

· 专题论坛 ·

针叶树基因组特征及其序列资源挖掘进展

许晨璐, 孙晓梅*, 张守攻

中国林业科学研究院林业研究所, 国家林业局林木培育重点实验室, 北京 100091

摘要 针叶树是裸子植物中最大也是分布最广的一支。作为多年生木本植物, 针叶树不仅为工业提供建筑、造纸等重要原料和其它可再生能源, 而且在北半球的生态平衡中也起着重要作用。因其独特的分类地位、重要的经济价值和生态价值, 针叶树序列资源挖掘备受重视。然而其庞大且复杂的基因组阻碍了这一进程, 截至2013年4月, 尚无获得全基因组序列的针叶树种。随着第2代测序技术的出现及生物信息学的快速发展, 针叶树序列资源挖掘也从转录组过渡到全基因组测序, 后者已在松属(*Pinus*)、云杉属(*Picea*)和黄杉属(*Pseudotsuga*)部分树种中开展。该文首次归纳了针叶树基因组特征, 回顾了针叶树序列资源挖掘进程, 并重点介绍了火炬松(*Pinus taeda*)、欧洲云杉(*Picea abies*)和白云杉(*Picea glauca*)的全基因组测序项目。

关键词 针叶树, 基因组, 研究进展, 转录组, 全基因组测序

许晨璐, 孙晓梅, 张守攻 (2013). 针叶树基因组特征及其序列资源挖掘进展. 植物学报 48, 684–693.

球果目(本文称之为针叶树)是裸子植物(Gymnosperm)四大分支中最大也是分布最广的一支[其它3支为苏铁目(Cycadales)、银杏目(Ginkgoales)和买麻藤目(Gnetales)], 包含6科[松科(Pinaceae)、南洋杉科(Araucariaceae)、罗汉松科(Podocarpaceae)、金松科(Sciadopityaceae)、柏科(Cupressaceae)和红豆杉科(Taxaceae)]69属605种(Christenhusz et al., 2011)。从与被子植物分离算起, 针叶树已独立进化了约3亿年(Troitsky et al., 1991), 其分类地位十分重要。针叶树虽然树种不多, 却主宰着北半球温带和寒带大部分陆地生态系统, 在全球碳循环中扮演着重要角色。同时, 针叶树多为重要造林树种, 经济价值巨大。鉴于针叶树重要的分类地位和生态、经济价值, 其序列资源挖掘研究受到生物进化学家、生态学家以及林木育种学家的高度重视, 获取针叶树全基因组序列对于更好地认识针叶树的起源与进化、维护针叶林生态健康(Neale, 2007)及加速基因组学辅助育种(genomics-assisted breeding)进程(许晨璐等, 2012)均具有重要意义。

针叶树基因组的进化机制及其组成、结构与被子植物显著不同(Kinlaw and Neale, 1997)。由于在现有

测序技术条件下, 基因组序列中基因分布特征、重复序列的数量、复杂度和分布直接影响到测序数据结果的输出, 基因组特征成为针叶树全基因组测序项目必须考虑的内容。本文总结了针叶树基因组特征, 依次介绍了目前已开展的针叶树转录组和全基因组测序项目, 以期为其它针叶树全基因组测序项目的开展提供参考。

1 针叶树的基因组特征

早期对针叶树基因组特征的研究多集中于其巨大的基因组本身。Gymnosperm DNA Cvalues数据库显示, 141个松科植物单倍体DNA含量高达9.5–36 Gb(Murray et al., 2012), 平均含量为23.68 Gb, 是拟南芥(*Arabidopsis thaliana*)的190倍, 毛果杨(*Populus trichocarpa*)的49倍。Zonneveld(2012)测定的64个属、172个针叶树种的基因组值为8.1–72 Gb, 仅罗汉松科和落叶松属(*Larix*)基因组相对较小(平均含量均为13.2 Gb)。与被子植物不同, 针叶树基因组没有表现出明显的多倍性(柏科除外)(Cui et al., 2006; Nystedt et al., 2013), 但不排除古多倍性(paleopolyploidy)。

收稿日期: 2013-01-26; 接受日期: 2013-08-05

基金项目: “十二五”国家科技支撑计划(No.2012BAD01B01)

* 通讯作者。E-mail: xmsun@caf.ac.cn

ploidy)的可能。如白云杉(*Picea glauca*)古基因(被子植物和裸子植物共有基因)复制事件出现的频率是针叶树特有基因复制事件出现频率的8倍(Pavy et al., 2012); 欧洲云杉(*Picea abies*)基因组中可能发生过古全基因组复制(ancient whole genome duplication)事件(Nystedt et al., 2013)。针叶树核型较被子植物保守, 其体细胞染色体基数介于7–19之间, 多集中于10–12条(Zonneveld, 2012), 如松科约有240个种的染色体基数为12, 仅北美黄杉(*Pseudotsuga menziesii*)为13。

1.1 针叶树的基因特征

1.1.1 基因序列含量

针叶树中真正编码蛋白质的序列仅占全基因组的很小一部分。欧洲云杉中仅有不足3%的序列与基因序列同源(Ingvarsson, 2012); 火炬松(*Pinus taeda*)中仅有1%是编码蛋白质的序列(Kovach et al., 2010)。由于基因组巨大, 针叶树基因密度较小, 对白云杉2条细菌人工染色体(bacterial artificial chromosome, BAC)序列(长度分别为172 kb和94 kb)进行测定, 各发现1个基因, 基因密度仅为拟南芥、水稻(*Oryza sativa*)、毛果杨和葡萄(*Vitis vinifera*)的十分之一(Hamberger et al., 2009); Kovach等(2010)对总长接近1 Mb的10个火炬松富集基因(BAC文库)进行测序, 也仅发现18个编码蛋白质的基因(其中15个可能是假基因)。与拟南芥等基因组含量相对较小的被子植物相比, 基因富含区域(gene-rich regions)在针叶树基因组中并不常见, 即使在基因富含区域, 基因密度也仅比非基因富含区域高出1倍(Pavy et al., 2012)。

尽管许多基因同时存在于针叶树和被子植物中, 但并不能据此准确推算出针叶树中编码蛋白质基因的数量。有研究表明, 针叶树中表达的基因数量可能与许多被子植物非常相似。如Rigault等(2011)估测白云杉核基因组含32 720个表达基因, 序列总长为47.3 Mb; 欧洲云杉中被强力支持的基因(well-supported gene)数为28 354个(Nystedt et al., 2013), 这与拟南芥和水稻的转录组大小相似。

1.1.2 基因特征

针叶树单个基因的长度大于被子植物中的同源基因(Kinlaw and Neale, 1997), 其原因可能是内含子较

长, 如最长的内含子为68 kb(Nystedt et al., 2013)。除转录因子外(Rigault et al., 2011), 针叶树拥有相对较大的基因家族(Kinlaw and Neale, 1997; Futamura et al., 2008; Morse et al., 2009), 这使得针叶树在应对环境变化时更具优势。假基因在针叶树中较为常见(García-Gil, 2008; Kovach et al., 2010; Nystedt et al., 2013), 它们可能参与大基因家族的形成(Kinlaw and Neale, 1997)。

针叶树蛋白质编码区域的进化速率较慢, 如松科不同种间的核苷酸替换速率仅为被子植物的十分之一(Buschizzo et al., 2012; Chen et al., 2012)。松科不同种间基因的同线性(syteny)和共线性(collinearity)较强(Krutovsky et al., 2004; Pavy et al., 2012), 如糖松(*Pinus lambertiana*)和火炬松虽隶属不同亚属, 但共线性却高达93%(Jermstad et al., 2011)。

1.2 针叶树重复序列特征

1.2.1 重复序列类型

针叶树中的重复序列包括串联重复序列(tandemly repeated sequence)和散布重复序列(dispersed repeated sequence)两种。前者包括SSR(simple sequence repeat)、小卫星DNA(minisatellite DNA)、卫星DNA(satellite DNA)、核糖体RNA基因和端粒重复序列等类型; 后者主要包括可移动元件(transposable element), 亦称转座子(transposon)。Ingvarsson (2012)研究发现, 欧洲云杉基因组中超过99%的序列是功能未知的中度或高度重复序列。也有研究认为, 针叶树基因组中约有75%是由重复序列构成, 20%–30%是由单拷贝序列构成(Kriebel, 1985), 按此比例计算, 单拷贝序列含量超过3 Gb, 远远超出基因表达所需要的100 Mb。后续研究证实, 这部分单拷贝序列可能也是重复序列(Elsik and Williams, 2001), 只不过由于进化时间过长, 导致分化严重而表现为单拷贝序列特征。如松属(*Pinus*)基因组中低拷贝非基因序列部分主要由极其发散的可移动元件构成(Morse et al., 2009; Kovach et al., 2010)。

转座子是针叶树中主要的重复序列类型。根据其转座机制不同, 转座子可分为DNA转座子(DNA transposon)和反转座子(retrotransposon)两类。DNA转座子是以DNA-DNA方式转座的转座子, 可通过

DNA复制或直接切除两种方式获得可移动片段,重新插入基因组DNA中,导致基因突变或重排,但一般不改变基因组的大小。反转座子则首先转录成RNA,然后以RNA为模板反转录合成新的转座子拷贝,再插入新的DNA序列中。反转座子可分为两类,即长末端重复序列反转座子(long terminal repeat retrotransposon, LTR-RTs)和非长末端重复序列反转座子(non-LTR-RTs)。其中, LTR-RTs包括*Ty1/Copia*和*Ty3/Gypsy*两个亚家族; non-LTR-RTs包括长散在重复(long interspersed element, LINEs)、短散在重复(short interspersed element, SINEs)和微型反向重复转座子(miniature inverted repeat transposable elements, MITEs)。研究发现,反转座子(特别是LTR-RTs)在针叶树中广泛存在且数量众多(Friesen et al., 2001; Ahuja and Neale, 2005; Nystedt et al., 2013),原因可能是针叶树缺乏有效的反转座子消除机制(Nystedt et al., 2013)。Hamberger等(2009)发现白云杉BAC文库中转座子占全部序列的20%; *Ty3/Gypsy*类反转座子的全部序列总长甚至超过拟南芥整个基因组的长度(Bennett et al., 2003)。欧洲云杉等6种针叶树中, *Ty3/Gypsy*亚家族的数量远大于*Ty1/Copia*(Nystedt et al., 2013)。相比数量庞大的反转座子,针叶树中DNA转座子的数量有限。

1.2.2 重复序列特征

与被子植物不同,针叶树基因组中的重复序列较为古老,彼此之间差异较大,且形式多样(Morse et al., 2009; Kovach et al., 2010; Magbanua et al., 2011; Liu et al., 2011)。例如Kovach等(2010)的研究发现,在总长接近1 Mb的火炬松BAC序列中包含近600种重复元件,包括SSR、反转座子和DNA转座子等,且任一重复序列占基因组的比例均不超过2%,欧洲云杉中超过86%的LTR-RTs以单拷贝形式存在,表明分化十分严重(Nystedt et al., 2013)。Parchman等(2010)的研究发现,扭叶松(*Pinus contorta*)中的反转座子具有很高的转录活性;也有研究认为所有的重复序列似乎累积了大量的突变、插入/缺失位点以及重排变异,表明这些重复序列可能已失去功能。针叶树转座子结构比较复杂,如火炬松中*PtIFG7-2*可能以巢式存在于众多特征未知的*Copia*类元件中(Kovach et al., 2010)。反转座子具有种属特异性,如某些*Ty3/*

*Gypsy*类LTR-RTs仅在云杉属中存在(Friesen et al., 2001)。

虽然有研究发现重复序列拷贝数量与基因组大小无显著相关性(Sarri et al., 2011),但散布于整个基因组的串联重复和反转座子扩增(retrotransposon expansion)的确构成了针叶树基因组的复杂性(Morse et al., 2009)。此外,转座子插入还被认为与长的内含子和假基因产生有关(Nystedt et al., 2013)。

2 转录组测序项目

鉴于EST测序可以快捷且高效地捕获基因,现已成为全基因组测序最有效的替代方法(Rudd, 2003)。自1998年首例高通量EST测序项目在火炬松中开展以来,多数针叶树种开展了EST或FL-cDNAs(full length cDNAs)测序项目(表1)。随着测序通量的增加,研究思路也从挖掘组织特异表达基因过渡到获取基因目录(gene catalogues)(Ralph et al., 2008; Rigault et al., 2011),相应的cDNA文库构建也从单组织过渡到多组织。为了尽可能获得更多的转录本,有些甚至对cDNA文库进行了均一化(normalization)处理(Rigault et al., 2011)。测序方法逐渐从传统的Sanger测序技术过渡到第2代测序技术,如Roche 454技术和Illumina HiSeq技术等。

此外,旨在利用第2代测序技术完成约1 000种植物转录组测序和组装的“千种植物转录组研究计划”(The 1000 Plant Transcriptome Project)已经对70个针叶树种进行了转录组测序。这些转录组测序项目极大地丰富了针叶树基因组资源。目前,DFCI(Dana Farber Cancer Institute)数据库中收录的针叶树转录本已达156 735条(其中包括松属77 326条和云杉属79 409条)。TreeGenes数据库中收录的松属和云杉属EST序列达1 021 388条。NCBI(National Center for Biotechnology) dbEST数据库中收录针叶树EST序列已达1 113 993条。这些EST序列的获得为针叶树全基因组测序的开展奠定了坚实的基础。

3 全基因组测序项目

尽管第1张林木全基因组序列草图早在7年前就已获得,然而针叶树全基因组测序却进展缓慢。2004年,美国农业部林务总局林木遗传研究所(Institute of

表1 近年来针叶树转录组测序的成果

Table 1 Conifer transcriptome sequencing programs reported since 1998

树种	主要完成单位	组织	通量	测序方法	参考文献
火炬松(<i>Pinus taeda</i>)	北卡州立大学(美国); 明尼苏达大学(美国)	木质部	833	Sanger	Allona et al., 1998
火炬松(<i>Pinus taeda</i>)	北卡州立大学(美国); 明尼苏达大学(美国)	木材形成组织	20 377	Sanger	Kirst et al., 2003
日本柳杉(<i>Cryptomeria japonica</i>)	林业及林产品研究所(日本)	种子和小孢子叶球	1 173	Sanger	Ujino-Ihara et al., 2003
火炬松(<i>Pinus taeda</i>)	乔治亚大学(美国)	根	6 202	Sanger	Lorenz et al., 2006
白云杉(<i>Picea glauca</i>)	拉瓦尔大学(加拿大); 明尼苏达大学(美国)等	各种组织	16 578	Sanger	Pavy et al., 2005
北美云杉(<i>Picea pungens</i>)等	英属哥伦比亚大学(加拿大)	各种组织	46 745	Sanger	Ralph et al., 2008
日本柳杉(<i>Cryptomeria japonica</i>)	林业及林产品研究所(日本)	雄球花	36 011	Sanger	Futamura et al., 2008
辐射松(<i>Pinus radiata</i>)	CSIRO Plant Industry(澳大利亚)	木质部	3 304	Sanger	Li et al., 2009
扭叶松(<i>Pinus contorta</i>)	怀俄明大学(美国)	针叶和小球果	303 450	454测序	Parchman et al., 2010
海岸松(<i>Pinus pinaster</i>)	马拉加大学(西班牙)	多个组织	55 322	Sanger/454测序	Fernández-Pozo et al., 2011
白云杉(<i>Picea glauca</i>)	Gytle公司(加拿大); 拉瓦尔大学(加拿大)等	各种组织	27 720	Sanger/454/RNA-seq	Rigault et al., 2011
南方红豆杉(<i>Laxus malrel</i>)	大连交通大学(中国)等	叶、干和根	36 493	Illumina	Hao et al., 2011
火炬松(<i>Pinus taeda</i>)等 ^①	乔治亚大学(美国); 俄亥冈州立大学(美国)等	枝条、针叶和茎尖	—	454测序	Lorenz et al., 2012
北美黄杉(<i>Pseudotsuga menziesii</i>)	霍恩海姆大学(德国); 多伦多大学密西沙加分校(加拿大)等	针叶和木材	170 859	454测序	Müller et al., 2012
欧洲云杉(<i>Picea abies</i>)	乌普萨拉大学(瑞典); 瑞典农业科学大学(瑞典)等	针叶	38 419	Illumina	Chen et al., 2012
扭叶松(<i>Pinus contorta</i>)和因蒂里厄云杉(<i>Picea engelmani</i> x <i>Picea glauca</i>)	英属哥伦比亚大学(加拿大)	顶芽、根和干	33 746; 40 862	第2代测序技术	Yeaman et al., 2013
北美黄杉(<i>Pseudotsuga menziesii</i>)	美国农业部林务总局西北太平洋研究站等	针叶	90 730	RNA-seq	Cronn et al., 2013
油松(<i>Pinus tabulaeformis</i>)	北京林业大学(中国)	多个组织	46 584	454测序	Niu et al., 2013
冷杉(<i>Akjes fabri</i>)	马尔堡大学(德国)等	一年生苗	25 149	454测序	Roschanski et al., 2013

① 包括糖松、北美黄杉、欧洲云杉、大西洋雪松、粗榧、长叶松、罗汉松、金松、北美红杉、欧洲红豆杉和瓦勒迈杉。
① Other species include *Pinus lambertiana*, *Pseudotsuga menziesii*, *Picea abies*, *Cedrus atlantica*, *Cephalotaxus harringtonia*, *Pinus palustris*, *Podocarpus macrophyllus*, *Sciadopitys verticillata*, *Sequoia sempervirens*, *Taxus baccata* and *Wollemia nobilis*.

Forest Genetics, USDA Forest Service)和加州大学戴维斯分校(UC Davis)联合发起了世界上第1个针叶树基因组测序项目——火炬松基因组项目(The Loblolly Pine Genome Project)。该项目以获得火炬松EST序列及全长蛋白质编码序列资源为首要目标,也

包含构建BAC文库、完整的物理图谱和高密度的遗传图谱,由此奠定了火炬松作为针叶树基因组学研究模式物种的地位。这一项目现已进行到松树参考序列项目阶段(表2)。现阶段,已经开展全基因组测序的针叶树种还包括欧洲云杉、白云杉、糖松、海岸松(*Pinus*

表2 针叶树全基因组测序项目

Table 2 Conifer whole genome sequencing projects

项目名称	树种	预期目标(基因组测序部分)	参与项目的科研机构	测序策略	测序方法
松树参考序列项目	火炬松(<i>Pinus taeda</i>)	(1) 获得火炬松高质量参考基因组序列	加州大学戴维斯分校、马里兰大学和印第安纳大学等	BAC文库和全基因组鸟枪法	Sanger/Illumina/454测序
	糖松(<i>Pinus lambertiana</i>)	(2) 为基因挖掘、注释和组装服务的转录组测序			
	北美黄杉(<i>Pseudotsuga menziesii</i>)	(3) 充实Dendrome和TreeGenes数据库			
	湿地松(<i>Pinus elliotii</i>)				
云杉基因组项目	欧洲云杉(<i>Picea abies</i>)	(1) 获得一个基因型的全基因组序列 (2) 获得不同组织基因的表达谱信息, 确定其中部分基因的功能	瑞典植物科学中心和SciLifeLab(瑞典)等	Fosmid pool和全基因组鸟枪法	Illumina
	白云杉(<i>Picea glauca</i>)	(1) 获得白云杉基因组序列草图 (2) 对基因组序列进行注释, 确认基因编码序列及针叶树特有基因			
SMarT-Forests	海岸松(<i>Pinus pinaster</i>) 欧洲赤松(<i>Pinus sylvestris</i>)	(1) 获得海岸松和欧洲赤松的参考基因组序列	阿尔卡拉大学(西班牙)和INIA-CIFOR(法国)等	基因组重测序	第2代测序技术

pinaster)、欧洲赤松(*Pinus sylvestris*)和北美黄杉。这些项目均由多家科研机构合作开展(表2)。

3.1 全基因组测序项目简介

由加州大学戴维斯分校联合另外6家科研机构共同实施的松树参考序列项目(Pine Reference Sequences)是由火炬松基因组项目发展而来。它以火炬松为主要研究树种, 采用Illumina双末端测序技术(Pair-end sequencing)对混合片段(大小包括500 bp、5 kb和40 kb)文库进行测序, 预期先发布覆盖度为21倍的全基因组鸟枪法(whole genome shotgun sequencing, WGS)序列和初始组装结果(包括富集基因的部分和全基因组), 随后2年内再发布具更高覆盖度(180倍)的组装结果, 之后再对糖松、湿地松(*Pinus elliotii*)和北美黄杉进行全基因组测序。这一项目的最新研究成果将在<http://pinegenome.org/pinerefseq/>网站上及时发布(其最新的组装版本(V1.0)已于2013年8月6日发布)。作为较早开展的全基因组测序项目, 松树参考序列项目也把探索适合其它针叶树的全基因组测序方法、技术平台和生物信息学软件, 及提高针叶树全基因组测序的速度和降低其成本作为预期目标。

瑞典植物科学中心(Umea Plant Science Center)联合10家科研机构(包括瑞典、加拿大、意大利和比利时等)共同开展的云杉基因组项目(The Spruce Genome Project)以获取欧洲云杉全基因组序列(包

括核基因组和叶绿体及线粒体基因组)为主要目标, 其它目标包括获得各组织基因的表达谱信息, 确定其中部分基因的功能、重要表型变异及与基因(或分子标记)的关联等信息, 以期为瑞典欧洲云杉的改良提供强大的基于基因组学的育种工具(genomics-assisted breeding tools)。该项目主要采用第2代测序技术对来自同一基因型的单倍体和双倍体进行同步测序, 最新的组装结果已于2013年5月22日在Nature上发表。

由加拿大3所高校共同发起的SMarTForests项目(<http://www.smartforests.ca/>)是在原有的Arborea项目(由拉瓦尔大学主持)和Treenomix项目(由英属哥伦比亚大学主持)的基础上开展的。它以获得白云杉的全基因组序列为首要目标, 同时评估全基因组的基因拷贝数量变异(copy number variation, CNV)和存在缺失变异(presence-absence variation, PAV), 以期为比较基因组学和关联研究打下基础。此外, 该项目还将开展白云杉与欧洲云杉的属内比较基因组学, 以及云杉与火炬松、糖松和北美黄杉的属间比较基因组学研究。目前, 借助Illumina HiSeq2000平台, SMarTForests项目已对来自白云杉的双倍体基因型(PG29)的21个paired-end文库进行了测序, 覆盖度高达63倍。借助ABYSS软件进行序列组装, 共获得600万条大于500 bp的重叠群(contig), 重叠群最大值高达78 855 bp, N50为4 234 bp, 序列组装总长度

为15.2 Gb。序列数据和首次组装结果(V1.0)已发布在NCBI(Bioproject PRJNA83435, Accession ALWZ-0000000000)网站上(Birol et al., 2013), 更新的组装结果将会陆续发布。

由欧盟组织和西班牙牵头, 联合法国、瑞典和比利时多家科研机构实施的ProCoGen项目(Promoting a Functional and Comparative Understanding of the Conifer Genome-implementing Applied Aspects for More Productive and Adapted Forests), 以挖掘对气候变化和林木生产力有影响的基因及基因调控网络为主要目标(Díaz-Sala and Cervera, 2011), 但目前公开发表的信息不多。

3.2 全基因组测序策略选择

针叶树全基因组测序首要的问题是选择合适的测序策略。鸟枪法测序(shotgun sequencing)以目的DNA随机打断成的大小不等的片段进行测序, 之后再利用计算机将这些片段序列连接起来。该方法现已成为大基因组测序项目普遍采用的测序策略。Green(2001)将该策略又细分为逐步克隆法(clone-by-clone shotgun sequencing)和全基因组鸟枪法两种途径。实际应用时, 也可以采取基于这两种方法的混合策略(mixed strategy sequencing)。

逐步克隆法(亦称之为hierarchical shotgun sequencing或BAC-based shotgun sequencing)是拟南芥、水稻和玉米(*Zea mays*)全基因组测序采用的方法。该方法遵守“先作图, 后测序”的原则, 需要精细的物理图谱作支撑。构建物理图谱的克隆系统包括酵母人工染色体载体(yeast artificial chromosome, YAC)、BAC和PAC(P1 artificial chromosome)等。其中BAC由于其插入片段大、稳定性好、嵌合体比例低和操作简单等优点已被广泛应用到人、动植物及微生物的基因组学研究, 成为基因组学研究的一个非常重要的工具。针叶树BAC文库的构建始于2007年, 到目前为止已对海岸松(Bautista et al., 2007)、火炬松(Morse et al., 2009; Kovach et al., 2010; Magbanua et al., 2011)、落羽杉(*Taxodium distichum*)(Liu et al., 2011)和白云杉(Hamberger et al., 2009)进行了BAC文库的构建, 并以此为基础进行了基因组序列特征的研究。然而, 鉴于针叶树的基因组非常巨大, 其不适合使用传统的BAC-based测序方法。以白云杉(基因

组大小为23 Gb)为例, 如果要构建一个基因组覆盖率为1倍(95%的可能性)、插入片段为107 kb的白云杉BAC文库, 则需要643 941个克隆, 工作量巨大。目前, 针叶树BAC克隆的构建都是为了分析针叶树基因组的结构特征(Bautista et al., 2007; Hamberger et al., 2009; Kovach et al., 2010), 而不是以全基因组测序为目的。

全基因组鸟枪法最早被用来对含重复序列较少的微生物基因组进行测序。该方法由于理论上不需要再构建物理图谱, 故迅速被多个植物全基因组测序项目所采用(Feuillet et al., 2011)。最近发布的植物基因组序列(玉米除外)都是借助全基因组鸟枪法获得的。经不断改进, 全基因组鸟枪法也成为适合针叶树全基因组测序的方法(Kovach et al., 2010; Birol et al., 2013)。针叶树应用此方法时, 一般是对100–800 bp的DNA片段双末端进行测序, 由于800 bp的DNA片段不足以覆盖针叶树中大部分的重复序列, 故需借助片段长度更长(10–20 kb)的跨步文库(jumping libraries)。对于重复序列更长的序列, 还可辅以Fosmid和BAC克隆末端测序(BAC-end sequence)或对Fosmid混合池(fosmids pools)的测序方法, 但WGS更依赖强大的数据处理软件支持(Miller et al., 2010)。

4 研究展望

尽管大基因组物种的序列拼接和组装面临很大的困难(Claros et al., 2012), 但大豆(*Glycine max*)(1.1 Gb)、玉米(2.73 Gb)、大麦(*Hordeum vulgare*)(5.1 Gb)和小麦(*Triticum aestivum*)(A、D基因组及部分B基因组)全基因组测序的完成, 以及最近大熊猫(*Ailuropoda melanoleuca*)(2.4 Gb)全基因组序列的发表对针叶树全基因组测序项目的开展具有很强的借鉴意义。第2代测序技术不仅使DNA测序成本7年间降为万分之一(Shendure and Aiden, 2012), 而且让基因组测序这项以前专属大型测序中心的“特权”能够被众多研究人员享用。目前, 第2代测序技术(如Illumina公司推出的读长高达2×300 bp的MiSeq测序仪)已使WGS很容易地实现对针叶树全基因组50–100倍的覆盖。第3代单分子测序技术在测序过程中省略了克隆步骤, 具有样品准备过程和分析流程简单、测序读长较长、测序精度高且能同时进行多

个样品分析等优点,极大地提高了测序速度,测序成本也降低了几个数量级(Delseny et al., 2010; Thudi et al., 2012)。由于读长更长,获得的重叠群长度是第2代测序技术的2倍(Schatz, 2012),这势必对未来其它针叶树全基因组测序项目的开展产生重要影响。

与基因组的巨大性相比,高复杂性对全基因组测序造成的困难更大。如针叶树中广泛存在的转座子对序列的组装构成了极大的障碍,特别是在短序列测序技术条件下,准确发现并处理这些高复杂度或低复杂度的重复区域成为全基因组测序至关重要的问题(Schatz et al., 2010)。重复序列会导致比对及组装时出现空缺(gap),或出现模棱两可的情况甚至无法直接比对和组装,而简单删除重复序列也不可取,因为这有可能删除有重要生物学功能的区域。

相对这些困难,针叶树基因组共线性强的特点成为测序及组装的有利条件,如多个物种的序列联合组装(joint assembly)可以提升组装的能力。对单倍体组织测序可以避免等位基因差异对序列组装的影响,起初人们认为单倍体DNA含量较少,可能不足以支持测序,但后来发现这一问题有可能通过组织培养技术获得成功解决。针叶树中有相当比例的重复序列比较古老,分化较为严重(Kovach et al., 2010),故针叶树基因组中重复序列较多可能并不是序列组装困难的原因,真正的困难可能是高昂的测序经费。

针叶树基因组资源挖掘已进入一个重要和多产的时期,首幅针叶树全基因组序列草图已于今年发布。这一里程碑事件将使针叶树基因组研究进入后基因组时代,同时也为主要针叶树树种的重测序及其它针叶树全基因组测序项目的开展打开了大门。相信在不久的将来,会有更多的针叶树树种列入全基因组测序项目中。

致谢 瑞典植物科学中心的Par K. Ingvarsson教授在成文过程中给予了大力支持,特此致谢!

参考文献

- 许晨璐, 张守攻, 孙晓梅 (2012). 针叶树基因组资源及其在遗传育种中的作用. 浙江农林大学学报 29, 768–777.
- Ahuja MR, Neale DB (2005). Evolution of genome size in conifers. *Silvae Genet* 54, 126–137.
- Allona I, Quinn M, Shoop E, Swope K, St Cyr S, Carlis J,

- Riedl J, Retzel E, Campbell MM, Sederoff R, Whetten RW (1998). Analysis of xylem formation in pine by cDNA sequencing. *Proc Natl Acad Sci USA* 95, 9693–9698.
- Bautista R, Villalobos DP, Díaz-Moreno S, Cantón FR, Cánovas FM, Claros MG (2007). Toward a *Pinus pinaster* bacterial artificial chromosome library. *Ann For Sci* 64, 855–864.
- Bennett MD, Leitch IJ, Price HJ, Johnston JS (2003). Comparisons with *Caenorhabditis* (~100 Mb) and *Drosophila* (~175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ~157 Mb and thus ~25% larger than the *Arabidopsis* genome initiative estimate of ~125 Mb. *Ann Bot* 91, 547–557.
- Biról I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MMS, Keeling CI, Brand D, Vanderwalk BP, Kirk H, Pandoh P, Moore RA, Zhao Y, Mungall AJ, Jaquish B, Yanchuk A, Ritland C, Boyle B, Bousquet J, Ritland K, MacKay J, Bohlmann J, Jones SJM (2013). Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. *Bioinformatics* 29, 1492–1497.
- Buschizzo E, Ritland C, Bohlmann J, Ritland K (2012). Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol Biol* 12, 8.
- Chen J, Uebbing S, Gyllenstrand N, Lagercrantz U, Lascoux M, Källman T (2012). Sequencing of the needle transcriptome from Norway spruce (*Picea abies* Karst L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. *BMC Genomics* 13, 589.
- Christenhusz MJM, Reveal JL, Farjon A, Gardner MF, Mill RR, Chase MW (2011). A new classification and linear sequence of extant gymnosperms. *Phytotaxa* 19, 55–70.
- Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N (2012). Why assembling plant genome sequences is so challenging. *Biology* 1, 439–459.
- Cronn R, Knaus BJ, Dolan P, Denver D, Clair BS (2013). Transcriptome dynamics of the dormancy-growth transition in Douglas-fir needles. Plant and Animal Genome XXI Conference. San Diego, CA. <https://pag.confex.com/pag/xxi/webprogram/Paper7999.html>
- Cui LY, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW

- (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res* **16**, 738–749.
- Delseny M, Han B, Hsing YI** (2010). High throughput DNA sequencing: the new sequencing revolution. *Plant Sci* **179**, 407–422.
- Díaz-Sala C, Cervera M** (2011). Promoting a functional and comparative understanding of the conifer genome—implementing applied aspects for more productive and adapted forests (ProCoGen). *BMC Proceedings* **5**, 158.
- Elsik CG, Williams CG** (2001). Families of clustered microsatellites in a conifer genome. *Mol Genet Genomics* **265**, 535–542.
- Fernández-Pozo N, Canales J, Guerrero-Fernández D, Villalobos DP, Díaz-Moreno SM, Bautista R, Flores-Monterroso A, Guevara MÁ, Perdiguero P, Collada C, Cervera MT, Soto Á, Ordás R, Cantón FR, Avila C, Cánovas FM, Claros MG** (2011). EuroPineDB: a high-coverage web database for maritime pine transcriptome. *BMC Genomics* **12**, 366.
- Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K** (2011). Crop genome sequencing: lessons and rationales. *Trends Plant Sci* **16**, 77–88.
- Friesen N, Brandes A, Heslop-Harrison JS** (2001). Diversity, origin, and distribution of retrotransposons (*gypsy* and *copla*) in conifers. *Mol Biol Evol* **18**, 1176–1188.
- Futamura N, Totoki Y, Toyoda A, Igasaki T, Nanjo T, Seki M, Sakaki Y, Mari A, Shinozaki K, Shinohara K** (2008). Characterization of expressed sequence tags from a full-length enriched cDNA library of *Cryptomeria japonica* male strobili. *BMC Genomics* **9**, 383.
- García-Gil MR** (2008). Evolutionary aspects of functional and pseudogene members of the phytochrome gene family in scots pine. *J Mol Evol* **67**, 222–232.
- Green ED** (2001). Strategies for the systematic sequencing of complex genomes. *Nat Rev Genet* **2**, 573–583.
- Hamberger B, Hall D, Yuen M, Oddy C, Hamberger B, Keeling CI, Ritland C, Ritland K, Bohlmann J** (2009). Targeted isolation, sequence assembly and characterization of two white spruce (*Picea glauca*) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defence reveal insights into a conifer genome. *BMC Plant Biol* **9**, 106.
- Hao DC, Ge GB, Xiao PG, Zhang YY, Yang L** (2011). The first insight into the tissue specific *Taxus* transcriptome via Illumina second generation sequencing. *PLoS One* **6**, e21220. doi:10.1371/journal.pone.0021220.
- Ingvarsson PK** (2012). A first look at the large and complex genome of Norway spruce (*Picea abies*). Plant and Animal Genome XX Conference. San Diego, CA. <https://pag.confex.com/pag/xx/webprogram/Paper2665.html>
- Jermstad KD, Eckert AJ, Wegrzyn JL, Delfino-Mix A, Davis DA, Burton DC, Neale DB** (2011). Comparative mapping in *Pinus*: sugar pine (*Pinus lambertiana* Dougl.) and Loblolly pine (*Pinus taeda* L.). *Tree Genet Genomes* **7**, 457–468.
- Kinlaw CS, Neale DB** (1997). Complex gene families in pine genomes. *Trends Plant Sci* **2**, 356–359.
- Kirst M, Johnson AF, Baucom C, Ulrich E, Hubbard K, Staggs R, Paule C, Retzel E, Whetten R, Sederoff R** (2003). Apparent homology of expressed genes from wood-forming tissues of loblolly pine (*Pinus taeda* L.) with *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **100**, 7383–7388.
- Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE, Loopstra CA, Hartigan J, Yandell M, Langley CH, Korf I, Neale DB** (2010). The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. *BMC Genomics* **11**, 420.
- Kriebel HB** (1985). DNA-sequence components of the *Pinus strobus* nuclear genome. *Can J For Res* **15**, 1–4.
- Krutovsky KV, Troggio M, Brown GR, Jermstad KD, Neale DB** (2004). Comparative mapping in the Pinaceae. *Genetics* **168**, 447–461.
- Li XG, Wu HX, Dillon SK, Southerton SG** (2009). Generation and analysis of expressed sequence tags from six developing xylem libraries in *Pinus radiata*, D. Don. *BMC Genomics* **10**, 41.
- Liu WX, Thummasuwan S, Sehgal SK, Chouvarine P, Peterson DG** (2011). Characterization of the genome of bald cypress. *BMC Genomics* **12**, 553.
- Lorenz WW, Ayyampalayam S, Bordeaux JM, Howe GT, Jermstad KD, Neale DB, Rogers DL, Dean JFD** (2012). Conifer DBMagic: a database housing multiple *de novo* transcriptome assemblies for 12 diverse conifer species. *Tree Genet Genomes* **8**, 1477–1485.
- Lorenz WW, Sun F, Liang C, Kolychev D, Wang HM, Zhao X, Cordonnier-Pratt MM, Pratt LH, Dean JFD** (2006). Water stress-responsive genes in loblolly pine (*Pinus taeda*) roots identified by analyses of expressed sequence tag libraries. *Tree Physiol* **26**, 1–16.
- Magbanua ZV, Ozkan S, Bartlett BD, Chouvarine P, Saski CA, Liston A, Cronn RC, Nelson CD, Peterson DG** (2011). Adventures in the enormous: a 1.8 million clone BAC library for the 21.7 Gb genome of loblolly pine. *PLoS*

- One **6**, e16214.
- Miller JR, Koren S, Sutton G** (2010). Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315–327.
- Morse AM, Peterson DG, Islam-Faridi MN, Smith KE, Magbanua Z, Garcia SA, Kubisiak TL, Amerson HV, Carloson JE, Nelson CD, Davis JM** (2009). Evolution of genome size and complexity in *Pinus*. *PLoS One* **4**, e4332.
- Müller T, Ensminger I, Schmid KJ** (2012). A catalogue of putative unique transcripts from Douglas-fir (*Pseudotsuga menziesii*) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings. *BMC Genomics* **13**, 673.
- Murray BG, Leitch IJ, Bennett MD** (2012). Gymnosperm DNA C-values database (release 5.0, Dec. 2012). <http://www.kew.org/cvalues/>
- Neale DB** (2007). Genomics to tree breeding and forest health. *Curr Opin Genet Dev* **17**, 539–544.
- Niu SH, Li ZX, Yuan HW, Chen XY, Li Y, Li W** (2013). Transcriptome characterisation of *Pinus tabulaeformis* and evolution of genes in the *Pinus* phylogeny. *BMC Genomics* **14**, 263.
- Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, Vicedomini R, Sahlin K, Sherwood E, Elfstrand M, Gramzow L, Holmberg K, Hällman J, Keech O, Klasson L, Koriabine M, Kucukoglu M, Käller M, Luthman J, Lysholm F, Niittylä T, Olson Å, Rilakovic N, Ritland C, Rosselló JA, Sena J, Svensson T, Talavera-López C, Theißen G, Tuominen H, Vanneste K, Wu ZQ, Zhang B, Zerbe P, Arvestad L, Bhalarao R, Bohlmann J, Bousquet J, Gil RG, Hvidsten TR, de Jong P, MacKay J, Morgante M, Ritland K, Sundberg B, Thompson SL, de Peer YV, Andersson B, Nilsson O, Ingvarsson PK, Lundeberg J, Jansson S** (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature*. doi: 10.1038/nature12211.
- Parchman TL, Geist KS, Grahn JA, Benkman CW, Buerkle CA** (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**, 180.
- Pavy N, Paule C, Parsons L, Crow JA, Morency MJ, Cooke J, Johnson JE, Noumen E, Guillet-Claude C, Butterfield Y, Barber S, Yang G, Liu J, Stott J, Kirkpatrick R, Siddiqui A, Holt R, Marra M, Seguin A, Retzel E, Bousquet J, MacKay J** (2005). Generation, annotation, analysis and database integration of 16 500 white spruce EST clusters. *BMC Genomics* **6**, 144.
- Pavy N, Pelgas B, Laroche J, Rigault P, Isabel N, Bousquet J** (2012). A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biology* **10**, 84.
- Ralph SG, Chun HJE, Kolosova N, Cooper D, Oddy C, Ritland CE, Kirkpatrick R, Moore R, Barber S, Holt RA, Jones SJM, Marra MA, Douglas CJ, Ritland K, Bohlmann J** (2008). A conifer genomics resource of 200 000 spruce (*Picea* spp.) ESTs and 6 464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). *BMC Genomics* **9**, 484.
- Rigault P, Boyle B, Lepage P, Cooke JEK, Bousquet J, MacKay JJ** (2011). A white spruce gene catalog for conifer genome analyses. *Plant Physiol* **157**, 14–28.
- Roschanski AM, Fady B, Ziegenhagen B, Liepelt S** (2013). Annotation and re-sequencing of genes from de novo transcriptome assembly of *Abies alba* (Pinaceae). *Applications in Plant Sciences* **1**(1), 1200179, doi: 10.3732/apps.1200179.
- Rudd S** (2003). Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci* **8**, 321–329.
- Sarri V, Ceccarelli M, Cionini PG** (2011). Quantitative evolution of transposable and satellite DNA sequences in *Picea* species. *Genome* **54**, 431–435.
- Schatz MC** (2012). De novo assembly of complex genomes using 3rd generation sequencing. Plant and Animal Genome XX Conference. San Diego, CA. <https://pag.confex.com/pag/xx/webprogram/Paper2444.html>
- Schatz MC, Delcher AL, Salzberg SL** (2010). Assembly of large genomes using second-generation sequencing. *Genome Res* **20**, 1165–1173.
- Shendure J, Aiden EL** (2012). The expanding scope of DNA sequencing. *Nat Biotechnol* **30**, 1084–1094.
- Thudi M, Li YP, Jackson SA, May GD, Varshney RK** (2012). Current state-of-art of sequencing technologies for plant genomics research. *Brief Funct Genomics* **11**, 3–11.
- Troitsky AV, Melekhovets YF, Rakhimova GM, Bobrova VK, Valiejo-Roman KM, Antonov AS** (1991). Angiosperm origin and early stages of seed plant evolution deduced from rRNA sequence comparisons. *J Mol Evol* **32**, 253–261.
- Ujino-Ihara T, Taguchi Y, Yoshimura K, Tsumura Y**

(2003). Analysis of expressed sequence tags derived from developing seed and pollen cones of *Cryptomeria japonica*. *Plant Biol* **5**, 600–607.

Yeaman S, Hodgins K, Nurkowski K, Rieseberg L, Aitken SN (2013). The genomics of adaptation to climate: de novo assembly and gene expression analysis of interior

spruce and Lodgepole pine. Plant and Animal Genome XXI Conference. San Diego, CA. <https://pag.confex.com/pag/xxi/webprogram/Paper7514.html>

Zonneveld BJM (2012). Conifer genome sizes of 172 species, covering 64 of 67 genera, range from 8 to 72 picogram. *Nordic J Bot* **30**, 490–502.

Characteristics of Conifer Genome and Recent Advances in Conifer Sequence Resources Mining

Chenlu Xu, Xiaomei Sun*, Shougong Zhang

Key Laboratory of Tree Breeding and Cultivation, State Forestry Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China

Abstract Conifers are the largest and most ubiquitous group of gymnosperms. They are woody perennials that shape many northern hemisphere ecosystems and support large industries through the provision of wood, fiber, and energy. Sequencing conifer genomes is relevant because of their taxonomic position, ecological significance, and economic importance. However, the large and complex genomes of conifers have hindered the mining process, and by April 2013, no whole-genome sequences for conifer have been obtained. With the emergence of next-generation sequencing and rapid development of bioinformatics, the conifer sequence resources mining project has transitioned from the transcriptome to whole-genome sequencing projects, initiated in *Pinus*, *Picea* and *Pseudotsuga* genus. This review summarizes the characteristics of the conifer genome, reviews the current status of genome sequencing of conifers and outlines the current state of knowledge concerning the genomes of conifer (*Pinus taeda*, *Picea abies* and *Picea glauca*).

Key words conifers, genome, research progress, transcriptome, whole genome sequencing

Xu CL, Sun XM, Zhang SG (2013). Characteristics of conifer genome and recent advances in conifer sequence resources mining. *Chin Bull Bot* **48**, 684–693.

* Author for correspondence. E-mail: xmsun@caf.ac.cn