◆濒危植物遗传多样性◆

厚叶木莲(*Manglietia pachyphylla*)基因组草图*

甘新军[1],宾　粤[2,3],陈焕锦[1],朱韦光[2,3],熊露桥[1],余恩萍[2,3,4],王峥峰[2,3**],徐凤霞[3,5],曹洪麟[2,3]
(1.广东从化陈禾洞省级自然保护区管理处,广东广州　510950;2.中国科学院华南植物园,广东省应用植物学重点实验室,中国科学院退化生态系统植被恢复与管理重点实验室,广东广州　510650;3.华南国家植物园,广东广州　510650;4.中国科学院大学,北京　100049;5.中国科学院华南植物园,中国科学院植物资源保护与可持续利用重点实验室,广东广州　510650)

摘要:厚叶木莲(*Manglietia pachyphylla*)为木兰科(Magnoliaceae)木莲属(*Manglietia*)的木本植物,零星分布于我国广东省和广西壮族自治区,为国家二级重点保护野生植物。了解濒危物种基因组信息及其遗传多样性有助于合理地保护和利用濒危物种,实现濒危物种的解濒和复壮。为此,本研究通过高通量测序方法对厚叶木莲基因组进行测序,并利用测序数据开展厚叶木莲基因组草图的组装;之后,基于组装的基因组预测其中的重复序列和基因,进行系统发育和基因家族分析。结果表明,组装的厚叶木莲基因组大小为2 092 298 891 bp,包含676个组装序列,N50(将组装的序列按照长度由大到小进行累加,当累加到某个序列时,累加的值为基因组50%的长度时,此序列的长度即为N50)为7 961 115 bp;利用BUSCO (Benchmarking Universal Single-Copy Orthologs),针对"eudicots"和"embryophyta"这两个BUSCO单拷贝基因库,对基因组组装的完整性进行评估,组装的厚叶木莲基因组完整性分别为96.6%和98.8%。厚叶木莲基因组有76.5%的序列为重复序列,共有37 900个基因,这些基因编码了41 675个蛋白质序列。系统发育分析发现厚叶木莲与望春玉兰(*Magnolia biondii*)聚在一起,两者分化时间大致为10 500 000年前。厚叶木莲中与木质部/韧皮部、肌动蛋白丝、热、光合作用以及多种次生代谢相关的基因家族显著扩张,其中次生代谢相关基因在厚叶木莲基因组上呈串联和近端重复,这些基因的扩张和重复形成方式可能与厚叶木莲适应高海拔环境有关。本研究是国内外木兰科木莲属首个基因组报道,为更好地保护和开发厚叶木莲及木兰科其他物种的种质资源提供了遗传信息和参考。

关键词:厚叶木莲;木兰科;木莲属;濒危植物;基因组组装;基因预测;基因家族;基因重复

中图分类号:Q16　　**文献标识码**:A　　**文章编号**:1005-9164(2023)06-1079-12
DOI:10.13656/j.cnki.gxkx.20240125.006

收稿日期:2023-06-10　　修回日期:2023-07-22
*广东省重点领域研发计划项目(2022B1111230001)及其子课题(2022B1111230001-2-5),广东从化陈禾洞省级自然保护区"厚叶木莲种群调查、评价及扩繁"项目和广东省林业局生态林业建设专项资金项目"高水平专类园珍稀特有植物保护与建设"资助。

【第一作者简介】
甘新军(1970-),男,高级工程师,主要从事林业资源开发、保护与管理研究。
【**通信作者简介】
王峥峰(1973-),男,研究员,主要从事保护生物学和种群遗传学研究,E-mail:wzf@scib.ac.cn。

【引用本文】
甘新军,宾粤,陈焕锦,等.厚叶木莲(*Manglietia pachyphylla*)基因组草图[J].广西科学,2023,30(6):1079-1090.
GAN X J,BIN Y,CHEN H J,et al.Draft Genome of *Manglietia pachyphylla* [J].Guangxi Sciences,2023,30(6):1079-1090.

木兰科(Magnoliaceae)是最原始的被子植物,其属种丰富,类型多样,是研究被子植物起源和进化的重要类群[1-4]。我国是木兰科起源地和避难所,为木兰科植物分布中心,拥有很多古老、孑遗和特有物种[2,5-7]。我国的木兰科物种现主要分布于我国热带亚热带地区,如云南省、广西壮族自治区、贵州省和湖南省。木兰科中的很多物种树形优美,叶型秀丽、色泽鲜艳,花形态各异、花色丰富明艳且高贵典雅,观赏性强,是优良的绿化树种[8-11]。木兰科物种的枝、叶、花含有挥发性有机物,作为绿化树种可以净化空气,促进人们身心健康[12-15],而且这些挥发性有机物与木兰科产生的其他次生代谢物如生物碱,常被用作中药[16-18]。另外,木兰科的一些种类具有较强的光合能力、固碳能力以及土壤改良作用,可用于人工林改造,实现林业提质增效[19,20];而且一些木兰科植物树干通直,木材材质均匀、细密,是很好的用材树种[8,21,22]。同时,木兰科植物最早记载于我国秦汉时代,历经千年,寄托了人们对美好生活的追求和向往,富有深厚的文化沉淀[23]。

厚叶木莲(*Manglietia pachyphylla*)为木兰科木莲属(*Manglietia*)的木本植物,为中国特有种,目前零星分布于我国广东省和广西壮族自治区海拔 500 m 以上的常绿阔叶林中[24,25]。厚叶木莲的显著特征是有光泽的革质厚叶,可作为园林绿化和用材物种[24]。厚叶木莲种群少,个体数量小,结实率低,种子易被动物啃食,自然更新差[25,26],现被列为国家二级重点保护野生植物[24]。目前国内外已开展厚叶木莲群落学[25,26]、花粉形态[27]和光响应生理[28]等方面的研究,但还未有关于厚叶木莲的遗传多样性及其基因组的研究报道。基因组及基于基因组开展的物种遗传多样性研究,可以揭示物种进化过程,从而了解物种的适应性并进一步应用于品种改良[29]。为此,本研究采用二代和三代高通量测序手段,对厚叶木莲基因组进行测序,组装其基因组草图,为今后更好地开展其进化、遗传多样性研究提供参考。

1 材料与方法

1.1 材料

在广东省广州市从化区陈禾洞省级自然保护区内选择 1 株厚叶木莲成年大树(生长点地理位置为 113°55′31.84″E,23°45′1.25″N),其胸高直径(Diameter at Breast Height,DBH)为 22.4 cm,采集其无虫咬和病斑的树叶 3 片。采集的树叶用剪刀剪碎后,用锡箔纸包裹后立刻投入液氮罐中,之后将样品送至武汉未来组生物技术有限公司进行高通量测序。

1.2 方法

测序包括 3 个方面的内容:一是采用 Nanopore PromethION 测序平台对厚叶木莲进行三代基因组测序,二是采用 MGI DNBSEQ-T7 测序平台对厚叶木莲进行二代基因组测序,三是采用 MGI DNBSEQ-T7 测序平台对厚叶木莲进行转录组测序,具体实验流程参考 Wang 等[30]的研究。针对所得的测序数据,利用不同程序开展数据处理和基因组组装、分析。在数据处理过程中主要使用程序的默认参数,如在程序运行过程中对默认参数进行改动,则会在文中具体说明。

1.2.1 测序数据前处理

本研究利用 Sickle v1.33(https://github.com/najoshi/sickle)对厚叶木莲二代基因组测序数据进行过滤,去除测序数据中碱基质量小于 30、片段长度小于 80 bp 的测序数据。过滤后的二代测序数据用 RECKONER v1.1[31]进行纠错。对于三代基因组测序数据,由于低质量数据在交付前测序公司已过滤,无需再过滤,本研究只利用 Porchop 0.2.4(https://github.com/rrwick/Porechop)对厚叶木莲三代基因组测序数据进行接头过滤。

1.2.2 基因组大小预测与组装

针对厚叶木莲二代测序数据,利用 GenomeScope 2.0[32]进行基因组大小预测(参数"-k 21")。针对厚叶木莲三代测序数据,本研究选择大于 10 kb 测序序列,利用 NextDenovo 2.3.1(https://github.com/Nextomics/NextDenovo)进行基因组组装,组装的基因组利用 Pseudohaploid(https://github.com/schatzlab/pseudohaploid)和 Purge_Dups v1.2.6[33]去除冗余序列(如杂合导致的拼接序列),之后利用 racon v1.5.0[34],hapo-G v1.3.2[35]和 polypolish v0.5.0[36]进行组装序列纠错。针对组装完成的基因组,利用 BUSCO(Benchmarking Universal Single-Copy Orthologs)v5.4.6[37]对照"eudicots_odb10.2020-09-10"和"embryophyta_odb10.2020-09-10"两个单拷贝基因库进行组装完整性的评估。

1.2.3 重复序列和基因的预测、注释

本研究利用 EDTA(Extensive De-novo TE Annotator)v2.1.0[38]和 RED(REpeat Detector)v2.0[39]预测厚叶木莲基因组中的重复序列。利用 BEDTools v2.29.2[40]中的"merge"命令将两个重复

序列预测结果合并,再利用"maskfasta"命令将预测的重复序列屏蔽掉。针对屏蔽了重复序列的基因组,首先利用 BRAKER2[41],同时结合转录组数据和 9 个物种的蛋白质序列(表 1)预测厚叶木莲基因组组装序列中的基因;然后将结果输入 funannotate pipeline v1.8.13(https://github.com/nextgenusfs/funannotate)中,同样结合转录组数据和 9 个物种的蛋白质序列,进一步预测厚叶木莲基因组组装序列中的基因,这一过程包括 3 个步骤和命令:"funannotate train""funannotate predict"和"funannotate update",在"predict"步骤中使用的参数为"-max_intronlen 100,000 -busco_db embryophyta -organism other"。

基因预测结束后,利用 8 个不同的蛋白质注释平台对预测的基因进行功能分析,包括:dbCAN (DataBase for automated Carbohydrate-active enzyme ANnotation)v10.0[42],eggNOG-mapper (Evolutionary genealogy of genes:Non-supervised Orthologous Groups-mapper) v5.0.2[43], GO (Gene Ontology)[44,45],KEGG (Kyoto Encyclopedia of Genes and Genomes)[46],InterPro (The Integrated Resource of Protein Domains and Functional Sites) v5.60-92.0[47], MEROPS v12.0[48], Pfam (The protein families database)v35.0[49]和 SignalP 5.0b[50]。

考虑到注释的基因中有可变剪切的状况,在进行比较基因组的研究中去除了各物种基因中的可变剪切产生的基因,只保留其中最长的基因序列进行分析。

表 1　预测厚叶木莲基因组基因的参考物种

Table 1　Reference species used for gene prediction of *Manglietia pachyphylla* genome assembly

物种 Species	中文名 Chinese name	数据来源 Data source
Arabidopsis thaliana	拟南芥	GCF_000001735.4(GenBank)
Aristolochia fimbriata	卷毛马兜铃	GCA_019845555.1(GenBank)
Chimonanthus salicifolius	柳叶蜡梅	http://xhhuanglab.cn/data/Chimonanthus_salicifolius.html
Cinnamomum micranthum	沉水樟	https://ngdc.cncb.ac.cn/gwh/ncbi_assembly/55517/show
Corymbia citriodora		GCA_014858505.1(GenBank)
Liriodendron chinense	鹅掌楸	https://doi.org/10.5061/dryad.s4mw6m947
Magnolia biondii	望春玉兰	https://doi.org/10.5061/dryad.s4mw6m947
Nymphaea colorata	蓝星睡莲	GCF_008831285.2(GenBank)
Persea americana	鳄梨	https://genomevolution.org/CoGe/SearchResults.pl?s=29305&p=genome

1.2.4　基因家族和系统发育分析

利用 OrthoFinder 3.0.0[51,52]并结合其他 9 个物种(表 1)的蛋白质序列进行基因家族分析。在 OrthoFinder 分析过程中,程序自动分析获得物种间的单拷贝基因,并用这些单拷贝基因进行系统发育树构建。利用构建的系统发育树,使用 MCMCTree[53]进行物种间分化时间的估算。在这一分析过程中需要参考已有物种间的分化时间,本研究从 http://timetree.org/获得这些信息,并在表 2 中列出。在得到有分化时间的系统发育树后,再利用 cafe (Computational Analysis of gene Family Evolution)v5[54]进行基因家族的扩张和收缩分析;对其中显著扩张和收缩的基因家族,则利用 TBtools v1.115[55]进行 GO 和 KEGG 的富集分析。

表 2　从 http://timetree.org/获得的物种对间分化时间

Table 2　Species pairs and their estimated divergence times from http://timetree.org/

物种对 Species pair	估测的分化时间/百万年前 Estimated divergent time/(million years ago,MYA)
Nymphaea colorata-Arabidopsis thaliana	168.4-91.6
Corymbia citriodora-Arabidopsis thaliana	100.0-112.2
Magnolia biondii-L. chinense	27.8-50.0
Chimonanthus salicifolius-P. americana	105.0-115.9
L. chinense-P. americana	114.0-130.3

1.2.5　基因重复(Gene duplications)

利用 wgd v1.1.2[56]进行基因组的全基因组重

复(Whole genome duplication)事件分析,该过程是在基因组中进行基因间的同源性分析,找到两两同源基因,并对两两同源基因进行同义突变率(synonymous substitution rate,*Ks*)计算,之后查看 *Ks* 值的密度分布图,其峰值出现的地方提示有全基因重复事件发生;进一步利用 Doubletrouble v0. 99. 1 (https://github. com/almeidasilvaf/doubletrouble) 开展全基因组重复基因、串联重复(Tandem duplications)基因、近端重复(Proximal duplications)基因、转座重复(Transposed duplications)基因、散在重复(Dispersed duplications)基因的分析[55]。其中,串联重复基因是指彼此连续或中间间隔不超过 5 个其他基因的近缘基因;近端重复基因是连续分布、中间间隔不超过 10 个其他基因的近缘基因;转置重复基因是由转座子介导的重复基因;散在重复基因是随机分布、彼此间不靠近的重复基因[57]。利用 TBtools 分别对全基因组重复基因、串联重复基因、近端重复基因进行 GO 和 KEGG 富集分析。

2 结果与分析

2.1 厚叶木莲基因组测序与组装

本研究中三代基因组测序共获得大约 118.1 Gb 测序数据,二代基因组测序获得约 264.1 Gb 测序数据,转录组测序获得约 31.4 Gb 测序数据。基因组和转录组原始测序数据已上传至 GenBank,三代基因组测序数据序列号为 SRR24423593、SRR24423594、SRR24423595;二代基因组测序数据序列号为 SRR24390471、SRR24390472、SRR24390473;转录组测序数据序列号为 SRR24415003。利用 GenomeScope 2. 0 分析厚叶木莲基因组大小为 1 969 269 649 bp。

本研究利用 NextDenovo 2. 3. 1 进行厚叶木莲基因组组装,得到基因组组装大小为 2 350 821 062 bp,包含 1 436 个拼接序列(contig),N50(将组装的序列按照长度由大到小进行累加,当累加到某个序列时,累加的值为基因组 50%的长度时,此序列的长度即为 N50)为 6 839 193 bp,最长拼接序列为 26 073 671 bp,最短拼接序列为 16 675 bp,平均为 1 637 062.0 bp。去除冗余序列和纠错后,最终的基因组组装大小为 2 092 298 891 bp,包含 676 个拼接序列,N50 为 7 961 115 bp,最长拼接序列为 26 180 362 bp,最短拼接序列为 27 281 bp,平均为 3 095 117 bp,组装的其他统计信息见表 3。组装的厚叶木莲基因组数据上传至 GenBank,序列号为 JASAUF000000000。

表 3 厚叶木莲基因组组装结果

Table 3 Statistics of genome assembly of *Manglietia pachyphylla*

原始组装结果 Initial assembly		最终组装结果 Final assembly	
组接序列长度/bp Length of sequence/bp	组接序列次序 Order of sequence length	组接序列长度/bp Length of sequence/bp	组接序列次序 Order of sequence length
N10 = 17 193 138	L10 = 12	N10 = 17 470 745	L10 = 11
N20 = 12 987 901	L20 = 28	N20 = 14 043 218	L20 = 24
N30 = 10 742 051	L30 = 48	N30 = 11 767 464	L30 = 40
N40 = 9 305 499	L40 = 71	N40 = 9 970 083	L40 = 60
N50 = 6 839 193	L50 = 101	N50 = 7 961 115	L50 = 83
N60 = 5 187 776	L60 = 141	N60 = 6 110 540	L60 = 113
N70 = 3 522 166	L70 = 195	N70 = 4 743 283	L70 = 152
N80 = 2 135 406	L80 = 282	N80 = 3 234 931	L80 = 206
N90 = 653 184	L90 = 489	N90 = 1 764 892	L90 = 292
N100 = 16 675	L100 = 1 436	N100 = 27 281	L100 = 676

Note:in the table,"N" and "L" before "N10","N20" ,…,"N100" and "L10","L20",…,"L100" refer to the length of the contig and the number of contigs when the cumulative length and number of contig account for 10%,20%,…,100% of the total length and number of contigs;The accumulating process is performed from the longest contig to the shortest contig in order.

BUSCO 对组装的基因组完整性的评估显示,比对“eudicots_odb10. 2020-09-10”单拷贝基因库,全部 2 326 个 BUSCO 单拷贝基因,96.6% BUSCO 单拷贝基因在厚叶木莲基因组中完整匹配。其中,能完整

匹配到BUSCO单拷贝基因库而且在厚叶木莲基因组中也是单拷贝的基因占90.8%,能完整匹配到BUSCO单拷贝基因库但在厚叶木莲基因组中为多拷贝的基因占5.8%;不完整匹配到BUSCO单拷贝基因库的基因有27个,占1.2%;有54个BUSCO单拷贝基因未在厚叶木莲基因组的基因中匹配到,占2.3%。比对"embryophyta_odb10.2020-09-10"单拷贝基因库,全部1 614个BUSCO单拷贝基因中,98.8% BUSCO单拷贝基因在厚叶木莲基因组中完整匹配,能完整匹配到BUSCO单拷贝基因库而且在厚叶木莲基因组中也是单拷贝的基因占95.3%,能完整匹配到BUSCO单拷贝基因库但在厚叶木莲基因组中为多拷贝的基因占3.5%;不完整匹配到BUSCO单拷贝基因库的基因有12个,占0.7%;有7个BUSCO单拷贝基因未在厚叶木莲基因组的基因中匹配到,占0.4%。

表4 EDTA检测的重复序列

Table 4 Repeat type and sequence detected by EDTA

种类 Class		数量 Count	长度/bp Length/bp	重复序列占基因组的比例/% Percentage of repeat sequences in the genome/%
Long terminal repeats	Copia	267 375	254 612 532	12.17
	Gypsy	501 489	475 508 695	22.73
	Unknown	606 245	353 364 188	16.89
Terminal inverted repeats	CACTA	234 311	86 471 640	4.13
	Mutator	419 086	103 230 948	4.93
	PIF_Harbinger	181 913	44 020 258	2.10
	Tc1_Mariner	14 856	5 113 434	0.24
	hAT	347 915	96 850 898	4.63
nonTIR	Helitron	109 850	28 264 246	1.35
Total		2 683 040	1 447 436 839	69.17

2.2 厚叶木莲基因组注释

在厚叶木莲基因组中,RED和EDTA分别检测到1 342 604 397 bp (64.16%)和1 447 436 839 bp (69.17%)重复序列,两者合并后得到1 601 108 919 bp的重复序列,占基因组的76.52%。EDTA检测结果表明,厚叶木莲基因组中重复序列最多的是Gypsy类的长末端重复序列(Long terminal repeat),有475 508 695 bp,占基因组的22.73%(表4)。

本研究对厚叶木莲基因组进行基因预测共获得37 900个基因,这些基因共编码了41 675种蛋白质。利用数据库对这些蛋白质序列进行功能注释,其中32 249个(77.4%)蛋白质序列注释到了8个蛋白质注释平台数据库中的一个,具体如下:17 886个蛋白质序列注释到GO数据库,21 430个蛋白质序列注释到InterPro数据库,31 146个蛋白质序列注释到eggNOG-mapper数据库,23 203个蛋白质序列注释到Pfam数据库,1 160个蛋白质序列注释到dbCAN数据库,1 016个蛋白质序列注释到MEROPS数据库,3 078个蛋白质序列注释到SignalP数据库,15 752个蛋白质序列注释到KEGG数据库。基因注释文件已上传至中国科学院华南植物园数据存储库(https://doi.org/10.57841/casdc.0001214)。

2.3 厚叶木莲基因家族与基因组进化分析

本研究对包括厚叶木莲在内的10个物种的蛋白质序列进行了基因家族分析,结果共得到24 616个基因家族。从表5统计结果可以看出,厚叶木莲基因组中有34 319个基因归为其中一个基因家族,占总基因的90.6%。对于木兰科其他两个物种鹅掌楸和望春玉兰,它们分配到基因家族的基因比例分别为95.7%和93.4%,均高于厚叶木莲;但厚叶木莲基因分布于15 894个基因家族中,占全部基因家族的64.6%;而鹅掌楸和望春玉兰的基因出现在13 858和14 935个基因家族中,占全部基因家族的56.3%和60.7%,低于厚叶木莲。24 616个基因家族中有710个厚叶木莲特有的基因家族,这些家族包含了3 373个厚叶木莲基因。鹅掌楸和望春玉兰特有的基因家族分别是437和999个,虽然一个低于厚叶木莲特有基因家族数目,一个高于厚叶木莲特有基因家族数目,但是它们所包含的基因数(分别为6 845和4 543)均高于厚叶木莲,而且这些特有基因数占两个

物种各自所有基因的比例也高于厚叶木莲,特别是鹅掌楸,为 19.4%(表 5,图 1)。3 种木兰科植物共同出现的基因家族有 345 个(图 1)。

对厚叶木莲特有的基因家族进行 GO 和 KEGG 富集分析,结果表明这些基因家族所包含的基因的主要功能与细胞内稳态(GO:0055067、GO:0006885、GO:0030433、GO:0036503)、葡聚糖代谢(Glucan metabolic process)(GO:0044042、GO:0016762、GO:0046527)、原核生物抗性(Prokaryotic defense system)、苯丙氨酸代谢(Phenylalanine metabolism)和转录(Transcription)等有关(https://doi.org/10.6084/m9.figshare.23690001.v3,表 S1 和 S2)。

表 5　基因家族及其基因信息统计

Table 5　Statistics of gene families and its related gene information

物种 Species	基因数目 Number of genes	在基因家族中的基因数目 Number of genes in ortho-groups	未归到基因家族的基因数目 Number of unassigned genes	在基因家族中的基因比例/% Percentage of genes in ortho-groups/%	未归到基因家族中的基因比例/% Percentage of unassigned genes/%	包含有该物种的基因家族数目 Number of ortho-groups containing species	包含有该物种基因的基因家族占所有基因家族的比例/% Percentage of ortho-groups containing species/%	特有基因家族的数目 Number of specie-sspecific ortho-groups	特有基因家族所包含的基因数目 Number of genes in specie-sspecific ortho-groups	特有基因家族所包含基因数目的比例/% Percentage of genes in species-specific ortho-groups/%
Arabidopsis thaliana	27 562	24 951	2 611	90.5	9.5	12 973	52.7	944	4 649	16.9
Aristolochia fimbriata	21 751	19 099	2 652	87.8	12.2	12 751	51.8	339	1 747	8.0
Chimonanthus salicifolius	36 651	31 717	4 934	86.5	13.5	14 639	59.5	1 356	5 694	15.5
Cinnamomum micranthum	26 680	25 255	1 425	94.7	5.3	13 392	54.4	273	1 224	4.6
Corymbia citriodora	35 628	32 099	3 529	90.1	9.9	14 115	57.3	1 198	6 981	19.6
L. chinense	35 269	33 740	1 529	95.7	4.3	13 858	56.3	437	6 845	19.4
Magnolia biondii	41 181	38 481	2 700	93.4	6.6	14 935	60.7	999	4 543	11.0
Manglietia pachyphylla	37 900	34 319	3 581	90.6	9.4	15 894	64.6	710	3 373	8.9
N. colorata	20 488	19 359	1 129	94.5	5.5	12 187	49.5	371	1 632	8.0
P. americana	22 440	21 163	1 277	94.3	5.7	13 280	53.9	158	414	1.8

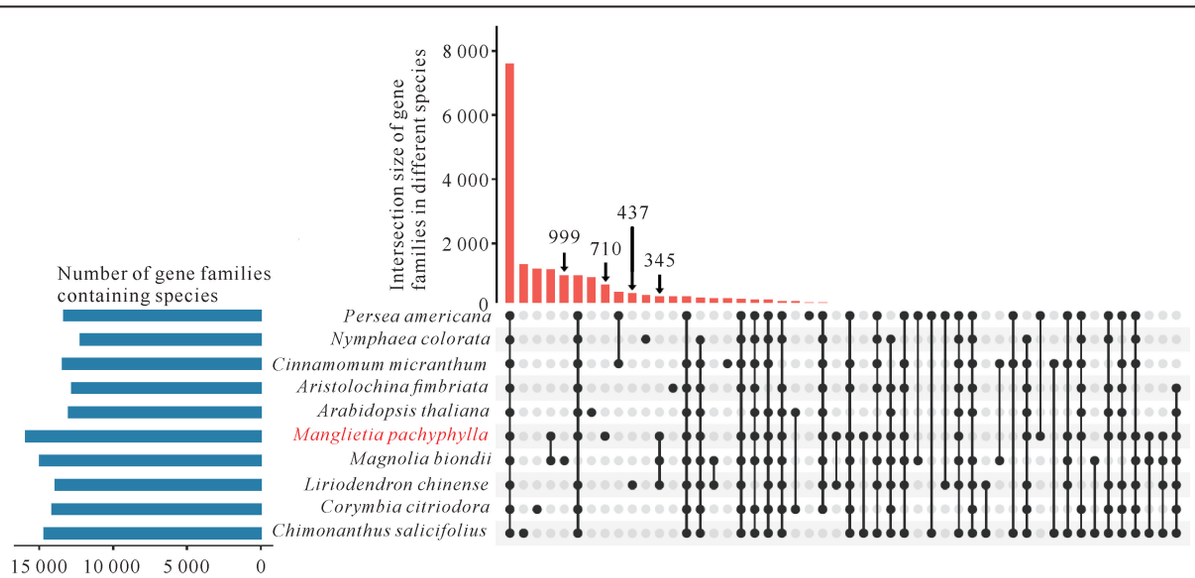
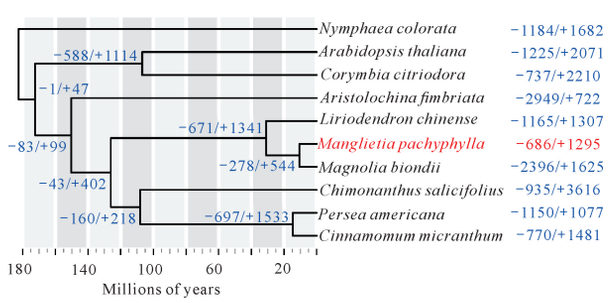


图 1　厚叶木莲与其他物种基因家族交集图

Fig. 1　Upset plot showing gene families in *Manglietia pachyphylla* and other species

系统发育分析表明厚叶木莲与望春玉兰聚在一起(图 2),两个物种是在约 10 500 000 年前(95%CI:4 989 810－17 055 600 年)从共同的祖先开始分化。对基因家族扩张和收缩分析结果表明,厚叶木莲基因组中有 686 个基因家族表现为收缩,1 295 个基因家族表现为扩张,其中有 136 个基因家族表现为显著收缩,417 个基因家族表现为显著扩张。GO 和 KEGG 富集分析表明,厚叶木莲基因组中显著扩张的基因家族与木质部、韧皮部发育(GO:0010088,GO:0010087)、肌动蛋白丝(Actin filament,GO:0030837、GO:0030833、GO:0061572、GO:0051017、GO:0030832、GO:0008064、GO:0030041)、细胞运动(Cell motility)、油菜素甾醇生物合成(Brassinosteroid biosynthesis)、类黄酮生物合成(Isoflavonoid biosynthesis)、维生素 B6 代谢(Vitamin B6 metabolism)、硫代葡萄糖苷(Glucosinolate biosynthesis)和萜类物质生物合成有关(https://doi.org/10.6084/m9.figshare.23690001.v3,表 S3 和表 S4);厚叶木莲基因组中显著收缩的基因家族主要与木质素(Lignin)代谢(GO:0009808、GO:0046274)、类黄酮生物合成,以及二苯乙烯、二芳基庚酸类和姜酚生物合成(Stilbenoid,diarylheptanoid and gingerol biosynthesis)相关(https://doi.org/10.6084/m9.figshare.23690001.v3,表 S5 和表 S6)。



Divergence times are shown below the tree; The "－/＋" and the numbers beside the tree nodes and species represent the number of contracted and expanded gene families in *Manglietia pachyphylla* and other species.

图 2　厚叶木莲与其他物种的系统发育树

Fig. 2　Inferred phylogenetic tree in *Manglietia pachyphylla* and other species

2.4　厚叶木莲全基因组和基因重复分析

全基因组重复事件分析表明,厚叶木莲和鹅掌楸、望春玉兰表现出同样的 *Ks* 峰型(图 3),因此它们近期共同经历了一次全基因组重复事件。对厚叶木莲基因组的基因重复研究表明,厚叶木莲基因组中有 4 769 个基因与其全基因组重复有关(12.6%),3 317 个基因为串联重复(8.8%),3 124 个基因为近端重复(8.2%),136 个基因为转置重复(0.4%),18 522 个基因为散在重复(48.9%)。GO 和 KEGG 富集分析表明,厚叶木莲基因组中全基因组重复基因主要功能与 DNA 内复制调控(Regulation of DNA endoreduplication)、植物昼夜节律(Circadian rhythm-plant)、鸟嘌呤核苷酸结合蛋白(GTP-binding proteins)、柠檬酸循环[Citrate cycle (TCA cycle)],蛋白酶体(Proteasome)、转录因子(Transcription factors)(https://doi.org/10.6084/m9.figshare.23690001.v3,表 S7 和表 S8)相关。GO 和 KEGG 富集分析表明,厚叶木莲基因组中串联重复基因与 NADH 氧化(NADH oxidation,GO:0006116)、寡肽转运(Oligopeptide transport,GO:0006857)、RNA 脱帽(RNA decapping,GO:0110154)、谷胱甘肽代谢过程(Glutathione metabolic process,GO:0006749)、过氧化氢分解代谢过程(Hydrogen peroxide catabolic process,GO:0042743、GO:0042744)、类黄酮生物合成、牛磺酸和低牛磺酸代谢(Taurine and hypotaurine metabolism)、苯并恶唑嗪酮类化合物生物合成(Benzoxazinoid biosynthesis)、玉米素生物合成(Zeatin biosynthesis)、萜类生物合成(Terpenoid biosynthesis)、苯丙烷类生物合成(Phenylpropanoid biosynthesis)、生物碱合成等相关(https://doi.org/10.6084/m9.figshare.23690001.v3,表S9和表

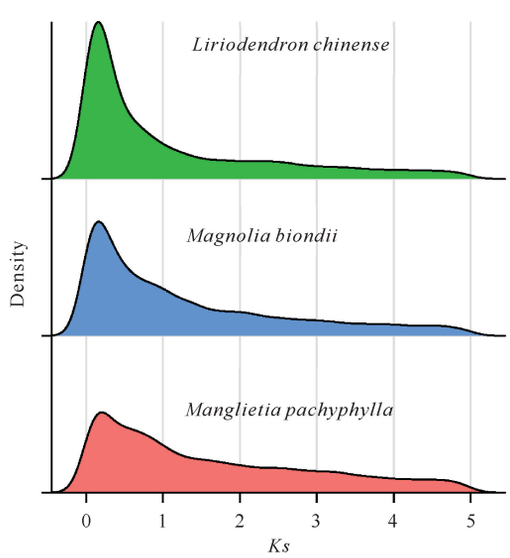


图 3　全基因组重复分析中同义突变率(*Ks*)的密度分布

Fig. 3　Density distribution of synonymous substitution rate (*Ks*) in whole-genome duplication

S10)。厚叶木莲基因组中近端重复基因主要和防御(GO:0043207、GO:0098542、GO:0051707、GO:0009607、GO:0050832、GO:0009605、GO:0006952)、次生代谢(GO:0044550、GO:0019748)、各类生物碱生物合成(Biosynthesis of various alkaloids)、花青素生物合成(Anthocyanin biosynthesis)、聚酮生物合成(Polyketide biosynthesis)、苯并恶唑嗪酮类化合物生物合成、萜类生物合成、硫代葡萄糖苷生物合成和油菜素甾醇生物合成等相关(https://doi.org/10.6084/m9.figshare.23690001.v3,表 S11 和表 S12)。

3 讨论

3.1 厚叶木莲的基因组组装

高通量测序已成为研究物种基因组和遗传多样性的主要手段。目前高通量测序分为短片段测序和长片段测序量两种方式,前者测序平台主要包括 Illumia 和 MGI 测序公司一系列仪器设备,后者测序平台主要包括 Nanopore 和 PacBio 测序公司一系列仪器设备。短片段高通量测序的优势是测序的准确性较高、数据量大和价格便宜;长片段测序的优势是测得的序列长度长、完整性好,有利于后续基因组组装的连续性,如 Nanopore 测序平台可测 100 kb 或更长的序列片段。但采用 Nanopore 和 PacBio 测序平台测得的序列错误率较高,为此 PacBio 公司推出了采用 Hifi(High fidelity reads)模式的测序方式,使得测序错误率大为降低,但是测序长度有所限制(15-20 kb)。为得到高质量基因组,可采用不同测序平台测序,再进行基因组组装,以实现优势互补。本研究采用 Nanopore 长片段测序平台的测序结果进行厚叶木莲的基因组初步组装,再用 MGI 短片段测序平台的测序结果对组装的基因组进行纠错,得到了较完整和准确的组装结果,但是还没有组装到染色体级别,后期本研究将继续采用 Hi-C(High-throughput chromosome conformation capture)测序方式进一步对厚叶木莲基因组进行测序,测得的结果可用于进一步组装以完善厚叶木莲基因组。

在采用 Nanopore 测序数据进行厚叶木莲基因组组装过程中,本研究采用 NextDenovo 2.3.1 软件进行序列拼接,这一软件采用先纠错再组装(Correction then assembly)的模式进行基因组组装,保证了组装序列的连续性和正确性,优于其他组装软件。同时本研究利用多种方法去除了原始组装序列中的冗余序列并进行后期纠错:其中 racon v1.5.0 和 hapo-G v1.3.2 分别使用 Nanopore 测序获得的长片段和 MGI 测序获得的短片段进行组装基因组内碱基和插入缺失错误的纠错,polypolish v0.5.0 主要针对组装序列中的同聚体长度(Polypolish-length,如“AAAAAA”这种重复序列的长度)进行纠错,3 个纠错软件的使用可以很好地保证组装基因组的准确性。

在基因组注释环节,本研究利用 EDTA 和 RED 软件相互组合的方法进行基因组中重复序列的查找。EDTA 主要用于基因组中转座子的查找,而 RED 查找包括转座子在内的所有重复序列类型,如微卫星体(Microsatellite)等,两者结合保证了重复序列查找的全面性,但 RED 分析的结果并没有对不同重复序列进行归类。本研究首先使用 BRAKER2,结合转录组数据和其他物种的蛋白质序列预测厚叶木莲基因组中的基因,得到初步基因序列注释结果。由于这一结果还存在较多错误预测,本研究进一步利用 funannotate 软件对初步预测的结果进行整合。Funannotate 是个软件整合工具,被编译为一个管段(Pipeline)流程。该软件可以进行不同方式的基因预测(包括 de novo 注释、转录组注释和同源蛋白质注释),再结合其他分析软件的结果(本研究中为 BRAKER2 基因预测结果),可以获得统一、可靠的高质量基因预测结果。

3.2 厚叶木莲的比较基因组学研究

本研究中厚叶木莲的基因组组装大小为 2 092 298 891 bp(约 2.09 Gb),大于木兰科的日本厚朴(*Magnolia hypoleuca*,修正名为 *Houpoea obovata*)基因组(约 1.64 Gb)[58]、厚朴(*Magnolia officinalis*,修正名为 *H. officinalis*)的基因组(约 1.68 Gb)[59]和鹅掌楸的基因组(约 1.74 Gb)[60],小于望春玉兰的基因组(约 2.22 Gb)[29]。基于“embryophyta_odb10.2020-09-10”单拷贝基因库,厚朴基因组的组装完整性评估为 86.20%的 BUSCO 单拷贝基因能被完整匹配到,日本厚朴基因组的组装完整性为 98.6%,望春玉兰基因组的组装完整性为 95.7%(上述数值为本研究利用望春玉兰基因组分析的结果),鹅掌楸基因组的组装完整性为 98.8%(此值为本研究利用鹅掌楸基因组重新分析的结果),而厚叶木莲的 BUSCO 组装完整性为 98.8%,与鹅掌楸的结果相同,但高于厚朴、日本厚朴和望春玉兰 3 个物种。从重复序列比例看,厚叶木莲基因组中重复序列占基因组的 76.5%,高于望春玉兰(66.48%)[29]、日

本厚朴(64.54%)[58]以及鹅掌楸(63.81%)[60],但小于厚朴(81.44%)[57],厚叶木莲基因组重复序列比例在已有木兰科物种基因组重复序列比例的范围内。上述结果说明,虽然本研究报道的厚叶木莲基因组还是草图状态,没有组装到染色体级别,但是组装已经很完整,结果可靠。

对基因家族的研究表明,厚叶木莲中与次生代谢相关的基因家族显著扩张,如萜类物质,这与日本厚朴、望春玉兰基因家族分析结果相似[29,58]。基因表达分析表明,望春玉兰花中萜类相关基因的表达高于其叶子,说明萜类物质是望春玉兰花香的主要原因[29],厚叶木莲花是否也有类似的结果还需结合基因组进一步研究确定。除了次生代谢物质相关的基因家族在厚叶木莲基因组中有显著扩张外,本研究还发现与木质部、韧皮部发育相关的基因家族,以及与光合作用[光合作用光反应(Photosynthesis light reaction,GO:0019684)、光合作用光收获(Photosynthesis light harvesting,GO:0009765)以及KEGG中的光蛋白(Photosynthesis proteins)]、温度[热反应(Response to heat,GO:0009408)、温度刺激反应(Response to temperature stimulus,GO:0009266)]等相关的基因家族也有扩张。厚叶木莲生长在较高海拔的地区,木材致密通直[26],这些扩张的基因家族是否与其适应高海拔的温度、光照、强风等生境有关还需进一步研究。厚叶木莲基因组中的扩张基因家族中的一部分与肌动蛋白丝(也称"微丝")相关,而肌动蛋白是植物细胞骨架的主要元素,相关研究表明它与植物抗真菌功能密切相关[61,62]。

对日本厚朴、望春玉兰的研究均表明木兰科都共同经历了一次近期的全基因组重复事件[29,58]。厚叶木莲有着与日本厚朴相似的全基因组重复基因、串联重复基因和近端重复基因比例。与全基因组重复相关的基因占厚叶木莲基因组基因的比例为12.6%,日本厚朴为13.4%[58];串联重复基因在厚叶木莲基因组基因的比例为8.8%,日本厚朴为7.6%[58];近端重复基因在厚叶木莲基因组基因的比例为8.2%,日本厚朴为9.4%[58]。基因串联重复和近端重复是基因形成的重要方式,与物种适应性密切相关[57]。日本厚朴中串联重复和近端重复基因的功能主要与苯丙烷类、萜类、类黄酮的生物合成等有关[58],这与厚叶木莲这两类基因的研究结果相同,但厚叶木莲可产生更多的次生代谢合成物,如苯并恶唑嗪酮类化合物、生物碱、聚酮等。厚朴基因组中,萜类合成相关的基因也呈串联重复状况[59]。值得注意的是日本厚朴原产日本,具有很好的抗寒性,研究发现苯丙烷类生物合成与日本厚朴抗寒性相关。

对厚叶木莲中与全基因组重复相关基因的富集分析表明,这些基因与植物最基本的生长发育密切相关,如植物昼夜节律相关基因[63]、与碳代谢相关的柠檬酸循环基因[64]、参与细胞的多种生命活动(如细胞通讯、核糖体与内质网的结合、小泡运输、蛋白质合成等)的GTP结合蛋白基因[65,66]等,说明全基因组重复在木兰科植物适应性方面具有重要作用。

4 结论

本研究首次报道了木兰科木莲属物种的基因组,这一结果对全面深入了解木兰科物种以及厚叶木莲的进化、适应具有重要作用。本研究组装的厚叶木莲基因组大小为2 092 298 891 bp,基因组序列中76.5%为重复序列。通过基因预测,在组装的厚叶木莲基因组中共注释到37 900个基因,它们编码了41 675种蛋白质。对厚叶木莲基因组研究发现,全基因组重复与木兰科植物进化适应性密切相关,很多与植物基本生长发育相关的基因通过全基因组重复在木兰科得到加强;厚叶木莲富含有与次生代谢相关的多种基因,由此产生的次生代谢物质有利于厚叶木莲在高海拔生长以及抵御病虫害,但其(种子)香味也使得其易受啃食伤害。因此,厚叶木莲基因组的组装为从遗传角度深入了解其濒危机制提供了可能,也为科学合理利用厚叶木莲以及提取其次生物质作为生物医药提供了参考。

参考文献

[1] 刘玉壶,夏念和,杨惠秋.木兰科(Magnoliaceae)的起源、进化和地理分布[J].热带亚热带植物学报,1995,3(4):1-12.

[2] 陈涛,张宏达.木兰科植物地理学分析[J].武汉植物学研究,1996,14(2):141-146.

[3] 张冰.木兰科(Magnoliaceae)植物区系分析[J].广西植物,2001,21(4):315-320.

[4] ROMANOV M S,DILCHER D L. Fruit structure in Magnoliaceae s. l. and Archaeanthus and their relationships [J]. American Journal of Botany,2013,100(8):1494-1508.

[5] 闫双喜,李永华,位凤宇.中国木兰科植物的地理分布[J].武汉植物学研究,2008,26(4):379-384.

[6] 王献溥,蒋高明.中国木兰科植物受威胁的状况及其保护措施[J].植物资源与环境学报,2001,10(4):43-47.

[7] XIE H H,TANG Y G,FU J,et al. Diversity patterns

and conservation gaps of Magnoliaceae species in China [J]. Science of the Total Environment, 2022, 813: 152665.
[8] 徐健,杨冬英. 木兰科植物在园林绿化上的应用[J]. 现代园艺,2009(11):46-47.
[9] 江世勇. 群芳独美木兰花[J]. 园林,1995(3):25.
[10] 易劲扬. 木兰科植物在园林绿化中的应用[J]. 现代园艺,2015(23):137-139.
[11] YIN Q,SHI X D,ZHU Z,et al. *Magnolia wufengensis* 'Jiaohong No. 2': a new *Magnolia* cultivar with bright red lotus-shaped flowers [J]. HortScience, 2022, 57(1):110-111.
[12] 钟瑞敏,张振明,肖仔君,等. 华南五种木兰科植物精油成分和抗氧化活性(英文)[J]. 云南植物研究,2006,28(2):208-214.
[13] 马惠芬,司马永康,张达,等. 木兰科含笑属含笑组 3 种植物叶的挥发性化学成分研究[J]. 西北林学院学报,2019,34(4):212-216.
[14] 王金凤,周琦,陈卓梅. 4 种木兰科景观树种挥发性有机物(VOCs)排放清单及其保健作用评价[J]. 林业与环境科学,2022,38(3):101-110.
[15] LIAO T Z,CAO J W,YANG Z Y,et al. Leaf and flower extracts of six *Michelia* L.: polyphenolic composition,antioxidant,antibacterial activities and in vitro inhibition of α-amylase and α-glucosidase [J]. Chemistry & Biodiversity,2022,19(3):e202100894.
[16] 宋万志,崔建芳,章观德. 木莲属土厚朴的研究[J]. 药学学报,1989,24(4):295-299.
[17] HE K Y,ZHANG S Q,LI X C,et al. Chemical composition and free radicals restraining activity of extracts from three *Manglietia* species leaves [J]. Journal of Forestry Research,2007,18(3):193-198.
[18] CHEN S X,WEI B C,FU Y. A study of the chemical composition and biological activity of *Michelia macclurei* dandy heartwood: new sources of natural antioxidants,enzyme inhibitors and bacterial inhibitors [J]. International Journal of Molecular Sciences, 2023, 24(9):7972.
[19] 戚嘉敏,许逸林,张鹏,等. 3 种木兰科珍稀濒危树种的光合及固碳特性[J]. 华南农业大学学报,2018,39(3):90-95.
[20] 陆湘云,刘奎,陈凯,等. 3 个木兰科树种对杉木人工林土壤肥力改良效果的初步评价[J]. 西部林业科学,2020,49(3):29-35.
[21] 杨世火,刘从达. 木兰科野生植物资源现状及开发利用初探[J]. 安徽农学通报,2009,15(12):20-21.
[22] XIA H,YANG L C,TU Z H,et al. Growth performance and G×E interactions of *Liriodendron tulipifera* half-sib families across ages in eastern China [J]. European Journal of Forest Research,2022,141(6):1089-1103.
[23] 杨梅. 植物及植物文化:以木兰科植物为例[J]. 现代园艺,2017(21):87-88.
[24] 缪绅裕,罗宇谦,蓝扬辉,等. 国家保护植物厚叶木莲资源调查研究[J]. 亚热带植物科学,2020,49(3):243-246.
[25] 曾庆文,周仁章,刘银至,等. 濒危植物厚叶木莲的群落学特征及其保护[J]. 热带亚热带植物学报,1999,7(2):109-119.
[26] 杨晓丽,邢福武,陈树钢,等. 广东省南昆山自然保护区厚叶木莲的群落特征研究[J]. 热带亚热带植物学报,2013,21(4):356-364.
[27] 徐凤霞,胡晓颖,徐信兰. 木莲属(木兰科)5 种植物的花粉形态[J]. 热带亚热带植物学报,2004,12(4):313-317.
[28] 孙谷畴,赵平,曾小平,等. 不同光强下生长的厚叶木莲(*Magnolietia pachyphylla*) 光合作用光响应的变化[J]. 应用与环境生物学报,2001,7(3):213-218.
[29] DONG S S,LIU M,LIU Y,et al. The genome of *Magnolia biondii* Pamp. provides insights into the evolution of Magnoliales and biosynthesis of terpenoids [J]. Horticulture Research,2021,8(1):38.
[30] WANG Z F,ROUARD M,DROC G,et al. Genome assembly of *Musa beccarii* shows extensive chromosomal rearrangements and genome expansion during evolution of Musaceae genomes [J]. GigaScience, 2022, 12: giad005.
[31] DLUGOSZ M,DEOROWICZ S. RECKONER: read error corrector based on KMC [J]. Bioinformatics,2017,33(7):1086-1089.
[32] VURTURE G W,SEDLAZECK F J,NATTESTAD M,et al. GenomeScope: fast reference-free genome profiling from short reads [J]. Bioinformatics, 2017, 33(14):2202-2204.
[33] GUAN D F,MCCARTHY S A,WOOD J,et al. Identifying and removing haplotypic duplication in primary genome assemblies [J]. Bioinformatics, 2020, 36(9): 2896-2898.
[34] VASER R,SOVIĆ I,NAGARAJAN N,et al. Fast and accurate de novo genome assembly from long uncorrected reads [J]. Genome Research,2017,27(5):737-746.
[35] AURY J M,ISTACE B. Hapo-G,haplotype-aware polishing of genome assemblies with accurate reads [J]. NAR Genomics and Bioinformatics, 2021, 3(2): lqab034.
[36] WICK R R,HOLT K E. Polypolish: short-read polishing of long-read bacterial genome assemblies [J]. PLoS Computational Biology,2022,18(1):e1009802.
[37] MANNI M,BERKELEY M R,SEPPEY M,et al. BUSCO: assessing genomic data quality and beyond [J]. Current Protocols,2021,1(12):e323.
[38] OU S J,SU W J,LIAO Y,et al. Benchmarking transposable element annotation methods for creation of a streamlined,comprehensive pipeline [J]. Genome Biology,2019,20(1):275.

[39] GIRGIS H Z. Red:an intelligent,rapid,accurate tool for detecting repeats de-novo on the genomic scale [J]. BMC Bioinformatics,2015,16:227.

[40] QUINLAN A R,HALL I M. BEDTools:a flexible suite of utilities for comparing genomic features [J]. Bioinformatics,2010,26(6):841-842.

[41] BRŮNA T,HOFF K J,LOMSADZE A,et al. BRAKER2:automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database [J]. NAR Genomics and Bioinformatics,2021,3(1):lqaa108.

[42] ZHANG H,YOHE T,HUANG L,et al. dbCAN2:a meta server for automated carbohydrate-active enzyme annotation [J]. Nucleic Acids Research,2018,46(W1):W95-W101.

[43] HUERTA-CEPAS J,FORSLUND K,COELHO L P,et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper [J]. Molecular Biology and Evolution,2017,34(8):2115-2122.

[44] The Gene Ontology Consortium. The gene ontology resource:20 years and still going strong [J]. Nucleic Acids Research,2019,47(D1):D330-D338.

[45] ASHBURNER M,BALL C A,BLAKE J A,et al. Gene ontology:tool for the unification of biology. The Gene Ontology Consortium [J]. Nature Genetics,2000,25(1):25-29.

[46] KANEHISA M,SATO Y,KAWASHIMA M,et al. KEGG as a reference resource for gene and protein annotation [J]. Nucleic Acids Research,2016,44(D1):D457-D462.

[47] MITCHELL A L,ATTWOOD T K,BABBITT P C,et al. InterPro in 2019:improving coverage,classification and access to protein sequence annotations [J]. Nucleic Acids Research,2019,47(D1):D351-D360.

[48] RAWLINGS N D,BARRETT A J,THOMAS P D,et al. The MEROPS database of proteolytic enzymes,their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database [J]. Nucleic Acids Research,2018,46(D1):D624-D632.

[49] EL-GEBALI S,MISTRY J,BATEMAN A,et al. The Pfam protein families database in 2019 [J]. Nucleic Acids Research,2019,47(D1):D427-D432.

[50] ARMENTEROS J J A,TSIRIGOS K D,SØNDERBY C K,et al. SignalP 5.0 improves signal peptide predictions using deep neural networks [J]. Nature Biotechnology,2019,37(4):420-423.

[51] EMMS D M,KELLY S. OrthoFinder:solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy [J]. Genome Biology,2015,16(1):157.

[52] EMMS D M,KELLY S. OrthoFinder:phylogenetic orthology inference for comparative genomics [J]. Genome Biology,2019,20(1):238.

[53] DOS REIS M,ZHU T Q,YANG Z H. The impact of the rate prior on Bayesian estimation of divergence times with multiple loci [J]. Systematic Biology,2014,63(4):555-565.

[54] HAN M V,THOMAS G W C,LUGO-MARTINEZ J,et al. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3 [J]. Molecular Biology and Evolution,2013,30(8):1987-1997.

[55] CHEN C J,CHEN H,ZHANG Y,et al. TBtools:an integrative toolkit developed for interactive analyses of big biological data [J]. Molecular Plant,2020,13(8):1194-1202.

[56] ZWAENEPOEL A,VAN DE PEER Y. Wgd-simple command line tools for the analysis of ancient whole-genome duplications [J]. Bioinformatics,2019,35(12):2153-2155.

[57] QIAO X,LI Q H,YIN H,et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants [J]. Genome Biology,2019,20(1):38.

[58] ZHOU L J,HOU F X,WANG L,et al. The genome of *Magnolia hypoleuca* provides a new insight into cold tolerance and the evolutionary position of magnoliids [J]. Frontiers in Plant Science,2023,14:1108701.

[59] YIN Y P,PENG F,ZHOU L J,et al. The chromosome-scale genome of *Magnolia officinalis* provides insight into the evolutionary position of magnoliids [J]. iScience,2021,24(9):102997.

[60] CHEN J H,HAO Z D,GUANG X M,et al. Liriodendron genome sheds light on angiosperm phylogeny and species-pair differentiation [J]. Nature Plants,2019,5(1):18-25.

[61] PORTER K,DAY B. From filaments to function:the role of the plant actin cytoskeleton in pathogen perception,signaling and immunity [J]. Journal of Integrative Plant Biology,2016,58(4):299-311.

[62] QIN L,LIU L J,TU J Y,et al. The ARP2/3 complex,acting cooperatively with Class Ⅰ formins,modulates penetration resistance in *Arabidopsis* against powdery mildew invasion [J]. The Plant Cell,2021,33(9):3151-3175.

[63] VENKAT A,MUNEER S. Role of circadian rhythms in major plant metabolic and signaling pathways [J]. Frontiers in Plant Science,2022,13:836244.

[64] STRZYZ P. Alternative cycle for citrate [J]. Nature Reviews Molecular Cell Biology,2022,23(5):305.

[65] BISCHOFF F,MOLENDIJK A,RAJENDRAKUMAR C S V,et al. GTP-binding proteins in plants [J]. Cellular and Molecular Life Sciences CMLS,1999,55:233-256.

[66] BENDER K W,ZIPFEL C. Plant G-protein activation:connecting to plant receptor kinases [J]. Cell Research,2018,28(7):697-698.

Draft Genome of *Manglietia pachyphylla*

GAN Xinjun[1],BIN Yue[2,3],CHEN Huanjin[1],ZHU Weiguang[2,3],XIONG Luqiao[1],
YU Enping[2,3,4],WANG Zhengfeng[2,3**],XU Fengxia[3,5],CAO Honglin[2,3]

(1. Administrative Office of Guangdong Conghua Chenhedong Provincial Nature Reserve,Guangzhou,Guangdong,510950,China;
2. Guangdong Provincial Key Laboratory of Applied Botany,Key Laboratory of Vegetation Restoration and Management of Degraded Ecosystems,South China Botanical Garden,Chinese Academy of Sciences,Guangzhou,Guangdong,510650,China;3. South China National Botanical Garden,Guangzhou,Guangdong,510650,China;4. University of Chinese Academy of Sciences,Beijing,100049,China;5. Key Laboratory of Plant Resource Conservation and Sustainable Utilization,South China Botanical Garden,Chinese Academy of Sciences,Guangzhou,Guangdong,510650,China)

Abstract: *Manglietia pachyphylla* is in the family of Magnoliaceae,one of the earliest divergent lineages of angiosperms. Magnoliaceae are valued for their large and fragrant colorful flowers,and medicine and timber sources. *Manglietia pachyphylla* is endemic to China,and only distributes in Guangdong Province and Guangxi Zhuang Autonomous Region. It is now national class Ⅱ key protected wild species. In this study,we sequenced the genome of *Manglietia pachyphylla* by high-throughput sequencing,and the sequencing data were used to assemble a sketch of the genome,based on which,we predict its repeat sequences and genes,and perform phylogentic and gene familes analyses. The assembly size of *Manglietia pachyphylla* was 2 092 298 891 bp,comprised 676 contigs with N50 (the contig length,included in the sum length of contigs from the longest to the shortest,covers ~50% of the total genome length) of 7 961 115 bp. Benchmarking Universal Single-Copy Orthologs (BUSCO) assessment revealed 98.8% and 96.6% completeness of the assembly using "embryophyta" and "eudicots" BUSCO database,respectively. Up to 76.5% of the genome is repetitive,with 37 900 genes encoding 41 675 proteins. Phylogenetic analyses of relative 10 species indicated that *Manglietia pachyphylla* was clustered in the same clade with *Magnolia biondii*,and divergence time between them were estimated to be 10.5 million years ago. We investigated genes related to phloem/xylem,actin filament,heat,photosynthesis and various plant secondary metabolites were significantly expanded,in which the secondary metabolites were mainly derived from tandem and proximal duplications. The gene expansions and duplications should play important roles in the adaption of *Manglietia pachyphylla* to the high altitude. This study is the first genomic report of the genus *Manglietia* at home and abroad,which provides genetic information and reference for better conservation and development of the germplasm resources of *Manglietia pachyphylla* and other species in the family of Magnoliaceae.

Key words: *Manglietia pachyphylla*; Magnoliaceae; *Manglietia*; endangered plant; genome assembly; gene prediction; gene family; gene duplication

责任编辑:米慧芝