遗传 HEREDITAS (Beijing) 2012 年 2 月, 34(2): 145－156
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2012.00145

木本植物全基因组测序研究进展

施季森, 王占军, 陈金慧

南京林业大学林木遗传与生物技术省部共建教育部重点实验室, 南京 210037

摘要: 近年来, 植物全基因组测序的结果正如雨后春笋般涌现, 木本植物全基因组测序也在紧锣密鼓地展开。但由于木本植物通常基因组较大, 基因组结构较为复杂, 在测序、测序后的组装、注释、功能分析等均存在较大的困难。在基因组测序分析的经费预算方面也存在着较大的压力。因此, 有必要对这方面的研究进展及其存在问题进行分析比较, 以提高林木全基因组研究方面的效率。文章在比较分析已经发展起来的 3 代基因测序技术(Sanger 测序法、合成测序法和单分子测序法)的基础上, 选择 4 种已经公布的木本植物(杨树、葡萄、番木瓜、苹果), 从全基因组测序的研究背景、测序结果及应用的研究进展和存在问题等方面进行了述评, 对未来要开展的木本植物全基因组测序前的准备工作(材料选择、遗传图谱和连锁图谱的构建、测序技术的选择), 全基因组测序结果的生物信息学分析和应用进行了讨论。

关键词: 木本植物; 测序技术; 全基因组

Progress on whole genome sequencing in woody plants

SHI Ji-Sen, WANG Zhan-Jun, CHEN Jin-Hui

Key Laboratory of Forest Genetics & Biotechnology of Ministry of Education, *Nanjing Forestry University*, *Nanjing* 210037, *China*

Abstract: In recent years, the number of sequencing data of plant whole genome have been increasing rapidly and the whole genome sequencing has been also performed widely in woody plants. However, there are a set of obstacles in investigating the whole genome sequencing in woody plants, which include larger genome, complex genome structure, limitations of assembly, annotation, functional analysis, and restriction of the funds for scientific research. Therefore, to promote the efficiency of the whole genome sequencing in woody plants, the development and defect of this field should be analyzed. The three-generation sequencing technologies (i.e., Sanger sequencing, synthesis sequencing, and single molecule sequencing) were compared in our studies. The progress mainly focused on the whole genome sequencing in four woody plants (Populus, Grapevine, Papaya, and Apple), and the application of sequencing results also was analyzed. The future of whole genome sequencing research in woody plants, consisting of material selection, establishment of genetic map and physical map, selection of sequencing technology, bioinformatic analysis, and application of sequencing results, was discussed.

Keywords: woody plants; sequencing technology; whole genome

收稿日期: 2011–09–07; 修回日期: 2011–11–02

基金项目: 国家自然科学基金重点项目(编号：30930077), 国家自然科学基金青年项目(编号：30901156), 国家林业局 948 引进项目(编号：2009-4-24), 江苏高校自然科学基金项目(编号：09KJA220001)和江苏省高校优势学科建设工程项目(PAPD)资助

作者简介: 施季森, 教授, 研究方向：林木遗传育种与林木基因组学。Tel: 025-85428711; E-mail: jshi@njfu.edu.cn

网络出版时间: 2012-1-11 14:35:52
URL: http://www.cnki.net/kcms/detail/11.1913.R.20120111.1435.004.html

林木不仅为人类提供了建筑、造纸等重要原料和其他可再生能源，而且在保水防沙和提高空气质量等领域起着重要的作用[1,2]。针对木本植物的全基因组测序研究不仅有助于了解树木的基因组结构和功能，而且对于探索木本植物的起源与进化、开展重要功能基因的定位和克隆、分子标记辅助选择(Marker-assisted selection, MAS)育种等均具有重要的指导意义。众所周知，木本植物生命周期较长、基因组的杂合度较高、基因组较大，且多数发生过基因组的复制，遗传背景不清晰，这些瓶颈限制了林木全基因组测序研究的进程[3,4]。2000 年，人们采用传统的Sanger测序技术完成拟南芥全基因组测序，揭开了植物全基因组研究的序幕[5]。2006 年，又利用Sanger测序技术完成了杨树的全基因组计划，开启了木本植物全基因组学研究的大门[6]。2011 年 2 月，继杨树[6]、葡萄[7]、番木瓜[8]、苹果[9]和桃树[10]基因组草图完成后，第 6 种木本植物麻风树的基因组草图也公布于世[11]，6 种木本植物全基因组测序的研究为人类进行其他木本植物的全基因组测序研究提供了大量的参考信息(表 1)。

作为植物全基因组研究的重要手段，测序技术经历了需要PCR扩增的第一代Sanger测序法和第二代合成测序法，已经发展到了无需PCR扩增，且具有高通量、低成本为特点的第 3 代单分子测序阶段(表 2) [14~18]。目前，第一代测序技术已经规模化，具有测序读长较长、测序精确度较高等优点，但测序成本高、速率等不足，因而制约其进一步扩大应用。与第一代测序技术相比，第二代测序技术降低了测序成本，提高了测序速率，且测序覆盖度更高。其中，美国的罗氏 454 测序技术测序读长较长(约在 400~1 000 bp/reads之间)，便于准确地序列拼接；但同聚物核苷酸测序时，存在较高的误差和价格较高的问题[18]；Solexa和SOLiD测序均属短读长技术，序列拼接繁琐且存在一定误差。相对于SOLiD而言，Solexa测序的错误率要高一些，SOLiD测序技术每个碱基都经过两次测序可检出和校正错误识别的碱基。第 3 代测序技术在测序过程中省略了克隆步骤，极大地提高了测序速度，测序成本也降低了好几个数量级，具有样品准备过程和数据分析流程简单、测序读长较长、测序精度非常高，能同时进行多个样品分析等优点[17]；但第 3 代测序技术的规模化应用尚需时日。

目前，第二代测序技术已被较多的应用于测序研究中，美国加利福尼亚大学利用罗氏 454 测序的最新技术Genome Sequencer FLX Titanium(GS FLX Titanium)[19]进行了海洋宏基因组测序，研究发现了固氮蓝藻新品种。2010 年*Nature*公布了采用Sanger和罗氏 454 两种测序法进行的苹果全基因组测序结果[9]，Illumina公司采用的低成本、高效率、快速从头组装测序短序列(35～100 bp)的测序方法，已成功

表 1 几种重要植物全基因组测序方法比较

物种	测序材料	测序方法	预测全长 (Mb)	覆盖深度	基因总数 (条)	基因均长 (bp)
拟南芥[5] (*Arabidopsis thaliana*)	哥伦比亚野生型	传统 Sanger (构建 BAC、TAC 和质粒文库)	125	—	31 114	2 373
水稻[12] (*Oryza sativa*)	籼稻 9311	Sanger	466	4.2×	37 544	5 000
	粳稻日本晴	Sanger	420	6.0×	40 844	—
黄瓜[13] (*Cucumis sativus*)	自交系 Chinese long 9930	Sanger 和 Illumina GA	243.5	72.2×	26 682	1 046
杨树[6] (*Populus trichocarpa*)	毛果杨雌株 Nisqually 1	Sanger	480	8.5×	45 555	2 300
葡萄[7] (*Vitis vinifera*)	自交纯系 PN 40024	Sanger	487	8.4×	30 434	3 399
番木瓜[8] (*Carica papaya*)	转基因雌株 SunUp	Sanger	372	3.0×	13 311	2 373
苹果[9] (*Malus domestica*)	优良品种金冠	Sanger 和罗氏 454	742.3	16.9×	57 386	—
桃树[10] (*Prunus persica*)	双单倍体系 Lovell	Sanger	227	7.7×	27 852	—
麻风树[11] (*Jatropha curcas*)	—	Sanger 和罗氏 454	285.9	—	40 929	—

表 2　3 代测序技术的特点[14~18]

测序技术	测序过程	产量和成本	优缺点
第一代　传统 Sanger 测序法 (ABI 公司)	PCR 扩增 DNA 片段合成随机末端分子——毛细管电泳分离——输出序列	6 Mb/day 约 500 $/Mb	测序读长较长、测序精确度较高，已经规模化，但测序成本高、速率低
第二代　合成测序法 (Roche 公司的 454, Illumina 公司的 Solexa, 和 ABI 公司的 SOLiD)	PCR 扩增 DNA 片段、序列组装——每次加单核酸，记录结果后洗脱、重复——输出序列	750～5 000 Mb/day 约 0.5～20 $/Mb	高速率、测序读长较长，同核苷酸聚物测序时误差较高，测序试剂价格较高(454)；低成本、高速率，测序读长较短，序列拼接繁琐(Solexa 和 SOLiD)；测序错误率较高(Solexa)；可确定错误识别碱基(SOLiD)
第三代　单分子测序法 (Helicos Biosciences 公司的 Heliscope, Pacific Biosciences 公司的 SMRT(single-molecule real-time), Life Technologies 公司的 FRET, Oxford Nanopore Technologies 公司的纳米孔单分子技术)	锚定 DNA 片段——每次加单核酸，记录结果后洗脱、重复或者实时测序——输出序列	5 000 Mb/day <0.5～20 $/Mb	低成本、高速率、长测序读长，简化了样品准备过程和数据分析流程，测序精度非常高

应用于黄瓜(Sanger 和 Illumina GA(Genome Analyzer))[13](表 1)、蚂蚁(*Camponotus floridanus* 和 *Harpegnathos saltator*)[20]、大熊猫(*Ailuropoda melanoleura*)[21]和土豆(*Solanum tuberosum* L.)[22]的全基因组测序研究中；第二代测序技术在遗传背景不清晰的木本植物全基因组研究中也将具有广阔的应用前景。新一代测序技术(New-generation sequencing technologies，NGSTs)的迅猛发展加速了植物全基因组，尤其是木本植物全基因组研究的进程[18]。但木本植物的测序研究策略和技术有许多方面仍然需要探索。本文通过分析其中 4 种较为典型的木本植物全基因组测序研究结果，围绕木本植物全基因组测序研究需要开展的前期基础工作，测序结果的生物信息学分析和应用进行分析和讨论。

1　杨树全基因组测序研究及结果利用

1.1　杨树全基因组测序的研究背景

杨树为杨柳科(Salicaceae)杨属(*Populus*)植物，染色体数为 2*n*=38。最古老的杨树发现于距今约 6 000 万年前的化石标本中，目前主要分布于北半球[23]。已经测序的毛果杨(*Populus trichocarpa*)的基因组为 480 Mb，系中等大小基因组物种。杨树生长迅速，易于进行常规育种和遗传转化等研究的实验操作[24]，表型遗传多样性丰富，遗传转化体系稳定[25]。已通过种间杂交建立了遗传作图群体，构建了标记有与生长速率、树高生长和木材材性等重要性状相关的遗传图谱[26]。因此，杨树被称为木本植物基因组研究的模式物种[23]。遗传图谱(Genetic map)和物理图谱(Physical map)的比较研究，更利于阐明基因组信息[27]，两种图谱的遗传标记相结合有助于测序结果的拼接和分析[28]。所以，在进行杨树全基因组测序研究之前，美国能源部橡树岭国家实验室联合田纳西州大学等组成的杨树全基因组课题组开展了大量的遗传图谱和物理图谱构建等前期基础性工作。

遗传图谱是通过遗传重组率计算得到的基因线性排列距离的图谱，该图谱的绘制依赖于DNA多态性标记的开发[29,30]。它不仅是研究遗传结构的有力工具，而且在重要经济和生物性状的QTL(Quantitative trait locus)定位、分子标记辅助选择[29]，后基因组时代编码基因的功能发掘，以及为进一步开展基因组保守序列和基因组测序提供基础信息[31]等研究也有重要价值。美国杨树课题组先后构建了一批遗传图谱，第一个为覆盖了 410 cM基因的遗传连锁图谱，总计有 356 个SSR标记，被标注于 155 个Scaffolds中。该图谱的标记有 91%为共显性的SSR标记[32]。Yin等[26]利用 544 个标记(439 个AFLP标记和 105 个SSR标记)构建了高密度遗传图谱，约覆盖了杨树基因组的 2 300～2 500 cM。该图谱为综合分析杨树基因组数据以及对将来开展其它重要树种的基因组结构、功能、进化和遗传改良研究等提供了研究基础。

物理图谱可确定被克隆的基因或DNA标记在染

色体上的精细位置。高密度的物理图谱为调控重要性状的基因进行定位、克隆和植物功能基因组研究提供了重要的信息[27]，并且有利于多位点、多个Contigs的共定位[3]。杨树基因组课题组采用大范围指纹图谱BAC文库法构建了包括 2 802 个Contigs的杨树物理图谱，估计覆盖了全基因组的 9.5 倍[6]。Tuskan等[6]结合遗传图谱和物理图谱信息，将 410 cM的拼接序列中接近 385 cM的序列锚定于杨树不同的连锁群。

1.2 杨树全基因组测序结果及利用

杨树全基因组课题组结合杨树遗传图谱信息，选择毛果杨雌株“Nisqually 1”为材料，采用全基因组鸟枪测序法绘制出毛果杨基因组草图，共获得4.2×10^9个高质量的核酸序列(Phred > 20)，基因组全长约 480 Mb，约覆盖杨树基因组的 8.5 倍。该研究共鉴定出 45 555 条蛋白编码基因，基因均长约 2 300 bp(http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html)(表 1)[6]。其中，包括来自于全长cDNA文库的 4 664 条基因全长序列。这些基因全长序列信息将有利于对杨树全基因组测序结果进行基因注释[6]。

在获得了杨树全基因组数据后，杨树全基因组课题组以基因的不同结构为研究单元，进行单核苷酸多态性(Single-nucleotide polymorphisms, SNPs)密度研究和进化分析；以拟南芥中基因组信息为参考，鉴定出杨树的纤维素合成相关基因和木质素合成酶基因。杨树纤维素合成相关基因家族中 93 个成员，34 个杨树木质素合成酶基因。通过进化分析研究发现，代表性的肉桂醇脱氢酶(Cinnamyl alcohol dehydrogenase, CAD)基因在毛果杨中是由单基因编码的，而在以往拟南芥的研究中发现是由 2 个基因编码的，这一发现对于开展杨树木质素的遗传操作带来极大的方便；他们还对黄酮类生物合成和类苯基丙烷类等次生代谢物合成酶基因、植物抗病基因(R基因)、生长素应答因子(Auxin response factor, ARF)、赤霉素(Gibberellins, GAs)和细胞分裂素(Cytokinins, CK)等植物激素相关基因进行了研究[6]。Woolbright等[33]参考杨树全基因组信息，在AFLP标记图谱和 4 000 个SSR标记(http://www.ornl.gov/sci/ipgc/ssr_resource. htm)基础上，利用 541 个AFLP标记和 111 个SSR标记加密了杂交杨的连锁遗传图谱。

2 葡萄全基因组测序研究

2.1 葡萄全基因组测序的研究背景

葡萄(*Vitis vinifera*)属于葡萄科葡萄属植物，染色体数为 2*n*=38，是重要的酿酒原料和水果[34]。葡萄具有悠久的栽培历史,早在新石器时代便有其记载[34]。在葡萄全基因组测序之前，遗传图谱和物理图谱构建已完成。Troggio等[35]以 94 个Syrah和Pinot Noir杂交F_1代群体为材料，综合SNPs、SSRs和AFLP分子标记，构建了葡萄遗传图谱。所有标记共覆盖了基因组的 1 245 cM的基因组大小，两个标记间的平均距离为 1.3 cM。该遗传图谱被锚定在了含有 994 个位点的BAC物理图谱上。

2.2 葡萄全基因组测序结果及利用

2007 年 9 月，*Nature*杂志报道了第 4 种显花植物(继拟南芥、水稻和杨树之后)，也是第二种木本植物和第一种果树——葡萄的全基因组测序结果[7]。研究表明，葡萄的基因组具有高度杂合的特点，有 13%的等位基因的序列之间存在明显的差异，高度杂合的特点严重阻碍了其测序后序列的拼接[7]。Jaillon等以通过连续自交方式获得的接近纯系(93%)的葡萄品系“PN 40024”(源于Pinot Noir)作为材料，采用全基因组鸟枪测序法进行测序，获得的基因组全长约为 487 Mb，约覆盖葡萄基因组的 8.4 倍。该研究共鉴定出 30 434 条蛋白编码基因，基因均长约 3 399 bp(http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/)(表 1)[7]，比杨树(45 555 条)[6]和水稻的蛋白编码基因(37 544 条—籼稻 9 311)[12]都要少(表 1)。

葡萄也是进行开花植物基因结构和起源进化研究较为理想的材料。Jaillon等[7]根据葡萄全基因组测序结果，进行了系统进化分析。按照蛋白质序列相似性原理分析了全基因组复制事件。结果表明，现有的单倍体的葡萄基因组在近代未发生全基因组复制事件(Genome-wide duplications, GWD)，而是经历了全基因组 3 倍化复制。这一事件也被视为古六倍体化进化机制的实例[7]。比较葡萄与杨树、拟南芥古六倍体化发生时间发现，葡萄的古六倍体化发生在杨树和拟南芥之后。这一结果说明，蔷薇类植物(Rosids)都起源于一个共同的古六倍体化祖先；葡萄

与拟南芥、杨树的比较显示, 拟南芥经历了两次全基因组复制事件后, 从真蔷薇I(Eurosid I clade)进化枝中分离出来, 演化为真蔷薇II(Eurosid II clade)进化枝, 而葡萄和杨树仍属于真蔷薇I(Eurosid I clade)进化枝。葡萄基因组序列与拟南芥、杨树和水稻基因组进行比对后发现, 葡萄与杨树的亲缘关系最近。项目组还利用蛋白质组学方法分析了葡萄的重要功能基因。他们发现葡萄基因组中存在两类与葡萄酒风味和保健作用直接相关的高拷贝数的基因, 这是在先前已经分析的植株中尚未发现的。葡萄酒的保健作用是由于天然植保素白藜芦醇的存在, 它直接与促进红酒消费市场的葡萄发育有关。在葡萄中已确认, 编码驱动白藜芦醇合成的对苯乙烯苷合酶(Stilbene synthases, STSs)基因家族中的基因已扩展到了 43 个, 而先前报道的仅有 20 个[7]; 葡萄酒的香气直接与驱动萜烯类(树脂、芳香精油类次生代谢物)合成的萜类合酶(Terpene synthases, TPSs)基因有关。在葡萄的基因组中, 发现有 89 个与萜烯类合成相关的功能基因和 27 个拟功能基因; 与拟南芥、水稻和杨树相比, 葡萄基因组中TPS基因家族的成员扩展了两倍[7]。葡萄基因组中还有许多重要发现, 限于篇幅不一一赘述。葡萄基因组中这些重要功能基因的发现, 不仅对于酿酒葡萄品质改良, 而且对于其它经济植物的品质改良也将起重要借鉴作用。

3　番木瓜全基因组测序研究

3.1　番木瓜全基因组测序的研究背景

番木瓜为番木瓜科(Caricaceae)番木瓜属(*Carica*)植物, 染色体数为 2*n*=18, 是世界第 5 大热带和亚热带的经济植物作物之一。番木瓜由最早仅在美国本土人工栽培, 现在已在世界许多地方栽培。番木瓜的生长周期短(仅为 9～15 个月), 周年开花[36]; 二倍体植株具有 9 对染色体, 且具有较为原始的性染色体(性染色体Mm为雄株, MM为雌株, M^hm为雌雄同株)， 基因组相对较小（372 Mb）; 而且番木瓜转基因体系已经十分成熟[37],是系统研究热带果树功能基因组的重要材料[38]。大量遗传图谱和物理图谱构建工作为番木瓜全基因组测序研究奠定了基础。Chen等[39]利用番木瓜AU9 和SunUp(旭日)的F_2代群体构建了高密度遗传图谱, 707 个标记被标注于 9 个主连锁群和 3 个小连锁群中, 包括 706 个微卫星标记和 1 个果色相关的标记。这些连锁标记覆盖了基因组的 1 037.1 cM的基因组大小, 标记间的平均间距为 1.45 cM。他们应用高信息含量的指纹系统(High-information-content fingerprinting system)构建的BAC文库中共有 39 168 个BAC克隆。

3.2　番木瓜全基因组测序结果及利用

2008 年 4 月, *Nature*杂志报道了美国夏威夷农业研究中心等组织完成的第一个用转基因(转抗环斑病毒基因)木本植物番木瓜(*Carica papaya* Linnaeus)的一个名为“旭日”(“SunUp” , 雌株)的品种为材料, 进行全基因组测序[8]。番木瓜基因组测序全长为 372 Mb, 覆盖番木瓜基因组的 3 倍, 约为拟南芥基因组大小的 3 倍, 与其他已完成测序的被子植物(拟南芥、水稻、杨树、葡萄)基因组相比, 番木瓜含有的基因数量最少, 仅含有 13 311 个基因, 基因均长约 2 373 bp(http://asgpb.mhpcc.hawaii.edu/papaya/)(表 1)[8], 具有的功能基因数量也最少, 仅含有少量与抗病相关的基因。Ming等[8]推测在长期的人工栽培过程中, 番木瓜的防御机制可能发生了特殊的进化。全基因组测序结果为在形态学、生理学、药学和营养学层面上开展番木瓜遗传育种研究提供了重要基础[8]。

系统发育研究表明, 番木瓜在近代也未发生基因组复制事件。番木瓜的一些片段分别与拟南芥中的 2～4 个片段呈共线性关系, 表明拟南芥的 1 次或 2 次基因组复制使得它与番木瓜之间的线性关系产生了分化。据推测, 拟南芥近代发生的α基因组复制事件可能仅影响到十字花科的一个成员, 而早先在双子叶植物出现早期发生的β基因组复制事件是拟南芥和番木瓜分化的真正原因[40]。基因组和亚基因组比对揭示番木瓜的γ基因组复制与拟南芥和杨树的γ基因组复制一致, 都是发生在靠近于被子植物最初发生分化的阶段[41]。番木瓜全基因组测序结果, 提供了一些具有重要科学价值的结果。虽然番木瓜基因组中既没有近代的基因组复制事件发生, 也不如其它已经测序的被子植物典型, 又是迄今为止已经测序的木本植物中功能基因数量最少的一个物种。但一个有趣的现象是在一些特定的功能群中, 基因的数量产生了十分明显的扩增。这些扩增基因是否有助于阐明番木瓜在长期的人工栽培中进化成

具有树木那样的习性还有待进一步研究，但明显看到测序品种基因扩增后在抗环纹病毒病能力的增加。另外，基因扩增对于番木瓜的淀粉的积累和运输，种子传播媒介的吸引，对于热带长日照的适应性，抗病基因(R基因)、转录因子、纤维素和木质素合成相关基因、生物钟调控基因、性别决定基因等功能研究方面，均有重要作用。项目组在测序材料的选择上与其他物种有特别之处，选择转基因材料进行测序研究，还可以对植物转基因后的目的基因的插入位置、插入拷贝数、目的基因的表达等重要科学问题进行研究。他们发现转番木瓜基因植株在核基因组中有 3 个位置与叶绿体的插入紧密关联，同时具有拓扑异构酶I的识别位置，这对于了解目的基因在转基因植株中的插入、表达和功能也具有十分重要的作用[8]。

4 苹果全基因组测序研究

4.1 苹果全基因组测序的研究背景

苹果属于蔷薇科(Rosaceae)、梨亚科(Pomoideae)、苹果属(*Malus Mill*)植物，染色体数为 2*n*=34，是大家所熟知的重要的水果之一。在苹果全基因组测序之前，也开展了遗传图谱和物理图谱的构建等研究工作。Han等[42]通过构建苹果基因组物理图谱，来研究复杂性状的遗传基础。该图谱最初约覆盖了单倍体基因组 10.5 倍，BAC文库中有 74 281 个克隆，包括 2 702 个Contigs，物理长度约 927 Mb。苹果基因组范围内物理图谱为染色体区域标记基因的开发、基因分离、QTL定位，比较基因组学分析植物染色体及全基因组测序等研究提供了基础。Han等[43]进一步分析了苹果基因组物理图谱，选择出 3 744 个高质量的BAC末端序列(BAC end sequences, BESs)进行标记开发。在大约 8.5%的BESs序列中发现有SSRs标记。将苹果BESs数据与拟南芥蛋白质组数据，拟南芥、杨树和水稻基因组数据进行了比较分析后，发现苹果与杨树的亲缘关系较拟南芥更近。2009 年，Han等[44]利用苹果BESs数据，筛选出了候选的SNPs标记，综合物理图谱和遗传图谱信息将候选SNPs定位到遗传图谱上。

4.2 苹果全基因组测序结果及利用

由意大利、美国和新西兰等国组成的项目组，以苹果(*Malus. domestica*)栽培品种“金冠(Golden Delicious)”为材料，采用Sanger和罗氏 454 测序法绘制了苹果基因组草图，同时讨论了苹果的起源及进化事件[9]。全基因组测序结果表明，苹果的基因组大小约为 742.3 Mb，覆盖苹果基因组的 16.9 倍，约含有 57 386 条蛋白编码基因(表 1)。

多数蔷薇科植物单倍体的染色体数在*n*=7～9 之间，而苹果属单倍体染色体数为 17。早先有报道认为苹果是由于发生了异源多倍体化(Allopolyploid)造成苹果染色体数目的非整倍性增多。新的研究发现，全基因组复制造成苹果染色体间存在大量的相似片段，树木个体内染色体间共线性分析表明，至少存在 4 对共线性长片段和 7 对共线性短片段。通过系统发育学分析发现苹果(*n*=17)与美吐根(*Gillenia*, *n*=9)的淀粉粒结合型淀粉合成酶*Wx*基因在两个种之间存在显著的线性关系。这很可能说明苹果属*n*=17 的单倍体的染色体数源于同源多倍体化(Auto- polyploidisation)。全基因组复制存在的古老的全基因组复制(Old genome-wide duplications, Old GWD)和近代的全基因组复制(Recent genome-wide duplications, Recent GWD)两种方式。苹果染色体间重组研究发现，其全基因组复制主要以近代的全基因组复制方式为主[9]。同时，部分证据支持真双子叶植物祖先的古代六倍体是单一起源的假说[9]。通过代表蔷薇科分类群中占大多数的梨亚科和苹果属的系统发生进化树的重构还表明，现代栽培苹果与新疆野苹果(*Malus sieveesii*)的亲缘关系比欧洲苹果(*Malus sylvestris*)、山荆子(*M. baccata*)、西府海棠(*M. micromalus*)、楸子(*M. prunifolia*)更近。Velasco等[9]认为，现代栽培苹果是由新疆野苹果进化而来，而不是早先认为的现代栽培苹果起源于欧洲苹果的观点。分子证据还支持这样的看法，现代栽培苹果(*M. domestica*)和新疆野苹果(*M. sieversii*)是同一个种，用*M. pumila* Mill.作为苹果的拉丁名可能更恰当。目前苹果基因组公布的仅仅是草图，但对于揭示苹果的起源和进化，寻找与许多重要性状相关的基因，开展分子植物育种等领域研究提供了重要突破口[9]。由于苹果自交不育、二倍体高度杂合，含有大量重复基因，苹果基因组测序和后续的序列拼接，基因组精细图的绘制等均面临着巨大的挑战[43]。

5　结　语

5.1　木本植物全基因组测序的前期基础

木本植物的全基因组测序研究面对的最大障碍是木本植物的基因组相对较大，而且较为复杂。测序材料的选择、测序结果的分析和应用等方面均要经过慎重考虑。在林木基因组项目启动之前，首先要考虑的是目标树种在理论方面可以阐述什么重要的科学问题，或者在应用方面有什么重要价值；其次，是要把测序材料的基本生物学背景了解清楚。大多数木本植物并未像水稻、玉米和大豆等农作物经历了较长时间驯化和近交繁殖[3,4]，可以较容易获得纯系。树木多数为异交物种，杂合性较强。经过反复杂交或回交的实验材料，基因组更为复杂。因此，在满足科学研究价值的前提下，应尽量选择基因组相对较小、分类地位上处于重要位置的原种作为测序材料，尤其是要优先选用单倍体材料或二倍体材料[7,10]，而不宜采用多倍体材料，以免木本植物高度杂合带来的测序和数据拼装的困难。由于经验不足或考虑不周,最近开始实施的一些木本植物的基因组计划中选择了天然的或人工的多倍体，或者是无意中选择了进化历史中发生过基因组复制事件的材料，造成了测序后基因组拼装和数据分析的困难，造成了不必要的人力、物力和时间的浪费。因此，为减少盲目性，在大规模深度测序之前，可以先作目标材料的基本生物学背景特征评估，如染色体倍性分析，或先作低覆盖度的初测序，了解基因组的复杂程度，以确定是否适合作复杂基因组的深度测序，以及能否开发出复杂基因组的拼装技术等均要一一作出决断。已经有遗传图谱和物理图谱的树种虽对基因组测序结果的分析有一定的帮助，但总体而言，遗传背景不一致、基因组高度杂合、进化和物种形成过程中发生过基因组复制事件等特征，均是木本植物遗传图谱、物理图谱的构建和全基因组的测序研究的限制因素。对于遗传背景不清晰的木本植物，可以先采用第二代测序技术进行转录组测序，以获得的大量Unigene信息构建遗传图谱和物理图谱，为待测序的物种提供遗传背景信息。

通量相对较高的第二代测序技术加速了木本植物全基因组测序的研究，但Solexa和SOLiD测序的读长较短，序列拼接繁琐且存在一定误差；454 测序读长较长，有利于大基因组和复杂基因组的拼装，但测序试剂的价格相对较高，而传统Sanger测序法具有测序读长较长、准确率高的优点，但存在效率相对较低的问题。因此有学者提出，将第一代和第二代测序方法或多种二代测序方法有效结合以提高测序的准确率，也便于测序结果的分析。例如黄瓜的全基因组测序是将Sanger 和 Illumina GA技术联合使用[13]、苹果和麻风树同为Sanger 和罗氏 454 技术结合[9] [11]，而土豆则是Illumina GA和罗氏 454 技术联用[22]，两种或多种测序方法的互补和联用进行测序研究，不失为当前一种可取的策略。2011 年 7 月，*Nature*杂志报到了利用ABI新推出的Ion Torrent PGM芯片式测序仪，在德国爆发的大肠杆菌的快速测序中应用，引起了业界的关注。“芯片就是测序仪”的概念逐渐开始流行。去年推出的 314 型芯片测序读长超过 100 bp，今年推出的 316 芯片读长达到 200 bp，2012 年面世的 318 芯片标称的读长要达到 400 bp，认为可以同罗氏的GS Junior相媲美，同时测序的费用也大大降低[45~47]。当然，随着单分子或其他更为先进的测序技术的问世和逐渐发展成熟，成本低、速率高、精度准的第 3 代测序技术在木本植物全基因组研究中必将有广阔的应用前景。

5.2　全基因组测序结果的生物信息学分析

第二代测序技术在木本植物全基因组测序研究中已有较多的应用。但是在第二代测序技术较多应用和测序设备、测序技术提高了测序效率的同时，也带来许多其他问题；如近来有研究发现，第二代测序技术的测序错读率相对较高。第二代测序结果的分析依赖于序列的拼接方法，随着拼接质量的提高，将会获得高质量的Contigs且增加了测序结果的覆盖率，二代测序技术中常用的序列拼接软件包括：GS Assembler(罗氏 454 商业化), Velvet[48], SOAP[49,50], ABySS[51], Edena[52], Euler-SR[53], CAP3[54], NextGENe[55], SHARCGS[56], FuzzyPath[57], QSRA[58], SeqCons[59], SHORTY[60], SOPRA[61], SSAKE[62], Taipan[63], VCAKE[64], MIRA, OASES和CLC等[55,65]。其中华大基因自行开发的SOAP软件已成功应用于大熊猫 [21]和土豆 [22]全基因组测序结果的拼接中，获得了较好的拼接结果；应用较广泛的Velvet软件具有能获得高覆盖度Contigs优点。然而单个拼接软件在不同物种测序结果的拼接中可能存在一定的局限，

Feldmeyer等[55]采用 3 种转录组拼接方法(Velvet, OASES和NGen)进行蜗牛转录组测序结果分析，结果表明NextGENe软件拼接的Contigs较长、Blast比对质量最高。Chen等先使用了 3 种软件(ABySS 1.2.1, Velvet 1.12 和Edena 2.1.1)进行了杯萼海桑转录组拼接[66]，再使用CAP3 对拼接结果进行二次拼接。多种序列拼接软件的比较研究能够提高序列拼接的精确度，对于非模式植物的全基因组测序研究具有重要的指导意义。复杂基因组测序及其第二代测序结果的准确率问题，已经引起了国内外科学家们的高度重视，并在努力攻关之中。

5.3 全基因组测序结果的应用

经过全基因组测序和生物信息学分析后，获得了大量的全基因组测序结果，测序结果能得到尽可能的全面应用，是对于树木基因组测序各种投入的最佳回馈。目前，可从以下方向开展应用研究：(1)利用生物信息学方法对全基因组测序结果进行基因注释，开展木本植物的重要性状相关基因的发现、克隆、功能验证和进化分析，例如控制开花、木材形成、树木生长习性、休眠、耐寒力、病虫害抗性(R基因)、果实发育、果实性状和品质等的相关基因；(2)进行比较基因组学研究，深入比较分析两种植物基因组序列的同线性关系，分析研究植物的起源和进化关系，同时探索控制植物重要性状的重要染色体片段或基因群(基因簇)，为重要基因的发现及克隆提供重要参考信息；(3)应用NGS获得的木本植物全基因组结果制作商品化基因芯片，使用Microarray方法进行基因表达研究；也可以直接采用NGS进行转录组测序，参考全基因组信息，利用转录组测序结果检测不同细胞或组织间基因的表达水平，为特定时间内基因表达的时空性提供信息；(4)基因组序列信息提供了大量的SSRs和SNPs等分子标记，有利于高密度遗传图谱和物理图谱的构建，高密度遗传图谱加速了分子标记与优良性状之间的连锁研究，有益于QTL研究平台的创建和染色体范围内研究自然群体基因渐渗；(5)探索全基因组关联分析(Genome-wide association studies, GWAS)，寻找个体基因组中表型相关的DNA序列的变异，评价决定个体基因型的成千上万单核苷酸多态性(SNPs)，由于GWAS使用的材料是自然群体，省去产生后代群体的过程，缩短了分子标记辅助选择育种的周期、降低了育种成本，并且自然群体能更准确地反映重组事件的发生[28,67]；(6)在全基因组测序结果的基础上，进行同种木本植物的不同时期、不同组织材料，野生型与突变体材料，未胁迫处理与各种胁迫处理(低温、干旱、高盐和ABA)材料[68]之间的转录组学(Transcriptomics)、代谢组学(Metabolomics)、蛋白质组学(Proteomics)和降解组学(Degradomics)的比较研究也亟待开展。

综上所述，选择合适的测序材料，与丰富的遗传背景、先进的测序技术、多种拼接方法综合使用及测序结果全面应用的有效结合，必然能促进木本植物全基因组测序研究的持续健康发展。

参考文献(References):

[1] Food and Agriculture Organization of the United Nations (FAO). The state of the world's forest. Food and Agriculture Organization of the United Nations, Rome, 2003, ISBN 92-5-104865-7. DOI

[2] Jansson S, Douglas CJ. *Populus*: a model system for plant biology. *Annu Rev Plant Biol*, 2007, 58: 435–458. DOI

[3] Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin TM, DiFazio SP, Ali J, Asano JK, Chan S, Cloutier A, Girn N, Leach S, Lee D, Mathewson CA, Olson T, O'connor K, Prabhu AL, Smailus DE, Stott JM, Tsai M, Wye NH, Yang GS, Zhuang J, Holt RA, Putnam NH, Vrebalov J, Giovannoni JJ, Grimwood J, Schmutz J, Rokhsar D, Jones SJM, Marra MA, Tuskan GA, Bohlmann J, Ellis BE, Ritland K, Douglas CJ, Schein JE. A physical map of the highly heterozygous *Populus* genome: integration with the genome sequence and genetic map and analysis of haplotype variation. *Plant J*, 2007, 50(6): 1063–1078. DOI

[4] Sterck L, Rombauts S, Jansson S, Sterky F, Rouzé P, Van de Peer Y. EST data suggest that poplar is an ancient polyploid. *New Phytol*, 2005, 167(1): 165–170. DOI

[5] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 2000, 408(6814): 796–815. DOI

[6] Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q,

Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, 2006, 313(5793): 1596–1604. DOI

[7] Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P; French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 2007, 449(7161): 463–467. DOI

[8] Ming R, Hou SB, Feng Y, Yu QY, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen CX, Qian WB, Shen JG, Du P, Eustice M, Tong E, Tang HB, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan PZ, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang JM, Wang JP, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang XY, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei HR, Irikura B, Paidi M, Jiang N, Zhang WL, Presting G, Windsor A, Navajas-Pérez R, Torres MJ, Feltus FA, Porter B, Li YJ, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang JM, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 2008, 452(7190): 991–996. DOI

[9] Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen ZT, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu CJ, Chu VT, King ST, Vick J, Tao QZ, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagné D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouzé P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel CE, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R. The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat Genet*, 2010, 42(10): 833–839. DOI

[10] International Peach Genome Initiative (IPGI). 2010. http://www.rosaceae.org/peach/genome. DOI

[11] Sato S, Hirakawa H, Isobe S, Fukai E, Watanabe A, Kato M, Kawashima K, Minami C, Muraki A, Nakazaki N, Takahashi C, Nakayama S, Kishida Y, Kohara M, Yamada M, Tsuruoka H, Sasamoto S, Tabata S, Aizu T, Toyoda A, Shin-i T, Minakuchi Y, Kohara Y, Fujiyama A, Tsuchimoto S, Kajiyama S, Makigano E, Ohmido N, Shibagaki N, Cartagena JA, Wada N, Kohinata T, Atefeh A, Yuasa S, Matsunaga S, Fukui K. Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res*, 2011, 18(1): 65–76. DOI

[12] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*, 2005, 436(7052): 793–800. DOI

[13] Huang SW, Li RQ, Zhang ZH, Li L, Gu XF, Fan W, Lucas WJ, Wang XW, Xie BY, Ni PX, Ren YY, Zhu HM, Li J, Lin K, Jin WW, Fei ZJ, Li GC, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia ZQ, Ren Y, Tian G, Lu Y, Ruan J, Qian WB, Wang MW, Huang QF, Li B, Xuan ZL, Cao JJ, Asan, Wu ZG, Zhang JB, Cai QL, Bai

YQ, Zhao BW, Han YH, Li Y, Li XF, Wang SH, Shi QX, Liu SQ, Cho WK, Kim JY, Xu Y, Heller-Uszynska K, Miao H, Cheng ZC, Zhang SP, Wu J, Yang YH, Kang HX, Li M, Liang HQ, Ren XL, Shi ZB, Wen M, Jian M, Yang HL, Zhang GJ, Yang ZT, Chen R, Liu SF, Li JW, Ma LJ, Liu H, Zhou Y, Zhao J, Fang XD, Li GQ, Fang L, Li YR, Liu DY, Zheng HK, Zhang Y, Qin N, Li Z, Yang GH, Yang S, Bolund L, Kristiansen K, Zheng HC, Li SC, Zhang XQ, Yang HM, Wang J, Sun RF, Zhang BX, Jiang SZ, Wang J, Du YC, Li SG. The genome of the cucumber, *Cucumis sativus* L. *Nat Genet*, 2009, 41(12): 1275–1281. DOI

[14] Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genom Human Genet*, 2008, 9(1): 387–402. DOI

[15] Kircher M, Kelso J. High-throughput DNA sequencing--concepts and limitations. *BioEssays*, 2010, 32(6): 524–536. DOI

[16] Munroe DJ, Harris TJR. Third-generation sequencing fireworks at Marco Island. *Nat Biotechnol*, 2010, 28(5): 426–428. DOI

[17] Delseny M, Han B, Hsing YI. High throughput DNA sequencing: The new sequencing revolution. *Plant Sci*, 2010, 179(5): 407–422. DOI

[18] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen ZT, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu PG, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 2005, 437(7057): 376–380. DOI

[19] Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, Tripp HJ, Affourtit JP. Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science*, 2008, 322(5904): 1110–1112. DOI

[20] Bonasio R, Zhang GJ, Ye CY, Mutti NS, Fang XD, Qin N, Donahue G, Yang PC, Li QY, Li C, Zhang P, Huang ZY, Berger SL, Reinberg D, Wang J, Liebig J. Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science*, 2010, 329(5995): 1068–1071. DOI

[21] Li RQ, Fan W, Tian G, Zhu HM, He L, Cai J, Huang QF, Cai QL, Li B, Bai YQ, Zhang ZH, Zhang YP, Wang W, Li J, Wei FW, Li H, Jian M, Li JW, Zhang ZL, Nielsen R, Li DW, Gu WJ, Yang ZT, Xuan ZL, Ryder OA, Leung FCC, Zhou Y, Cao JJ, Sun X, Fu YG, Fang XD, Guo XS, Wang B, Hou R, Shen FJ, Mu B, Ni PX, Lin RM, Qian WB, Wang GD, Yu C, Nie WH, Wang JH, Wu ZG, Liang HQ, Min JM, Wu Q, Cheng SF, Ruan J, Wang MW, Shi ZB, Wen M, Liu BH, Ren XL, Zheng HS, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie XY, Lu ZH, Zheng HC, Li YR, Steiner CC, Lam TT, Lin SY, Zhang QH, Li GQ, Tian J, Gong TM, Liu HD, Zhang DJ, Fang L, Ye C, Zhang JB, Hu WB, Xu AL, Ren YY, Zhang GJ, Bruford MW, Li QB, Ma LJ, Guo YR, An N, Hu YJ, Zheng Y, Shi YY, Li ZQ, Liu Q, Chen YL, Zhao J, Qu N, Zhao SC, Tian F, Wang XL, Wang HY, Xu LZ, Liu X, Vinar T, Wang YJ, Lam TW, Yiu SM, Liu SP, Zhang HM, Li DS, Huang Y, Wang X, Yang GH, Jiang Z, Wang JY, Qin N, Li L, Li JX, Bolund L, Kristiansen K, Wong GKS, Olson M, Zhang XQ, Li SG, Yang HM, Wang J, Wang J. The sequence and *de novo* assembly of the giant panda genome. *Nature*, 2010, 463(7279): 311–317. DOI

[22] Potato Genome Sequencing Consortium, Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, Orjeda G, Guzman F, Torres M, Lozano R, Ponce O, Martinez D, De la Cruz G, Chakrabarti SK, Patil VU, Skryabin KG, Kuznetsov BB, Ravin NV, Kolganova TV, Beletsky AV, Mardanov AV, Di Genova A, Bolser DM, Martin DM, Li G, Yang Y, Kuang H, Hu Q, Xiong X, Bishop GJ, Sagredo B, Mejía N, Zagorski W, Gromadka R, Gawor J, Szczesny P, Huang S, Zhang Z, Liang C, He J, Li Y, He Y, Xu J, Zhang Y, Xie B, Du Y, Qu D, Bonierbale M, Ghislain M, Herrera Mdel R, Giuliano G, Pietrella M, Perrotta G, Facella P, O'Brien K, Feingold SE, Barreiro LE, Massa GA, Diambra L, Whitty BR, Vaillancourt B, Lin H, Massa AN, Geoffroy M, Lundback S, DellaPenna D, Buell CR, Sharma SK, Marshall DF, Waugh R, Bryan GJ, Destefanis M, Nagy I, Milbourne D, Thomson SJ, Fiers M, Jacobs JM, Nielsen KL, Sønderkær M, Iovene M, Torres GA, Jiang J, Veilleux RE, Bachem CW, de Boer J, Borm T, Kloosterman B, van Eck H, Datema E, Hekkert BL, Goverse A, van Ham RC, Visser RG. Genome sequence and analysis of the tuber crop potato. *Nature*, 2011, 475(7355): 189–195. DOI

[23] Polle A, Douglas C. The molecular physiology of poplars: paving the way for knowledge-based biomass production. *Plant Biol*, 2010, 12(2): 239–241. DOI

[24] Tuskan GA, DiFazio SP, Teichmann T. Poplar genomics is getting popular: the impact of the poplar genome project on tree research. *Plant Biol*, 2004, 6(1): 2–4. DOI

[25] Stettler RF, Heliman PE, Bradshaw HD. Biology of

Populus And Its Implications for Management and Conservation. Ottawa: NRC Research Press, 1999: 1–7.

[26] Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA. Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. *Theor Appl Genet*, 2004, 109(3): 451–463. DOI

[27] Mozo T, Dewar K, Dunn P, Ecker JR, Fischer S, Kloska S, Lehrach H, Marra M, Martienssen R, Meier-Ewert S, Altmann T. A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nat Genet*, 1999, 22(3): 271–275. DOI

[28] Schneeberger K, Weigel D. Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci*, 2011, 16(5): 282–288. DOI

[29] Yin TM, Zhang XY, Huang MR, Wang MX, Zhuge Q, Tu SM, Zhu LH, Wu RL. Molecular linkage maps of the *Populus* genome. *Genome*, 2002, 45(3): 541–555. DOI

[30] Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol*, 2009, 27(9): 522–530. DOI

[31] Dinus RJ, Tuskan GA. Integration of molecular and classical genetics: a synergistic approach to tree improvement. In: Klopfenstein NB, Chun YW, Kim M-S, Ahuja MR, eds. Micropropagation, Genetic Engineering, and Molecular Biology of *Populus*. Fort Collins: General Technical Report RM-GTR-297, USDA Forest Service, 1997: 220–235. DOI

[32] Tuskan GA, Gunter LE, Yang ZK, Yin TM, Sewell MM, DiFazio SP. Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa*. *Can J Forest Res*, 2004, 34(1): 85–93. DOI

[33] Woolbright SA, Difazio SP, Yin T, Martinsen GD, Zhang X, Allan GJ, Whitham TG, Keim P. A dense linkage map of hybrid cottonwood (*Populus fremontii* × *P. angustifolia*) contributes to long-term ecological research and comparison mapping in a model forest tree. *Heredity*, 2008, 100(1): 59–70. DOI

[34] Marguerit E, Boury C, Manicki A, Donnart M, Butterlin G, Némorin A, Wiedemann-Merdinoglu S, Merdinoglu D, Ollat N, Decroocq S. Genetic dissection of sex determinism, inflorescence morphology and downy mildew resistance in grapevine. *Theor Appl Genet*, 2009, 118(7): 1261–1278. DOI

[35] Troggio M, Malacarne G, Coppola G, Segala C, Cartwright DA, Pindo M, Stefanini M, Mank R, Moroldo M, Morgante M, Grando MS, Velasco R. A dense single- nucleotide polymorphism-based genetic linkage map of grapevine (*Vitis vinifera* L.) anchoring Pinot Noir bacterial artificial chromosome contigs. *Genetics*, 2007, 176(4): 2637–2650. DOI

[36] Ming R, Yu QY, Moore PH. Sex determination in papaya. *Semin Cell Dev Biol*, 2007, 18(3): 401–408. DOI

[37] Fitch MMM, Manshardt RM, Gonsalves D, Slightom JL, Sanford JC. Virus resistant *papaya* plants derived from tissues bombarded with the coat protein gene of *papaya* ringspot virus. *Nat Biotechnol*, 1992, 10(11): 1466–1472. DOI

[38] Liu ZY, Moore PH, Ma H, Ackerman CM, Ragiba M, Yu QY, Pearl HM, Kim MS, Charlton JW, Stiles JI, Zee FT, Paterson AH, Ming R. A primitive Y chromosome in papaya marks incipient sex chromosome evolution. *Nature*, 2004, 427(6972): 348–352. DOI

[39] Chen CX, Yu QY, Hou SB, Li YJ, Eustice M, Skelton RL, Veatch O, Herdes RE, Diebold L, Saw J, Feng Y, Qian WB, Bynum L, Wang L, Moore PH, Paull RE, Alam M, Ming R. Construction of a sequence-tagged high-density genetic map of *papaya* for comparative structural and evolutionary genomics in *Brassicales*. *Genetics*, 2007, 177(4): 2481–2491. DOI

[40] Bowers JE, Chapman BA, Rong JK, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 2003, 422(6930): 433–438. DOI

[41] Schranz ME, Mitchell-Olds T. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell*, 2006, 18(5): 1152–1165. DOI

[42] Han YP, Gasic K, Marron B, Beever JE, Korban SS. A BAC-based physical map of the apple genome. *Genomics*, 2007, 89(5): 630–637. DOI

[43] Han YP, Korban SS. An overview of the apple genome through BAC end sequence analysis. *Plant Mol Biol*, 2008, 67(6): 581–588. DOI

[44] Han YP, Chagné D, Gasic K, Rikkerink EHA, Beever JE, Gardiner SE, Korban SS. BAC-end sequence-based SNPs and Bin mapping for rapid integration of physical and genetic maps in apple. *Genomics*, 2009, 93(3): 282–288. DOI

[45] Katsnelson A. DNA sequencing for the masses. *Nature*, 2010, doi:10.1038/news.2010.674. DOI

[46] Zakaib GD. Chip chips away at the cost of a genome. *Nature*, 2011, 475(7356): 278. DOI

[47] Sims PA, Greenleaf WJ, Duan HF, Xie XS. Fluorogenic DNA sequencing in PDMS microreactors. *Nat Methods*, 2011, 8(7): 575–580. DOI

[48] Zerbino DR, Birney E. Velvet: algorithms for de novo

short read assembly using de Bruijn graphs. *Genome Res*, 2008, 18(5): 821–829. DOI

[49] Li RQ, Li YR, Kristiansen K, Wang J. SOAP: Short oligonucleotide alignment program. *Bioinformatics*, 2008, 24(5): 713–714. DOI

[50] Li RQ, Zhu HM, Ruan J, Qian WB, Fang XD, Shi ZB, Li YR, Li ST, Shan G, Kristiansen K, Li SG, Yang HM, Wang J, Wang J. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 2010, 20(2): 265–272. DOI

[51] Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao YJ, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJM. *De novo* transcriptome assembly with ABySS. *Bioinformatics*, 2009, 25(21): 2872–2877. DOI

[52] Hernandez D, François P, Farinelli L, Østerås M, Schrenzel J. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res*, 2008, 18(5): 802–809. DOI

[53] Chaisson MJ, Pevzner PA. Short read fragment assembly of bacterial genomes. *Genome Res*, 2008, 18(2): 324–330. DOI

[54] Huang XQ, Madan A. CAP3: A DNA sequence assembly program. *Genome Res*, 1999, 9(9): 868–877. DOI

[55] Feldmeyer B, Wheat CW, Krezdorn N, Rotter B, Pfenninger M. Short read Illumina data for the *de novo* assembly of a non-model snail species transcriptome (*Radix balthica*, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC Genomics*, 2011, 12(1): 317. DOI

[56] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Res*, 2007, 17(11): 1697–1706. DOI

[57] Sudbery I, Stalker J, Simpson JT, Keane T, Rust AG, Hurles ME, Walter K, Lynch D, Teboul L, Brown SD, Li H, Ning ZM, Nadeau JH, Croniger CM, Durbin R, Adams DJ. Deep short-read sequencing of chromosome 17 from the mouse strains A/J and CAST/Ei identifies significant germline variation and candidate genes that regulate liver triglyceride levels. *Genome Biol*, 2009, 10(10): R112. DOI

[58] Bryant DW Jr, Wong WK, Mockler TC. QSRA: a quality-value guided de novo short read assembler. *BMC Bioinformatics*, 2009, 10: 69. DOI

[59] Rausch T, Koren S, Denisov G, Weese D, Emde AK, Döring A, Reinert K. A consistency-based consensus algorithm for *de novo* and reference-guided sequence assembly of short reads. *Bioinformatics*, 2009, 25(9): 1118–1124. DOI

[60] Hossain MS, Azimi N, Skiena S. Crystallizing short-read assemblies around seeds. *BMC Bioinformatics*, 2009, 10: S16. DOI

[61] Dayarian A, Michael TP, Sengupta MA. SOPRA: Scaffolding algorithm for paired reads via statistical optimization. *BMC Bioinformatics*, 2010, 11(1): 345. DOI

[62] Warren RL, Sutton GG, Jones SJM, Holt RA. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 2007, 23(4): 500–501. DOI

[63] Schmidt B, Sinha R, Beresford-Smith B, Puglisi SJ. A fast hybrid short read fragment assembly algorithm. *Bioinformatics*, 2009, 25(17): 2279–2280. DOI

[64] Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V, Mardis ER, Dangl JL, Jones CD. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 2007, 23(21): 2942–2944. DOI

[65] Paszkiewicz K, Studholme DJ. *De novo* assembly of short sequence reads. *Brief Bioinform*, 2010, 11(5): 457–472. DOI

[66] Chen SF, Zhou RC, Huang YL, Zhang M, Yang GL, Zhong CR, Shi SH. Transcriptome sequencing of a highly salt tolerant mangrove species *Sonneratia alba* using Illumina platform. *Mar Genomics*, 2011, 4(2): 129–136. DOI

[67] Nordborg M, Weigel D. Next-generation genetics in plants. *Nature*, 2008, 456(7223): 720–723. DOI

[68] Ishitani M, Xiong L, Stevenson B, Zhu JK. Genetic analysis of osmotic and cold stress signal transduction in *Arabidopsis*: interactions and convergence of abscisic acid-dependent and abscisic acid-independent pathways. *Plant Cell*, 1997, 9(11): 1935–1949. DOI