植物遗传资源学报 2018,19(1):143-149
Journal of Plant Genetic Resources
DOI:10.13430/j.cnki.jpgr.2018.01.016

新疆沙冬青基因组调查测序与基因组大小预测

王　雪,周佳熠,孙会改,禹瑞敏,高　飞,周宜君
(中央民族大学生命与环境科学学院,北京 100081)

摘要:新疆沙冬青是中国荒漠地区代表性常绿阔叶植物,属于第三纪孑遗植物。其极强的逆境耐受性受到了研究者的广泛关注,但由于缺乏基因组序列,分子生物学研究水平进展缓慢。本研究对新疆沙冬青进行了基因组调查测序,共得到 65 Gb 大小的双端测序数据。结合基于 K-mer 分析和流式细胞分析的方法,预测基因组大小、杂合率和 GC 含量等特征,估计基因组大小为 770 ~787 Mb。测序数据拼接构建得到 contigs 的 N50 为 684 bp,总读长为 0.538 Gb;进一步组装后 scaffolds 的 N50 为 12.09 kb,总读长为 0.602 Gb。对拼接数据进行 SSR 分子标记预测,共得到 151858 个 SSR,其中二核苷酸重复单元比例最高为 56.39%,在二核苷酸重复单元中,AT/TA 组成形式占多数。本研究首次报道了荒漠植物新疆沙冬青的基因组特征,为后续基因组学研究提供参考。

关键词:新疆沙冬青;基因组大小;流式细胞分析;K-mer 分析;SSR 分子标记

Genomic Survey Sequencing and Estimation of Genome Size of *Ammopiptanthus mongolicus*

WANG Xue, ZHOU Jia-yi, SUN Hui-gai, YU Rui-min, GAO Fei, ZHOU Yi-jun
(*College of Life and Environmental Sciences, Minzu University of China, Beijing 100081*)

Abstract: Ammopiptanthus *mongolicus* (Maxim. ex Kom.) Cheng f., one of the tertiary relict plants, is a rare evergreen broad-leaved shrub distributed in desert region of central Asia. *A. mongolicus* (Maxim. ex Kom.) Cheng f. grows in environment with extremely high level of drought and freezing stresses, and the plant species can be used as an ideal model for stress tolerance research in plant. Previous studies revealed that *A. mongolicus* (Maxim. ex Kom.) Cheng f. showed high levels of tolerance to various abiotic stresses. Several genes presumed to play important roles in stress tolerance of *A. mongolicus* (Maxim. ex Kom.) Cheng f. were cloned and characterized, including *AnGobs1*, *AnPLD*, *AnMCSD*, *AnBADH*, and *AnAFP*. However, research in *A. mongolicus* (Maxim. ex Kom.) Cheng f. at the molecular level was hindered by the lack of genome information. There were several genome size prediction methods, including Feulgen densitometry, field gel electrophoresis, quantitative reverse transcription PCR (qRT-PCR), flow cytometry analysis, and K-mer analysis. Flow cytometry was the standard method for estimating genome size. K-mer analysis not only predicted genome size accurately, but also provideed more information about genome, such as heterozygosity and GC percentage. In the present study, flow cytometry analysis was performed to predict the genome size of *A. mongolicus* (Maxim. ex Kom.) Cheng f.. Genomic survey sequencing of *A. mongolicus* (Maxim. ex Kom.) Cheng f. was then conducted, and a total of 65 Gb sequencing reads was obtained. K-mer analysis using these sequencing reads was conducted to predict the genome size, heterozygosity and GC content of *A. mongolicus* (Maxim. ex Kom.) Cheng f.. Based on the results of flow cytometry and K-mer analyses, the genome size of *A. mongolicus* (Maxim. ex Kom.) Cheng f. was estimated to be 770-787 Mb. The genome size of *A. mongolicus* (Maxim. ex Kom.)

收稿日期:2017-04-21　**修回日期**:2017-06-28　**网络出版日期**:2017-12-26
URL:http://kns.cnki.net/kcms/detail/11.4996.S.20171226.1514.024.html
基金项目:国家自然科学基金项目(31370356,31670335,31770363);北京市大学生创新训练计划(BEIJ2017110010)
第一作者研究方向为生物化学与分子生物学。E-mail:wangxue@muc.edu.cn
通信作者:高飞,研究方向为生物化学与分子生物学。E-mail:gaofei@muc.edu.cn

Cheng f. was close to that of *Cicer arietinum* L. , and was smaller than that of *Glycine max* L. . The total reads length of contigs was 0. 538 Gb, with N50 of 684 bp, and after further assembly, the N50 of scaffolds was 12. 09 Kb with a total reads length of 0. 602 Gb. The predicted heterozygosity was 0. 0844% and the GC content was 36. 51%. The low heterozygosity of *A. mongolicus* (Maxim. ex Kom.) Cheng f. will facilitate the whole genome sequencing of *A. mongolicus* (Maxim. ex Kom.) Cheng f. , and subsequent gene annotation and comparative genomics study. Simple sequence repeat (SSR) molecular markers were predicted using the assembled genome sequences and 151858 SSRs were obtained. Of all SSR categories, the dinucleotide repeat unit was the largest category, with a percentage of 56. 39% , and the AT/TA components was the dominant dinucleotide repeat unit. Our study reported the genome size prediction of *A. mongolicus* (Maxim. ex Kom.) Cheng f. for the first time and provided a large number of genomic sequences for further research in *A. mongolicus* (Maxim. ex Kom.) Cheng f. . The large number of SSR molecular markers identified in the present study will promote the study of genome mapping, evolutionary biology, and population genetics in *A. mongolicus* (Maxim. ex Kom.) Cheng f. .

Key words: *Ammopiptanthus mongolicus* (Maxim. ex Kom.) Cheng f. ; genome size; flow cytometry; K-mer; SSR molecular marker

新疆沙冬青(*Ammopiptanthus mongolicus* (Maxim. ex Kom.) Cheng f.),又名小沙冬青、矮沙冬青、矮黄花木,源于第三纪,属于豆科(Leguminosae)沙冬青属(*Ammopiptanthus* S. H. Cheng),是中亚温带荒漠区仅有的两种常绿阔叶灌木之一。其自然分布区域狭窄,仅分布于我国的新疆昆仑山和西天山的结合部,克孜勒苏柯尔克孜自治州的乌恰县境内,少量还延伸到苏联[1]。新疆沙冬青起源于古地中海热带地区,是地中海退缩、气候旱化过程中幸存的第三纪豆科孑遗植物,对于研究中亚地区古地理、古气候的变化,以及古植物区系的变迁,均具有相当重要的科学价值[2]。新疆沙冬青适应性非常强,能够耐受-30 ~ 50 ℃的温度条件、干旱胁迫、盐胁迫、风沙侵蚀,保持着正常的生长发育过程和生物、生态学特性[3-4],在维护荒漠地区脆弱的生态系统中起着重要作用,也是用于干旱地区绿化观赏和防风固沙的优秀材料。乌恰县本地的克尔柯孜族人民还将其入药,它的枝叶味苦,含多种生物碱,可供药用,具有祛风湿、活血散痕的效用;还可作杀虫剂和燃料[5]。同时它的根系发达,可在海拔 1800 ~ 2600 m 石质坡地、溪流冲刷沟边及石砾质的河漫滩上生存,逆境耐受性很强,是我国荒漠地区特有的资源植物,也是研究植物逆境调控机制的极佳材料。

对于缺乏基因组数据的非模式资源植物来说,基因组特征的研究是分子机理研究和植物基因资源开发的前提。人类基因组测序计划完成后高通量测序技术飞速发展,利用基于 NGS 的基因组调查测序(GSS, genomic survey sequencing)成为获得未知基因组信息的最佳选择。豆科植物中许多重要经济作物,包括大豆、绿豆、苜蓿、鹰嘴豆等均已完成了全基因组测序[6-8],但对于荒漠地区资源植物的新疆沙冬青的基因组研究还未见报道。植物基因组包含了编码生物性状的全部遗传信息,研究基因组的特征是全面了解基因组信息的基础。基因组大小是指生物单倍体中的 DNA 含量,是基因组多样性的基本参数,具有物种特异性。杂合率和 GC 含量等参数在基因组后续的拼接组装过程中,直接影响了拼接的准确性。本研究利用流式细胞分析和基于K-mer 分析的方法,预测了新疆沙冬青的基因组特征,包含基因组大小、杂合率和 GC 含量等参数,这些数据的获得完善了对于新疆沙冬青基因组的认识,并促进了优秀基因资源的挖掘、新疆沙冬青进化过程的分析以及沙冬青属特异性逆境调控机制的揭示。

此外,开发 DNA 分子标记是构建基因图谱,筛选重要基因和数量性状位点 QTLs 资源的有效途径。相比于 RFLP、AFLP 和 RAPD 等分子标记,简单重复序列(SSR, simple sequence repeat)分子标记具有突变率高、稳定性好且特异性强等优势,是目前最常用的分子标记[9]。SSR 分子标记在基因组中分布广泛,在不同物种间具有高度的多态性,因此在研究植物的进化关系中有重要作用。本研究对新疆沙冬青的 SSR 序列进行了筛选和类型统计,为进一步开发全基因水平的 SSR 分子标记,开展功能基因组研究奠定了基础。

1 材料与方法

1.1 试验材料

新疆沙冬青的种子采自新疆维吾尔自治区乌恰县,其种皮坚硬,经由 65% 的浓硫酸处理 20 min 后,

萌发并种植于中央民族大学生命与环境科学学院植物培养室中生长备用。玉米品种郑 58 的种子取自中国农业科学院生物技术研究所,经双氧水消毒 10 min 后浸泡在饱和硫酸钙溶液中,待到萌芽后转入实验室土壤环境中生长备用。

1.2　试验方法

1.2.1　流式细胞分析　采用美国 BD 公司生产的 FACSCalibur 流式细胞仪进行基因组大小检测,并使用 CellQuest 软件捕捉荧光信号数据,使用 ModFit 软件分析结果。

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、过滤、收集滤液,1000 r/min 离心 5 min 后,弃上清,收集细胞沉淀。将玉米和新疆沙冬青样品分别进行细胞核提取后,混匀用 PI(Propidium iodide,碘化丙啶)染液对细胞核 DNA 进行荧光标记,在避光染色 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植物材料作为对照,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基因组大小判断待测样基因组大小。

1.2.2　K-mer 分析　采用改良 CTAB 法提取新疆沙冬青新鲜叶片的基因组 DNA,将样品随机打断为 350 bp 的插入片段,构建测序 DNA 文库,利用 Illumina Hiseq Xten PE 150 平台进行高通量双末端(Paired-End)测序,得到的测序结果进行过滤,去除低质量 reads,后通过基于 K-mer 分析的方法进行基因组特征预测,包括基因组大小和杂合率等。选取 K 值为 17 进行预测分析,即对测序数据进行 17 nt 的连续分割,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,即可从所有测序数据中统计 K-mer 频数分布,计算获得 K-mer 深度估计值,作 K-mer 分布曲线。用公式基因组大小 = K-mer 总数/K-mer 期望深度估计基因组大小,并可通过曲线的拖尾现象进行基因组重复序列的估计。

利用软件 genomescope 进行基因组杂合率的分析,利用参考基因组中特异区域的杂合 K-mer 数和纯合 K-mer 数、重复序列的 K-mer 覆盖度等负二项因素的综合模型来描绘 K-mer 分布曲线,通过杂合峰值和纯合峰值比例来确定基因组的杂合率。

1.2.3　基因组组装拼接　利用 soapdenovo 对高通量测序数据进行拼接,为寻找合适的 K-mer 取值,首先设定 K 值为 30、40 和 45 后进行数据拼接测试,选取拼接结果 N50 长度适宜的 40 作为 K-mer 值。去除 reads 小于 100 bp 的低质量测序序列,将过滤后的 clean reads 构建 contig 并进一步拼接组装为 scaffold,获得含有 N 的初级基因组序列。

1.2.4　SSR 分析　利用 Perl 语言脚本 MISA 进行基因组的 SSR 分析,找到基因组序列中的所有微卫星重复序列,统计 SSR 的类型、数量、长度及在 scaffold 的位置、起止位点等特征。在运行 MISA 之前需要对寻找的微卫星重复单元进行参数自定义设定,编辑 misa. ini 文件,参数设定如下:单核苷酸单元的重复数≥16,二核苷酸单元的重复数≥6,三核苷酸单元和四核苷酸单元重复数≥5,五核苷酸单元的重复数≥4,六核苷酸单元的重复数≥3。

2　结果与分析

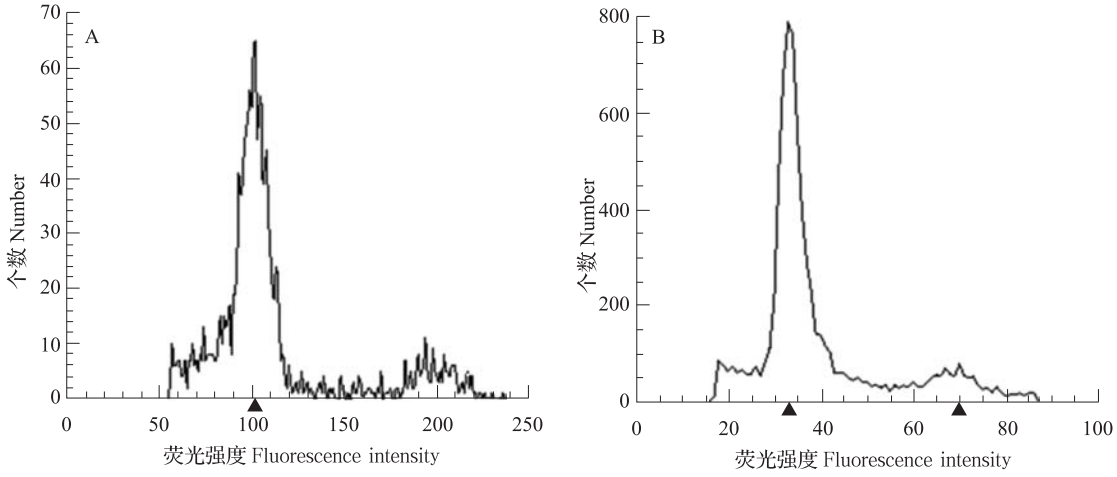
2.1　流式细胞分析预测基因组大小

将新疆沙冬青叶片样品的细胞核进行 PI 染色后,采用流式细胞仪获取荧光信号,荧光染料 PI 能够均匀地嵌入 DNA 双链中,且其嵌入量与 DNA 含量成正比,因此可以通过荧光信号强度对样品中的 DNA 进行相对定量。对玉米和新疆沙冬青基因组进行单独检测,发现玉米基因组大小测定峰(图 1)与新疆沙冬青基因组大小测定峰(图 1)并无重叠,表明两种混合样品的区分度高,不会出现荧光信号混淆现象,也表明利用玉米品种郑 58 作为内标基因组计算新疆沙冬青基因组大小是可行的。

对玉米和新疆沙冬青混合样品的 PI 发射荧光强度进行测定分析可知,玉米 78.75% 的细胞处于 G1 期,G2 期细胞占 6.53%(图 2),新疆沙冬青 76.53% 的细胞为 G1 期,7.81% 的细胞处于 G2 期(图 2)。已知玉米基因组大小为 2.3 Gb,根据流式细胞检测结果比较玉米和新疆沙冬青样品 G1 期细胞荧光强度的倍数关系,可计算出新疆沙冬青的基因组大小为 770 Mb。

2.2　测序数据的统计

采用 Illumina Hiseq 高通量测序平台对新疆沙冬青进行基因组调查测序,过滤掉低质量数据后,共得到 65 Gb 大小的双端测序数据(表 1)。其中经高通量测序后共得到 68 Gb 的测序数据,去除低质量测序序列、接头污染序列等低质量数据后,得到去冗余的 clean reads 共占到原始测序数据的 97.7%,后利用这 65 Gb 的高质量数据进行新疆沙冬青基因组的初步组装和分析。



A:玉米样品结果;B:新疆沙冬青样品结果

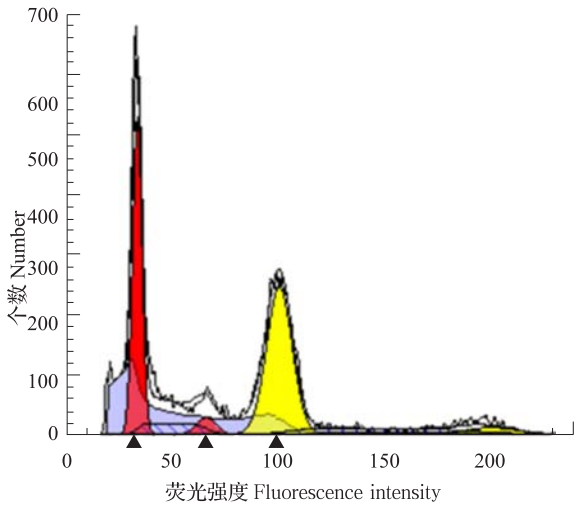
横坐标是 DNA 染料 PI 收到波长为 488 nm 的激光激发后所放出的荧光强度,即通过线性关系可检测细胞中 DNA 含量;纵坐标为细胞数,下同

A:Mazie sample,B:*A. mongolicus*(Maxim. ex Kom.) Cheng f. sample

The abscissa is the fluorescence intensity emitted by the DNA dye PI at the wavelength of 488 nm after laser excitation, measuring the DNA content in cells by linear relationship, the ordinate is the cell number, the same as below

图 1 玉米样品郑 58 和新疆沙冬青样品流式细胞分析结果

Fig. 1 The *C*-value measurement of mazie Zheng 58 and *A. mongolicus*(Maxim. ex Kom.) Cheng f. sample



红色为待测样新疆沙冬青;黄色为标样玉米

In red as the test sample *A. mongolicus*(Maxim. ex Kom.) Cheng f., in yellow as the standard sample maize

图 2 玉米和新疆沙冬青混合样品流式细胞分析结果

Fig. 2 The *C*-value measurement of the mixed samples of *A. mongolicus*(Maxim. ex Kom.) Cheng f. and maize

表 1 GSS 数据量统计

Table 1 The statistic of GSS

统计项 Statistical item	数据量 Data
Raw reads 数 Raw reads number	492112308
Raw bases 数 Raw bases number	71233011700
Clean reads 数 Clean reads number	480797482
Clean bases 数 Clean bases number	69654988450
Clean reads 占比(%) Clean reads rate	97.70
低质量 reads 占比(%) Low-quality reads rate	1.75
接头污染 reads 占比(%) Adapter polluted reads rate	0.55
含 N 比例大于 5% reads 占比(%) Content of N reads rate >5%	0

2.3 K-mer 分析预测基因组大小和杂合率

将全部的 65 Gb 测序数据利用 jellyfish 进行 K-mer 分析,K 值取 17,得到结果如图 3。可见在深度为 52 时分布曲线有一峰值,对应横坐标即为 K-mer 的期望深度。统计可知 K-mer 总数为 39.99 Gb,可以利用公式:基因组大小 = K-mer 总数/K-mer 期望深度,从而估算基因组大小为 787 Mb。此结果

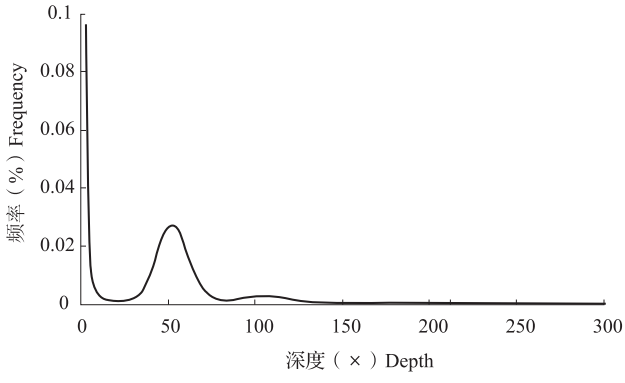


图 3 K-mer 分布曲线

Fig. 3 K-mer(K = 17) analysis for estimating the genome size of *A. mongolicus*(Maxim. ex Kom.) Cheng f.

与流式细胞分析预测出来的基因组大小结果基本一致,说明新疆沙冬青的基因组大约为 770 ~ 787 Mb,基因组调查测序深度约可达 84.4 ×。通过 Genomescope 软件对测序数据的分析,得到新疆沙冬青的基因组杂合率为 0.0844%,重复序列含量为 62.36%,杂合率很低,数据拼接难度较小。

2.4　基因组测序的组装和拼接

采用 Soap denovo 软件对所有测序的 clean reads 进行 denovo 拼接组装,选择 K-mer 值为 40 得到最优的拼接效果,即 N50 值最恰当(N50 为将 reads 按照从长到短排列后依次相加,当为总长度一半时最后加上的 reads 长度)。组装结果如下表 2 所示,构建得到 contigs 的 N50 为 684 bp,总读长为 0.538 Gb;进一步组装后 scaffold 的 N50 为 12.09 kb,总读长为 0.602 Gb。新疆沙冬青基因组高通量测序数据经过拼接组装后共得到 617 Mb 数据,与预测到的基因组大小 770 Mb 相差小,拼接能够覆盖 78.40% ~ 80.13% 的基因组,表明数据组装效果较好。

新疆沙冬青基因组中 GC 含量为 36.51%,与拟南芥中 GC 含量(35.97%)最为接近。与苹果、百脉根、苜蓿等 GC 平均含量低于 30% 的植物相比 GC 含量较高,同时其值也低于与水稻、玉米、高粱、短柄草等的 GC 平均含量,这些植物的 GC 平均含量高于 40%,因此,与多数基因组数据已释放的植物相比属于 30% ~47% 这个中间范围[10]。而研究表明过高(>65%)或者过低(<25%)的 GC 含量会造成高通量测序时的测序偏差错误,甚至影响拼接数据的准确性[11]。

2.5　新疆沙冬青的 SSR 序列分析

利用 MISA 脚本在新疆沙冬青的基因组中共找到 151858 个 SSR,将得到的分析结果按照重复单元的核苷酸数量进行分类统计:单核苷酸 SSR 有 20718 个,占总数的 13.64%;二核苷酸 SSR 有 85630 个,占总数的 56.39%;三核苷酸 SSR 有 27184 个,占总数的 17.90%;四核苷酸 SSR、五核苷酸 SSR 和六核苷酸 SSR 分别占总数的 2.04%、2.24%、7.79%。可知,二核苷酸重复是新疆沙冬青基因组中出现的主要形式,同时四核苷酸重复占比率最小。在二核苷酸重复单元中,多数为 AT/AT,占核苷酸重复总数的 39.84%,而 CG/GC 仅占总数的 0.04%。在含有单核苷酸重复单元 SSR 中,以 A/T 的重复组成为主(表 3)。

表 2　新疆沙冬青基因组 *de novo* 拼接结果

Table 2　Statistics of the assembled genome sequences

统计量 Statistical item	重叠群 Contigs	拼接群 Scaffolds
序列数 Number of sequences	1662250	356137
总长度(bp) Total length	577998698	646676084
N50 长度(bp) N50 length	684	12381
N90 长度(bp) N90 length	131	995
大于 500 bp 序列数 Number of sequences ≥500 bp	227122	111153
大于 1 kb 序列数 Number of sequences ≥1 kb	114998	80231
大于 10 kb 序列数 Number of sequences ≥10 kb	134	16975
拼接群中重叠群数 Number of contigs in scaffolds	1441573	—
ACGT 总占比(%) Total rate of ACGT	—	95.64
CG 含量(%) CG content	—	36.51

表 3　新疆沙冬青中 SSR 统计结果

Table 3　Simple sequence repeat types deteacted in *A. mongolicus* sequences

SSR 重复单元 SSR repeat type		SSR 数量 Number of SSR	重复单元的比例(%) Rate of repeat type	SSR 重复单元 SSR repeat type		SSR 数量 Number of SSR	重复单元的比例(%) Rate of repeat type
单核苷酸	A/T	18683	12.30		ATC/ATG	1744	1.15
Mononucleotide	C/G	2035	1.34		CCG/CGG	136	0.09
二核苷酸	AC/GT	7848	5.17	四核苷酸	AAAC/GTTT	89	0.06
Dinucleotide	AG/CT	17218	11.34	Tetranucleotide	AAAG/CTTT	312	0.21
	AT/AT	60497	39.84		AAAT/ATTT	1170	0.77
	CG/CG	67	0.04		AACT/AGTT	153	0.10
三核苷酸	AAC/GTT	2214	1.46		AAGG/CCTT	74	0.05
Trinucleotide	AAG/CTT	5048	3.32		AATC/ATTG	25	0.02
	AAT/ATT	13176	8.68		AATG/ATTC	108	0.07
	ACC/GGT	1804	1.19		AATT/AATT	249	0.16
	ACG/CGT	85	0.06		ACAT/ATGT	343	0.23
	ACT/AGT	461	0.30		ACTC/AGTG	64	0.04
	AGC/CTG	486	0.32		AGAT/ATCT	246	0.16
	AGG/CCT	2030	1.34		ATCC/ATGG	28	0.02

3 讨论

3.1 基因组大小的预测方法

基于检测手段的基因组大小预测方法有很多,包含孚尔根光密度测量法、脉冲场凝胶电泳、实时荧光定量 PCR 技术、流式细胞分析等。自 1983 年 D. W. Galbraith等[12] 利用流式细胞术测定植物的 DNA C 值以来,流式细胞分析已经成为测定植物基因组大小的标准方法。随着高通量测序技术 NGS 的高速发展,研究者们探索未知物种基因组的步伐加快,需要更加便捷高效的方法来探索基因组特征。而基于 K-mer 分析的生物信息学预测基因组特征方法的出现,对于探索未知物种的基因组具有很大帮助。这种方法已在许多物种中成功应用,包括玉米[13]、刺梨[14]、龙须菜、栽培黄瓜、蜡果杨梅等,发现利用基于 K-mer 分析预测基因组特征的方法与传统的基因组大小鉴定方法的预测结果基本一致。本试验所获得的流式细胞分析结果为 770 Mb,同时基于K-mer 分析的结果为 787 Mb,表明这两种方法结果一致可信,可适用于新疆沙冬青等非模式植物的基因组大小预测。

3.2 基因组大小与豆科植物的比较

基因组大小是指生物单倍体中 DNA 的总含量,是植物最基本最重要的生物多样性参数。在每个物种中细胞的染色体数和 DNA 含量即 C 值是固定的,可用于评估生物体的生物学特征,是比较和进化基因组学研究的基础,在物种的鉴定、分类和进化等方面有重要意义[15]。

对所有的陆生植物的基因组大小进行统计,研究者们发现不同物种的基因组大小相差很大。在被子植物中已知最小基因组的植物为螺旋狸藻[16],基因组大小为 63 Mb,而百合科的四倍体贝母[17]有最大基因组为 127 Gb,两者相差约为 2000 倍。新疆沙冬青所属的豆科植物的基因组大小被报道为 472 ~ 1100 Mb,目前豆科的百脉根基因组最小。本试验的研究材料新疆沙冬青基因组大小大约为 770 Mb,与鹰嘴豆的数据较接近[18]。研究基因组大小对了解生物的性状特征、生理规律具有重要意义。新疆沙冬青基因组大小测定的完成,为研究豆科植物基因组大小变化规律提供了参考依据。

据研究基因组大小的变化与基因组中基因数目的变化关联不大,而是由于物种间反转录子类序列和非编码序列的大小变化导致的。这些非编码区域的变化包括内含子大小、转座因子的拷贝数、SSR 数和假基因数等,据报道在苹果与梨的基因组比较研究中发现影响基因组大小不同的主要因素是转座子类重复序列的变化,而两个物种中编码序列相似度很高,不是基因组大小的影响因素[19]。本试验共鉴定到新疆沙冬青中的 SSR 含量为 151858 个,且以二核苷酸重复单元为主,而刺梨中鉴定到 SSR 有 167859 个,且以单核苷酸重复单元为主,但两种植物的基因组大小却相差 1.88 倍,推测基因组大小变化与 SSR 的类型组成关系更密切。

3.3 基因组杂合率

参考测序数据的杂合度有利于寻找合适的基因组拼接方法,根据杂合度大小通常将基因组划分为微杂合基因组(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%)[20],杂合率过高会影响拼接质量。本试验中预测到的基因组杂合率为 0.0844%,杂合率较低,适用于 WGS 拼接策略。新疆沙冬青测序数据经组装拼接后为 617 Mb,依据基因组大小预测结果,拼接覆盖率可以达到 78.40% ~80.13%。新疆沙冬青基因组的杂合程度低,可能与其作为第三纪孑遗植物的种群大小和分布地区有关。由于生长环境极端恶劣,地理区域与其他亲缘物种的隔离效应,导致了新疆沙冬青种群小,结构简单,仅分布于我国新疆乌恰县境内,分布地区狭窄。种群内及种群间基因交流程度低,造成了基因资源的稳定性,表明新疆沙冬青的基因资源较为古老珍贵,并有极强的环境适应能力,是进一步筛选优秀基因资源,研究植物抗逆机制的重要植物材料。

3.4 SSR 标记特征

SSR 简单重复序列作为以 PCR 为基础的分子标记,具有成本低廉、特异性强的优点,广泛应用于基因定位、遗传作图、品种鉴定及辅助育种等方面[21]。对于非模式植物新疆沙冬青 SSR 标记的开发,有利于全面地了解基因组结构,研究重要基因功能。本试验中发现新疆沙冬青中 SSR 主要以二核苷酸重复单元为主,其次为单核苷酸和三核苷酸。R. Chakraborty 等[22]学者在人类基因组序列的突变频率研究中发现二核苷酸重复单元的 SSR 突变率最高。因此新疆沙冬青中检测到的丰富且突变率高的二核苷酸 SSR 是发展基因组分子标记的重要资源,利于建立非模式植物基因组全面的 SSR 标记系统。

本研究首次报道了荒漠代表植物新疆沙冬青的基因组基本特征,而豆科作为在被子植物中仅次于菊科及兰科的第三大科,分布于全世界,其中许多代

表性种属包括大豆、国槐、紫荆等均已经得到了深入的研究,并广泛应用于日常生活中。随着研究者们将目光转向环境恶劣地区的资源植物,新疆沙冬青的独特适应机制则引起了研究者们的兴趣,但目前关于基因组分子水平的特征研究未见报道。本试验中共拼接得到 617 Mb 基因组序列并鉴定到 151858 个 SSR 分子标记,对新疆沙冬青的基因组特征有了初步的探索,是进一步构建高密度遗传图谱以及研究逆境条件下新疆沙冬青特殊调控机制的基础。

参考文献

[1] 潘伯荣,余其立,严成. 新疆沙冬青生态环境及渐危原因的研究[J]. 植物生态学报,1992,16(3):276-282

[2] 杨期和,葛学军,叶万辉,等. 矮沙冬青种子特性和萌发影响因素的研究[J]. 植物生态学报,2004,28(5):651-656

[3] 尹林克,王烨. 沙冬青幼苗生长规律初步分析[J]. 干旱区研究,1990,7(1):61-64

[4] 尹林克,王烨. 沙冬青属植物花期生物学特性研究[J]. 植物学报,1993,10(2):54-56

[5] 姬玉英,刘红波. 新疆特有濒危珍稀植物——矮沙冬青[J]. 新疆林业,2007(4):41

[6] Schmutz J, Cannon S B, Schlueter J, et al. Genome sequence of the palaeopolyploid soybean[J]. Nature, 2010, 463(7278): 178-183

[7] Young N D, Oldroyd G E D, Geurts R, et al. The Medicago genome provides insight into the evolution of rhizobial symbioses[J]. Nature, 2011, 480(7378): 520-524

[8] Yang J K, Kim S K, Kim M Y, et al. Genome sequence of mungbean and insights into evolution within *Vigna* species[J]. Nat Commun, 2014, 5(5543): 5443

[9] Song Q J, Shi J R, Singh S, et al. Development and mapping of microsatellite(SSR) markers in wheat[J]. Theor Appl Genet, 2005, 110(3): 550-560

[10] Shangguan L, Han J, Kayesh E, et al. Evaluation of genome sequencing quality in selected plant species using expressed sequence tags. [J]. PLoS One, 2013, 8(7): e69890

[11] 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): 18

[12] Galbraith D W, Harkins K R, Maddox J M. Rapid flow cyto-metric analysis of the cell cycle in intact plant tissues[J]. Science, 1983, 220(4601): 1049-1051

[13] Mikel M A. Genome assembly and analysis of the maize(*Zea mays* L.) iodent inbred PH207[C]// International Plant and Animal Genome Conference Xxii. San Diego, CA: Plant & Animal Genome, 2014

[14] Lu M, An H, Li L. Genome survey sequencing for the characterization of the genetic background of *Rosa roxburghii* Tratt and leaf ascorbate metabolism genes[J]. PLoS One, 2016, 11(2): e0147530

[15] Wang G, Meng Y, Yang Y. Genome size variation among and within Ophiopogoneae species by flow cytometric analysis[J]. Braz J Bot, 2017, 40(2): 1-9

[16] Greilhuber J, Borsch T, Muller K, et al. Smallest angiosperm genomes found in Lentibulariaceae, with chromosomes of bacterial size[J]. Plant Biol, 2006, 8(6): 770-777

[17] Leitch I J, Soltis D E, Soltis P S, et al. Evolution of DNA amounts across land plants(Embryophyta)[J]. Ann Bot, 2005, 95(1): 207-217

[18] Varshney R K, Song C, Saxena R K, et al. Draft genome sequence of chickpea(*Cicer arietinum*) provides a resource for trait improvement[J]. Nat Biotechnol, 2013, 31(3): 240-246

[19] Wu J, Wang Z, Shi Z, et al. The genome of the pear(*Pyrus bretschneideri* Rehd.)[J]. Genome Res, 2013, 23(2): 396

[20] 伍艳芳,肖复明,徐海宁,等. 樟树全基因组调查[J]. 植物遗传资源学报,2014,15(1):149-152

[21] 蔡斌,李成慧,姚泉洪,等. 葡萄全基因组 SSR 分析和数据库构建[J]. 南京农业大学学报,2009,32(4):28-32

[22] Chakraborty R, Kimmel M, Stivers D N, et al. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci[J]. Proc Natl Acad Sci USA, 1997, 94(3): 1041-1046