植物遗传资源学报 2014,15(1):149-152
Journal of Plant Genetic Resources
DOI:10.13430/j.cnki.jpgr.2014.01.020

樟树全基因组调查

伍艳芳,肖复明,徐海宁,章　挺,江香梅
(江西省林业科学院/国家林业局樟树工程技术研究中心,南昌 330032)

摘要:樟树是我国特有的珍贵用材和经济树种,富含多糖多酚、萜类等次生代谢物质,是香精香料、油脂化工和医药等的重要原料树种。本研究采用高通量测序技术(Illumina Hiseq™ 2000)首次测定了樟树基因组大小,并利用生物信息方法估计樟树杂合率、重复序列情况和 GC 含量等基因组信息,为全基因组测序策略的选择提供依据。主要结论如下:(1)樟树基因组大小粗略估计为 760 Mb 左右;(2)樟树基因组有较高的杂合率和一定的重复,杂合率约为 0.65%;(3)由于樟树杂合率较高,全基因组鸟枪法策略不适合该基因组测序分析,可尝试使用 BAC-to-BAC 策略或 fosmid 策略,有利于樟树基因组的序列拼接和组装。

关键词:基因组大小;樟树;杂合率;GC 含量

Genome Survey in *Cinnamomum camphora* L. Presl

WU Yan-fang, XIAO Fu-ming, XU Hai-ning, ZHANG Ting, JIANG Xiang-mei
(*Jiangxi Academy of Forestry/Comphor Engineering Technology Research Center for State Forestry Administration, Nanchang 330032*)

Abstract: *Cinnamomum camphora* L. Presl is a kind of precious wood species and the main no-timber trees of special local product in our country and possess abundant the secondary metabolism, such as polysaccharides, polyphenols, oleicacid, and terpenoid. *C. camphora* is a metallic element used in many important industries. In this study, the genome size of *C. camphora* was determined by new-generation sequencing technologies (Illumina Hiseq™ 2000). Hybridity percentage, repeats, and GC depth were also estimated by bioinformatics. The main results were showed that the average genome size and hybridity percentage of *C. camphora* was about 760 Mb and 0.65%, respectively. Moreover, whole-genome shotgun sequencing should not be used to *C. camphora* genome sequencing, and the BAC-to-BAC or fosmid could be used.

Key words: genome size; *Cinnamomum camphora*; hybridity percentage; GC depth

樟树(*Cinnamomum camphora* L. Presl)是我国特有的珍贵用材和经济树种,也是樟科樟属常绿乔木植物中经济价值最大的树种之一。其叶、树皮、枝茎、根系、花、果和种子等中皆含精油[1-2],是医药、食品、日用化工、香精香料和生物农药等天然物质的主要来源,具有较大的开发利用潜力。

樟树染色体数目和核型前人已有研究[3-4],陈成彬等[3]曾报道樟树染色体数目为 2n = 2x = 24,由 20 m + 4 sm 构成,按 Stebbins 核型分类标准[5],属于 2A 型。然而,有关樟树基因组大小和全基因组测序研究均未见报道。目前,常用的测定基因组大小的方法为 Feulgen 分光光度法[6]和流式细胞术(FC, flow cytometry)[7-8]。新一代测序技术(NGSTs, new-generation sequencing technologies)的迅猛发展加速了植物全基因组,尤其是木本植物全基因组研究的进程[9]。杨树[10]、葡萄[11]、番木瓜[12]、苹果[13]、桃树[14]和麻风树[15]等木本植物基因组草图的完成为人类进行其他木本植物的全基因组测序研究提供了大量的参考信息。但树木多数为异交物种,杂合性较强,基因组相对较大,而且较为复杂。因此,为减

收稿日期:2013-04-24　**修回日期**:2013-06-25　**网络出版日期**:2013-12-19
URL:http://www.cnki.net/kcms/detail/11.4996.S.20131219.1116.015.html
基金项目:"赣鄱英才 555 工程"领军人才培养计划项目
第一作者研究方向为林木遗传育种。E-mail:yanfangwu2012@163.com
通信作者:江香梅,研究方向为林木遗传育种。E-mail:zjiang2013@126.com

少盲目性,在大规模深度测序之前,可以先作低覆盖度的初测序,了解基因组的复杂程度,以确定该植物的测序研究策略和拼装技术[16]。本研究采用新一代高通量测序技术,对樟树的全基因组大小进行测定和评估,旨在为樟树全基因组测序方案的制定提供重要依据。樟树的全基因组测序,将为掌握和利用樟树潜在的基因资源,阐明樟树油脂和精油合成途径及其调控机制,从而利用分子生物学手段对樟树进行定向遗传改良奠定基础。

1 材料与方法

1.1 试验材料

试验材料为采自江西省林业科学院树木园中树龄约 30 年生的樟树。于 2012 年 4 月初选取生长状态良好的成年植株,剪取顶端新萌发的幼叶,液氮速冻后置-70℃低温冰箱保存备用。

1.2 样品提取及检测

采用 CTAB 法[17]提取樟树叶片基因组 DNA,Quant-iT™ dsDNA BR Assay Kit 试剂盒检测浓度,琼脂糖凝胶电泳检测完整性,检测参数为:胶浓度 0.5%,电压 3.5 V,电泳时间 960 min;以 1 kb DNA Extension Ladder (Invitrogen) 和 λ-Hind Ⅲ digest (Takara)作为 Marker。

1.3 测序

将提取的 DNA 样品送到深圳华大基因研究中心有限公司进行测序分析。首先将樟树的 DNA 样品进行随机打断,构建 170 bp、350 bp 和 500 bp 的小片段测序文库,然后采用 Illumina Hiseq™ 2000 测序平台进行双末端(Pair-End)测序,过滤掉低质量数据后,得到的数据用于基因组大小、杂合度和 GC 含量等信息的后续分析。参考其他木本植物的基因组大小,结合樟树的自身特性,按 600 Mb 左右的樟树基因组大小来估算测序覆盖度。

1.4 17-mer 分析以及基因组大小估计[18]

在基因组组装前,为了用测序所得的 reads 信息估计基因组特征,采用基于 K-mer 的分析方法来估计基因组大小和杂合率等,取 K 为 17 来进行分析。假设从 reads 中逐碱基取出的所有 K-mer 能够遍历整个基因组,且 K-mer 深度频率分布服从泊松分布,即可从所有测序 reads 中统计 K-mer 频数分布,计算获得 K-mer 深度分布曲线和深度乘积曲线。根据曲线获得 K-mer 深度估计值,用于估计基因组大小。同时,K-mer 分布图还被用来判断基因组的重复含量,如果这个基因组含有高比例的重复,那么其分布图将显示出粗的拖尾现象。

1.5 杂合率估计

采用模拟数据拟合的方式来进行基因组杂合率评估。选用大小为 116 Mb 的拟南芥基因组序列,随机生成 29X, 3464964747 bp reads, 读长错误率 0.0015%。因为二倍体复杂基因组进一步分为微杂合基因组(0.5% ≤杂合率 < 0.8%)、高杂合基因组(杂合率≥ 0.8%)以及高重复基因组(重复序列比例≥ 50%),针对不同类型的基因组,采用的测序方法及组装软件不一样,因此分别加入杂合率 0.5%、0.65% 和 0.8% 的模拟数据进行拟合,将所得到的模拟数据分别进行 17-mer 分析。

1.6 GC 含量及分布分析

利用高质量数据进行 SOAP denovo 组装[18-19],采用 K = 47 bp 构建 Contig 和 Scaffold,得到 Scaffold 序列后没有补洞,直接拼接组装获得原始基因组序列,这是一个最初的组装版本。用 soap 将过滤后的 reads 比对到该组装序列上,获得碱基深度。以 10 kb 为窗口,在序列上无重复前进,计算每个窗口的平均深度与 GC 含量,做出 GC_depth 点图。根据 GC_depth 分布图可以看出测序是否有明显的 GC 偏向,也可以判断是否存在细菌污染等情况。同时,可以根据 GC 聚成块的分层来判断基因组的杂合率和重复的分布情况。

2 结果与分析

2.1 测序数据量统计

采用高通量测序技术进行本次测序,过滤掉低质量数据后,得到的总测序量为 31.1 Gb 用于后续分析,若樟树基因组大小如预计的 600 Mb,那么测序覆盖度将为 44.5 X(表 1)。

表 1 数据量统计

Table 1 Data statistics

文库 ID Lib ID	读长(bp) Read length	插入片段大小(bp) Insert size	数据量 (Mb) Data	测序深度 (X) Sequence depth
CINverDABDBAAPE	100	170	14406.2	24.0
CINverDABDFAPE	100	350	8957.6	14.9
CINverDABDIAPEI-52	100	500	7764.7	12.9
合计 Total			31128.6	44.5

2.2 17-mer 分析以及基因组大小估计

使用樟树 30 Gb 的数据用于 17-mer 分析,其频率分布如图 1。横坐标表示 17-mer 出现的次数,纵

坐标表示出现的频率。从图中可以观察到,17-mer 分布曲线成峰情况较好。在 32 附近有一个峰值,即 K-mer 的期望深度。从表 2 可以知道 K-mer 的总数是 24 Gb,从而可以通过公式(基因组大小 = K-mer 的总数/K-mer 的期望深度)估算基因组大小为 759 Mb。从图 1 还可以观察到,在期望深度的 1/2 处有一个明显的凸峰。据此,可以判断樟树基因组具有较高杂合率的可能性。K-mer 曲线呈现明显拖尾,说明樟树基因组重复序列含量较高。

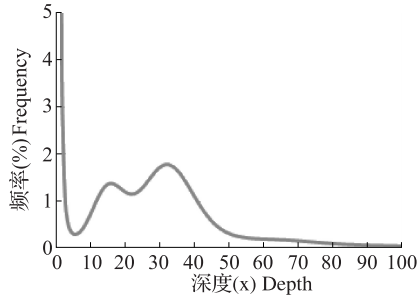

图 1　17-mer 分布曲线

Fig. 1　The distribution curve of 17-mer

表 2　17-mer 分析数据统计

Table 2　Data statistics and analysis of 17-mer

K-mer 总数 (Mb) K-mer sum	K-mer 深度 K-mer depth	基因组大小 (Mb) Genome size	所用碱基数 (Mb) Used base	所用读长 (Mb) Used read	测序深度 X
24614	32	759	30175	347	43.1

2.3　杂合率估计

由图 2 可看出,真实曲线(*C. camphora*)的主峰和杂合峰在杂合率 0.65% 时形成的峰最接近,可以大致地认为该物种的杂合率处于 0.65% 水平,说明樟树杂合率较高,可能影响组装效果。

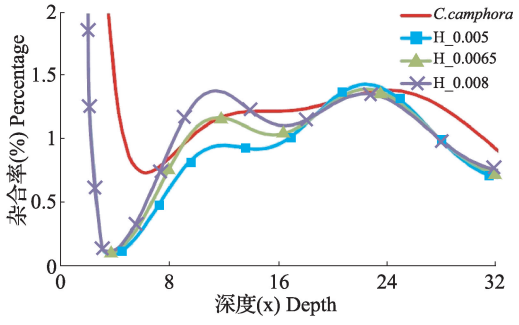

图 2　17-mer 杂合率估计图

Fig. 2　The statement of hybridity percentage of 17-mer

2.4　GC 含量及分布分析

GC_depth 分析显示(图 3),樟树样品无明显异常,测序无明显偏向,樟树 GC_depth 深度分布分成了 2 层,将较低的一层对应的序列提取出来和 nt 库进行比对,结果显示该样品没有细菌等其他基因组污染。GC 聚成的块分成了 2 层,推测是由杂合引起。因为杂合会使两条同源染色体杂合的部位只装出了 1 条,或 2 条都有装出,同时该部位以上的 read 乘数是整个基因组乘数的一半,导致 GC 含量图中出现较低的一层[18]。

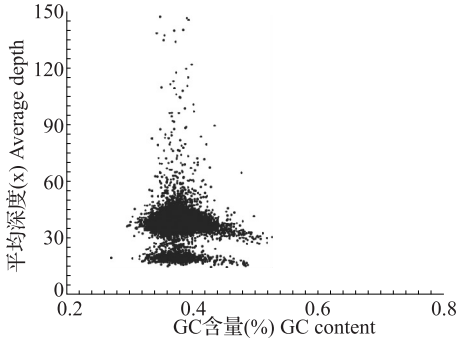

图 3　GC_depth 分布图

Fig. 3　The distribution figure of GC_depth

3　讨论

基因组大小又称基因组含量或 DNA 1C 值,是指物种配子染色体组所含 DNA 的量。基因组大小是比较和进化基因组学研究的基础,对不同物种基因组大小进行比较分析,可以掌握基因组大小变化规律。程蛟文等[20]为掌握蔬菜基因组含量变化规律,利用植物 DNA 1C 值数据库和相关文献收集整理了主要蔬菜作物的基因组含量信息,通过统计比较分析认为:(1)主要蔬菜种类中,石蒜科(19.08 pg)蔬菜平均基因组含量最高,十字花科(0.78 pg)和葫芦科(0.78 pg)蔬菜最低;(2)多年生和单子叶蔬菜平均基因组含量分别极显著高于非多年生蔬菜和双子叶蔬菜。本文对樟树基因组大小进行测定,为木本植物基因组大小变化规律提供参考。

流式细胞术是目前应用较多的基因组大小测定方法,在毛竹(*Phyllostachys edulis*)[21]、五节芒(*Miscanthus floridulus*)[7]等植物中都有应用。除此之外,还有 Feulgen 分光光度法[22]和脉冲场凝胶电泳法[23]等。高通量测序技术的迅速发展为基因组评估提供了更迅捷的方法。李西文等[24]应用 454 高通量测序技术对荷花玉兰叶绿体全基因组进行测序,解析了其基因组结构,并与近缘物种基因组进行了比较分析,得到了很好的研究结果。本文采用的 K-mer 分析法是基于全基因组测序片段的 K-mer 深度分布[25]估计物种基因组大小的方法,得到了樟树基因组大小、杂合度、GC 含量等结果,为该物种的

进一步研究提供详细的遗传背景。

本研究首次测定了樟树基因组大小,并对基因组的杂合率、GC 含量及分布等进行了评估,主要结论如下:(1)樟树基因组大小粗略估计为 760 Mb 左右;(2)樟树基因组有较高的杂合和一定的重复,杂合性约为 0.65%,对组装效果影响较大。用 WGS 策略进行组装有一定的风险和难度;(3)由于樟树杂合率较高,全基因组鸟枪法策略不适合该基因组分析。可尝试使用 BAC-to-BAC 策略或 fosmid 策略进行组装,这种组装策略对杂合较高的基因组拼接帮助较大。上述樟树全基因组调查分析结果将会对樟树全基因组图谱绘制方案的制定提供重要依据。

参考文献

[1] 戴宝合. 野生植物资源学[M]. 北京:农业出版社,1993

[2] 肖复明,江香梅,熊彩云,等. 樟树种子油中脂肪酸成分分析[J]. 江西林业科技,2008(1):60-61

[3] 陈成彬,李秀兰,孙成仁,等. 中国樟科 5 属 9 种植物的核型研究[J]. 武汉植物学研究,1998,16(3):219-222

[4] Okada H,Karvologieal R T. Studies in some speeies of Lauraeeas [J]. Taxon,1975,24(2-3):271-280

[5] Stebbins G L. Chromosomal evolution in higher plants[M]. London:Edward Arnold LTD,1971:87-89

[6] Du B,Wang D. C-values of seven marine mammal species determined by flow cytometry[J]. Zool Sci,2006,23(11):1017-1020

[7] 邓果特,刘清波,蒋建雄,等. 五节芒基因组大小测定[J]. 植物遗传资源学报,2013,14(2):339-341

[8] Kikuchi S,Tanaka H,Shiba T,et al. Genome size,karyotype,meiosis and a novel extra chromosome in *Torenia fournieri*,*T. baillonii* and their hybrid[J]. Chromosome Res,2006,14(6):665-672

[9] Margulies M,Egholm M,Altman W E,et al. Genome sequencing in microfabricated high-density picolitre reactors[J]. Nature,2005,437(7057):376-380

[10] Tuskan G A,Difazio S,Jansson S,et al. The genome of black cottonwood,*Populus trichocarpa* (Torr. & Gray) [J]. Science,2006,313(5793):1596-1604

[11] Jaillon O,Aury J M,Noel B,et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla[J]. Nature,2007,449(7161):463-467

[12] Ming R,Hou S B,Feng Y,et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)[J]. Nature,2008,452(7190):991-996

[13] Velasco R,Zharkikh A,Affourtit J,et al. The genome of the domesticated apple (*Malus* × *domestica* Borkh.)[J]. Nat Genet,2010,42(10):833-839

[14] Wu J,Wang Z,Shi Z,et al. The genome of the pear (*Pyrus bretschneideri* Rehd.)[J]. Genome Res,2013,23:396-408

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

[16] 施季森,王占军,陈金慧. 木本植物全基因组测序研究进展[J]. 遗传,2012,34(2):145-156

[17] Xu M,Zang B,Yao H S,et al. Isolation of high quality RNA and molecular manipulations with various tissues of *Populus*[J]. Russian J Plant Physiol,2009,5:716-719

[18] Li R Q,Fan W,Tian G,et al. The sequence and *de novo* assembly of the giant panda genome[J]. Nature,2009,463:311-317

[19] Li R Q,Zhu H M,Ruan J,et al. *De novo* assembly of human genomes with massively parallel short read sequencing[J]. Genome Res,2010,20:265-272

[20] 程蛟文,吴智明,崔竣杰,等. 主要蔬菜作物基因组含量统计与比较分析[J]. 园艺学报,2013,40(1):135-144

[21] 李潞滨,武静宇,胡陶,等. 毛竹基因组大小测定[J]. 植物学通报,2008,25(5):574-578

[22] Ha S H,Kim J B,Park J S,et al. A comparison of the carotenoid accumulation in *Capsicum varieties* that show different ripening colours:Deletion of the capsanthin-capsorubin synthase gene is not a prerequisite for the formation of a yellow pepper[J]. J Exp Bot,2007,58(12):3135-3144

[23] 刘艳鸣,张奇亚. 利用脉冲场凝胶电泳测定东湖浮游病毒基因组的大小[J]. 武汉大学学报:理学版,2005,51(S):238-240

[24] 李西文,高欢欢,王一涛,等. 荷花玉兰叶绿体全基因组高通量测序及结构解析[J]. 中国科学:生命科学,2012,42(12):947-956

[25] Havlak P,Chen R,Durbin K J,et al. The atlas genome assembly system[J]. Genome Res,2004,14(4):721-732

(上接第 148 页)

[16] Xu L,Menard R,Bert A,The E2 ubiquitin-conjugation enzymes AtUBC1 and AtUBC2,play redundant role and are involved in activation of FLC expression and repression of flowering in *Arabidopsis thaliana*[J]. Plant J,2009,57:279-288

[17] Xiong L,Yang Y. Disease resistance and abiotic stress tolerance in rice are inversely modulated by an abscisic acid-inducible mitogen-activated protein kinase[J]. Plant Cell,2003,15:745-759

[18] Lee J S,Wang S,Sritutim S,et al. *Aradidopsis* mitogen-activated protein kinase MPK12 interacts with the MAPK phosphatase IBR5 and regulates auxin signaling[J]. Plant J,2009,57:975-985

[19] Nakagami H,Pitzschke A,Hirt H. Emerging MAP kinase pathways in plants stress signaling[J]. Trends Plant Sci,2005,10:339-346

[20] Jammes F,Song C,Shin D,et al. MAP kinases MPK9 and MPK12 are preferentially expressed in guard cells and positively regulate ROS-mediated ABA signaling [J]. PNAS, 2009, 106: 20520-20525

[21] Cho S K,Ryu M Y,Song C,et al. *Arabidopsis* PUB22 and PUB23 are homologous U-Box E3 ubiquitin ligases that play combinatory roles in response to drought stress[J]. Plant Cell, 2008, 7: 1899-1914

[22] Huang C,Ding S,Zhang H,et al. CIPK7 is involved in cold response by interacting with CBL1 in *Arabidposis thaliana*[J]. Plant Sci,2011,181:57-64

[23] Fukao T,Yeung E,Bailey-Seres J. The submergence tolerance regulator SUB1A mediates crosstalk between submergence and drougeht tolerance in rice[J]. Plant Cell,2011,23:412-427

[24] Xu Z S,Ni Z Y,Liu L,et al. Characterization of the TaAIDFa gene encoding a CRT/DRE-binding factor responsive to drought,high-salt,and cold stress in wheat[J]. Mol Genet Genomics,2008,280:497-508