基于 Illumina 高通量测序技术的草豆蔻基因组研究

郑 燕[1]，何志凯[1]，姚梦鹃[1]，郭云端[1]，陈曦荣[1]，涂梦薇[2]，张勇洪[1]，李 琛[1, 3*]

1. 湖北医药学院基础医学院 药用植物实验室，湖北 十堰 442000
2. 湖北医药学院第一临床学院，湖北 十堰 442000
3. 湖北医药学院生物医药研究院 武当特色中药研究湖北省重点实验室，湖北 十堰 442000

摘 要：目的 对药食同源植物草豆蔻 *Alpinia katsumadai* 进行基因组调研分析，完善草豆蔻基因组遗传信息。**方法** 研究采用 Illumina 高通量测序技术，基于 K-mer 分析手段来研究草豆蔻基因组的大小及杂合度，利用 MISA 方法分析可能的 SSR 分子标记。**结果** 草豆蔻的基因组大小为 1.60 Gb，基因组杂合度为 0.44%，重复序列比例为 72.72%；在基因组序列中，进行 SSR 分子标记分析，共鉴定出了 364 395 个 SSR，其中，单、双、三核苷酸重复模体的比例较高，分别占比 64.25%、24.05%、10.31%；从测序得到的 350 bp 文库中，随机取 10 000 条单端 reads，与 NT 库进行 BLAST 比对，发现草豆蔻的近缘物种艳山姜 *Alpinia zerumbet* 和小豆蔻 *Elettaria cardamomum* 上的 reads 数分别占比对上 NT 库 reads 数的 12.89%和 12.36%。**结论** 通过对草豆蔻植物的基因组大小、杂合度调查及 SSR 分子标记分析表明，草豆蔻物种基因组属于高重复、大基因组的复杂基因组，这为草豆蔻的资源保护和遗传多样性分析及品种选育提供了遗传信息支撑。

关键词：高通量测序；草豆蔻；基因组大小；杂合度；简单重复序列

中图分类号：R282.12 **文献标志码：**A **文章编号：**0253 - 2670(2020)13 - 3530 - 05

DOI: 10.7501/j.issn.0253-2670.2020.13.022

Genome survey study of *Alpinia katsumadai* based on Illumina high throughput sequencing

ZHENG Yan[1], HE Zhi-kai[1], YAO Meng-juan[1], GUO Yun-duan[1], CHEN Xi-rong[1], TU Meng-wei[2], ZHANG Yong-hong[1], LI Chen[1, 3]

1. Laboratory of Medicinal Plant, Institute of Basic Medical Sciences, Hubei University of Medicine, Shiyan 442000, China
2. The First Clinical College, Hubei University of Medicine, Shiyan 442000, China
3. Hubei Key Laboratory of Wudang Local Chinese Medicine Research, Biomedical Research Institute, Hubei University of Medicine, Shiyan 442000, China

Abstract: Objective To analyze the genome survey of medicinal and edible plant *Alpinia katsumadai* and complete its genome genetic information. **Methods** This study was based on high throughput sequencing platform Illumina, and K-mer analysis was applied to estimate the genome size and heterozygosity rate of *A. katsumadai*. Meanwhile, simple sequence repeat (SSR) loci that were suitable as markers were identified by MISA software. **Results** The estimated genome size of *A. katsumadai* was 1.60 Gb, with a 0.44% heterozygosity rate and 72.72% repeats; In the genome sequence, 364 395 simple sequence repeats (SSRs) were detected by SSR molecular marker analysis, among which mono-nucleotide, di-nucleotide and tri-nucleotide repetitive motifs ranked the higher percentages of 64.25%, 24.05% and 10.31%, summed up to 98.61%; From the 350 bp library obtained by sequencing, 10 000 single-end reads were randomly selected and blasted with NT bank, the results showed that its genetically close species *Alpinia zerumbet* and *Elettaria cardamomum* were blasted with the reads of 12.89% and 12.36% in NT bank. **Conclusion** The genome size, heterozygosity rate and SSR molecular marker analysis' genome survey study on *A. katsumadai* indicated that the genome of *A.*

收稿日期：2019-12-03

基金项目：国家自然科学基金资助项目（31701294）；湖北医药学院高层次人才启金资助项目（2017QDJZR26）；湖北省卫生健康委员会项目（ZY2019Q004）；药用资源与天然药物化学教育部重点实验室开放基金资助项目（2019005）；湖北重点实验室建设基金资助项目（WLSP201905）

作者简介：郑 燕（1993—），女，湖北十堰人，硕士在读，研究方向为中药资源及药用植物。Tel: 18772954983 E-mail: 573641723@qq.com

***通信作者** 李 琛（1984—），男，湖北黄石人，博士，副教授，主要从事中药资源及药用植物方面研究。Tel: 15897849905 E-mail: 172723301@qq.com

katsumadai species was a complex, highly repetitive and large genome, which provided genetic information support for the resource protection, genetic diversity analysis and variety breeding of *A. katsumadai*.

Key words: high throughput sequencing; *Alpinia katsumadai* Hayata; genome size; heterozygosity rate; simple sequence repeat

草豆蔻 *Alpinia katsumadai* Hayata 为姜科山姜属植物，又名草蔻、豆蔻等，主要分布于我国海南、广东、广西、云南等地[1]。草豆蔻以干燥近成熟种子入药，为传统的药食同源常用中药，收载于《中国药典》2015 年版，性温、味辛，归脾、胃经，具有燥湿行气、温中止呕功效，主治寒湿内阻、脘腹胀满冷痛、嗳气呕逆、不思饮食等症[2]。此外，草豆蔻茎杆麻皮可制作保健产品和工艺品。草豆蔻的主要药效成分是挥发油类、黄酮类、二苯庚烷类[3-5]，现代药理学研究表明，草豆蔻有抗胃溃疡、保护胃黏膜[6]、抗氧化[7]、抗菌[8]、抗肿瘤[9]等多种药理活性。

本草基因组是利用基因组学的方法和手段来研究传统中草药的遗传信息及调控机制，从基因组水平来阐明中药的作用机制及分子育种等方面的新兴学科。本草基因组学主要涉及中药的基因组、转录组、蛋白质组、代谢组等理论与技术[10-13]。目前，已有灵芝、丹参、铁皮石斛等重要中草药的基因组被解析[14-16]。利用本草基因组学的研究手段来评价草豆蔻的基因组大小及复杂程度，有助于后续草豆蔻全基因组的测序策略。基于基因组调研开发的分子标记即简单重复序列（simple sequence repeat，SSR），具有高度的多态性，有助于作为分子标记应用于遗传图谱及种质资源鉴定的研究[17]。

本研究采用 Illumina 二代测序技术，针对草豆蔻进行了基因组调研，分析了其基因组的 K-mer 分布曲线，GC 含量比例以及基于基因组调研的 SSR 分析，完善草豆蔻基因组遗传信息，为草豆蔻的资源保护和遗传多样性分析及品种选育提供了遗传信息支撑。

1 材料与仪器

1.1 材料

草豆蔻取自中国热带农业科学院热带作物品种资源研究所苗圃，由中国热带农业科学院热带作物品种资源研究所于福来副研究员鉴定为为姜科山姜属草豆蔻 *Alpinia katsumadai* Hayata。选取 1 株草豆蔻植物的幼嫩叶片提取 DNA 进行基因组 Illumina 二代测序。

1.2 仪器与试剂

Molecular Imager ChemiDOC XRS＋凝胶成像仪（美国 BIO-RAD 公司）；Eppendorff 高速冷冻离心机（德国 Eppendorf 公司）；PowerPac™ Basic 电泳仪（美国 BIO-RAD 公司）；DYCP-31E 型电泳槽（北京六一仪器厂）；Nano Drop 2000（美国 Thermo Fisher 公司）；植物基因组 DNA 快速提取试剂盒（上海生工公司）。

2 方法

实验流程按照 Illumina 公司提供的标准流程执行，包括 DNA 文库制备实验和测序实验、基因组调研图信息分析流程，见图 1。

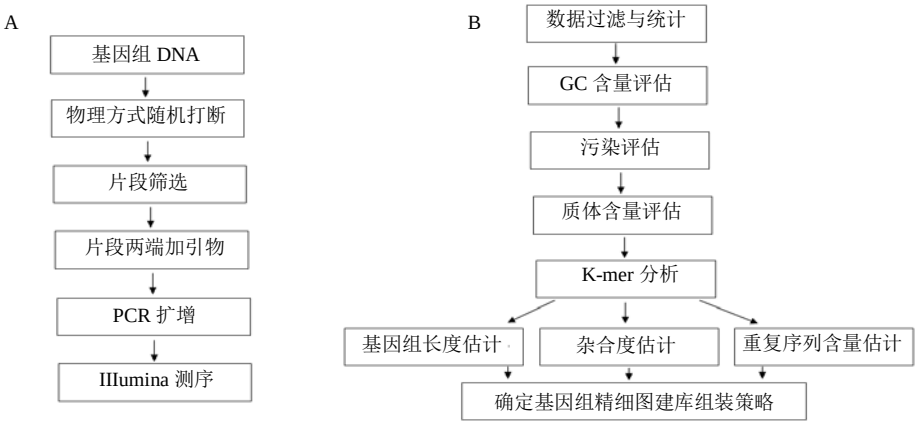


图 1 实验流程 (A) 和基因组调研信息分析流程 (B)

Fig. 1 Experiment procedures (A) and genome survey procedures (B)

2.1　基因组 DNA 的提取

采用 CTAB 法进行基因组 DNA 的提取，应用 Nano Drop 2000 进行检测，测得 $A_{260\ nm}/A_{280nm}$≥1.8。将提取的 DNA 样品送至百迈客公司，构建插入片段为 350 bp 文库 1 个，使用 Illumina HiSeq 进行双端测序。

2.2　测序数据质控

采用 FastQC（http://www.bioinformatics.babraham.ac.uk/projects/fastqc）对数据进行质量控制。采用 AdapterRemoval（version 2.1.7）去除 3’端的接头污染[18]。得到的高质量的数据用于后续基因组大小、杂合度、GC 含量及 SSR 分析。

2.3　21-mer 分析及基因组大小估计

将测序得到的 reads 经过滤后，采用基于 K-mer 的方法进行草豆蔻基因组大小和杂合率的预测，进行 21-mer 分析。对这些序列进行频率做图，获得 K-mer 分布曲线。K-mer 深度分布曲线属于标准的泊松分布曲线。采用百迈客自主研发的软件“kmer_freq_stat”计算得到每一个 K-mer 的深度分布图，并计算出杂合率[19]。

2.4　SSR 分析

采用微卫星识别工具（microsatellite identification tool，MISA，http://pgrc.ipk-gatersleben.de/misa/）在所有序列中搜索 SSR 位点，搜索所采用的参数：mono-10、di-6、tri-5、Tetra-5、penta-5、hexa-5，复合序列中 2 个不同 SSR 之间允许的最大间隔设置为 100 bp。

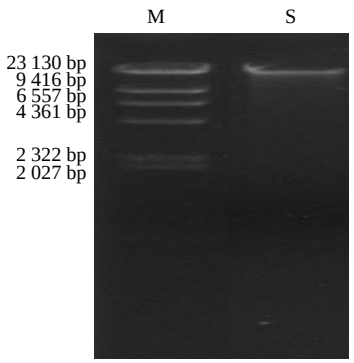
3　结果与分析

3.1　草豆蔻基因组 DNA 的提取

选取中国热带农业科学院热带作物品种资源研究所苗圃内草豆蔻的幼嫩叶片，采用 CTAB 法提取其基因组 DNA。电泳图表明所提取的基因组 DNA 质量良好，见图 2。

3.2　草豆蔻测序数据和拼接

利用 Illumina Hiseq 高通量测序技术测序获得原始数据后，采用 FastQC 来过滤去除低质量数据，从单碱基质量分布图、Base content 分布图、GC 含量分布图、Sequence base quality 分布图来分析测序的质量。结果表明测序质量很好，GC 含量的分布正常，见图 3。数据过滤后，获得 54.22 Gb 的高质量数据，总测序深度约为 33.97X，Q20 比例达到 97.94%以上，Q30 比例达到 93.89%以上，基因组的 GC 含量约 39.68%。



M-λ-Hind III digest DNA Marker　S-提取的草豆蔻基因组 DNA
M-λ-Hind III digest DNA Marker　S-genomic DNA from sample extract

图 2　草豆蔻基因组 DNA 电泳图

Fig. 2　Genomic DNA electrophoresis of *A. katsumadai*

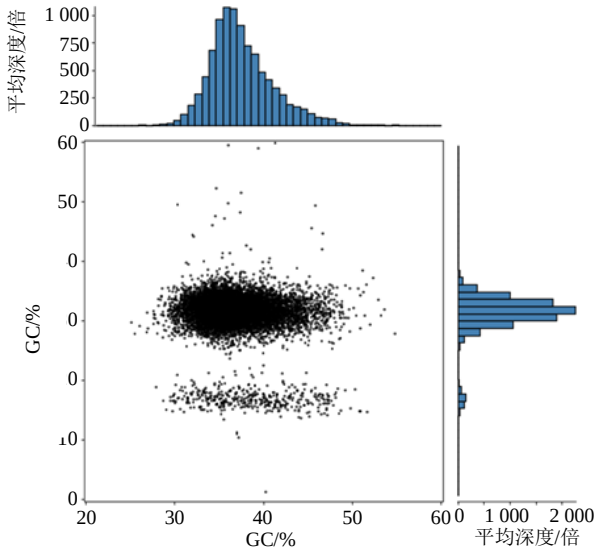


图 3　GC 含量和平均测序深度分布图

Fig. 3　GC content and average sequencing depth distribution

3.3　草豆蔻基因组大小及杂合度估计

使用百迈客自主研发的软件“kmer_freq_stat”进行 54.22 Gb 数据的 21-mer 分析，见图 4。横坐标表示 21-mer 深度，纵坐标表示出现的频率。根据该软件[19]计算估计出草豆蔻基因组大小为 1.60 Gb，杂合度为 0.44%，重复序列比例为 72.72%。结果表明草豆蔻基因组属于高重复、大基因组的复杂基因组。

3.4　草豆蔻基因组 SSR 分析

采用微卫星识别工具 MISA 在所有序列中搜索 SSR 位点，总共搜索到 364 395 个 SSR；在所有具

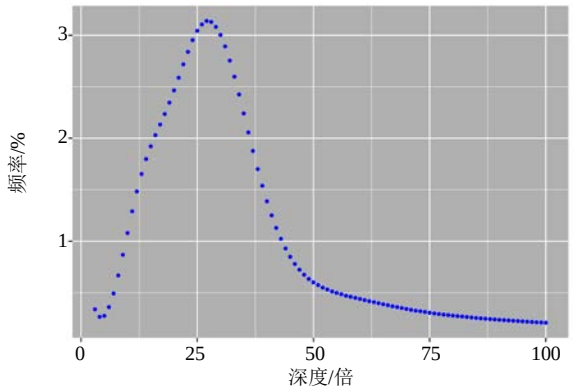


图 4 21-mer 深度分布图

Fig. 4 21-mer depth distribution

有 SSR 的序列中，49 128 条序列包含 1 个以上 SSR；以复合形式存在的 SSR 数量为 28 825 个。分别对不同类型的 SSR 模体进行统计，单核苷酸重复模体、二核苷酸重复模体、三核苷酸重复模体、四核苷酸重复模体、五核苷酸重复模体和六核苷酸重复模体分别有 234 121、87 655、37 582、3 529、883、625 个，其分别占总重复模体数的 64.25%、24.05%、10.31%、0.97%、0.24%、0.17%，见图 5。随后，进一步对每一种 SSR 重复模体按照序列组成进行细分，单、双、三核苷酸重复模体中 A/T、AT/AT、和 AAT/ATT 的含量最高，见表 1。

3.5 草豆蔻物种比对分析

为了调查草豆蔻物种的多样性，我们从测序得到的 350 bp 文库中，随机取 10 000 条单端 reads，与 NT 库进行 BLAST 比对。BLAST 使用 ncbi-blast+ 2.2.29 版本，参数设置为 -num_descriptions 100-num_alignments 100-evalue 1e-05。能够比对上 NT 库的 reads 占提取 reads 数的 9.46%，其中比对到艳山姜 *Alpinia zerumbet* (Pers.) Burtt. et Smith 和小豆蔻 *Elettaria cardamomum* Maton 上的 reads 数分

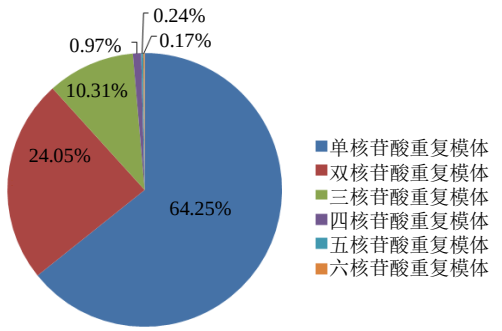


图 5 不同类型的 SSR 模体分布图

Fig. 5 Distribution of SSR based on repeat type

表 1 单、双、三核苷酸重复模体结果统计

Table 1 Summary for repeat type in mononucleotide repeat motifs, dinucleotide repeat motifs and trinucleotide repeat motifs

重复类型	数目
单核苷酸重复模体	234 121
A/T	213 620
C/G	20 501
双核苷酸重复模体	40 877
AC/GT	3 605
AG/CT	10 683
AT/AT	26 510
CG/CG	79
三核苷酸重复模体	14 072
AAC/GTT	835
AAG/CTT	4 280
AAT/ATT	6 041
ACC/GGT	143
ACG/CGT	130
ACT/AGT	52
AGC/CTG	413
AGG/CCT	1 215
ATC/ATG	444
CCG/CGG	519

别占比对上 NT 库 reads 数的 12.89%和 12.36%，这 2 个物种皆为草豆蔻的近缘物种，且比对结果中未发现动物等异常比对，具体比对统计见表 2。

表 2 350 bp 文库 NT 库比对详表

Table 2 Blast of 350 bp library with NT bank

物种	匹配序列比例/%
艳山姜 *Alpinia zerumbet*	12.89
小豆蔻 *Elettaria cardamomum*	12.36
小果野蕉 *Musa acuminata*	4.43
甜瓜 *Cucumis melo*	4.01
黄花姜黄 *Curcuma flaviflora*	3.80
胡萝卜 *Daucus carota*	3.06
野生种番茄 *Solanum pennellii*	2.95
可可 *Theobroma cacao*	2.85
栽培种番茄 *Solanum lycopersicum*	2.43
姜 *Zingiber officinale*	2.43
葡萄 *Vitis vinifera*	2.32
小麦 *Triticum aestivum*	2.32
海枣 *Phoenix dactylifera*	1.58
赤豆 *Vigna angularis*	1.37
观音姜 *Curcuma roscoeana*	1.26
蒺藜苜蓿 *Medicago truncatula*	1.16
距花山姜 *Alpinia calcarata*	1.16
陆地棉 *Gossypium hirsutum*	1.05
大豆 *Glycine max*	1.05
雷蒙德氏棉 *Gossypium raimondii*	1.05
其他	33.66

4 讨论

中药是我国传统医学的宝库，其中以人参、灵芝、何首乌、枸杞等最为著名。新中国成立以来，国家加大了中医药的研究，进一步促进了中药的现代化，其中，以屠呦呦为代表的中医药研究者从复合花序植物黄花蒿茎叶中提取得到青蒿素，这种药品可以有效降低疟疾患者的死亡率，屠呦呦也因为发现青蒿素获得了诺贝尔医学奖，可以说植物药为我国及世界医药学发展作出了巨大贡献[20]。但是，这些中草药由于其遗传信息如基因组的缺乏，导致现代生命科学特别是分子生物学的前沿技术很难应用到中药的研究。基于此，陈士林等[9-12]提出了本草基因组学的概念，使得中药学的研究深入到分子水平的基因组学相关研究。高通量二代测序技术的发展能够让人们分析所关注中药的基因组信息，是当前基因组评估的重要手段。

目前，针对二代测序的结果，主要采用 K-mer 分析的方法对基因组进行预估，并能获取如物种杂合率，GC 含量，重复序列比例等相关信息。基于基因组调研开发的 SSR 具有高水平的多态性，是遗传研究中可靠的分子标记。本研究采用 Illumina 二代测序，利用生物信息学的方法，首次对传统南药草豆蔻的基因组大小、杂合度及 SSR 进行了分析。预测草豆蔻的基因组大小为 1.60 Gb，基因组杂合度为 0.44%，重复序列比例为 72.72%，结果表明草豆蔻属于高重复、大基因组的复杂基因组。同时，本研究开发出一系列 SSR 中，单、双、三核苷酸重复模体占主导地位，这些 SSR 标记的进一步开发有助于研究草豆蔻的物种进化、遗传多样性、种质资源鉴定等方面。为了调查草豆蔻物种的多样性，本课题组从测序得到的 350 bp 文库中，随机取 10 000 条单端 reads，与 NT 库进行 BLAST 比对，其中比对到艳山姜和小豆蔻上的 reads 数分别占比对上 NT 库 reads 数的 12.89%和 12.36%，这 2 个物种皆为草豆蔻的近缘物种。因此，以 Illumina 高通量技术为基础的草豆蔻基因组调研，完善了草豆蔻基因组遗传信息，为阐明草豆蔻道地性形成和维持的遗传机制以及道地药材的资源保护和品种选育提供了信息支撑。

参考文献

[1] 肖培根. 新编中药志 [M]. 北京: 化学工业出版社, 2002.

[2] 中国药典 [S]. 一部. 2015.

[3] 晏小霞, 王茂媛, 王祝年, 等. 草豆蔻不同部位挥发油化学成分 GC-MS 分析 [J]. 热带作物学报, 2013, 34(7): 1389-1394.

[4] 谢 鹏, 秦华珍, 谭喜梅, 等. 草豆蔻化学成分和药理作用研究进展 [J]. 辽宁中医药大学学报, 2017, 19(3): 60-63.

[5] 李晓鹏, 孙爱玲, 柳仁民, 等. 高速逆流色谱分离纯化草豆蔻中山姜素和小豆蔻明 [J]. 中草药, 2011, 42(4): 687-690.

[6] 吴 珍, 陈永顺, 杜士明, 等. 草豆蔻挥发油对大鼠醋酸性胃溃疡的影响 [J]. 中国医院药学杂志, 2010, 30(7): 560-563.

[7] 吴 珍, 陈永顺, 王启斌. 草豆蔻总黄酮抗氧化活性研究 [J]. 医药导报, 2011, 30(11): 1406-1409.

[8] 黄文哲, 戴小军, 刘延庆, 等. 草豆蔻中黄酮和双苯庚酮的抑菌活性 [J]. 植物资源与环境学报, 2006, (1): 37-40.

[9] 王 萍, 石海莲, 吴晓俊. 中药草豆蔻抗肿瘤化学成分和作用机制研究进展 [J]. 中国药理学与毒理学杂志, 2017, 31(9): 880-888.

[10] 尉广飞, 董林林, 陈士林, 等. 本草基因组学在中药材新品种选育中的应用 [J]. 中国实验方剂学杂志, 2018, 24(23): 18-28.

[11] 陈士林, 宋经元. 本草基因组学 [J]. 中国中药杂志, 2016, 41(21): 3881-3889.

[12] 陈士林, 孙永珍, 徐 江, 等. 本草基因组计划研究策略 [J]. 药学学报, 2010, 45(7): 807-812.

[13] 陈士林, 何 柳, 刘明珠, 等. 本草基因组方法学研究 [J]. 世界科学技术—中医药现代化, 2010, 12(3): 316-324.

[14] Chen S, Xu J, Liu C, *et al*. Genome sequence of the model medicinal mushroom *Ganoderma lucidum* [J]. *Nat Comm*, 2012, 3(2): 913-919.

[15] Xu H, Song J, Luo H, *et al*. Analysis of the genome sequence of the medicinal plant *Salvia miltiorrhiza* [J]. *Mol Plant*, 2016, 9(6): 949-952.

[16] Liang Y, Xiao W, Hui L, *et al*. The genome of *Dendrobium officinale* Illuminates the biology of the important traditional Chinese orchid herb [J]. *MolPlant*, 2015, 8(6): 922-934.

[17] Gábor T, Zoltán G, Jerzy J. Microsatellites in different eukaryotic genomes: Survey and analysis [J]. *Genom Res*, 2000, 10(7): 967-972.

[18] Lindgreen S. Adapter Removal: Easy cleaning of next-generation sequencing reads [J]. *BMC Res Notes*, 2012, 5(1): 337-342.

[19] Guillaume M, Carl K. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers [J]. *Bioinformatics*, 2011, 27(6): 764.

[20] 张铁军, 王于方, 刘 丹, 等. 天然药物化学史话: 青蒿素——中药研究的丰碑 [J]. 中草药, 2016, 47(19): 3351-3361.