

•研究报告•

基于系统基因组学分析揭示早期 陆生植物的复杂网状进化关系

舒江平^{1,2} 刘 莉^{1,2} 沈 慧² 戴锡玲¹ 王全喜^{1,4} 严岳鸿^{2,3,4*}

1 (上海师范大学生命与环境科学学院, 上海 200234)

2 (上海辰山植物园, 中国科学院上海辰山植物科学研究中心, 上海 201602)

3 (国家林业局华东野生濒危资源植物保育中心, 上海 201602)

4 (上海市资源植物功能基因组学重点实验室, 上海 201602)

摘要: 植物由水生走向陆生的进化过程中经历了非常复杂的演化, 期间产生的大量基因的进化路线可能互不相同, 因此仅仅使用系统发育树无法呈现真实的演化关系。系统发育网络图能够清楚地展示包括垂直演化和水平演化在内的复杂网状进化关系。本文选取莱茵衣藻(*Chlamydomonas reinhardtii*)和4种陆生植物, 利用系统基因组学的方法, 筛选得到1,668个一对一直系同源基因, 重新构建了陆生植物的系统发育网状进化关系。结果发现, 使用不同的分析策略所得到的系统发育树不同; 对1,668个基因单独分析, 发现存在15种不同的拓扑结构; 对5个物种筛选得到的直系同源基因进行系统发育网络分析显示, 在非常稳健的系统发育网络图中, 仅仅5个物种就存在9个不同的分离支, 暗示着非常复杂的网状进化关系; 而且藻类植物与苔藓植物和石松类植物的分离支之间差异很小, 这可能是产生系统发育树冲突的原因之一, 也暗示着早期陆生植物发生了复杂的辐射演化。

关键词: 网状进化; 基因树冲突; 系统发育基因组学; 陆生植物演化; 转录组从头测序

The complex reticulate evolutionary relationships of early terrestrial plants as revealed by phylogenomics analysis

Jiangping Shu^{1,2}, Li Liu^{1,2}, Hui Shen², Xiling Dai¹, Quanxi Wang^{1,4}, Yuehong Yan^{2,3,4*}

1 College of Life and Environmental Sciences, Shanghai Normal University, Shanghai 200234

2 Shanghai Chenshan Plant Science Research Center, Chinese Academy of Sciences; Shanghai Chenshan Botanical Garden, Shanghai 201602

3 East China Wild Endangered Resources Plant Conservation Center, State Forestry Administration, Shanghai 201602

4 Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai 201602

Abstract: Plants from aquatic to terrestrial ecosystems have undergone a very complex evolution, and their evolutionary pathways of large numbers of genes may be different from one another, so that traditional phylogenetic trees cannot show true evolutionary relationships. The phylogenetic network graph is a good solution to show the complex relationships of reticulate evolution, including vertical evolution and horizontal evolution. In this paper, we selected *Chlamydomonas reinhardtii* and four terrestrial plants, and screened 1,668 one-to-one orthologous genes to reconstruct the phylogenetic relationship of terrestrial plants based on phylogenomics. Results showed that phylogenetic trees were different based on different analysis strategies. The 1,668 genes were analyzed separately and 15 different topologies were found. The phylogenetic network of the orthologous genes obtained from the five species was analyzed, and the results showed that in a very robust phylogenetic network map, only five species have nine different split branches, suggesting a very complex evolutionary relationship network. Furthermore, the difference in split branches between algae and bryophytes or lycophytes is very small, which may be one of the reasons influencing the phylogenetic tree conflict, and implies that early terrestrial plants underwent a complex radiate evolution.

Key words: reticulate evolution; gene tree conflict; phylogenomics; terrestrial plant evolution; *de novo* transcriptome

收稿日期: 2017-02-18; 接受日期: 2017-06-20

基金项目: 上海市绿化和市容局 2015 年科学技术项目(G152420)

* 通讯作者 Author for correspondence. E-mail: yan.yh@126.com

在距今约4.8亿年前的奥陶纪(Ordovician)时期,藻类植物(绿藻)由水生向陆生发展产生了早期陆地植物。这一重大的进化历史事件(李承森, 1994; Kenrick & Crane, 1997; Wellman et al, 2003; Steemans et al, 2009; Rubinstein et al, 2010; Wickett et al, 2014)引起了包括现存植物和化石植物在形态、生理、生殖和生态等性状上的极大多样化。而重建早期陆生植物的系统演化关系, 一直以来都是植物系统发育学研究的基本内容, 对于理解植物胚胎、维管组织、种子和花等的发育和演化都至关重要(Wickett et al, 2014)。

虽然现在利用大量的分子数据研究早期陆生植物的演化已得到相对普遍认同的观点, 但是石松类植物(lycophytes)和苔藓植物(bryophytes)的关系在早期研究中饱受争议。在志留纪(Silurian)和早泥盆纪(Lower Devonian)时期, 出现了大量结构非常简单的石松类植物的祖先——裸蕨类(psilophytes)植物。基于现存形式和化石形式的形态比较研究, 得出了两种假设: (1)石松类植物直接起源于藻类植物; (2)石松类植物和苔藓植物的关系更为紧密。

支持第一种假设的研究者认为苔藓植物和石松类植物的相似性是由于平行进化引起的, 而不是由于二者之间的系统演化所关联(Scott, 1900; Eames, 1936)。石松类植物可能起源于一群复杂的藻类, 这些藻类的生长接近于从海洋转运而来的潮汐的上限, 很可能是在潮间带生长旺盛的具有游动生殖细胞的褐藻类(phaerophyta)植物(Church, 1919; Arnold, 2013)。另一种观点认为这些藻类可能是绿藻(chlorophyta)。这些绿藻被认为是一类细丝状的胶毛藻类型(Bohlin, 1901; Fritsch, 1916), 但是其祖先绿藻是具有同形生活史的直立薄壁细胞形式, 其生活史由两种相似的无性世代交替进行(Fritsch, 1945)。

支持第二种假设的研究者认为石松类植物和苔藓植物之间存在很多相似的特征, 暗示二者之间存在某种系统演化关系。一些研究认为石松类植物和苔藓植物是由一种非常原始的陆生颈卵器植物演化而来的两条不同进化路线(Bower, 1935; Zimmermann, 1938); 另一些研究认为石松类植物是由苔藓植物演化而来(Campbell, 1899)。

系统发育学(phylogenetics)试图利用各种性状(包括化石、形态、分子等)重新构建地球上所有生物的演化历史, 并以系统发育树的形式呈现。随着

分子数据的增加, 特别是基因组时代的来临, 各个不同类群的系统发育树得以重建并有非常高的统计学支持(Dunn et al, 2008; Regier et al, 2010; Misof et al, 2014; Wickett et al, 2014)。然而, 与此同时, 不同基因构建的系统发育树之间各不一致, 即基因树冲突(conflicting gene trees)现象也日益突出(邹新慧和葛颂, 2008)。在物种演化历程中, 不同基因可能有不同的进化路线, 主要分为两种: 垂直演化(vertical evolutionary processes)和水平演化(horizontal evolutionary processes)。垂直演化, 即遗传信息直接由亲代传递给子代, 主要包括核酸替换(DNA substitutions)、插入(insertions)、删除(deletions)和基因复制(genic duplications)、易位(translocations)、倒位(inversions); 水平演化主要包括重组(recombination)、杂交(hybridization)、基因渐渗(introgression)、基因水平转移(horizontal gene transfer)和基因组融合(genome fusion)等。系统发育树只能呈现垂直演化过程, 而对于相对复杂的水平演化过程, 系统发育树则无法展示。此外, 即使是垂直演化, 如果存在不完全的谱系分选(incomplete lineage sorting)等进化历史事件, 基因演化树和物种演化树可能截然不同。

随着系统发育学研究的深入, 系统发育网络(phylogenetic network)可以很好地展现复杂的网状进化(reticulate evolution)关系(Huson & Bryant, 2006; Nakhleh, 2010; Huson & Scornavacca, 2011)。理论上, 当进化过程中存在网状进化事件时, 有根的系统发育网络可以清楚地呈现物种演化, 但是这种系统发育网络分析所涉及的计算很困难, 至今仍然没有一种成熟和广泛使用的分析工具(Tofigh et al, 2011)。值得庆幸的是, 已经有大量的分析工具用于计算无根的系统发育网络, 以此来呈现系统发育分析和谱系地理学研究中的网状进化事件(Bandelt et al, 1999; Clement et al, 2000; Bryant & Moulton, 2004; Huson & Bryant, 2006)。

本文以早期陆生植物演化分析为例, 利用系统基因组学和系统发育网络分析早期陆生植物演化过程中可能存在的网状进化关系。

1 材料与方法

1.1 转录组测序

转录组测序样品取自上海辰山植物园的蕨类

温室内单个福建观音座莲(*Angiopteris fokiensis*)植株的营养叶, 取样后迅速将样品放到液氮中冷冻。利用TRIzol[®] Reagent试剂(Invitrogen, 上海)提取总RNA, 然后利用Plant RNA Purification Reagent试剂(Invitrogen, 上海)对提取的总RNA进行纯化, 之后用Agilent 2100检测RNA提取质量。利用带有Oligo dT的磁力架(Invitrogen, 上海)与polyA进行A-T碱基配对, 从总RNA中分离出mRNA。利用金属离子将获得的mRNA随机断裂成200 bp左右的小片段。以这些小片段mRNA为模板, 用随机引物合成第一条cDNA链, 然后加入缓冲液、dNTPs、RNase H和DNA polymerase I合成第二条cDNA链。对双链cDNA进行加工修饰, 末端加poly (A)并连接测序接头adaptor, 得到完整的cDNA。利用PCR进行cDNA扩增, 构建测序文库, 用Illumina HiSeq 2500进行测序。

1.2 序列组装和ORF预测

利用Illumina双末端测序(paired-end, PE)方法进行高通量测序, 得到的原始图像数据经过Base Calling转化为序列数据, 即原始序列读取片段(raw reads), 以FASTQ格式储存。对原始序列进行质量评估和可信度分析, 并去除制备文库时产生的接头序列、两端低质量序列($Q < 20$ bp)、含N > 10%的序列以及修剪后长度小于20 bp的序列, 得到有效数据(clean data)。利用Trinity (<http://trinityrnaseq.sourceforge.net/>, 版本号 trinityrnaseq-r2013-02-25) (Grabherr et al, 2011)对所有有效数据进行从头组装。由于Trinity组装结果中含有大量的冗余序列, 利用CD-HIT (Fu et al, 2012)去除重复序列, 得到unigene序列。利用Trinity中的transcoder工具对unigene进行开放阅读框(open reading frame, ORF)预测, 得到序列中最佳的ORF区域, 使用pfam (<http://pfam.xfam.org/>)数据库对预测结果进行校正, 将比对结果保留到pfam数据库的蛋白序列中。

1.3 转录组数据完整性评估

使用转录组数据进行系统发育分析, 评估测序组装得到的转录序列是否覆盖所有可能的转录本是非常重要的。这直接关系到系统发育分析中数据的大小和完整性, 从而影响系统发育树的拓扑结构。我们选取由30个植物基因组筛选得到的1,440个普遍存在于植物中的单拷贝直系同源基因组成的数据集, 使用BUSCO v2 (Simão et al, 2015)软件评估福建观音座莲转录组数据的完整性。

1.4 直系同源基因筛选

从EnsemblPlants (<http://plants.ensembl.org/index.html>)数据库网站上下载得到莱茵衣藻(*Chlamydomonas reinhardtii*)、小立碗藓(*Physcomitrella patens*)、江南卷柏(*Selaginella moellendorffii*)的基因组数据, 从congenie (<http://congenie.org/>)网站获得欧洲云杉(*Picea abies*)的基因组数据。对5个物种(莱茵衣藻、小立碗藓、江南卷柏、福建观音座莲和欧洲云杉)基因的蛋白序列使用OrthoMCL (Li et al, 2003)软件筛选同源基因($E \text{ value} \leq 1e-10$, inflation = 2.5), 运用Agalma (Dunn et al, 2013)系统发育基因组学分析流程筛选一对一直系同源基因。其大致流程如下: (1)对每个同源基因聚类组中的序列使用MUSCLE (Edgar, 2004)软件对齐; (2)使用GBLOCKS (Talavera & Castresana, 2007)软件筛选基因序列的保守区域; (3)使用RAxML (Stamatakis, 2014)软件对筛选得到的保守序列构建最大似然树, 自举检验100次; (4)使用DendroPy (Sukumaran & Holder, 2010)软件对单基因树进行修剪, 得到每个类群只保留一条序列的一对一直系同源基因矩阵。

1.5 系统发育与网状进化分析

系统发育分析分别使用串联和联合分析。串联矩阵的方法: (1)对每个筛选得到的直系同源基因使用MUSCLE软件比对, 形成序列矩阵; (2)对每个矩阵使用GBLOCK软件进行序列保守区域的筛选; (3)将筛选之后的保守序列矩阵串联成多基因超级矩阵(supermatrix); (4)使用RAxML软件进行最大似然法分析, 选择执行快速支持度分析并以Prottest 3 (Darriba et al, 2011)软件搜索的PROTCATJTT为最佳模型自举检验1,000次。联合基因树的方法: (1)对每个直系同源基因单独使用RAxML软件进行最大似然法分析, 自举检验100次; (2)将单基因树根据拓扑结构进行分类, 并统计不同拓扑结构的基因数量; (3)使用ASTRAL (Mirarab et al, 2014)软件将多个单基因树整合成超级物种树(supertree)。

对筛选得到的蛋白超级矩阵, 使用SplitsTree4软件(Huson & Bryant, 2006)推测5个类群的系统发育网络关系。

2 结果

2.1 转录组组装和ORF预测

采用Illumina HiSeq 2500高通量测序平台对福

建观音座莲进行转录组测序, 得到4.5 Gb原始数据 (raw data), 原始数据已上传至NCBI的SRA数据库 (SAMN03575909)。为了保证后续生物信息分析的准确性, 对原始数据进行过滤去杂, 获得了4.4 Gb有效数据(clean data)。对有效数据进行de novo拼接组装, 共得到63,262个转录本。去除冗余序列后得到57,637条单基因簇(unigene), N50为1,629 bp。从unigene中预测得到35,972条ORF, 占总unigene的62.41%。详细结果见附录1。

2.2 转录组数据完整性评估

使用植物直系同源基因数据集对福建观音座莲的转录组数据完整性进行评估, 在由30个植物基因组筛选得到的1,440个植物单拷贝直系同源基因中, 完全匹配到的单拷贝直系同源基因(complete)有956个, 占总BUSCO的66.4%; 部分片段匹配到的单拷贝直系同源基因(fragment)有85个, 占总BUSCO的5.9%; 没有匹配到的植物单拷贝直系同源基因(missing)有399个, 占总BUSCO的27.7% (表1)。

2.3 系统发育分析

我们筛选得到1,668个100%类群覆盖的一对一

单拷贝直系同源基因。首先, 使用串联矩阵的方法, 将每个单基因蛋白序列矩阵串联成多基因超级矩阵, 包含324,365个保守的氨基酸位点; 以莱茵衣藻为外类群, 使用最大似然法重建得到5个类群的系统发育关系为: (莱茵衣藻, (江南卷柏, (小立碗藓, (福建观音座莲, 欧洲云杉)))) (图1)。然后, 使用联合基因树的方法, 以莱茵衣藻为外类群, 使用最大似然法对每个单基因蛋白序列矩阵构建系统发育树, 得到1,668棵单基因树, 包括15种树形(图2)。对所有单基因树联合得到超级物种树为: (莱茵衣藻, (小立碗藓, (江南卷柏, (福建观音座莲, 欧洲云杉)))) (图1)。

2.4 网状进化

不同的基因可能存在不同的演化历史, 我们使用1,668个基因构建5个植物的系统发育网络, 结果发现存在9个分离支(split branch), 除了物种分离支外, 还存在4种网络分离支, 除了最短的分离支(图3箭头标示)支持率仅为21.6%外, 其他支都是100%的支持率(图3)。结果显示, 小立碗藓和莱茵衣藻亲缘关系更近一些, 而江南卷柏更远, 然而这种结果也不太稳定(最短分离支的支持率低, BS = 21.6%)。

表1 转录组和基因组组装完整性评估结果统计
Table 1 The assessment results of assembly completeness of transcriptome and genome

物种 Species	分类 Classification	BUSCO评估结果 BUSCO results
福建观音座莲 <i>Angiopteris fokiensis</i>	真蕨类植物 Monilophytes	C: 66.4% [S: 43.2%, D: 23.2%], F: 5.9%, M: 27.7%, n: 1440
欧洲云杉 <i>Picea abies</i>	种子植物 Spermatophytes	C: 34.0% [S: 28.9%, D: 5.1%], F: 7.4%, M: 58.6%, n: 1,440
江南卷柏 <i>Selaginella moellendorffii</i>	石松类植物 Lycophytes	C: 63.2% [S: 10.0%, D: 53.2%], F: 4.7%, M: 32.1%, n: 1,440
小立碗藓 <i>Physcomitrella patens</i>	苔藓植物 Bryophytes	C: 70.1% [S: 46.0%, D: 24.1%], F: 2.6%, M: 27.3%, n: 1,440
莱茵衣藻 <i>Chlamydomonas reinhardtii</i>	藻类植物 Thallophytes	C: 18.8% [S: 17.9%, D: 0.9%], F: 1.7%, M: 79.5%, n: 1,440

C: 完全匹配的BUSCOs; S: 完全匹配的单拷贝BUSCOs; D: 完全匹配的有副本的BUSCOs; F: BUSCOs部分片段; M: 缺失的BUSCOs; n: 总BUSCOs数目。
C, Complete BUSCOs; S, Complete and single-copy BUSCOs; D, Complete and duplicated BUSCOs; F, Fragmented BUSCOs; M, Missing BUSCOs; n, Total BUSCO groups searched.

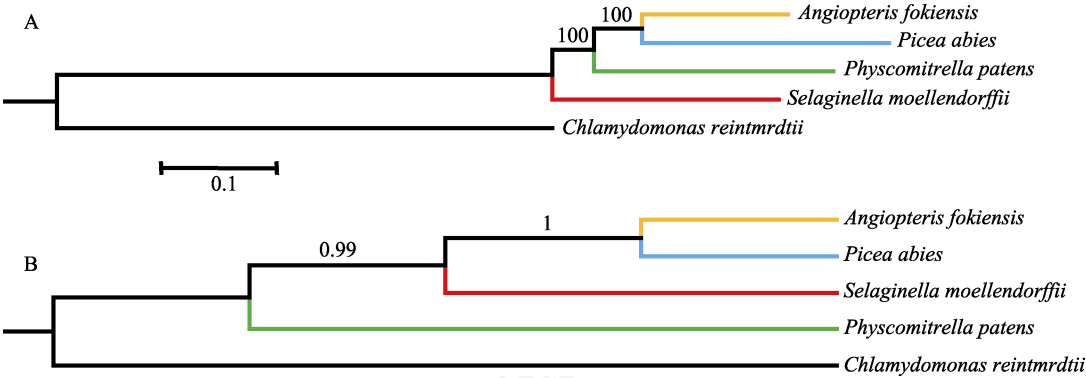


图1 基于串联和联合的方法分析得到的系统发育树。(A)使用串联矩阵构建的最大似然树;(B)使用联合基因树构建的物种树。
Fig. 1 The phylogenetic trees based on concatenation and coalescence methods. (A) The maximum likelihood tree based on concatenation method; (B) The species tree based on coalescence method.

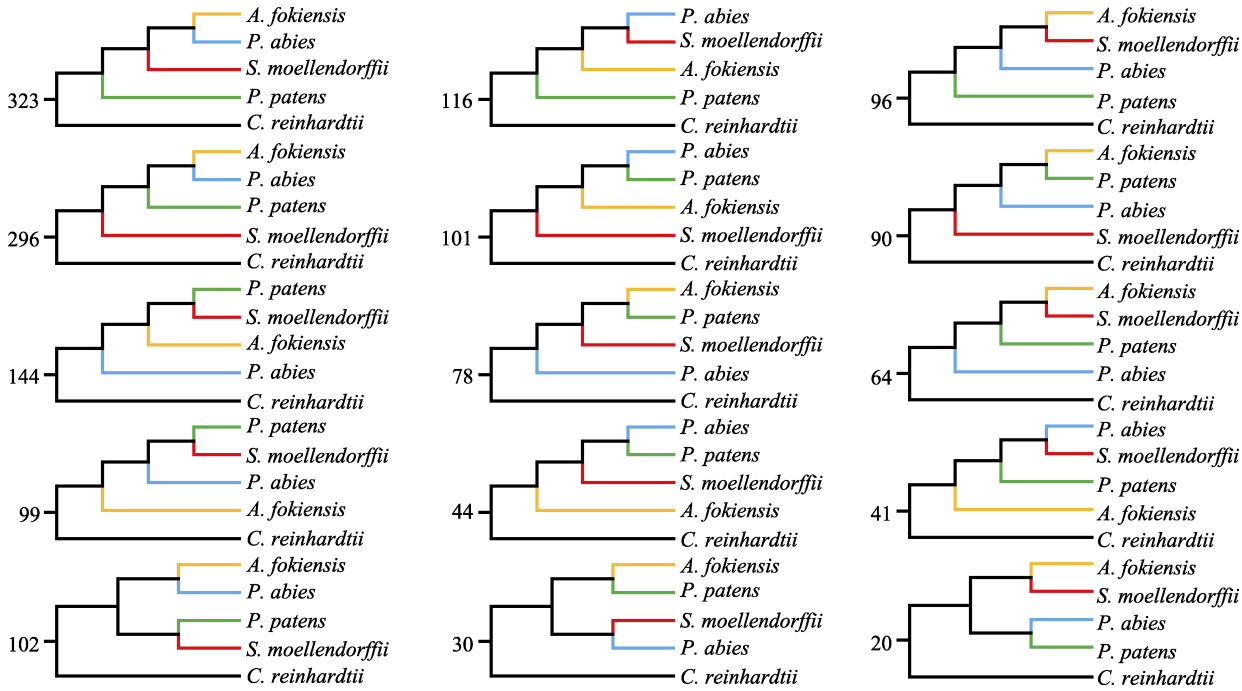


图2 使用最大似然法构建的15种拓扑结构的基因树。数字表示每种拓扑结构的数量。
Fig. 2 Fifteen topological structure of gene trees based on maximum likelihood method. The numbers mean the amount of the topological structure.

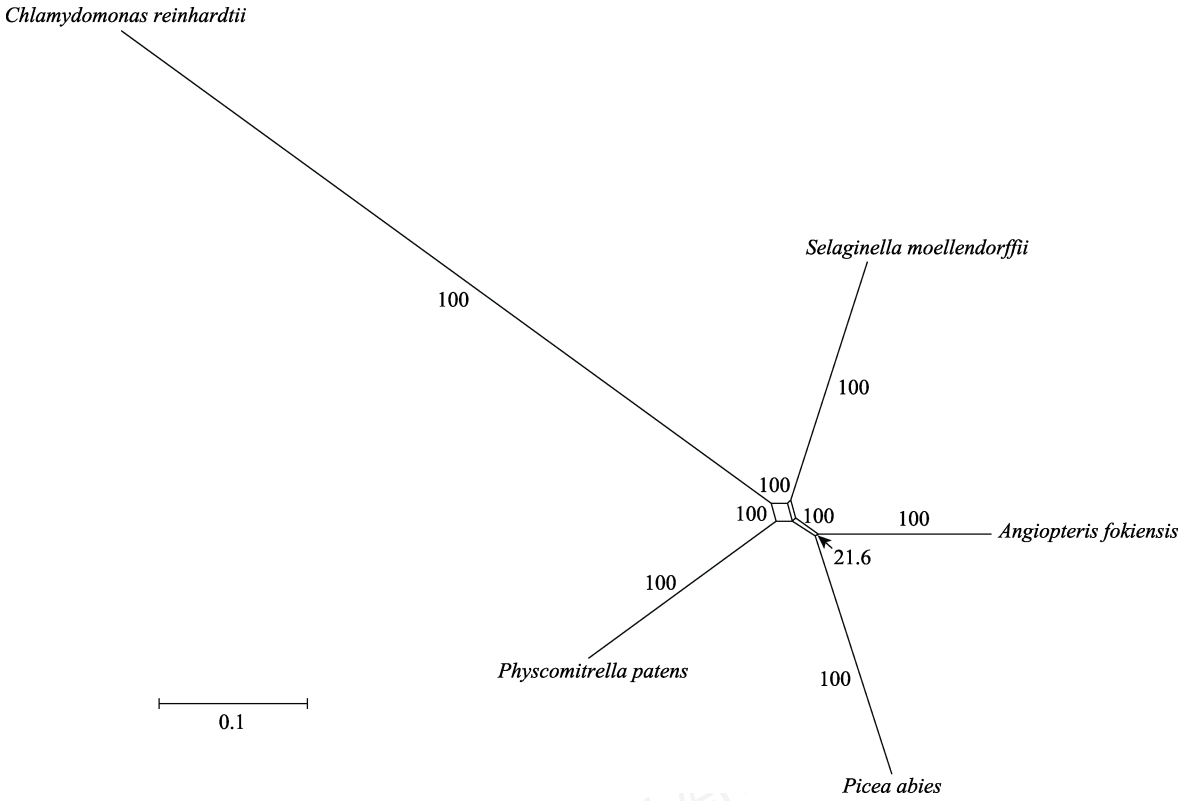


图3 基于1,668个基因构建的早期陆生植物的系统发育网络。数字表示每个分离支的支持率, 除了最短分离支(箭头)之外, 其他分离支的支持率都为100%; 平行的分离支为同一种分离支。
Fig. 3 The phylogenetic network of early land plants based on 1,668 genes. The numbers mean the bootstrap support of each split branch. In addition to the shortest split branch (arrow), the bootstrap support of other split branches is 100%. The parallel split branches are the same type of split branch.

3 讨论

早在1859年达尔文在《物种起源》一书中就提出,地球上所有的生物都起源于一个共同祖先。并且其曾使用简单的系统树来表示地球上物种的演化过程。随着系统发育分析的发展,人们开始使用各种性状来构建地球上不同类群生物的演化历史,试图实现构建整个地球生命之树的伟大理想。随着分子生物学的迅猛发展,越来越多的分子数据得到积累,使得分子数据成为系统发育分析的主要性状。人们开始联合使用多个基因和DNA片段进行系统发育分析,然而随着研究的深入,不同的研究者使用不同的基因序列对同一类群进行系统发育分析所得到的系统演化树却互不相同。Cranston等(2009)在6个水稻物种中筛选得到307个直系同源基因,研究发现这些基因之间存在高度的不一致性,仅有8个基因树与串联分析的物种树一致(Cranston et al, 2009); Salichos和Rokas (2013)从23种酵母基因组中筛选得到1,070个直系同源基因,其中没有一种基因树与串联分析的物种树一致(Salichos & Rokas, 2013)。

本研究中,我们从5个物种的基因组和转录组中筛选得到1,668个一对一直系同源基因,将这些基因串联建成包含324,365个保守氨基酸位点的超级矩阵,运用最大似然法重建了植物的演化关系。结果和以往研究不同的是,石松类植物首先在藻类植物之后演化出来。随后,使用联合基因树的方法,首先用最大似然法构建1,668个单基因树,然后整合成为超级物种树。和串联矩阵的结果不同,苔藓植物首先在藻类植物之后演化出来。由此可见,即使是相同的数据,不同的分析策略所得到的系统发育树也是自相矛盾的。将构建得到的1,668个基因树进行分类,得到15种不同的拓扑结构,即这些基因可能存在15种不同的进化过程。进一步统计发现,支持串联矩阵方法的分析结果的基因有296个,少于支持联合基因树方法的分析结果的基因(323个)。由此可见,这1,668个基因数据本身就存在不确定性,这些经历不同进化过程的基因的信息含量没有明显的差异,这可能也造成了两种分析策略所得到的结果有所不同。

不同的基因分析得到不同的系统树,这种不一致性可能是由于基因进化历史重建过程分析错误

造成,更有可能是由于存在复杂的生物学过程引起(Woodhams et al, 2016)。这些进化事件在生物演化过程中时常发生,特别是杂交物种形成事件, Mallet (2005)估计平均大约25%的植物和10%的动物与另外至少一个物种发生了杂交。随着研究的深入,人们已经开始尝试在不完全谱系分选(Degnan & Rosenberg, 2009; Nakhleh, 2013; Mirarab et al, 2016)、杂交(Nakhleh, 2010; Woodhams et al, 2016)、基因水平转移(Szöllősi et al, 2012)等进化事件存在的情况下重建物种演化关系。更多的研究者认为在存在复杂的网状进化事件时,应该使用系统发育网络来重建物种演化过程中基因网络关系,从而理解基因树冲突的原因和物种演化关系(Sneath, 1975; Syvanen, 1985; Delwiche & Palmer, 1996; Griffiths & Marjoram, 1996; Rieseberg, 1997; Doolittle, 1999; Huson & Bryant, 2006; Huson & Scornavacca, 2011)。本研究中,我们的取样类群相对较少,因此1,668个基因只得到了15种不同的演化路线,这在一定程度上预示着早期陆生植物在进化过程中存在着非常复杂的网状演化关系。随后我们对这1,668个基因进行系统发育网络分析,发现早期陆生植物之间确实存在复杂的网状进化关系。仅仅在5个类群之间就出现9个分离支,除了最短的分离支(图3箭头标示, BS = 21.6%)外,其他分离支都得到100%的统计学支持。虽然系统发育网络图中,物种间的分离支差异不大,但总体上仍能看出在藻类植物演化之后苔藓植物先分化出来。

参考文献

- Arnold CA (2013) An Introduction to Paleobotany. Read Books Ltd., New York.
- Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, 16, 37–48.
- Bohlin K (1901) Utkast till de gröna algernas och arkegoniaternas fylogeni. Almqvist & Wiksells Boktryckeri AB, Uppsala.
- Bower FO (1935) Primitive land plants. *Science*, 81, 537–539.
- Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, 21, 255–265.
- Campbell D (1899) Lectures on the Evolution of Plants. Kessinger Publishing, London.
- Church AH (1919) Thalassiophyta and the Subaerial Transmigration. Oxford University Press, Oxford.

- Clement M, Posada D, Crandall KA (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, 9, 1657–1659.
- Cranston KA, Hurwitz B, Ware D, Stein L, Wing RA (2009) Species trees from highly incongruent gene trees in rice. *Systematic Biology*, 58, 489–500.
- Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27, 1164–1165.
- Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24, 332–340.
- Delwiche CF, Palmer JD (1996) Rampant horizontal transfer and duplication of Rubisco genes in eubacteria and plastids. *Molecular Biology and Evolution*, 13, 873–882.
- Doolittle WF (1999) Phylogenetic classification and the universal tree. *Science*, 284, 2124–2128.
- Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452, 745–749.
- Dunn CW, Howison M, Zapata F (2013) Agalma: an automated phylogenomics workflow. *BMC Bioinformatics*, 14, 330.
- Eames AJ (1936) *Morphology of Vascular Plants*. McGraw-Hill Book Company, New York.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792–1797.
- Fritsch F (1945) Studies in the comparative morphology of the algae. IV. Algae and archegoniate plants. *Annals of Botany*, 9, 1–29.
- Fritsch FE (1916) The algal ancestry of the higher plants. *New Phytologist*, 15, 233–249.
- Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150–3152.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644–652.
- Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3, 479–502.
- Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23, 254–267.
- Huson DH, Scornavacca C (2011) A survey of combinatorial methods for phylogenetic networks. *Genome Biology and Evolution*, 3, 23–35.
- Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. *Nature*, 389, 33–39.
- Li CS (1994) Origin of land plants is an important event of life evolution. *Bulletin of National Natural Science Foundation of China*, (4), 8–14. (in Chinese with English abstract) [李承森 (1994) 生物进化的重大事件——陆地植物的起源及其研究的新进展. *中国科学基金*, (4), 8–14.]
- Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13, 2178–2189.
- Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20, 229–237.
- Mirarab S, Bayzid MS, Warnow T (2016) Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65, 366–380.
- Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30, 541–548.
- Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346, 763–767.
- Nakhleh L (2010) Evolutionary phylogenetic networks: models and issues. In: *Problem Solving Handbook in Computational Biology and Bioinformatics* (eds Lenwood SH, Naren R), pp. 125–158. Springer, New York.
- Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends in Ecology & Evolution*, 28, 719–728.
- Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, 463, 1079–1083.
- Rieseberg LH (1997) Hybrid origins of plant species. *Annual Review of Ecology and Systematics*, 28, 359–389.
- Rubinstein CV, Gerrienne P, de la Puente G, Astini R, Steemans P (2010) Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). *New Phytologist*, 188, 365–369.
- Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497, 327–331.
- Scott DH (1900) *Studies in Fossil Botany*. Adam & Charles Black, London.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212.
- Sneath PH (1975) Cladistic representation of reticulate evolution. *Systematic Zoology*, 24, 360–368.
- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312–1313.
- Steenmans P, Le Hérisse A, Melvin J, Miller MA, Paris F, Verniers J, Wellman CH (2009) Origin and radiation of the

- earliest vascular land plants. *Science*, 324, 353.
- Sukumaran J, Holder MT (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26, 1569–1571.
- Syvanen M (1985) Cross-species gene transfer, implications for a new theory of evolution. *Journal of Theoretical Biology*, 112, 333–343.
- Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences, USA*, 109, 17513–17518.
- Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, 56, 564–577.
- Tofigh A, Hallett M, Lagergren J (2011) Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8, 517–535.
- Wellman CH, Osterloff PL, Mohiuddin U (2003) Fragments of the earliest land plants. *Nature*, 425, 282–285.
- Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA (2014) Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences, USA*, 111, 4859–4868.
- Woodhams MD, Lockhart PJ, Holland BR (2016) Simulating and summarizing sources of gene tree incongruence. *Genome Biology and Evolution*, 8, 1299–1315.
- Zimmermann W (1938) Phylogenie. In: *Manual of Pteridology* (ed. Frans V), pp. 558–618. Springer, New York.
- Zou XH, Ge S (2008) Conflicting gene trees and phylogenomics. *Journal of Systematics and Evolution*, 46, 795–807. (in Chinese with English abstract) [邹新慧, 葛颂 (2008) 基因树冲突与系统发育基因组学研究. *植物分类学报*, 46, 795–807.]

(责任编辑: 王艇 责任编辑: 时意专)

附录 Supplementary Material

附录1 福建观音座莲转录组测序和组装结果统计

Appendix 1 The results of transcriptome sequence and assemble of *Angiopteris fokiensis*
<http://www.biodiversity-science.net/fileup/PDF/2017042-1.pdf>

舒江平, 刘莉, 沈慧, 戴锡玲, 王全喜, 严岳鸿. 基于系统基因组学分析揭示早期陆生植物的复杂网状进化关系. 生物多样性, 2017, 25 (6): 675–682.
<http://www.biodiversity-science.net/CN/10.17520/biods.2017042>

附录1 福建观音座莲转录组测序和组装结果统计
Appendix 1 The results of transcriptome sequence and assemble of *Angiopteris fokiensis*

原始数据 Rawdata (G)	4.5	Q30 值 Q30 (%)	91.85
有效数据 Cleandata (G)	4.4	单基因簇数量 Number of unigenes	57,637
总碱基含量 Total bases (bp)	3,415,761,579	N50 值 N50 (bp)	1,629
Q20 值 Q20 (%)	96.1	GC 含量 GC content	0.45