Asia-Pacific Journal of Pharmaceutical Sciences
亚太药学期刊
2019; 1(3): 24-27
http://www.sciencepublishinggroup.com/j/apjps

Genome Survey Sequencing of *Coptis chinensis*

Hou Feixia, Wang Qihao, Peng Cheng, Gao Jihai*

Pharmacy College, Chengdu University of Traditional Chinese Medicine, Chengdu, China

Email address:
houfeixia123@163.com (Hou Feixia), gaojihaiwuwei@163.com (Gao Jihai)

*Corresponding author

To cite this article:
Hou Feixia, Wang Qihao, Peng Cheng, Gao Jihai. Genome Survey Sequencing of *Coptis chinensis*. *Asia-Pacific Journal of Pharmaceutical Sciences*. Vol. 1, No. 3, 2019, pp. 24-27.

Received: December 2, 2019; **Accepted**: February 19, 2020; **Published**: March 6, 2020

Abstract: *Coptis chinensis* is a famous medicinal plant, which belongs to *Coptis* of Ranunculaceae. In China it has been applied for more than 2,000 years, and it is also one of the key medicinal plants worldwide with the whole plants and extract. The genome of *Coptis chinensis* was investigated in order to accelerate the completion of its genome sequencing, promote its research in molecular breeding, genetic evolution and the excavation of important medicinal plant gene resources. Based on high-throughput sequencing technology, two libraries with fragment sequences about 270 bp were constructed. After performing double-terminal 270 bp (PE 270) sequencing, 49.82 Gb raw data was obtained. It was estimated that *Coptis chinensis* genome size was 1.06 Gb, the repeat sequence content was about 71.46%, the heterozygosity was about 0.24%, and the GC content was about 39.50%. The study showed that *Coptis chinensis* had complex geome with repeat and large sequence data. Therefore, three generation sequencing technology is recommended to construcut a 20 Kb library and test 100× Pacbio data, and construcut the 270 bp small library and test 50× illumina data, which will contribute to the acquisiton of high-quality complete genome map of *Coptis chinensis*. This study also provides basic data for mining functional genes of *C. chinensis*.

Keywords: *Coptis chinensis*, Genome Sequencing, Genomic Trait Assessment

黄连基因组调研测序研究

侯飞侠，汪颀浩，彭成，高继海*

成都中医药大学药学院西南特色中药资源重点实验室, 成都, 中国

邮箱
houfeixia123@163.com（侯飞侠），gaojihaiwuwei@163.com（高继海）

摘要：黄连*Coptis chinensis*为著名的药用植物，归于毛茛科（Ranunculaceae）黄连属（*Coptis*），在中国药用历史已达两千余年，也是世界范围内研究开发的重点药用植物之一。为加速完成黄连物种的基因组测序，促进其在分子育种、遗传进化、特色功能基因资源的挖掘等方面的研究，对黄连基因组进行了调研。采用高通量测序技术，构建了2个270bp文库，进行双端270 bp (PE 270)测序，获得了49.82 Gb原始数据，估计黄连基因组大小为1.06 Gb，重复序列含量约71.46%，杂合率约0.24%，GC含量约39.50%。本研究表明黄连属于高重复大基因组的复杂基因组，建议后续采用三代测序技术，构建20 Kb文库，测100×的Pacbio数据；构建270bp小文库，测50×的illumina数据，有助于黄连高质量全基因组图谱的获得。本研究结果也为黄连功能基因的挖掘提供了基础数据。

关键词：黄连，基因组测序，基因组特征评估

1．引言

黄连为毛茛科植物黄连*Coptis chinensis* Franch.、三角叶黄连*Coptis deltoidea* C. Y. Cheng et Hsiao或云连*Coptis teeta* Wall.的干燥根茎，具有清热燥湿，泻火解毒之功效，临床主要用于湿热痞满、心火亢盛、血热吐衄、痈肿疔疮等症[1]。黄连疗效显著，需求量大，野生居群被长期过度采挖，导致生境不断遭到破坏，部分品种已处于灭绝的边缘。为了保护黄连野生品种，解决供需矛盾，更好的利用其药用价值，很多的科研工作已经对黄连进行了人工栽培，提高了黄连产量。目前对黄连药用植物的研究主要集中在分类学[2]、形态学[3]、生态学[4]和生物学特性[5]方面。近年来也有利用分子标记对黄连进行分子鉴定[6, 7]、遗传多样性[8, 9]和系统发育关系[10,11]研究的报道，涉及RAPD、ISSR、ITS以及多个叶绿体条形码标记等，对黄连的遗传育种具有一定的参考价值。鉴于黄连属物种的复杂性，中医药学者们还尝试开发具有更高特异性的标记序列，如针对叶绿体全基因组的测序发现了更多的SSR标记[12]，而针对更多黄连属物种的叶绿体基因组比对分析发现两段高变基因区[13]，为黄连鉴定的专一条形码寻找指出了方向。此外，黄连的转录组学也提供了较多的候选标记[14]。总体上，黄连在分子水平的研究仍然稀缺，尤其是还欠缺基因组数据库的共享平台，导致黄连系统性的现代科学研究乏力。

随着高通量核苷酸测序技术不断更新换代，目前多种药用植物已完成了全基因组测序，如灵芝[15]，丹参[16]，铁皮石斛[17]等，相关流程体系均已臻成熟。对黄连进行基因组测序，利于深入开展黄连的遗传和分子遗传研究，促进该药用植物的遗传育种，对分子改良和产量提高等也具有重要意义。本研究利用Illumina HiSeq 4000测序平台对该物种的基因组大小进行测定和评估，旨在为黄连全基因组测序方案提供参考依据和候选测序植株。

2．材料与方法

2.1．实验材料

本研究所用材料采于成都中医药大学药用植物园，由成都中医药大学卢先明教授鉴定为黄连*Coptis chinensis* Franch，活体材料继续种植保存于植物园，凭证标本保存于成都中医药大学国家中药种质资源库。剪取幼叶，液氮速冻后置于-70°C超低温冰箱保存备用。

2.2．方法

2.2.1．文库构建和测序

使用CTAB法提取黄连叶片基因组DNA，将DNA样品进行随机打断，构建2个270bp小片段文库。文库使用北京百迈客生物科技有限公司的Illumina HiSeq 4000测序平台进行双端270 bp (PE 270)测序。

2.2.2．样品污染评估

样品如果存在污染不仅会降低有效数据量，同时还会影响调研图分析结果的准确性，导致基因组大小、杂合率、重复序列比例和GC含量等基因组特征评估结果出现较大偏差，使得基因组组装建库策略出现偏差，最终影响后续的基因组组装效果。为了判断提取的样品DNA是否受到污染，本研究从测序得到的2个270 bp文库中，随机取10, 000条单端reads，与NT库进行BLAST[18]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。

2.2.3．基因组大小、重复序列比例和杂合率评估

测序结果经过过滤，去除低质量reads后，基于K-mer的分析方法对黄连的基因组大小、重复序列比例和杂合率进行评估。取K为21来进行分析，假设K-mer的深度频率服从泊松分布，且从reads中逐碱基取出的所有K-mer能够遍历整个基因组，即可从所有测序reads中统计K-mer频数分布，计算获得K-mer深度估计值，作K-mer深度分布曲线。用公式基因组大小=总碱基数/平均测序深度=K-mer总数/K-mer平均深度估计基因组大小。由于测序片段是随机打断的，标准的K-mer深度分布曲线呈正态分布，根据实际曲线偏离正态分布的程度，来估计基因组杂合度和重复比例。

2.2.4．GC含量评估

根据所有测序reads中GC量和所有测序reads中碱基含量，即可算出GC含量。

3．结果与分析

3.1．测序数据统计

本研究挑选黄连单株，使用基因组DNA构建的2个270 bp文库，在Illumina PE270测序平台测序并过滤后，获得49.82 Gb数据，总测序深度约为47×，测序数据Q20（测序质量值在20以上）比例均在95.83%以上，Q30（测序质量值在30以上）比例均在90.64%以上。测序结果见表1。

表1 样品测序结果统计表。

文库ID	插入片段大小(bp)	数据量(Gb)	测序深度(×)	Q20(%)	Q30(%)
1	270	27.23	25.74	95.83	90.64
2	270	22.59	21.36	95.84	90.69
Total		49.82	47.09		

3.2．样品污染评估

评估样品是否存在污染的标准为：如果有一定比例的reads比对上进化距离较远的物种如动物，微生物等，则判断样品可能存在污染，需要进一步检查原因。通过将从文库中随机挑取到的10, 000条单端reads与NT库进行BLAST比对，发现2个270 bp文库能够比对上NT库的reads分别占提取reads数的9.34%和9.44%。两个文库比对上NT库reads数占比最高的均为毛茛科的Aconitum chiisanense和Megaleranthis saniculifolia，这两个物种皆为黄连的近缘物种，且比对结果中未发现动物等异常比对。因此该样品测序数据不存在污染，可用于基因组调研图分析。

3.3．基因组大小、重复序列和杂合率评估

使用2个270 bp文库数据构建K=21的K-mer分布图（图1），进行基因组大小、重复序列比率和杂合率的评估。由图1可知，平均K-mer深度即主峰对应的K-mer深度为36。K-mer深度出现在主峰对应深度2倍以上的序列为重复序列，即深度大于72的K-mer序列为重复序列。K-mer深度出现在主峰对应深度一半处的序列为杂合序列，即深度出现在18附近的K-mer序列为杂合序列。从测序数据中得到的总K-mer数为4,315,375,083个，去除深度异常的K-mer后，共38,559,152,762个K-mer用于基因组长度估计，计算得到的基因组长度约1.06 Gbp。依据K-mer分布情况，估计重复序列含量约71.46%，没有明显杂合峰，估计杂合度约0.24%，杂合度较低。因此该物种基因组属于高重复，大基因组的复杂基因组。

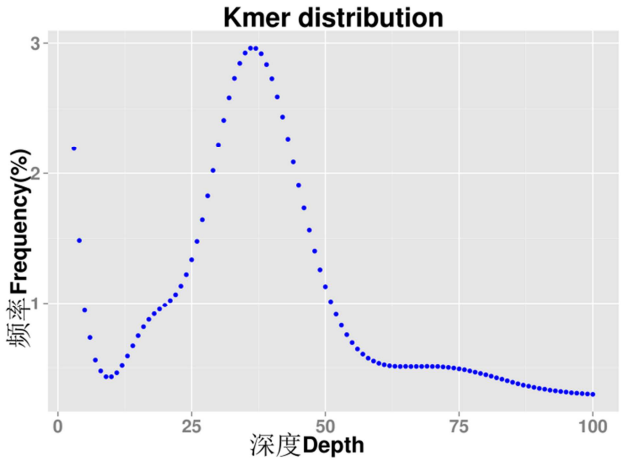

图1 K-mer分布曲线。

3.4．GC含量评估

基因组GC含量对二代基因组测序的随机性有较大影响。过高(>65%)或过低(<25%)的GC含量会导致测序偏向性，严重影响基因组分析结果。物种GC含量是评估调研图分析准确性和后续基因组组装难度的重要指标之一。通过对调研图文库测序数据分析，该物种基因组的GC含量约39.50%，较为适中，不会影响分析的准确性。

4．讨论

黄连在临床上具有世界公认的确切疗效，市场需求快速增大，而野生资源的过度开发严重破坏了黄连的生境，导致黄连资源逐渐紧缺，甚至有些品种濒临灭绝。近年来，黄连的人工栽培得到了国家和地方的大力支持，这在一定程度上保障了市场对黄连的需求量，但是黄连的分子生物学研究始终不足，这必然制约该物种的遗传改良和快速育种的研究和应用，阻碍黄连药用价值的进一步开发利用。

本研究使用Illumina HiSeq 4000测序平台对黄连基因组DNA 2个270bp文库进行双端270 bp测序，通过将测序结果与NT库比对，表明样品不存在污染，可用于基因组调研图分析；测序获得了49.82 Gb原始数据，估计黄连基因组大小为1.06 Gb，重复序列含量约71.46%，杂合率约0.24%，GC含量适中（39.50%），不会影响分析的准确性。从基因组基本结构特征上看，黄连属于高重复大基因组的复杂基因组，建议后续采用三代测序技术，构建20 Kb文库，测100×的Pacbio数据；构建270bp小文库，测50×的illumina数据，有助于黄连高质量全基因组图谱的获得。

孙全[19]等之前进行了黄连基因组勘测与分析，虽然获得了大量的数据，但是由于样品植株选择与测序深度（30×）的不足，导致其所得序列不能进行有效的高质量组装，其组装的结构也不能进行基因的预测和分析，且其实验材料的基因组杂合度较高（1.1%），故其作为完整基因组测序分析的价值有限。与孙全的测试结果相比，本研究通过精选实验材料，所获得基因组测序结果具有以下优势：数据结果更庞大，测试的深度更深，杂合率更低。本文与孙全等的研究结果为黄连品种研究提供一定的分子参考依据，为加速完成黄连物种的基因组测序，以启动该物种在分子水平、遗传进化、重要药用植物基因资源的挖掘等方面的大量研究奠定基础。

5．结论

本研究采用高通量测序技术对黄连基因组DNA 2个270bp文库进行了双端270 bp测序，获得了49.82 Gb原始数据，估计黄连基因组大小为1.06 Gb，重复序列含量约71.46%，杂合率约0.24%，GC含量约为39.50%。从基因组基本结构特征上看，黄连属于高重复大基因组的复杂基因组，建议后续采用三代测序技术，构建20 Kb文库，测100×的Pacbio数据；构建270bp小文库，测50×的illumina数据，有助于黄连高质量全基因组图谱的获得。该研究为加速完成黄连物种的基因组测序，促进其在分子育种、遗传进化、特色功能基因资源的挖掘等方面的研究奠定了基础。

致谢

感谢四川省中医药管理局科学技术研究专项(2018QN001，2016ZY008)，国家自然科学基金人才培养项目(J1310034)，中药学四川省科技厅创新团队(2017TD0001)和成都中医药大学“杏林学者”学科人才提升计划(QNXZ2018017)给予的资金支持。

参考文献

[1] 中国药典. 一部[S]. 2015: 303。

[2] 孙红祥, 张昌禧, 吴远文. 黄连属药用植物的数量分类学研究[J]. 中国药学杂志, 1995, 30(1): 7。

[3] 陈红, 王海洋. 重庆石柱黄连叶形态分析初探[J]. 西南大学学报(自然科学版), 2007, 29(1): 31。

[4] 陈萍萍, 刘学医. 宣黄连的生态学和栽培研究[J]. 中国野生植物资源, 2012, 31(5): 75。

[5] 汪建云, 刘经伦, 施晓春, 等.云南黄连生物学特性和栽培管理初探[J]. 保山学院学报, 2013, 32(2): 1。

[6] 孙涛, 孔德英, 滕少娜, 等. 基于ITS2序列的黄连及其伪混品的分子鉴定[J]. 贵州农业科学, 2013, 41(9): 20。

[7] 李波, 刘俊, 闵道长, 等. 黄连属植物DNA条形码研究[J]. 江西农业大学学报, 2017, 39(6): 1089。

[8] 张春平, 何平, 胡世俊, 等. 药用三角叶黄连遗传多样性的ISSR分析[J]. 中国中药杂志, 2009, 34(24):3176。

[9] 陈大霞, 王钰, 张雪, 等. 黄连属部分药用植物遗传多样性与亲缘关系的SCoT分析[J]. 中国中药杂志, 2017, 42(3): 473.

[10] He Y, Fan G, Gao J, et al. Correlation analyses between molecular perspective and phytochemical variations in Coptis chinensis, Franch [J]. Biochem Syst Ecol, 2015, 61: 143。

[11] 张春平，何平，何俊星，等. 药用峨眉野连遗传多样性的RAPD分析[J]. 中国中药杂志，2010, 35(2): 138-141。

[12] He Y, Xiao HT, Deng C, et al. Complete chloroplast genome sequence of Coptis chinensis Franch. and its evolutionary history [J]. BioMed Research International. 2017, doi: 10.1155/2017/8201836。

[13] 赵振宇. 黄连属植物物种鉴定及mini-barcode研究[M]. 2019, 中国中医科学院。

[14] Chen H, Deng C, Nie H, et al. Transcriptome analyses provide insights into the difference of alkaloids biosynthesis in the Chinese goldthread (Coptis chinensis Franch.) from different biotopes [J]. Peer J, 2007, DOI: 10.7717/peerj.3303。

[15] Chen S, Xu J, Liu C, et al. Genome sequence of the model medicinal mushroom Ganoderma lucidum [J]. Nat Commun, 2012, 3 (2): 913。

[16] 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。

[17] Liang Y, Xiao W, Hui L, et al.The genome of Dendrobium officinale Illuminates the biology of the important traditional Chinese orchid herb [J]. Mol Plant, 2015, 8 (6): 922。

[18] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool [J]. J Mol Biol, 1990, 215: 403。

[19] 孙全, 何晓红, 牟军, 等. 黄连基因组勘测与分析[J].四川大学学报(自然科学版), 2014, 51(06): 1330。