东乡野生稻叶绿体基因组拼接及系统进化分析

林张翔，王营营，付菲，叶楚玉，樊龙江
（浙江大学农业与生物技术学院农学系，浙江省作物种质资源重点实验室，杭州 310058）

摘要：叶绿体基因组序列对于研究植物物种的起源、进化演变及不同物种间的亲缘关系等具有重要意义。高通量测序技术的快速发展，推动了植物叶绿体基因组的测序工作。但传统的叶绿体基因组测序方法需要建立在分离纯化叶绿体 DNA 的基础上，操作繁琐，耗时较长。为了优化叶绿体基因组 DNA 序列的获取和拼接方法，以东乡野生稻（Oryza rufipogon）嫩绿叶为材料，不需分离叶绿体 DNA，利用高通量测序获得的全基因组短序列（reads）及叶绿体基因组高度保守的特性，与参考序列进行比对，从而组装拼接出叶绿体 DNA 序列，并同时利用生物信息学手段和 PCR 扩增进行补洞。最终获得东乡野生稻完整叶绿体基因组序列，大小为 134537bp，大（LSC）、小（SSC）单拷贝区和反向互补重复区（IR）大小分别为 80585bp、12346bp 和 20803bp，共注释叶绿体基因 152 个。基于获取的东乡野生稻及其他叶绿体基因组序列，通过构建进化树分析，结果显示在禾本科中水稻与麻竹（Dendrocalamus latiflorus）和黍亚科（Panicoideae）亲缘关系最近，粳稻与中国普通野生稻的亲缘关系较近，粳稻与籼稻并非同时驯化出现。

关键词：作物遗传学；东乡野生稻；叶绿体基因组；高通量测序；嫩绿叶
中图分类号：S511.9

Assembly and phylogenetic analysis of Dongxiang wild rice chloroplast genome

LIN Zhangxiang, WANG Yingying, FU Fei, YE Chuyu, FAN Longjiang
(Department of Agronomy, College of Agriculture and Biotechnology, Zhejiang Key Laboratory of Crop Germplasm, Zhejiang University, Hangzhou 310058, China)

Abstract: Complete chloroplast genome sequence is very useful for studying the evolution of species. The rapid development of high-throughput sequencing technology promotes the plant chloroplast genome sequencing. For traditional chloroplast genome sequencing method, it is necessary to isolate and purify the chloroplast DNA before sequencing. Due to low concentration of chloroplast DNA, it is difficult to separate it from nuclear genome DNA. Therefore, chloroplast DNA isolation-based method is tedious and time consuming. This study employed a simple and rapid method for chloroplast genome sequences acquisition without isolation of chloroplast DNA. Based on conservation of chloroplast genomes, the whole genome short reads generated by Illumina Hiseq 2000 were directly used to map against chloroplast reference genomes. Subsequently, the aligned reads were collected and further did de novo assembly. Finally, the chloroplast genome sequence of Dongxiang wild rice was obtained. The chloroplast genome is 134537bp in size, and has a typical quadripartite structure with the large (LSC, 80585bp) and small copy (SSC, 12346bp) regions separated by two copies of an inverted repeat (IRs, 20803bp each) region. In total, 152 chloroplast genes were successfully annotated. The phylogenetic tree of Dongxiang wild rice and 14 Poaceae chloroplast genomes shows that Dongxiang wild rice has a closer relationship with Dendrocalamus latiflorus and Panicoideae. Furthermore, we build a phylogenetic tree based on SNPs of Dongxiang wild rice and other 22 Oryza chloroplast genomes. The result illustrates that indica has a closer relationship with wild rice-I, while japonica are closer to wild rice-III, suggesting that indica and japonica were domesticated during different periods

Key words: crop genetics; donxiang wild rice; chloroplast genome; high-throughput sequencing; fresh green leaf

作者简介：林张翔(1991-),女,硕士,生物信息学
通信联系人：樊龙江(1965-),男,教授,生物信息学. E-mail: fanlj@zju.edu.cn

0 引言

叶绿体是具有半自主遗传体系的细胞器，是绿色植物进行光合作用的重要场所。叶绿体基因组是独立于核基因组外的器官基因组，具有单独的转录和转运系统，平均大小介于120～180kb，是一个环形的双链结构，其 DNA 序列高度保守[1]。叶绿体基因组的大小远小于核基因组，但其在每个细胞的拷贝数介于 1000～10 000，叶绿体基因结构和序列为研究物种进化起源及不同物种之间的亲缘关系提供了重要的资源和信息。

高通量测序技术的快速发展，推动了植物叶绿体基因组的研究。自 1986 年首次获得地钱（*Marchantia polymorpha*）[2]和烟草（*Nicotiana tabacum*）[3]的叶绿体基因组的完整序列以来，叶绿体基因组数据库不断增加充实。截至 2013 年 8 月 9 日，美国国家生物技术中心（The National Center for Biotechnology Information，NCBI）的细胞器基因组数据库（Organelle Genome Resources）（http://www.ncbi.nlm.nih.gov/genomes/）共收录了来自不同植物的 285 条叶绿体基因组。

虽然越来越多的植物叶绿体基因组测序完成，但传统的方法需要首先分离纯化叶绿体 DNA，再对其进行高通量测序或桑格法测序。而由于叶绿体基因组 DNA 含量低，难于和核基因组 DNA 分离，因此传统的方法难度较高并且耗时较长。2009 年，台湾科学家首次利用一种简便的基于 PCR 的方法，不用分离纯化叶绿体 DNA，获得了麻竹（*Dendrocalamus latiflorus*）和绿竹（*Bambusa oldhamii*）[4]的叶绿体基因组，之后又用该方法获得了文心兰（*Oncidium*）[5]的叶绿体基因组序列。随着第二代测序技术的发展，2010 年中国科学研究院利用 Roche GS FLX 测序平台从椰枣树（*Phoenix dactylifera L.*）[6]全基因组测序数据中获得了其叶绿体基因组序列。本研究以东乡野生稻（*Oryza rufipogon*）绿色嫩叶为材料，在前期开展东乡野生稻基因组研究基础上[7]，利用一种优化的叶绿体基因组 DNA 序列获取和拼接方法，获得了野生稻叶绿体基因组的完整序列，并据此序列进行了野生稻的系统进化分析。

1. 材料与方法

1.1 材料与数据

本实验以东乡野生稻（*Oryza rufipogon*）绿色鲜叶为研究材料。材料由中国水稻研究所提供，其 Illumian 高通量测序数据来自本课题组前期研究结果[7]。其它数据包括 NCBI（http://www.ncbi.nlm.nih.gov/）上公布的 7 条稻属不同种的叶绿体基因组序列（NC_017835、JN005833、NC_008155、NC_001320、NC_016927、NC_005973、GU592209），韩斌课题组发布的关于栽培稻起源研究的部分数据（http://www.ebi.ac.uk/ena/）[8]（ERX046456、ERX046479、ERX046720、ERX046828、ERX046846、ERX005522、ERX014265、ERX002911、ERX014455、ERX046332、ERX046915、ERX046902、ERX046903、ERX046905），以及羊茅（*Festuca arundinacea*；NC_011713）、黑麦草（*Lolium perenne*；NC_009950）、剪股颖（*Agrostis stolonifera*；NC_008591）、大麦（*Hordeum vulgare*；NC_008590）、小麦（*Triticum aestivum*；NC_002762）、绿竹（*Bambusa oldhamii*；NC_012927）、短柄草（*Brachypodium distachyon*；NC_011032）、麻竹（*Dendrocalamus latiflorus*；NC_013088）、柳枝稷（*Panicum virgatum*；NC_015990）、玉米（*Zea mays*；NC_001666）、薏苡（*Coix lacryma-jobi*；NC_013273）、甘蔗（*Saccharum offcinarum*；NC_006084）、高粱（*Sorghum bicolor*；NC_008602）、禾本科早期分化物种（*Anomochloa marantoidea*；NC_014062）等 14 个禾本科（Poaceae）不同属的叶绿体基因组序列。

1.2 方法

1.2.1 基因组文库构建及测序

提取足量的东乡野生稻叶片 DNA，用物理方法随机打断成不同长度的 DNA 片段，通过凝胶电泳技术得到长度为 500bp 的 DNA，构建 Illumina Hiseq 2000 平台测序文库，然后在单链 DNA 片段两端加上接头并对其进行扩增，最后对文库进行双末端测序。

1.2.2 原始数据预处理

高通量测序后得到图像文件（.TIF），检测集群强度并对每个集群进行定位，然后对背景噪音程度进行评价，根据评价结果，将图像信息转换为碱基序列，然后将原始数据（raw data）去除接头序列和低质量读序，过滤后得到干净读序（clean data）。

1.2.3 叶绿体基因组拼接

将已知的 7 条水稻叶绿体基因组序列作为参考序列（reference），将上一步得到的干净读序与参考序列联配。所用的软件为 Bowtie2（http://bowtie-bio.sourceforge.net/index.shtml）[9]，参数按软件默认设置，得到所需的 sam 文件。然后用 perl 语言编写脚本，提取干净读序中与参考序列联配上的短序列（reads），生成 fastq 格式文件，用于后续的拼接。

本研究选用 Velvet 软件[10]用于拼接，通过寻找短序列之间的重叠区域（overlap）将高质量的短序列拼接成重叠群序列（contig），然后将所有的短序列定位到拼好的重叠群序列（scaffold）上，再根据 PE（pair-end）关系将重叠群连接成 scaffold 序列。因参数设置对 Velvet 的运行结果有很大的影响，尤其是 K-mer 值和覆盖深度（coverage）的设置。所以实验设置了多个参数进行调试，K-mer 设为 31~61 共 13 个水平，覆盖深度设为 100~400 共 4 个水平，总共设置了 42 组参数，然后利用 perl 脚本计算出每一组参数拼接结果的 N50（覆盖 50%所有核苷酸的最大序列重叠群长度），选取 N50 最大的一组参数。实验最终选用 K-mer 值为 33，覆盖深度设为 100，此时的 N50 最大，为 69 539bp。

因为 scaffold 序列内存在一些未知序列片段（gap），所以实验选用补洞软件 Gapcloser[11](http://soap.genomics.org.cn/soapdenovo.html)进行补洞，得到一条共有序列（consensus），参数按软件默认设置。同时，根据洞两端的序列设计引物进行 PCR 扩增，将扩增得到的产物进行双末端测序，从而对生物信息学补洞的结果进行了验证。

为确保序列拼接的准确性，实验选取了 7 个变异位点进行 PCR 验证。以已知的 7 条水稻叶绿体基因组序列作为参考序列，利用 SAMTools[12]软件(http://samtools.sourceforge.net/）将联配得到的 sam 文件转为 vcf 文件，得到东乡野生稻所有的变异位点。实验选取其中的 7 个单核苷酸多态性位点（Single Nucleotide Polymorphism，SNP）进行验证，再将测序结果与拼接结果作对比，最终得到完整的东乡野生稻叶绿体基因组序列。

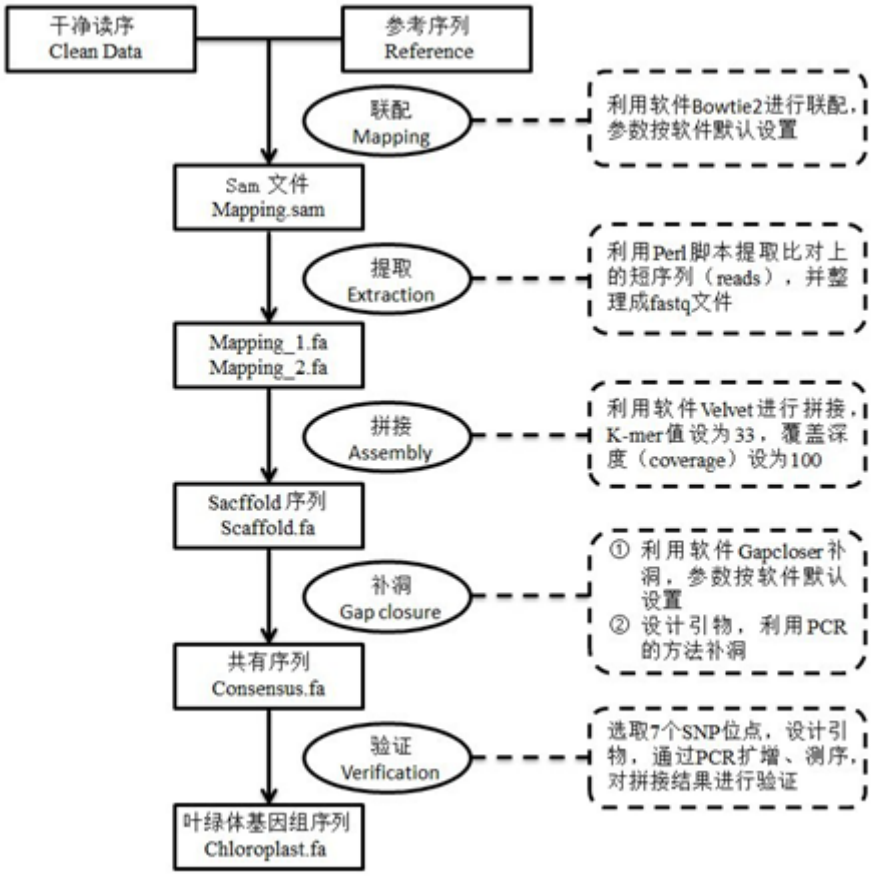


图 1　东乡野生稻叶绿体基因组拼接流程

Fig. 1 Assembly process of Dongxiang wild rice

1.2.4 叶绿体基因组注释

采用在线注释软件 DOGMA（http://dogma.ccbb.utexas.edu/）[13]对东乡野生稻叶绿体全基因组序列进行基因组注释。根据起始密码子和终止密码子序列手工调整 DOGMA 初步注释的编码蛋白基因范围，得到最终的注释结果并递交 Genbank（KF562709）。将东乡野生稻叶绿体基因组用在线绘图工具 GenomeVx（http://wolfe.gen.tcd.ie/GenomeVx/）[14]绘制叶绿体全基因组物理图谱。

1.2.5 进化树构建

实验将东乡野生稻与 14 条禾本科不同属的叶绿体基因组序列进行比对，运用邻接法（Neighbor-Joining，NJ）构建系统进化树，Bootstrap 运行 1000 次来检测各分支的置信度。再以东乡野生稻叶绿体基因组序列为参考序列，选取韩斌课题组发布的关于栽培稻起源的研究数据中的 15 个品种的全基因组测序数据，用 Bowtie2 将其分别与实验所得的东乡野生稻叶绿体基因组序列联配，利用 SAMTools 软件得到每一个品种的 SNP 位点信息，根据所有品种的 SNP 位点信息，每一个品种生成一条 SNP 序列。然后用 Clustal X2（http://www.clustal.org）将已知的 7 条水稻叶绿体基因组序列与东乡野生稻叶绿体基因组序列进行多序列联配，提取与上一步得到的 SNP 位点相对应的碱基，对应生成 7 条 SNP 序列，然后将所有的 SNP 序列合成一个 fasta 文件。最后用 MEGA5.0 软件[15]（Molecular Evolutionary

Genetics Analysis）对水稻叶绿体 SNP 进行分析，采用 Kimura-2-Parameter 模型计算核苷酸差异值，运用邻接法建树，Bootstrap 运行 1000 次。

2 结果与分析

2.1 叶绿体基因组拼接结果

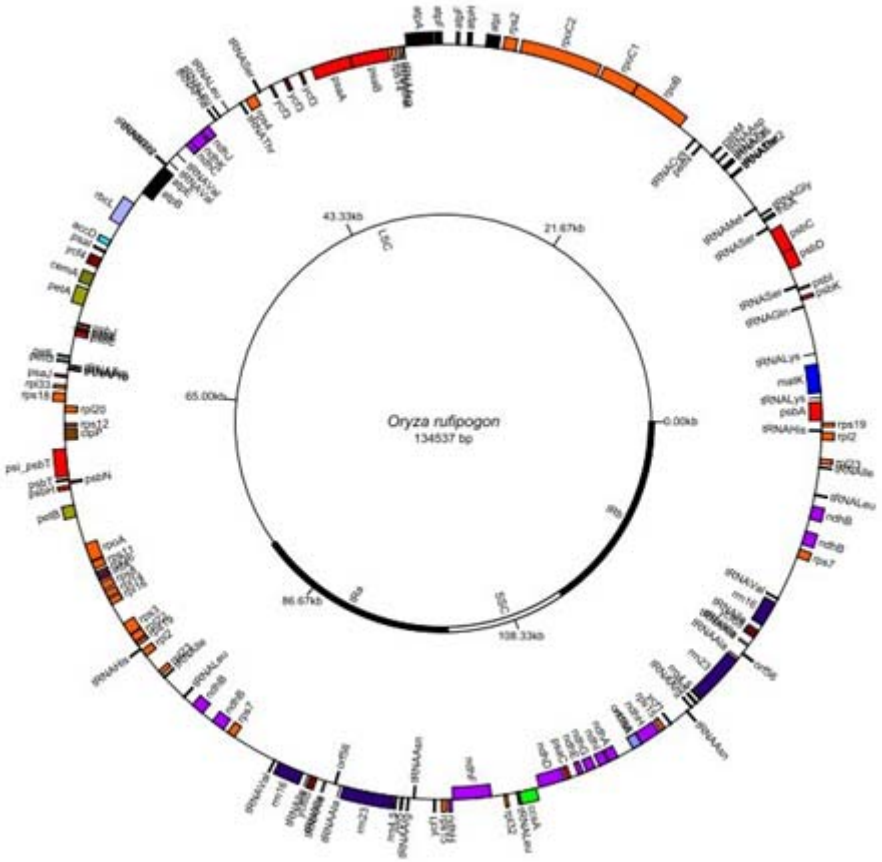
本实验选用已知的 7 条稻属叶绿体基因组序列作为参考序列，将东乡野生稻全基因组高通量测序所得到的短序列（reads）与其联配，根据叶绿体基因组的保守性，从中筛选出东乡野生稻的叶绿体基因组短序列，再通过拼接补洞得到完整的序列。

东乡野生稻的平均测序深度为 55，高通量测序得到的原始短序列中共有 4.36%能与参考序列联配，Velvet 拼接后共得到 scaffold4 条，N50 为 69 539bp，总长 134 469bp，对越南野生稻叶绿体基因组序列（NC_017835）的覆盖度达到 100%。通过 Gapcloser 补洞，得到一条完整的共有序列，总长共 134 537bp，GC 含量为 39.01%，利用 PCR 补洞的结果与其一致。此外，实验还选取了 7 个 SNP 位点进行 PCR 验证，其中有 5 个位点与拼接结果一致，2 个位点的验证结果与参考序列一致，且测序峰图清晰，可信度高，这对拼接结果提出了质疑。为验证这两个位点的可靠性，实验又采用 CLC Genomics Workbench 进行了重新拼接，拼接结果与 Velvet 拼接结果一致。将高通量测序得到的短序列联配到拼接得到的序列上，用 Tablet（http://bioinf.scri.ac.uk/tablet/）[16]对联配结果进行查看，发现该两个位点属于纯合位点，联配上的短序列均支持拼接结果。造成这种差异的原因还需进一步探讨，但实验最终选用了拼接的结果。

2.2 叶绿体基因注释

东乡野生稻叶绿体基因组为典型的环状结构，由四部分组成：大单拷贝区域（Large Single Copy，LSC），大小为 80 585bp，覆盖基因组 59.9%；小单拷贝区域（Small Single Copy，SSC），大小为 12 346bp，覆盖基因组 9.2%，两个反向重复区域（Inverted Repeats，IRs），大小为 20 803bp，覆盖基因组 15.5%（图 2）。东乡野生稻叶绿体总共编码 152 个基因，包括 95 个蛋白编码基因，49 个 tRNA 基因和 8 个 rRNA 基因，其中有 24 个基因在 IRs 区域存在 2 个拷贝，且所有的 rRNA 均在 IRs 区域。根据 CpBase（http://chloroplast.ocean.washington.edu/）将这 152 个基因进行功能分类，可分为与光合作用相关的基因 50 个，与转录和翻译有关的基因 87 个，与生物合成相关的基因 2 个，开放阅读框和其他蛋白编码基因共 13 个。

与东乡野生稻相比，栽培稻叶绿体基因组（NC_008155）总长 134 525bp，大、小单拷贝区和反向重复区域的长度分别为 80 592bp、12 335bp 和 20 799bp，覆盖基因组的比例基本与东乡野生稻一致。栽培稻叶绿体共编码基因 162 个（比东乡野生稻多 10 个），其中蛋白编码基因 114 个，tRNA 基因 40 个，rRNA 基因 8 个。根据 CpBase 功能分类，包括光合作用相关基因 32 个，转录翻译相关基因 80 个，开放阅读框和其他编码基因共 50 个，其中开放阅读框的个数远大于东乡野生稻。



内环黑色粗线表示反向重复区域（IRa 和 IRb），将基因组的大单拷贝区（LSC）和小单拷贝区（SSC）分开。外环上的基因按顺时针方向转录，内环上的基因按逆时针方向转录。

The thick black lines in inner ring indicates the extent of the inverted repeats (IRa and IRb), which separate the genome into the large single-copy (LSC) and small single-copy (SSC). Genes on the outside of the map transcribed in the clockwise direction and genes on the inside of the map are transcribed in the counterclockwise direction.

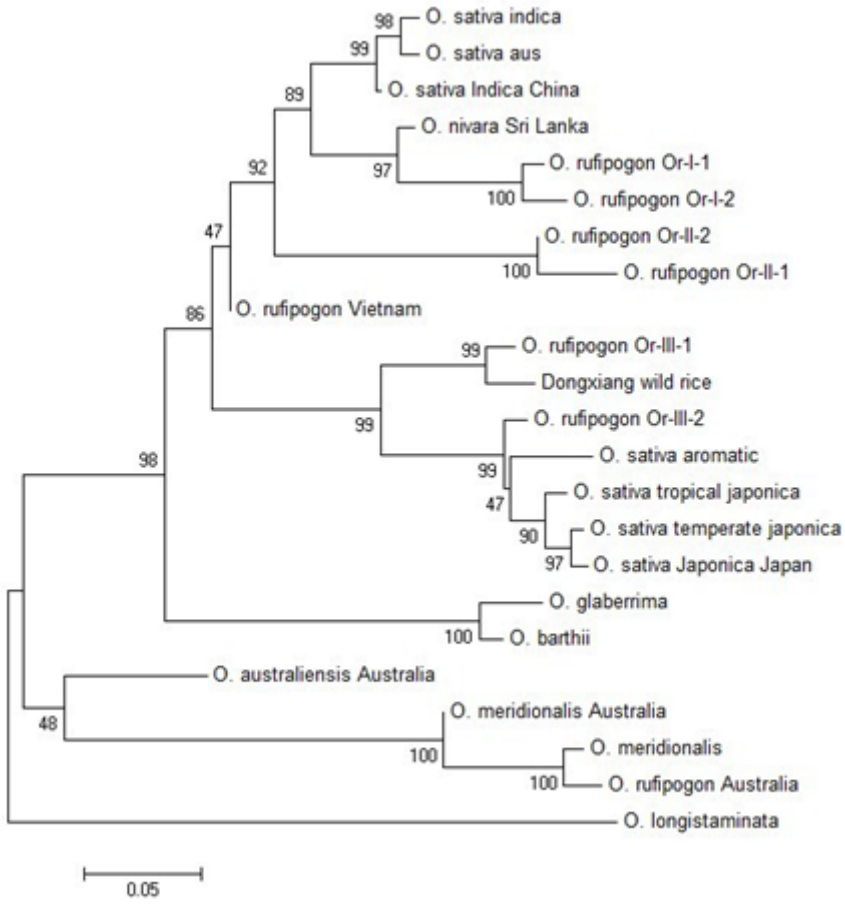
图 2 东乡野生稻叶绿体基因组注释图谱

Fig. 2 Gene map of Dongxiang wild rice chloroplast genome

2.3 系统发育分析

本实验将东乡野生稻与 14 个禾本科不同属叶绿体基因组序列利用邻接法建树，根据聚类结果，在禾本科中东乡野生稻与麻竹（*Dendrocalamus latiflorus*）和黍亚科（Panicoideae）的亲缘关系最近。

本实验选用了 23 个水稻叶绿体基因组数据，采用邻接法构建进化树（见图 3）。从进化树的结果可以看出，籼稻和 I 类野生稻亲缘关系较近，粳稻和 III 类野生稻亲缘关系较近（I 类、II 类和 III 类野生稻参照参考文献[8]中分类），因此结果支持韩斌课题组关于水稻驯化起源的假说[8]：粳稻首先由中国南方珠江流域的普通野生稻驯化而来，然后粳稻向东南和南亚传播，与当地的野生稻杂交，形成籼稻，因此粳稻和籼稻并非同时驯化形成的。



190 利用东乡野生稻与22条其它稻属物种叶绿体基因组序列的单核苷酸多态性位点构建的邻接法进化树。
The phylogenetic tree based on SNPs of Dongxiang wild rice and other 22 *Oryza* chloroplast genomes using NJ method.

图 3 基于叶绿体基因组序列构建东乡野生稻和其它禾本科物种的系统进化树

Fig. 3 The chloroplast genome-based phylogenetic trees of Dongxiang wild rice with other members from the
195 grass family (Poaceae)

3 讨论

叶绿体基因组结构和序列信息为转化技术和物种进化分析提供了重要的资源和信息。截至2013年8月13日，美国国家生物技术中心(The National Center for Biotechnology Information, NCBI)质体基因组数据库（http://www.ncbi.nlm.nih.gov/genomes/）（Organelle
200 Genome Resources）已经收录了285条植物界（Viridiplantae）叶绿体基因组完整序列，但相对于整个植物界的物种数量，这仅是很小的一部分。随着高通量测序技术的迅速发展，新的生物信息学软件的开发，将会有更多的植物叶绿体基因组测序成功。采用高通量测序技术最重要的环节之一是测序模板的制做，而对于传统的测序方法，叶绿体DNA的分离纯化是制做模板的主要限制因素。叶绿体基因组多为母系遗传，具有高度的保守性，本研究根据该特
205 性，利用生物信息学手段，优化出了一种简便快捷的方法，即从绿色叶片全基因组高通量测序数据中提取拼接叶绿体基因组序列。这种方法不需要单独分离纯化叶绿体DNA，降低了实验难度，缩短了实验时间，同时由于叶绿体基因组多拷贝的特性，降低了全基因组测序深度的要求。利用该方法本研究最终获得了东乡野生稻叶绿体全基因组序列，并通过比较稻属种间的叶绿体基因组，分析了水稻的驯化起源问题，研究结果支持2012年韩斌课题组在
210 《Nature》上发表的关于水稻驯化起源的假说[8]。

从全基因组数据中提取拼接叶绿体基因组的方法虽然简便快捷，但也有它不可避免的缺

点。提取叶绿体基因组序列是基于合适的参考序列的，因此该方法的关键是存在亲缘关系相近的叶绿体基因组序列，对于亲缘关系较远的参考序列是不适用的。此外，该方法是利用叶绿体基因组的保守性，提取与参考序列相似的短序列进行拼接，但在高等植物中叶绿体基因插入到核基因组的现象较为普遍，因此提取到的序列包含部分叶绿体基因插入到核基因组的片段，即使叶绿体 DNA 的拷贝数很高，这类片段对整体的拼接影响不大，但仍然不能排除核基因插入的可能性。

致谢

感谢中国水稻所魏新华和郭龙彪为实验提供研究材料，同时感谢叶小倩（浙江大学茶学系）在实验上给予的帮助。

[参考文献] (References)

[1] 李西文, 胡志刚, 林小涵, 等. 基于 454FLX 高通量技术的厚朴叶绿体全基因组测序及应用研究. 药学学报, 2012, 47(1): 124-130.

[2] Ohyama K, Fukuzawa H, Kohchi T, et al. Structure and organization of Marchantia polymorpha chloroplast genome: Ⅰ. cloning and gene identification. Journal of Molecular Biology, 1988, 203(2): 281-298.

[3] Shinozaki K, Ohme M, Tanaka M, et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. The EMBO Journal, 1986, 5(9): 2043-2029.

[4] Wu F H, Kan D P, Lee S B, et al. Complete nucleotide sequence of Dendrocalamus latiflorus and Bambusa oldhamii chloroplast genomes. Tree Physiology, 2009, 29(6): 847-856.

[5] Wu F H, Chan M T, Liao D C, et al. Complete chloroplast genome of Oncidium Gower Ramsey and evaluation of molecular markers for identification and breeding in Oncidiinae. BMC Plant Biology, 2010, 10: 68.

[6] Yang M, Zhang X, Liu G, et al. The complete chloroplast genome sequence of date palm (Phoenix dactylifera L.). PLOS ONE, 2010, 5(9): e12762.

[7] Wang Y, Bai X F, Yan C H, et al. Genomic dissection of small RNAs in wild rice (Oryza rufipogon) lessons for rice domestication. New Phytologist, 2012, 196(3): 914-925.

[8] Huang X H, Kurata N, Wei X H, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature, 2012, 490(25): 497-501.

[9] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 2009, 10(3): R25.

[10] Zerbino D R, Birney E. Velvet: algorithms for DE NOVO short read assembly using DE bruijn graphs. Genome Research, 2008, 18: 821-829.

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

[12] Li H, Handsaker B, Wysoker A, et al. The sequence Alignment/ Map format and samtools. Bioinformatics (Oxford, England), 2009, 25(16): 2078-2079.

[13] Wyman S K, Jansen R K, Boore J L. Automatic annotation of organellar genomes with DOGMA. Bioinformatics (Oxford, England), 2004, 20(17): 3252-3255.

[14] Conant G C, Wolfe K H. GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics (Oxford, England), 2008, 24(6): 861-862.

[15] Tamura K, Peterson D, Peterson N, et al. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution, 2011, 28(10): 2731-2739.

[16] Milne L, Bayer M, Cardle L, et al. Tablet—next generation sequence assembly visualization. Genome Research, 2010, 26(3): 401-402.