版纳园等自主开发出叶绿体全基因组分析比较基因组程序包

文章来源：西双版纳热带植物园 发布时间：2012-11-28 【字号： 小 中 大 】

高通量测序（High-throughput sequencing），又称“下一代”测序（Next-generation sequencing），是近年来在测序技术发展史中具有革命性改变的新突破，能一次并行对几十万到几百万条DNA分子同时测序，因此能对物种的转录组和基因组进行比以往较细致全貌的分析。

但是，由于“下一代”测序技术原始数据的读长（read length）只有几十个或一、两百个碱基，按照传统的分析流程，必须要通过生物信息学工具将这些短的碱基数据组装成较长的序列组（contigs）或基因组的框架，或者把这些序列比对到已有的参照基因组或者相近物种基因组序列上，才能进一步取得具有生物学意义的结果。对于没有参照基因组的非模式生物，要把这些海量的短序列数据组装的工作面临一定程度上的难度，制约了这类数据在非模式生物基因组研究的发展。

考虑到大部分生态学研究里的热带生物都是没有参照基因组的非模式生物，在中科院西双版纳热带植物园生态进化组Cannon研究员的领导下，版纳植物园、北京基因组所及德州理工大学的科研人员研发了直接分析高通量短序列数据的程序包，简化了高通量数据的比较基因组和转录组研究。由于此方法不需事先组装基因组，而以直接通过分析检测数据中的kmer片段是否存在及其出现频次，来探讨一定数量目标基因组中的序列差异，所以可以突破此类数据经常面临的生物信息学的分析瓶颈。通过筛选单个基因组独有或多个基因组共享的kmer片段及找出含这群kmer片段的数据后，此程序可以对这些数据进行组装，以取得较长的序列探讨下一步的生物学问题。

基于先前的工作基础（见已在*Molecular Ecology* 发表的论文，CANNON, C. H., KUA, C.-S., ZHANG, D. and HARTING, J.R. (2010), *Assembly free comparative genomics of short-read sequence data discovers the needles in the haystack. Molecular Ecology*, 19:147－161），研究人员进一步改善了非组装分析法，以比较174个叶绿体全基因组数据印证此程序包的功能和运行流程，并于*PlOS ONE*发表了题为*Reference-Free Comparative Genomics of 174 Chloroplasts*的论文。

由于这174个由低等植物和高等植物组成的叶绿体全基因组分析涉及的内容十分广泛，研究人员只能简洁的阐述几个发现，如虽然植物叶绿体基因组的基因结构和含量看起来十分保守，但是kmer片段分析可以把不同支流的植物清楚的分类。寄生植物的叶绿体基因组表现出预期的整体进化加速，而半寄生植物比全寄生植物的叶绿体基因组中含有较多的新基因序列，印证了基因组的演化机制受控于其功能。研究也发现了一段在被子植物里非常保守的基因序列。这分析里所有的成果都在该文章的补充材料部分。

此程序包内有4个不同功能的程序，可用LINUX和苹果操作系统以命令行运行。程序包已上传到全球最大开源软件开发平台sourceforge，下载网址为：http://sourceforge.net/projects/referencefree/

此研究得到了中国科学院知识创新工程重要方向项目和云南省高端科技人才引进计划项目的资助。