

Zhou Y, Zi H, Tong J, et al., 2019. A genome survey of *Rhododendron simsii* and *Rhododendron indicum*[J]. Mol Plant Breed, 17(15): 4928-4935.

Zou Z, Yang J, 2019. Genomic analysis of Dof transcription factors in *Hevea brasiliensis*, a rubber-producing tree[J]. Ind Crops Products, 134: 271-283.

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