川芎基因组 survey 测序及其特征分析

毛常清[1]，沙秀芬[1,2#]，黄 静[1]，陶 珊[1]，彭 芳[1]，李 群[2]，张 超[1]，袁 灿[1*]

1. 四川省农业科学院 经济作物育种栽培研究所，四川 成都 610300
2. 四川师范大学 生命科学学院，四川 成都 610101

摘 要：目的 利用流式细胞术和高通量测序技术，对川芎 *Ligusticum chuanxiong* 基因组大小和特征进行分析，为川芎全基因组精细测序和分子机制研究奠定基础。**方法** 以已知基因组大小的绿豆和陆地棉作为内标植物，用流式细胞术估算川芎基因组大小。利用高通量测序技术对川芎基因组进行 survey 分析，测算出川芎基因组大小、重复序列、杂合率和 GC 含量等，利用生物信息学对基因组进行预测、注释和基因家族鉴定。**结果** 估算出川芎基因组大小约为 3 058.37 Mb，重复序列为 79.79%，杂合率约 2.16%，GC 含量为 36.32%，川芎基因组呈现高重复、高杂合、基因组庞大等特征。共预测到 34 864 个基因，有 30 737 个基因在功能数据库中被注释，有 53 个基因注释到阿魏酸合成过程中。初步鉴定到 2058 个特异基因家族，2001 个单拷贝基因。**结论** 初步获得了川芎基因组大小、基因组特征、功能基因及基因家族等信息，为进一步川芎全基因组精细测序提供了参考，为阐明川芎阿魏酸等药效成分生物合成提供了基因资源。

关键词：川芎；流式细胞术；高通量测序技术；survery 分析；基因组大小；基因组特征

中图分类号：R286.12 **文献标志码：**A **文章编号：**0253 - 2670(2023)03 - 0907 - 08

DOI: 10.7501/j.issn.0253-2670.2023.03.025

Genome survey and characteristic analysis of *Ligusticum chuanxiong*

MAO Chang-qing[1], SHA Xiu-fen[1,2], HUANG Jing[1], TAO Shan[1], PENG Fang[1], LI Qun[2], ZHANG Chao[1], YUAN Can[1]

1. Industrial Crop Research Institute, Sichuan Academy of Agricultural Sciences, Chengdu 610300, China
2. College of Life Sciences, Sichuan Normal University, Chengdu 610101, China

Abstract: Objective To analyze the genome size and characteristics of the Chuanxiong (*Ligusticum chuanxiong* Hort.) by using flow cytometry and high-throughput sequencing technology, which provides a basis for the detailed sequencing of *L. chuanxiong* genome and the study of molecular mechanisms. **Methods** The genome size of *L. chuanxiong* was estimated by flow cytometry to use Lüdou [*Vigna radiata* (Linn.) Wilczek.] and Ludimian (*Gossypium hirsutum* L.) with known genome size as internal standard plants. High-throughput sequencing technology was used to conduct survey analysis of the *L. chuanxiong* genome. Then, bioinformatics was used to analyze the genome size, repeat sequence, heterozygosity, GC content, gene prediction, annotation and gene family identification and other information of *L. chuanxiong*. **Results** The genome size of *L. chuanxiong* was estimated to be about 3 058.37 Mb, the repeat sequence, heterozygosity rate and GC content were respectively about 79.79%, 2.16%, and 36.32%. It showed that the genome of *L. chuanxiong* was characterized by high repeatability, high heterozygosity and large genome. A total of 34 864 genes were predicted, and 30 737 genes were annotated in the functional database, and 53 genes were annotated in ferulic acid synthesis. A total of 2058 specific gene families and 2001 single-copy gene were identified. **Conclusion** The genome size, genome characteristics, functional genes and gene families of *L. chuanxiong* were initially obtained, which provides a reference for further whole-genome sequencing of *L. chuanxiong*, and also provide some gene resources for elucidating the biosynthesis of ferulic acid and other medicinal components in *L. chuanxiong*.

Key words: *Ligusticum chuanxiong* Hort.; flow cytometry; high-throughput sequencing; survey analysis; genome size; genomic characteristics

收稿日期： 2022-09-14

基金项目：国家中药材产业技术体系（CARS-21）；四川省科技厅重点研发项目（2019YFS0156）；四川省育种攻关（2021YFYZ0012）；四川省财政自主创新专项（2022ZZCX076）

作者简介：毛常清（1994—），女，硕士，研究实习员，主要从事药用植物种质资源保护和遗传育种研究。E-mail: mcq1937@163.com

***通信作者：**袁 灿，助理研究员，主要从事药用植物资源创新、遗传育种与栽培研究。E-mail: schkyjzsxy@163.com

#共同第一作者：沙秀芬（1993—），硕士，主要从事细胞工程与植物资源研究。E-mail: shaxiufen6@163.com

川芎 *Ligusticum chuanxiong* Hort.是伞形科藁本属草本植物，以干燥的根茎入药，始载于《神农本草经》，已有 1500 多年的种植历史，是我国传统的大宗中药材。川芎广泛分布于我国四川彭州、什邡、眉山等地，在云南、贵州、广西、湖北、江西等省也少量引进种植。川芎性温，味辛、微苦，具有活血行气、祛风止痛的功效，其主要化学成分包含挥发油、生物碱、有机酸和多糖等，在临床上广泛地用于治疗心脏病、脑梗死及尿路结石等疾病[1-3]。

基因组序列是研究一个物种遗传背景的基础，随着高通量测技术的逐渐成熟，许多植物基因组序列相继被发表，包括大量药用植物基因组。如通过高通量测序技术成功测算出灵芝、丹参、人参、三七、天麻、穿心莲、黄花蒿、广藿香、铁皮石斛等几十种重要药用植物的基因组大小和特征[4-5]。但我国药用植物种类丰富，约占中药材资源总数的 87%[6]，同大多数药用植物一样，现报道川芎的研究主要集中在化学成分[2]和药理药效机制上[7]，分子生物学方面仅开展了川芎转录组分析和利用通用引物分析其遗传多样性[8-10]，在分子遗传学系统研究上存在较大空白。虽然药用植物基因测序技术的应用为川芎全基因组测序提供了技术基础，但由于川芎基因组结构庞大，遗传背景复杂，直接进行全基因组测序存在一定困难，因此在进行全基因组测序之前，有必要对川芎基因组大小进行调研。

本研究采用流式细胞术和 Illumina Hiseq 2500 高通量测序技术相结合的方式对川芎基因组大小进行估算，对所得的基因组数据进行 K-mer 分析、基因组预测及注释，关注阿魏酸等成分合成途径的基因注释，为进一步全基因组精细测序和药效成分合成分子机制研究提供参考依据和基因资源。

1 材料

分别采集 2 个月左右的绿豆 *Vigna radiata* (Linn.) Wilczek.、陆地棉 *Gossypium hirsutum* L.幼嫩叶片和 1 个月左右川芎幼嫩叶片（种植于四川省农科院经济作物育种栽培研究所基地）用于流式细胞分析，并采集川芎叶片用于基因组测序。绿豆由四川省农业科学院经济作物育种栽培研究所叶鹏盛研究员鉴定为豆科豇豆属植物绿豆，陆地棉由中国农业科学院棉花研究所杜雄明研究员鉴定为锦葵科棉属植物陆地棉，川芎由四川农业大学陈兴福教授鉴定为伞形科藁本属植物川芎。

2 方法

2.1 样品的制备

分别选取绿豆、棉花、川芎幼嫩叶片各 2 份，每份 120 mg，洗净置于预冷的培养皿中，向培养皿中加入 1 mL 预冷的 OttoI 细胞裂解液，快速切碎叶片后，用移液枪上下吹打混匀（避免气泡），所得的提取液用 42 μm 尼龙膜滤过到离心管中，低速冷冻离心后，弃上清液，向沉淀中加入 1 mL 冰浴的 OttoII 缓冲液重悬细胞，放置 4 ℃备用。

采用改良 CTAB 法提取川芎基因组 DNA，使用 NanoDrop 2000C 超微量分光光度计和 1%琼脂糖凝胶电泳检测 DNA 浓度及完整性。

2.2 流式细胞术测定川芎基因组大小

将上述制备好的细胞悬浮液样品中加 50 μL 1 mg/mL RNase，50 μL、50 μg/mL PI（DNA 荧光染料碘化丙啶，预先经 0.22 μm 微孔滤膜滤过，−20 ℃保存），混匀，4 ℃避光染色 10 min。随后用 FACSCalibur 流式细胞仪检测 PI 在 488 nm 激发光下发出的荧光，CellQuest 软件捕捉荧光信号数据，ModFit 软件分析结果。测定基因组大小。

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

2.3 基因组测序和质量评估

构建 270 bp 和 500 bp 的小片段文库，利用 Hiseq2500 测序技术对文库进行双端测序。从文库中随机取 10 000 条单端 read 与 NCBI 数据库中的核苷酸数据库（NT）进行 BLAST[11]比对，判断样本是否被外源物种污染。数据测序由北京百迈客生物科技有限公司完成。

2.4 基因组大小、重复序列和杂合率预测

选取 *K* 值为 21 对基因组大小、重复序列、杂合率进行预测。用程序 Jellyfish 计算 *K*-mer 分布，并通过 *K*-mer 分布曲线初步评估基因组重复序列含量和杂合度。

基因组大小＝*K*-mer 总数/*K*-mer 期望深度值

2.5 基因组初步组装和 GC-Depth 分布分析

利用 SOAPdenovo 进行组装得到 contig，利用双末端信息进行 gap 填充，将无 overlap 关系的 contig 拼接组装成 scaffold，获得含有 N（重复序列）的初级基因组序列。过滤后的 read 比对到已组装好的基因序列上，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 图[12]。

2.6 重复序列分析

对测序所得的数据进行重复序列分析。使用 4 个互补程序 LTR_FINDER[13]、MITE-Hunter[14]，RepeatScout[15]和 PILER-DF[16]构建川芎重复序列文库，随后由 PASTEClassifier[17]分类，并与 Repbase[18]转座因子库结合起来作为最终的库，然后运行软件 Repeat-Masker[19]在最终文库中找到同源重复序列。

2.7 基因预测和注释

基因从头预测，在滤过掉小于 1000 bp 大小的 scaffold 后，用 Genscan 和 Augustus 软件通过拟南芥的训练集预测川芎基因。将预测到的基因比对到非冗余蛋白序列（non-redundant，Nr）、真核生物蛋白直系同源簇（clusters of euKaryotic orthologous groups，KOG）、基因本体（gene ontology，GO）、swiss-prot 和京都基因与基因组百科全书（Kyoto encyclopedia of genes and genomes，KEGG）等数据库进行 BLAST 分析来对预测基因进行注释。然后，通过 KEGG 在线网站（http://www.genome. jp/kegg/）检索 KEGG 途径，使用 Blast2GO 软件处理得到的 Nr 注释结果进行 GO 分类，使用在线网站（http://www.ncbi.nlm.nih. gov/COG/）处理 KOG 注释结果进行 KOG 分类。

2.8 基因家族鉴定及系统发育树构建

使用 ORTHOMCL v2.0.9 将预测到的川芎蛋白序列与丹参 *Salvia miltiorrhiza* Bunge、胡萝卜 *Daucus carota* var. *sativa* Hoffm.、葡萄 *Vitis vinifera* L.和拟南芥 *Arabidopsis thaliana* L.等 4 种植物中氨基酸数目大于 50 的序列汇集到一个蛋白数据库中，通过 blastp 比对获得所有物种蛋白序列之间的相似性关系，*E* 值为 1×10^{-5}，去掉序列一致度小于 30%，覆盖率小于 30%的序列；并对比对结果进行聚类，默认膨胀系数为 1.5。在 ORTHOMCL 结果中检索单拷贝基因家族，利用单拷贝同源基因基于最大似然法（maximum likelihood，ML）进行进化树构建。

3 结果与分析

3.1 流式细胞术检测基因组大小

采用绿豆和陆地棉作为内标植物，用流式细胞仪检测其混合样品的荧光强度。已知绿豆基因组大小为 579 Mb[20]，陆地棉基因组为 2173 Mb[21]，通过 2 个内标植物计算出川芎的基因组大小分别为 525/78×579＝3 897.12 Mb，变异系数（coefficient of variation，CV）为 3.40%（图 1）；613/439×2173＝3 034.28 Mb，CV 为 2.01%（图 2）。

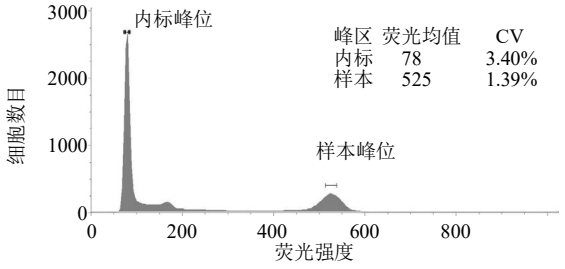


图 1 绿豆与川芎混合样品流式细胞术测定

Fig. 1 FCM determination for mixed samples of *V. radiata* and *L. chuanxiong*

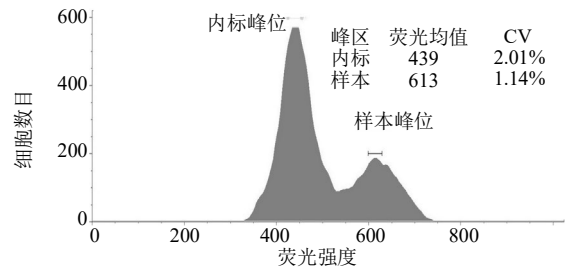


图 2 陆地棉花与川芎混合样品流式细胞术测定

Fig. 2 FCM determination for mixed samples of *G. hirsutum* and *L. chuanxiong*

3.2 川芎基因组测序数据统计及质量评估

使用川芎基因组 DNA 构建 270 bp 的文库，通过 Illumina Hiseq2500 测序平台测序并过滤得到 222.04 Gb 高质量的数据，测序深度为 72.59 X，测序数据 Q20 比例均在 97.59%以上，Q30 比例均在 94.74%以上。随机筛选的 1000 条单端 read 能够比对上 NT 核酸数据库的 read 占总 read 的 7.62%，其中比对上野胡萝 *Daucus carota* L.、细叶藁本 *Ligusticum tenuissimum* (Nakai) Kitagawa 的 read 数分别占比对上 NT 库 reads 数的 61.81%、3.01%，且未发现动物、微生物等异常比对，说明样本不存在污染。

3.3 川芎基因组特征

通过 270 bp 文库数据构建 *K*＝21 的 *K*-mer 分布图（图 3），*K*-mer 深度 62 X 为主峰（由于杂合度较高，本研究对主峰判断参考流式细胞结果），测序得到 *K*-mer 的总数为 189 618 848 178，估算出川芎基因组大小为 3 058.37 Mb，与流式细胞实验中以陆地棉为内标植物测定的结果相近。在主峰对应深度的 1/2 处出现明显的杂合峰，深度为 31 X，说明川芎基因组具有较高的杂合度。进一步根据变异数目

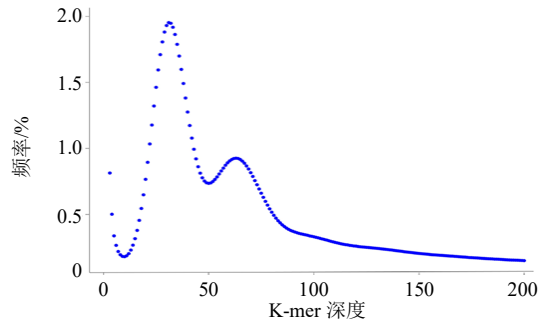


图 3　*K*-mer 分布估算基因组大小

Fig. 3　*K*-mer distribution to estimate genome size

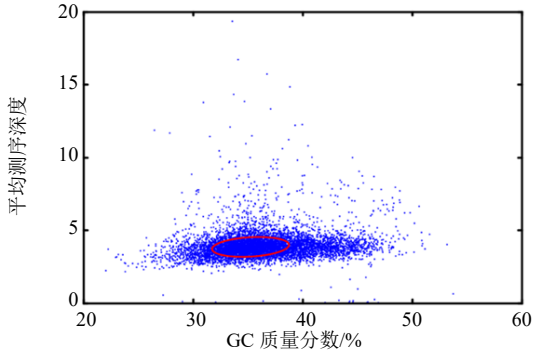


图 4　川芎基因组 GC 含量和平均测序深度

Fig. 4　GC content and average sequencing depth of *L. chuanxiong* genome

占基因组大小的比例即杂合度估算该基因组杂合率，在所有有效数据中检测到每 139 个碱基对中就有 3 个 SNP，估算出川芎基因组具有较高杂合率，约为 2.16%。*K*-mer 分布存在较长拖尾，暗示川芎基因组存在较高的重复率，估算重复序列的基因组大小估计为 2 440.13 Mb，约为川芎基因组的 79.79%。说明川芎属于高重复、高杂合、大基因组等基因组特征的复杂物种。

3.4　川芎基因组组装和 GC 含量分析

使用 222.04 Gb 高质量数据，基于 $K=41$ 组装产生 12 973 787 个 contig 和 8 119 089 个 scaffold。contig N50 为 286 bp，N90 为 131 bp，最大长度达到 26 842 bp，总长度为 3 198 598 874 bp；scaffold N50 为 493 bp，N90 为 191 bp，最大长度为 29 779 bp，总长度为 3 284 536 081 bp。其中，contig 和 scaffold 的 N50 值相对较短，可能是由于川芎高杂合率引起的。通过对组装的 contig 进行 GC 含量的统计，结果显示川芎基因组的 GC 含量约为 36.32%（图 4），说明测序不具有明显的 GC 偏向性，不影响测序分析的准确性。

3.5　川芎基因组重复序列分析

重复序列检测显示其总长度为 2 250 901 770 bp，约为基因组大小的 73.60%，低于 *K*-mer 分析估算的重复序列含量，原因可能是组装效果的限制，导致组装过程中重复序列损失 6.19%。注释上，能够找到明确重复序列元件的总长度约为 2 026.25 Mb。其中，长末端重复序列（long terminal repeated，LTR）是最丰富的重复元件，占基因组的 14.14%，其次是长散在重复元件（long interspersed nuclear elements，LINE），占基因组的 0.49%。SSR 重复序列总长度约 49.41 Mb，占基因组的 1.62%，占重复序列的 2.44%。

3.6　基因预测和注释

总共预测到 34 864 个基因，总长度为 27 532 625 bp。在预测的基因中，查找到 79 130 个外显子，总长度为 18 557 406 bp；79 129 个内含子，总长度为 8 975 219 bp。有 30 737 个基因在功能数据库中能比对到注释信息，Nr 具有最高的注释率（30 584，87.72%）KEGG 具有最低的注释率（9623，27.60%）。在预测的基因中，15 184 个基因被分类为 GO 功能类别；15 598 个基因被分类为 KOG 功能类别；9623 个基因注释到 125 个 KEGG 代谢途径。在 GO 功能类别中，包含分子功能、细胞组成和生物过程（图 5）；在 KOG 功能类别中，通用功能预测的基因最多，其次是翻译后修饰、蛋白质周转、分子伴侣和信号转导机制（图 6）；在参与基因数目最多的前 10 条 KEGG 代谢途径中，注释基因分别参与核糖体代谢、植物激素信号转导、内质网蛋白质、剪接体、碳代谢、氨基酸生物合成、RNA 转运、氧化磷酸化、植物-病原菌互作、淀粉和蔗糖代谢等途径。

3.7　基因家族鉴定及系统发育分析

通过与胡萝卜、丹参、葡萄和拟南芥等物种蛋白序列比对（图 7-A、B），川芎中常见的基因家族中的基因数量小于同科的胡萝卜，每个基因家族的平均基因数量与其他物种相当，但川芎独特基因家族的数量比其他物种的独特基因家族中的基因数量要大得多，共计 2058 个基因家族。所有 5 个物种共有的基因家族为 7112 个，其中 2001 个基因是单拷贝基因，即每个基因家族中只存在一个直系同源基因，可用于系统发育推断和发散时间估计。对 2001 个单拷贝基因利用 ML 构建系统发育树（图 7-C）。

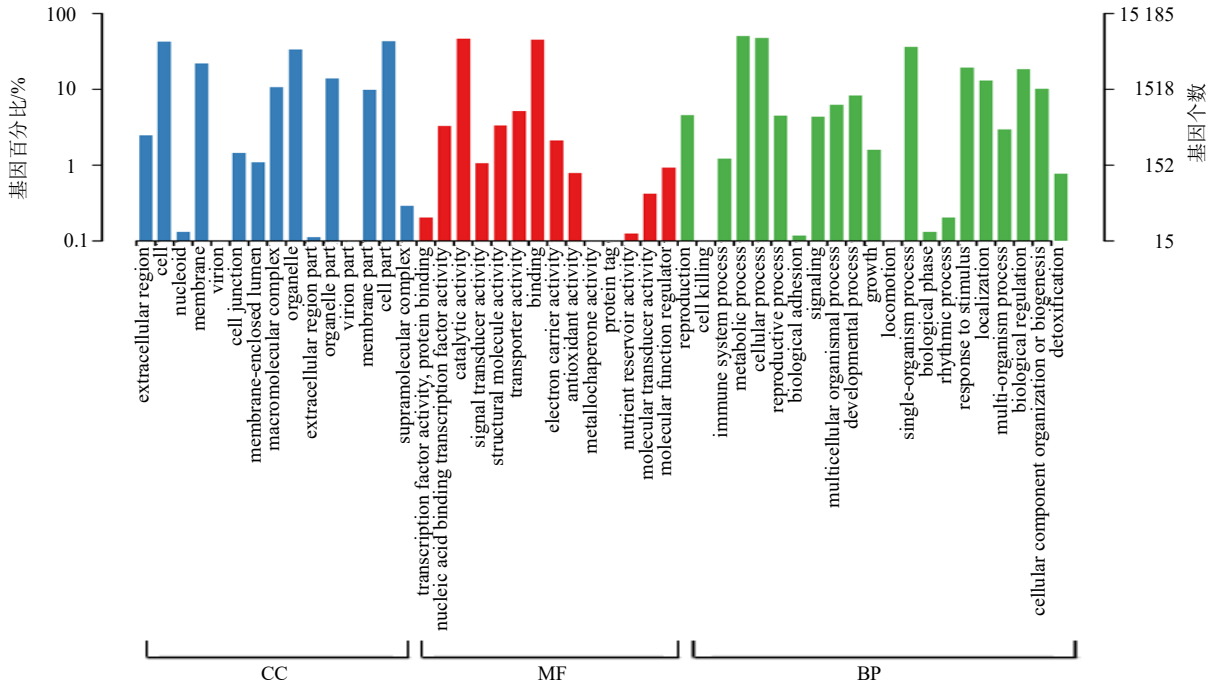


图 5 川芎基因的 GO 注释

Fig. 5 GO annotations of *L. chuanxiong* genes

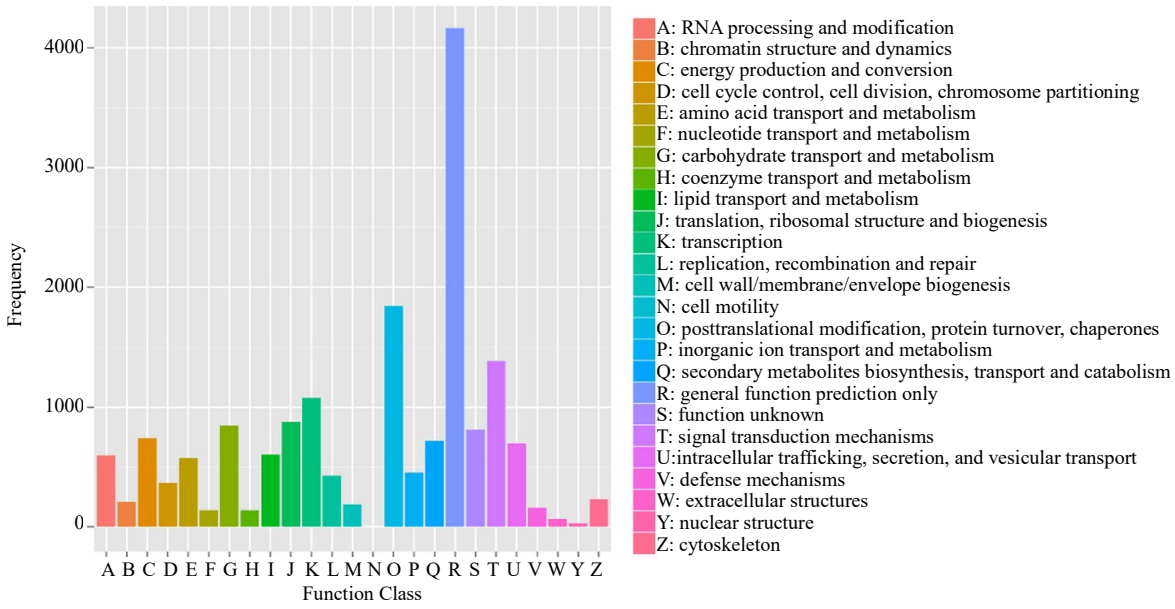


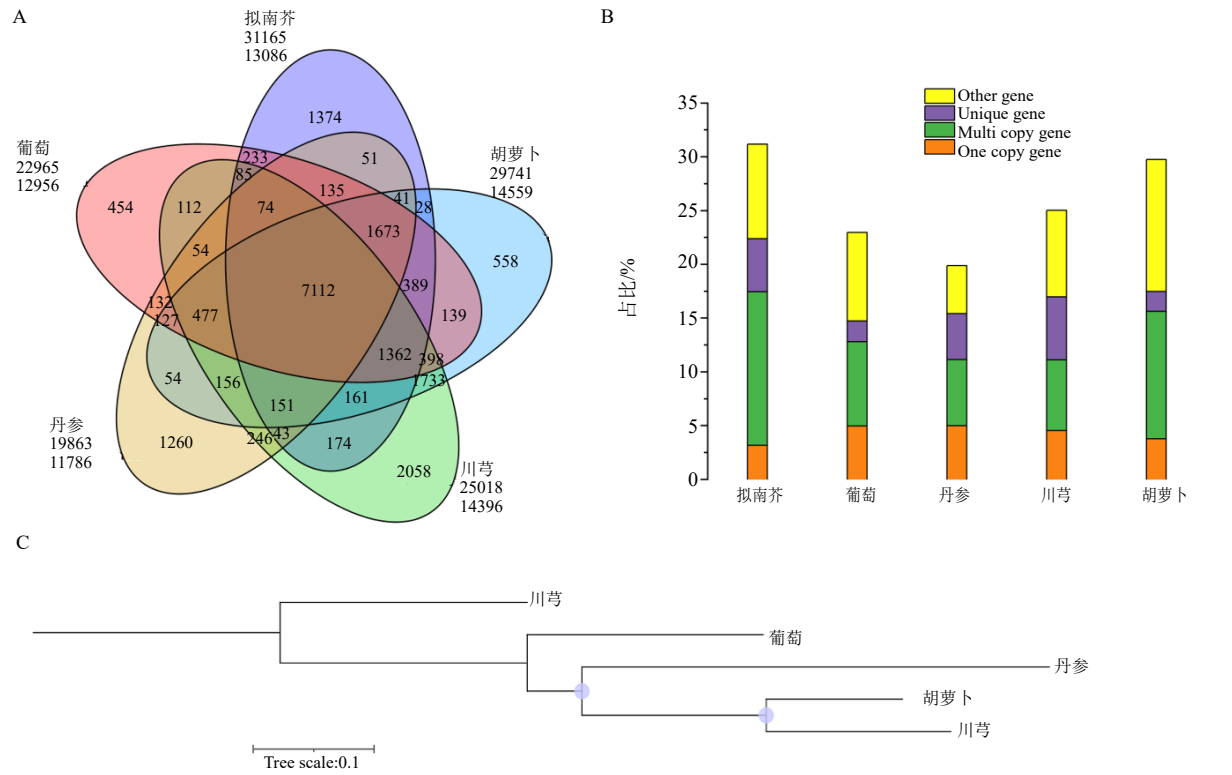
图 6 川芎基因功能注释 KOG 功能分类

Fig. 6 KOG functional classification of *L. chuanxiong* genes

从川芎系统发育树分析，川芎、胡萝卜、丹参、葡萄和拟南芥来自于共同祖先。其中，川芎与胡萝卜最晚与其他物种发生分歧，二者进化分支长度差异最小，分歧时间更短，亲缘关系更近。从遗传变异度上来看，胡萝卜所在的分支最短，遗传变异度最小，进化距离最近，川芎的遗传变异度和进化距离仅次于胡萝卜，丹参的遗传变异度最高，进化距离最远。

3.8 参与阿魏酸生物合成的基因

阿魏酸是川芎的主要药效成分，属于苯丙烷生物合成途径，是木质素合成的中间体。近年来研究者



A-来自 5 种相关植物物种的基因家族的韦恩图　B-不同类别的基因家族的组成　C-系统发育树（蓝色圆圈表示自展支持值 100%）

A-Venn diagram of gene families from five related plant species　B-composition of different categories of gene families　C-phylogenetic tree (blue circle means the bootstrap is 100%)

图 7　川芎基因家族鉴定及系统发育树

Fig. 7　Gene family identification and phylogenetic tree of *L. chuanxiong*

证实了在川芎、当归等伞形科物种中，阿魏酸的合成途径主要包括 COMT 途径和 CCoAMT 途径[9,22-24]，主要参与的酶有苯丙氨酸解氨酶、肉桂酸-4-羟化酶、香豆酸-3-羟化酶、咖啡酸-*O*-甲基转移酶、4-香豆酸辅酶 A 连接酶、莽草酸 *O*-羟基肉桂酰转移酶、奎宁 *O*-羟基肉桂酰转移酶、咖啡酰辅酶 A-*O*-甲基转移酶、肉桂酰辅酶 A 还原酶、醛脱氢酶家族成员 C4。为挖掘川芎中参与阿魏酸生物合成的主要基因，本研究从注释的 KEGG 代谢通路中提取苯丙烷生物合成参考途径（map00940）相关数据，对阿魏酸合成过程的相关基因进行了检索，共有 53 个功能基因覆盖到阿魏酸合成途径中（表 1）。其中，编码咖啡酸-*O*-甲基转移酶、4-香豆酸辅酶 A 连接酶、莽草酸 *O*-羟基肉桂酰转移酶、咖啡酰辅酶 A-*O*-甲基转移酶等酶的基因均有多个同源拷贝存在，预测川芎基因组在进化的过程中某一时间点发生过基因扩张。

4　讨论

基因组包含了一个生物体所有基因的总和，了解生物体基因组信息有助于深入了解生物体遗传、进化、生物合成、次生代谢等全部过程。通过基因组测序技术可以对特定物种基因组进行测序，利用生物信息学方法对测序序列进行拼接和组装，最终获得该物种基因组序列，进而了解其基因组信息[25]。2009 年，陈士林团队[6]首次提出本草基因组计划，此后，越来越多的学者开始在药用植物基因组学研究上投入大量精力。药用植物全基因组水平研究，有助于阐明药用植物活性成分生物合成和代谢调控途径之间的关系，为具有相似有效成分和药理活性的近缘物种间的系统发育关系研究奠定了基础，也为药用植物的遗传育种和基因资源保护提供了重要依据[26]。

本研究利用流式细胞术和高通量测序相结合的方式对川芎基因组进行测算。通过流式细胞术估算出川芎基因组大小分别为3 897.12 Mb 和3 034.28 Mb，通过高通量测序的 *K*-mer 分析，综合 2 个分析结果估算川芎基因组大小为 3 058.37 Mb，属于基因组较大的物种。本研究测得川芎基因组 GC 含量为 36.32%，处于植物基因组 GC 含量应介于 25%～

表 1 覆盖到阿魏酸合成过程的基因

Table 1 Genes covered in the synthesis of ferulic acid

参与的酶	KO 条目	基因数目	Gene ID
PAL	K10775	3	Final.GeMoMaRNA8150_R2.1、Final.GeMoMaGENE22547.1_R0.1、Final.GeMoMaGENE13790.1_R1.1
C4H	K00487	1	Final.GeMoMaGENE18800.1_R0.1
C3H	K09754	2	Final.GeMoMaRNA28548_R4.1、Final.GeMoMaGENE7004.1_R0.1
COMT	K13066	16	Final.GeMoMaRNA19647_R2.1、Final.GeMoMaRNA26573_R2.1、Final.GeMoMaGENE12592.1_R0.1、Final.GeMoMaRNA26573_R3.1、Final.GeMoMaRNA26573_R4.1、Final.GeMoMaRNA19647_R1.1、Final.GeMoMaGENE12592.1_R1.1、Final.GeMoMaRNA19611_R1.1、Final.GeMoMaRNA9578_R1.1、Final.GeMoMaGENE9264.1_R0.1、Final.GeMoMaRNA9577_R0.1、Final.GeMoMaRNA37499_R0.1、Final.GeMoMaGENE10450.1_R0.1、Final.GeMoMaGENE18725.1_R1.1、Final.GeMoMaGENE25.1_R0.1、Final.GeMoMaGENE19164.1_R3.1
4CL	K01904	6	Final.GeMoMaRNA36980_R0.1、Final.GeMoMaAT3G21240.1_R1.1、Final.GeMoMaGENE21063.1_R1.1、Final.GeMoMaGENE21063.1_R0.1、Final.GeMoMaRNA24793_R0.1、Final.GeMoMaRNA5565_R0.1
HCT	K13065	15	Final.GeMoMaGENE8816.1_R1.1、Final.GeMoMaGENE2205.1_R0.1、Final.GeMoMaGENE6989.1_R0.1、Final.GeMoMaGENE16353.1_R0.1、Final.GeMoMaGENE562.1_R1.1、Final.GeMoMaGENE5325.1_R0.1、Final.GeMoMaRNA3684_R2.1、Final.GeMoMaGENE30103.1_R0.1、Final.GeMoMaGENE3442.1_R1.1、Final.GeMoMaGENE562.1_R0.1、Final.GeMoMaGENE30103.1_R1.1、Final.GeMoMaGENE2204.1_R0.1、Final.GeMoMaGENE8247.1_R1.1、Final.GeMoMaGENE21397.1_R1.1、Final.GeMoMaGENE16353.1_R1.1
HQT	K13065	1	Final.GeMoMaRNA3684_R2.1
CCoAMT	K00588	6	Final.GeMoMaGENE1183.1_R0.1、Final.GeMoMaRNA24943_R2.1、Final.GeMoMaGENE1183.1_R1.1、Final.GeMoMaGENE16758.1_R1.1、Final.GeMoMaGENE31028.1_R0.1、Final.GeMoMaAT4G34050.1_R0.1
CCR	K09753	2	Final.GeMoMaRNA22068_R0.1、Final.GeMoMaGENE9615.1_R2.1
ALDH2C4	K12355	2	Final.GeMoMaGENE32557.1_R0.1、Final.GeMoMaGENE32557.1_R1.1

65%的合理范围[27]，说明川芎基因组测序的结果和组装是正确可靠的。通过 *K*-mer 分析，估算出川芎重复序列含量为 79.79%，杂合率约为 2.16%，与地黄[28]、黄芪[29]等药用植物类似，呈现高重复、高杂合的基因组特征，进一步说明川芎基因属于高重复、高杂合、基因组庞大的物种。

在基因预测、注释和基因家族的鉴定中，本研究共预测到川芎编码蛋白基因 34 864 个，远高于其伞形科亲缘关系较近的胡萝卜（32 113 个）[30]，可能是因为组装的都是短片段测序文库，川芎中的基因数量可能被高估了。此外，本研究中测序完成后从头组装中产生的 contig N50 为 286 bp，scaffold N50 为 493 bp，明显较预期短，这与广藿香[4]、罗汉果[12]等药用植物的全基因组调研结果一致。提示对于川芎这种具有复杂基因组特征的物种来说，利用二代高通量测序对其全基因组进行精确测序仍然存在技术难度。因此，提高川芎基因组的测序深度和组装质量，建议后续的研究可采用二代和三代测序技术相结合，并利用全基因组染色体构象捕获技术（high-through chromosome conformation capture，Hi-C），解析全基因组范围内整个染色质 DNA 在空间位置上的关系，获得完整准确的全基因组图谱[31]，得到高质量的川芎基因组序列。

本研究首次利用流式细胞技术和基因组 survey 分析，初步获得川芎基因组大小和结构特征，即基因组庞大、序列重复率高、序列杂合度高，为下一步进行全基因组精细测序奠定基础。本研究中组装产生的大量川芎基因组序列和注释基因为后续分子标记的开发和基因功能研究提供了大量的资源。同时，本研究挖掘了阿魏酸合成途径中的参与基因，对川芎阿魏酸生物合成途径潜在分子机制的初步研究，为进一步研究川芎生物学和选育具有优良药用性状的品种奠定了基础。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] 李霞. 川芎及其提取物的临床应用 [J]. 甘肃医药, 2017, 36(5): 344-346.

[2] 金玉青, 洪远林, 李建蕊, 等. 川芎的化学成分及药理作用研究进展 [J]. 中药与临床, 2013, 4(3): 44-48.

[3] Ran X, Ma L, Peng C, *et al*. *Ligusticum* chuanxiong Hort.: A review of chemistry and pharmacology [J]. *Pharm Biol*, 2011, 49(11): 1180-1189.

[4] He Y, Xiao H T, Deng C, *et al*. Survey of the genome of *Pogostemon cablin* provides insights into its evolutionary

history and sesquiterpenoid biosynthesis [J]. *Sci Rep*, 2016, 6: 26405.

[5] Yan L, Wang X, Liu H, *et al*. The genome of *Dendrobium officinale* illuminates the biology of the important traditional Chinese orchid herb [J]. *Mol Plant*, 2015, 8(6): 922-934.

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

[7] 邱芬, 刘勇, 张蓬勃, 等. 川芎嗪对成体大鼠局灶性脑缺血后皮质和纹状体半暗带细胞增殖的作用 [J]. 中药材, 2006, 29(11): 1196-1200.

[8] 王岚. 用 ISSR 分子标记研究川产道地药用植物川芎的遗传多样性 [D]. 成都: 四川大学, 2007.

[9] 宋涛. 川芎根茎、叶转录组测序及分析 [D]. 成都: 西南交通大学, 2015.

[10] 袁灿, 彭芳, 杨泽茂, 等. 川芎转录组 SSR 分析与 EST-SSR 标记的开发 [J]. 中国中药杂志, 2017, 42(17): 3332-3340.

[11] Altschul S F, Gish W, Miller W, *et al*. Basic local alignment search tool [J]. *J Mol Biol*, 1990, 215(3): 403-410.

[12] 唐其, 马小军, 莫长明, 等. 罗汉果全基因组 Survey 分析 [J]. 广西植物, 2015, 35(6): 786-791.

[13] Xu Z, Wang H. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons [J]. *Nucleic Acids Res*, 2007, 35(Web Server issue): W265-W268.

[14] Han Y J, Wessler S R. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences [J]. *Nucleic Acids Res*, 2010, 38(22): e199.

[15] Price A L, Jones N C, Pevzner P A. De novo identification of repeat families in large genomes [J]. *Bioinformatics*, 2005, 21(Suppl 1): i351-i358.

[16] Edgar R C, Myers E W. PILER: identification and classification of genomic repeats [J]. *Bioinformatics*, 2005, 21(suppl_1): i152-i158.

[17] Wicker T, Sabot F, Hua-Van A, *et al*. A unified classification system for eukaryotic transposable elements [J]. *Nat Rev Genet*, 2007, 8(12): 973-982.

[18] Jurka J, Kapitonov V V, Pavlicek A, *et al*. Repbase Update, a database of eukaryotic repetitive elements [J]. *Cytogenet Genome Res*, 2005, 110(1/2/3/4): 462-467.

[19] Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences [J]. *Curr Protoc Bioinformatics*, 2009, 4: 10.

[20] Arumuganathan K, Earle E D. Nuclear DNA content of some important plant species [J]. *Plant Mol Biol Report*, 1991, 9(3): 208-218.

[21] Li F G, Fan G Y, Lu C R, *et al*. Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution [J]. *Nat Biotechnol*, 2015, 33(5): 524-530.

[22] 耿飒, 徐存拴, 李玉昌. 木质素的生物合成及其调控研究进展 [J]. 西北植物学报, 2003, 23(1): 171-181.

[23] Barrière Y, Ralph J, Méchin V, *et al*. Genetic and molecular basis of grass cell wall biosynthesis and degradability. II. Lessons from brown-midrib mutants [J]. *C R Biol*, 2004, 327(9/10): 847-860.

[24] 刘敬, 李文建, 王春明, 等. 当归中有效成分阿魏酸的生物合成及调控 [J]. 中草药, 2008, 39(12): 1909-1912.

[25] 聂小军. 基于高通量测序技术的小麦和紫茎泽兰基因组学初步研究 [D]. 杨凌: 西北农林科技大学, 2013.

[26] 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.

[27] Aird D, Ross M G, Chen W S, *et al*. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries [J]. *Genome Biol*, 2011, 12(2): R18.

[28] 赵乐, 朱昀昊, 王敏, 等. 基于流式细胞术和基因组 survey 分析的地黄基因组研究 [J]. 中草药, 2021, 52(3): 821-826.

[29] 孙会改, 韦春香, 杨旻啸, 等. 基于流式细胞术和 K-mer 分析的黄芪基因组大小估测 [J]. 中草药, 2019, 50(6): 1448-1452.

[30] Iorizzo M, Ellison S, Senalik D, *et al*. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution [J]. *Nat Genet*, 2016, 48(6): 657-666.

[31] Xie T, Zheng J F, Yang Q Y, *et al*. De novo plant genome assembly based on chromatin interactions: A case study of *Arabidopsis thaliana* [J]. *Mol Plant*, 2015(3): 489-492.

[责任编辑 时圣明]