基于流式细胞术和 K-mer 分析的黄芪基因组大小估测

孙会改[1,2]，韦春香[1]，杨旻啸[1]，高　飞[1*]，周宜君[1*]

1. 中央民族大学生命与环境科学学院，北京　100081
2. 河北中医学院药学院，河北 石家庄　050200

摘　要：目的　使用流式细胞术与 K-mer 分析估测黄芪基因组的大小及复杂程度，为黄芪基因组测序奠定基础。**方法**　以番茄为内标，用流式细胞仪测定经碘化丙啶（PI）染色的番茄细胞核和黄芪细胞核混合样品的 PI 荧光强度，通过比较黄芪与番茄细胞 DNA 含量峰值的倍数关系，计算黄芪的基因组大小。采用高通量测序技术进行黄芪基因组调查测序，利用生物信息学方法估计黄芪杂合率、重复序列和 GC 含量等信息。**结果**　采用流式细胞术测得黄芪基因组大小约为 1 426 Mb。通过基因组调查测序，获得了 95 Gb 黄芪基因组 DNA 测序数据，K-mer 分析表明黄芪基因组约为 1 456 Mb，测序深度为 39×。K-mer 分布曲线有明显的杂合峰，基因组杂合率为 2.1%。**结论**　黄芪基因组大小为 1.45 Gb 左右，杂合率较高。这一结果可为黄芪基因组学研究提供参考。

关键词：黄芪；流式细胞术；基因组大小；K-mer 分析；杂合率

中图分类号：R282.12　　**文献标志码：**A　　**文章编号：**0253 - 2670(2019)06 - 1448 - 05

DOI: 10.7501/j.issn.0253-2670.2019.06.029

Estimation of genome sizes of *Astragalus membranaceus* based on flow cytometry and K-mer analysis

SUN Hui-gai[1,2], WEI Chun-xiang[1], YANG Min-xiao[1], GAO Fei[1], ZHOU Yi-jun[1]

1. College of Life and Environmental Sciences, Minzu University of China, Beijing 100081, China
2. Hebei University of Chinese Medicine, Shijiazhuang 050200, China

Abstract: Objective　To determinate the genome size and complexity of *Astragalus membranaceus* by using flow cytometry (FCM) and K-mer analysis, which can lay the foundation for the screening of functional genes of *A. membranaceus*. **Methods** *Lycopersicon esculentum* was served as an internal reference in this study. The mixed sample of *A. membranaceus* cell nucleus and *L. esculentum* cell nucleus was stained using propidium iodide (PI). The PI fluorescence intensities of the sample were measured by FCM. The genome size of *A. membranaceus* was calculated by comparing the multiple relationship between the peak of DNA content in the cells of *A. membranaceus* and *L. esculentum*. The genome of *A. membranaceus* was sequenced by using high-throughput sequencing technologies. The genome size of *A. membranaceus* was calculated by K-mer analysis. The hybridity percentage, repetitive sequence, and GC of *A. membranaceus* were estimated by bioinformatics analysis. **Results**　The genome size of *A. membranaceus* was about 1 426 Mb. For K-mer analysis, more than 95 Gb high quality data from the genome was generated. The average genome size and sequencing coverage depth of *A. membranaceus* was about 1 456 Mb and 39 times respectively. The genome of *A. membranaceus* had obvious hybridity peak by K-mer method, and the hybridity percentage as high as 2.1%. **Conclusion**　The genome size of *A. membranaceus* was about 1.45 Gb and the heterozygosity is high. These data would provide a reference for the genomic research in *A. membranaceus*.

Key words: *Astragalus membranaceus* (Fisch.) Bunge; flow cytometry; genome size; K-mer analysis; hybridity percentage

黄芪 *Astragalus membranaceus* (Fisch.) Bunge 为豆科（Leguminasae）多年生草本药用植物，作为我国一味传统中药，其药理作用明确，在我国已有 2 000 多年的药用历史，是我国 40 种常用大宗药材

收稿日期：2018-08-09

基金项目：国家自然科学基金项目（31770363）；国家自然科学基金项目（31670335）；国家大学生创新训练计划（GCCX2017110015）

作者简介：孙会改（1986—），女，河北中医学院讲师，研究方向为生物化学与分子生物学。Tel: (010)68932633　E-mail: sunhuigai66@163.com

***通信作者**　高　飞，男，博士，副教授，博士生导师，主要从事生物化学与分子生物学研究。Tel: (010)68932633　E-mail: gaofei@muc.edu.cn

周宜君，女，博士，教授，博士生导师，主要从事生物化学与分子生物学研究。Tel: (010)68932922　E-mail: zhouyijun@muc.edu.cn

品种之一。中医认为，黄芪的性质温和且味道甜，归脾经和肺经，具有补气固表、托毒消肿、利尿生肌、生津养血、增强免疫力之功效[1]，在古代被誉为“补药之长”。最初记载于《神农本草经》，位列上品，应用于药品食品等领域，黄芪经典名方包括玉屏风散、补中益气汤、黄芪建中汤和归脾汤等，以黄芪为原料生产的中药达 200 多种，是临床应用最为广泛的补益中药。

基因组大小是指真核生物单倍体基因组中 DNA 的含量，又称为 C 值，具有物种相对稳定性和特异性，是生物基因组多样性非常重要的基本参数[2-3]。基因组大小的度量，一般以质量计算，单位通常是 pg，也可以用核苷酸碱基对的数量表示（Mb 或 Mbp），1 pg 约等于 978 Mb[4]。测定生物体基因组大小可以估测物种的 DNA 含量，是基因组学测序的前提，不仅对该物种的分子遗传研究具有重要意义，而且对包括系统分类学、生物进化在内的多个研究领域有重要的推动作用[3]。然而目前仅有很少一部分药用植物 C 值被测定，有关 C 值在药用植物的系统演化和进化等方面的研究还没有引起重视。本研究采用流式细胞仪和 K-mer 分析对黄芪基因组大小进行估测，并且确定了基因组杂合度等指标，为黄芪基因组等相关研究提供了必要的基础数据。

1 材料

黄芪种子来自蒙古内蒙古自治区鄂尔多斯市，其植株经中央民族大学生命与环境科学学院刘博博士鉴定为膜荚黄芪 *Astragalus membranaceus* (Fisch.) Bunge。番茄 *Lycopersicon esculentum* Mill. 取自北京农林科学院蔬菜研究中心，选取 2 个月左右的幼嫩叶片用于流式细胞分析。

2 方法

2.1 流式细胞分析

采用美国 BD 公司的 FACSCalibur 流式细胞仪进行基因组大小检测，并使用 CellQuest（BD 公司）软件获取数据，使用 ModFit 软件（Yerity SoftwareHouse 公司）分析结果。

测定操作程序：取植株新鲜叶片 1 g，在 2 mL 细胞裂解液中用锋利的刀片切碎、滤过、收集滤液，1 000 r/min 离心 5 min 后，弃上清，收集沉淀细胞。用碘化丙啶（propidium iodide，PI）染液对细胞核 DNA 进行荧光标记，在暗处染色 20 min 后，用流式细胞仪进行待测样品基因组大小鉴定。

用已知基因组大小的材料作为对照，将对照材料横坐标固定，然后与待测样混匀后进行检测，根据峰的位置并参考对照样品的大小判断待测样基因组大小。

2.2 K-mer 分析

采用改良 CTAB 法[5]提取叶片基因组 DNA，将样品进行随机打断，构建插入 350 bp 的 DNA 文库，再用 Illumina Hiseq Xten PE 150 平台进行双末端（Paired-End）测序，滤过掉低质量数据，得到的高质量数据，采用模拟数据拟合的方式进行基因组杂合率评估；利用高质量数据进行 SOAP de novo 组装[6]，构建 Contig 和 Scaffold，对 Contigs 间空隙（“N”）进行局部组装，适当延长 Contigs。采用基于 K-mer 的分析方法来估计基因组大小，即从一段连续序列中迭代地选取长度为 K 个碱基的序列，取 K 为 17 来进行分析[7]。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 深度估计值，用以下公式估计基因组大小。

基因组大小＝K-mer 个数/K-mer 期望深度

3 结果与分析

3.1 流式细胞分析估测黄芪基因组大小

番茄是茄科的模式植物，其基因组为 0.9 Gb[8-9]。番茄和黄芪单独测定的结果见图 1、2。番茄基因组的测定峰与黄芪基因组测定峰无重叠，保证了采用番茄作为内标计算黄芪基因组大小的准确性。

荧光染料 PI 能均匀地插入到 DNA 碱基对中，并且插入量与 DNA 的含量呈正比，即荧光强度与

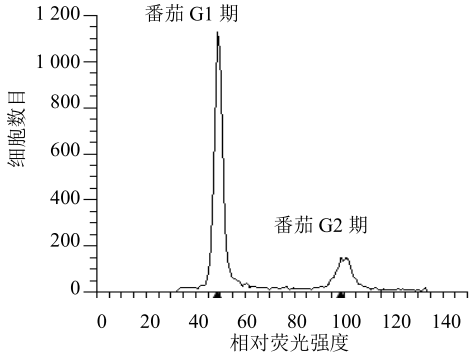

图 1 番茄基因组的相对荧光强度

Fig. 1 Relative fluorescence intensity of *Lycopersicon esculentum*

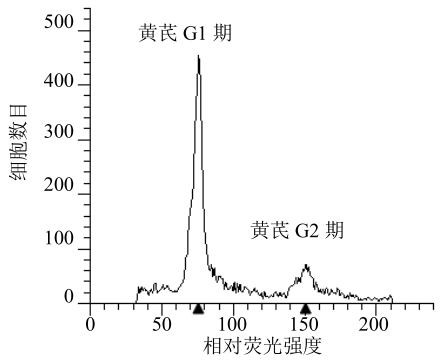

图 2　黄芪基因组的相对荧光强度

Fig. 2　Relative fluorescence intensity of *Astragalus membranaceus*

DNA 的含量呈正比[4]。基于此原理，可根据待测样品与对照品的荧光比值来算出待测样品的 DNA 含量。

待测样品 DNA 含量＝对照品 DNA 含量×待测样品的荧光强度/对照品的荧光强度

对番茄和黄芪的 PI 发射荧光强度的测定结果分析表明（图 3），番茄 73.77%的细胞在 G1 期，位置为图 3 中横坐标 46.13。黄芪 54.55%的细胞为 G1 期，位置在 73.14。推算黄芪的基因组大小约为 1 426 Mb。

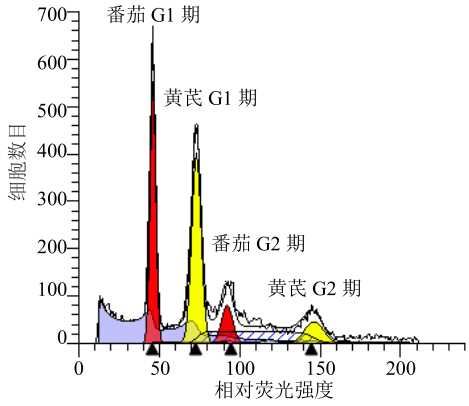

图 3　番茄与黄芪基因组的相对荧光强度

Fig. 3　Relative fluorescence intensity of *L. esculentum* and *A. membranaceus*

3.2　17-mer 分析测定黄芪基因组大小

对黄芪基因组的 Survey 分析采用全基因组鸟枪法（WGS）策略，利用第二代测序技术构建插入片段为 350 bp 的 DNA 文库，在 Illumina Hiseq Xten PE 150 上对这些片段两端进行双末端（Paired-End）测序，测序共得到 125 Gb 的 DNA 序列数据，过滤掉低质量数据后，得到的 95 Gb 数据用于后续 17-mer 分析。使用黄芪样本 95 Gb 的数据用于 17-mer 分析，其频率分布见图 4。横坐标表示 17-mer 出现的次数，纵坐标表示出现的频率。图 4 显示，17-mer 分布曲线成峰情况较好，17-mer 分布曲线为非正常泊松分布，呈现双峰分布，在 31 和 15 处各有一个峰值。推测 31 为主峰，即 K-mer 的期望深度，K-mer 总数是 45 139 Mb（约 45 Gb），结合公式估算黄芪基因组大小约为 1 456 Mb。

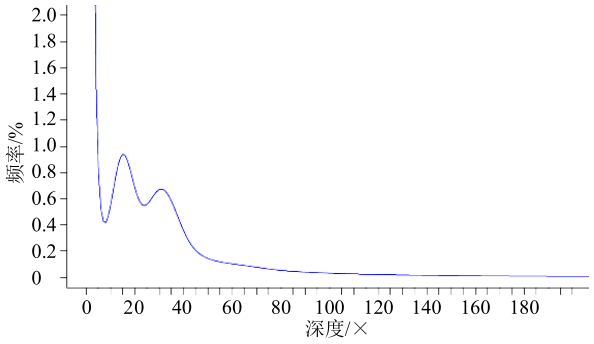

图 4　17-mer 分布曲线

Fig. 4　Distribution curve of 17-mer

从图 4 可以清楚地观察到，在期望深度的 1/2 位置处有明显的杂合峰，说明该基因组有一定的杂合度。为进一步预估黄芪样本的杂合率，用拟南芥基因组模拟对应深度的短片段数据，在杂合率不同梯度组合情况下进行 K-mer 曲线拟合。根据 K-mer 曲线拟合情况来估计黄芪的杂合率（图 5）。由图 5 可以看出，黄芪真实曲线的主峰和杂合峰在杂合率 2.1%时形成的峰与拟南芥的模拟数据最接近，因此可大致认为黄芪的杂合率处于 2.1%水平，说明黄芪的杂合率较高。

3.3　黄芪基因组杂合率和 GC 含量

利用 95 Gb 高质量 DNA 测序数据进行黄芪基因组的初步组装，将短 reads 打断构建 De-Brujin-Graph，构建 Contigs 和 Scaffold，对 Contigs

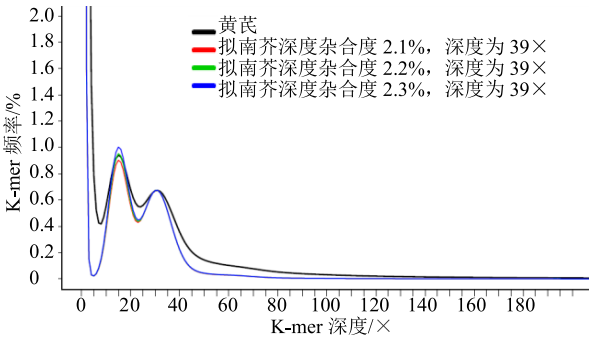

图 5　17-mer 杂合率估计图

Fig. 5　Genome heterozygosity estimation based on 17-mer curve

间空隙（“N”）进行局部组装，适当延长 Contigs。从表 1 可以看出，由于黄芪的杂合率高达 2.1%，组装的结构中 Scaffold 总长度约为 1 191 Mb，Scaffold N50 长度为 1 235 bp，Scaffold 总数为 2 192 376；Contig 总长度约为 893 Mb，Contig N50 长度为 234 bp，Contig 总数为 4 518 403。

表 1 组装结果统计

Table 3 Statistics of assembly results in *A. membranaceus*

指标	Contig		Scaffold	
	长度/bp	数目	长度/bp	数目
N50	234	832 952	1 235	224 162
N60	177	1 276 026	868	339 586
N70	143	1 842 406	590	506 859
N80	120	2 525 454	372	767 062
N90	102	3 335 110	166	1 246 895
最长	17 797	1	42 657	1
总数	893 662 368	4 518 403	1 190 659 687	2 192 376
总数＞100 bp	817 900 811	3 470 174	1 190 659 687	2 192 376
总数＞2 kp	46 459 363	15 230	424 405 250	114 410

GC_depth 分析显示（图 6 和 7），黄芪的 GC 分布 GC 测序无明显偏向，呈现出 2 部分的区域分布，GC 含量 30%～50%、测序深度 20×～80×这个区域分布比较集中，说明测序中不含有污染。对应图 6 的基因组 GC 含量分布图，在 GC 含量 30%位置的大峰为杂合峰，在 GC 含量 50%的位置为纯合峰。由基因组测序深度分布图（图 8）可以看出，主峰在 GC 含量 33%的位置，与前期为调查黄芪基因组的 GC 含量基本一致，且散点也分布在 GC 含量 35%附近（图 6）。GC 聚成的块分成了 2 层，平均测序深度为 31 ×，低深度分布区域的深度约为正常深度分布的 1/2，这可能是黄芪高度杂合的结果。因为杂合会使两天同源染色体杂合的部位只装出了 1 条，或 2 条都没有装出，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。

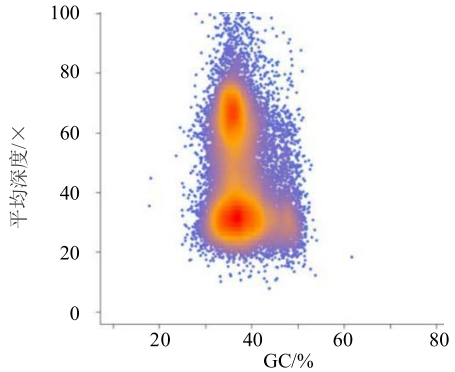

图 6 GC_depth 分布

Fig. 6 Distribution of GC_depth

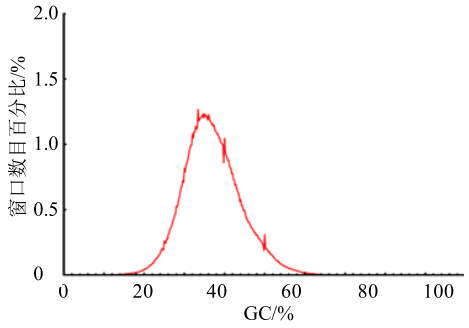

图 7 GC 含量分布

Fig. 7 Distribution curve of GC content

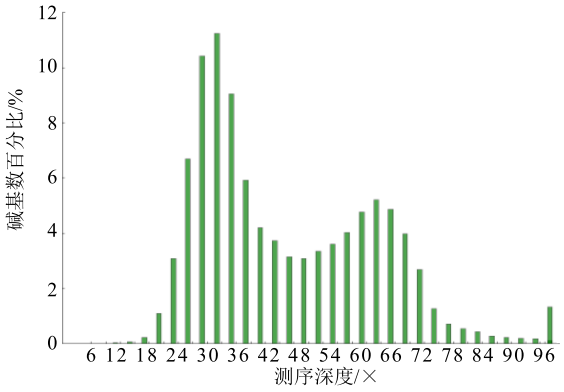

图 8 黄芪基因组测序深度分布

Fig. 8 Distribution diagram of sequencing depth

4 讨论

真核生物的体细胞染色体数目和 DNA 含量即 C 值对每种生物而言是一个常数，因此基因组大小可以看作是一个特性参数，可用来评估生物体的生物特性，是比较和进化基因组学研究的基础，研究

表明，基因组作为一种特殊的结构组分，决定细胞核的特性，同时影响生物学特征，如细胞的大小[10]、生物体的寿命[11]、物种濒危程度[12]以及细胞的代谢率[13]等，因此研究基因组大小对生物体的生理特性具有重要的意义。本实验对黄芪基因组大小进行测定，为豆科植物基因组大小变化规律提供参考。据报道，豆科中的菜豆 *Phaseolus vulgaris* Linn. 基因组为 587 Mb，大豆 *Glycine max* (Linn.) Merr. 基因组为 1 100 Mb，绿豆 *Vigna radiata* (Linn.) Wilczek. 基因组为 543 Mb。基于第二代高通量测序技术的黄芪基因组大小为 1 456 Mb，与大豆的基因组大小相近，比其他几种豆科植物基因组大。

目前基因组大小的检测方法有几种，包括实时荧光定量 PCR 法、显微分光光度计法、孚尔根光密度测量法、流式细胞术[14]。其中流式细胞术综合了激光技术、计算机技术、半导体技术、流体力学、细胞化学等各学科知识，能在短时间内作用于大量群体细胞，且操作方便、测定快速，是目前应用最为广泛的方法，如药用植物黄芩[15]、茅苍术[16]、鬼针草[17]等基因组大小的测定。然而，由于所采用的裂解液的化学成分、材料类型、内标的选择、处理环境等因素不同，同一物种的基因组大小测定结果可能会有所不同。Bennett 等[3]认为内标的选择对流式细胞仪测定结果有着重要的影响，应选取细胞核容易提取的植物，Bai 等[18]认为内标应选择应尽量不含有影响细胞核提取和 PI 染色的化合物，同时要求与待测物种的基因组大小接近，且较易区分[19]。具备这些条件的内标可以避免非线性误差的存在。

本研究采用流式细胞术和 K-mer 分析估测黄芪基因组大小、杂合度、GC 含量等结果，为该物种的全基因组测序研究提供参考数据。综合流式细胞术和 K-mer 分析可知，黄芪基因组属于大型高杂合基因组，这一因素导致基因组组装完整性不高，单纯用二代测序数据无法得到理想的组装结果，如本研究进行的初步组装得到的 Contig 数目在 450 万以上，Scaffold 数目 200 万以上，说明组装碎片太多，效果较差。目前三代测序发展迅速，测序成本不断下降，对于黄芪基因组的测序，可以考虑使用 2 种或多种测序方法互补和联用，或者使用纯三代测序策略。

参考文献

[1] 中国药典 [S]. 一部. 2015.

[2] Swift H. The constancy of desoxyribose nucleic acid in plant nuclei [J]. *Proceed Nat Acad Sci*, 1950, 36(11): 643-654.

[3] Bennett M D, Leitch I J. Nuclear DNA amounts in angiosperms: Progress, problems and prospects [J]. *Annals Bot*, 2005, 95(1): 45-90.

[4] Dolezel J, Bartos J, Voglmayr H, *et al*. Nuclear DNA content and genome size of trout and human [J]. *Cytom Part A*, 2003, 51(A): 127-128.

[5] Liu L, Ma X, Wei J, *et al*. The first genetic linkage map of Luohanguo (*Siraitia grosvenorii*) based on ISSR and SRAP markers [J]. *Genome Nat Res Council Canada*, 2011, 54(1): 19-25.

[6] Li R, Fan W, Tian G, *et al*. The sequence and de novo assembly of the giant panda genome [J]. *Nature*, 2010, 463(7279): 311-317.

[7] Marcis G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers [J]. *Bioinformatics*, 2011, 27(6): 764-770.

[8] Dolezel J, Greilhuber J, Suda J. Estimation of nuclear DNA content in plants using flow cytometry [J]. *Nat Protocols*, 2007, 2(9): 2233-2244.

[9] Michaelson M J, Price H J, Ellison J R, *et al*. Comparison of plant DNA contents determined by feulgen microspectrophotometry and laser flow cytometry [J]. *Ameri J Bot*, 1991, 78(2): 183-188.

[10] Gregory T R. Nucleotypic effects without nuclei: Genome size and erythrocyte size in mammals [J]. *Genome*, 2000, 43(5): 895-901.

[11] Monaghan P, Metcalfe N B. Genome size and longevity [J]. *Trends Genetics Tig*, 2000, 16(8): 331-332.

[12] Vinogradov A E. Selfish DNA is maladaptive: evidence from the plant Red List [J]. *Trends Genetics Tig*, 2003, 19(11): 609-614.

[13] Vinogradov A E. Genome size and extinction risk in vertebrates [J]. *Proceed Biol Sci*, 2004, 271(1549): 1701-1705.

[14] Gregory T R. Animal genome size database [J]. *Noncod DNA*, 2001, 29(8): 1297-1305.

[15] 张琳琳, 曹 博, 白成科. 应用流式细胞术测定药用植物黄芩基因组大小 [J]. 中国农学通报, 2013, 29(25): 130-135.

[16] 倪金菁, 贺 彬, 汪文杰, 等. 流式细胞法测定茅苍术基因组大小 [J]. 中药材, 2015, 38(6): 1153-1156.

[17] 逄洪波, 高 秋, 李玥莹, 等. 利用流式细胞仪测定鬼针草基因组大小 [J]. 基因组学与应用生物学, 2016(7): 1800-1804.

[18] Bai C, Alverson W S, Follansbee A, *et al*. New reports of nuclear DNA content for 407 vascular plant taxa from the United States [J]. *Annals Bot*, 2012, 110(8): 1623-1629.

[19] Doležel J, Bartoš J. Plant DNA flow cytometry and estimation of nuclear genome size [J]. *Annals Bot*, 2005, 95(1): 99-110.