植物科学学报 2021, 39(6): 681~691
Plant Science Journal http://www.plantscience.cn

DOI:10.11913/PSJ.2095-0837.2021.60681
谢玲娟, 叶楚玉, 沈恩惠. 植物基因组测序研究进展[J]. 植物科学学报, 2021, 39(6): 681-691
Xie LJ, Ye CY, Shen EH. Advances in plant genome construction [J]. *Plant Science Journal*, 2021, 39(6): 681-691

植物基因组测序研究进展

谢玲娟[1], 叶楚玉[1], 沈恩惠[1,2*]

(1. 浙江大学农业与生物技术学院, 杭州 310058; 2. 浙江大学新农村发展研究院, 杭州 310058)

摘 要: 从第一个模式植物拟南芥被测序, 植物基因组测序已经有21年的历史。随着科学技术的不断发展进步, 测序成本大幅降低, 基因组的组装质量显著提升。本文统计了2000 - 2020年间植物参考基因组从头测序的进展, 分析了植物基因组测序数量变化与测序技术发展之间的联系, 以及测序基因组大小与倍性及重复序列的关系, 总结了历年来测序物种的主要分类及其在系统发生树的分布, 最后讨论了未来植物基因组潜在的研究方向。

关键词: 植物基因组; 测序技术; 基因组大小; 物种分布; 物种分类

中图分类号: Q943 **文献标识码**: A **文章编号**: 2095-0837(2021)06-0681-11

Advances in plant genome construction

Xie Ling-Juan[1], Ye Chu-Yu[1], Shen En-Hui[1,2*]

(1. *College of Agriculture and Biotechnology*, *Zhejiang University*, Hangzhou 310058, China;
2. *Rural Development Academy*, *Zhejiang University*, Hangzhou 310058, China)

Abstract: Since the first model plant, *Arabidopsis thaliana* (L.) Heynh, was sequenced in 2000, significant advances have been made in the sequencing of plant genomes over the last 21 years. With continuous development of technology, the cost of sequencing has greatly reduced and genome quality has significantly improved. The tremendous information hidden in these sequences should provide valuable resources for biological research. Here, we summarized and discussed the advances in plant reference genome *de novo* sequencing that have occurred over the last 21 years. We analyzed dynamic changes between the annual amount of sequenced plant genomes and sequencing technology, explored the relationship between sequenced genome size and chromosome ploidy and repetitive sequences, and summarized the main species classifications and distributions of sequenced plant genomes in the species phylogenetic tree. Finally, potential research hotpots of plant genomes were discussed.

Key words: Plant genome; Sequencing technology; Genome size; Species distribution; Species classification

2000年, 第一个模式高等植物拟南芥的全基因组序列信息发表[1], 此后, 植物基因组相关研究如雨后春笋层出不穷。植物基因组测序的重要性不言而喻, 通过测序, 我们就像获得了植物生长发育、繁殖、适应性的"说明书"。浩瀚的基因组数据能够提升生物学研究的广度和深度[2], 为学者研究植物重要性状(如果实品质、抗性水平、开花时间等)提供了有力工具[3-7], 还可据此推测基因组演变的框架[8-11], 大大提升了科学家研究物种进化等问题及准确联系表型与变异的能力。随着测

收稿日期: 2021-06-02, 修回日期: 2021-07-08。
基金项目: 中央高校基本科研业务费(2021QN81013)。
This study was supported by a grant from the Fundamental Research Funds for the Central Universities (2021QN81013).
作者简介: 谢玲娟(1997-), 女, 硕士研究生, 研究方向为生物信息学(E-mail: 1303748955@qq.com)。
* 通讯作者(Author for correspondence. E-mail: enhuishen@zju.edu.cn)。

序技术日新月异的发展，测序速度和质量不断提升，测序成本大幅降低[12, 13]，陆续公布的植物基因组序列不仅有助于挖掘植物中丰富的基因遗传资源[14, 15]，促进对关键农艺性状候选基因的深入探索和分子标记的开发[16]，同时也为植物进化研究提供了数据基础[17-19]，已经成为作物育种改良的重要资源和工具。

本文回顾了植物参考基因组的从头测序工作，收集了2000-2020年的相关测序进展，讨论了历年测序植物基因组的文章数量变化、测序基因组大小、测序物种的系统关系等内容，有利于研究人员选择相关测序技术和把控未来基因组研究趋势。

1 植物基因组测序进展概况

截至2020年底，已有917篇文章报道了植物全基因组精细图和草图的绘制，涉及843个不同的物种(图1)。2000年*Nature*上发表了模式植物拟南芥(*Arabidopsis thaliana* (L.) Heynh.)的测序工作[1]，正式揭开了植物基因组测序的序幕。这项艰巨的任务由多个实验室合作完成，是植物基因组学研究的一个里程碑，极大地激发和加速了植物基因组研究[20]。两年后，基因组比拟南芥大3倍多的禾本科植物水稻(*Oryza sativa* L. ssp. *indica*)作为第一个单子叶模式作物被测序并发表[21]。一个是没有经过人工选择的野生植物，另一个是长期受到选择的栽培植物，这两者基因组测序的完成，给生物学家们的研究带来了新的可能性。

测序技术的不断发展与进步推动着植物基因组测序数量的快速增长和规模的不断扩大。最初拟南芥[1]、水稻[21]和单细胞绿藻(*Ostreococcus tauri*)[22]等物种的全基因组测序和组装利用的是第一代DNA测序技术，主要为桑格(Sanger)等提出的双脱氧链终止法[23]。但受限于Sanger测序技术时间长、成本高、通量低等缺点，植物基因组测序的进展缓慢。第二代测序技术原理多为边合成边测序，与第一代技术相比通量大大提高。2005年454 Life Sciences公司基于焦磷酸测序[24]推出了第一台二代测序系统Genome Sequencer 20 System，随后其它测序平台也陆续推出，包括世界上广泛采用的Illumina公司的HiSeq和NovaSeq系列。第二代测序技术的高通量、低成本以及在基因组组装中相对成熟的算法，使得植物全基因组测序迎来了爆发式增长[25-27]。然而某些复杂基因组具有较高的重复序列和杂合率，仅用二代测序的短读长(reads)进行基因组拼接会产生较高的错误率，尤其是在高重复序列区域。因此第三代测序技术应运而生，特别是由美国Pacific Biosciences (PacBio)公司开发的SMRT(single molecule real-time)和英国Oxford Nanopore Technology的纳米孔(nanopore)测序技术实现了长读长和单分子测序，大大降低了基因组拼接的难度。SMRT测序采用荧光标记的脱氧核苷酸(dNTP)和零模波导孔技术(zero-mode waveguides，ZMW)对单个DNA分子进行测序，平均读长在15 kb[28]，但其测序准确率仅为87.5%。2019年，PacBio更新了其平台，使用环形一致性测序(circular consensus sequencing，CCS)模式生成高精度的读序(high fidelity reads，HiFi)[29]，利用该技术，目前已有包括马铃薯

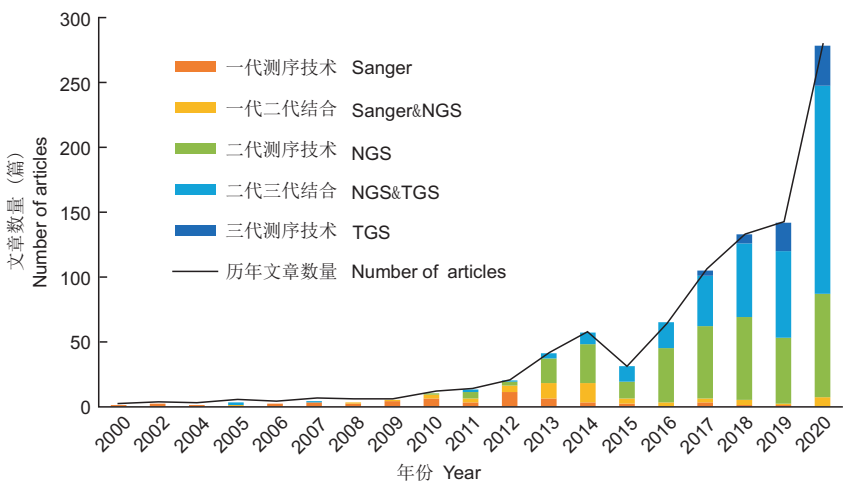


图1 近年来基于不同测序技术发表的植物基因组文章数量

Fig. 1 Number of published articles among different sequencing technologies on plant genomes in recent years

(*Solanum tuberosum* L.)[30]、紫花苜蓿(*Medicago sativa* L.)[31]、野苹果(*Malus domestica* cv. Gala)[32]、紫果西番莲(*Passiflora edulis* Sims.)[33]等在内的 6 个物种完成了组装，展示出了较高的准确率和完整性。纳米孔测序技术的基本原理是当 DNA 或 RNA 分子通过纳米孔时，由于碱基大小不同，引起孔内电阻变化，通过检测这些电信号从而识别碱基排列，平均读长在几十至数百 kb[34]。这些长读长的序列可以直接跨过大多数重复序列区域，使组装准确率得到大幅提升[35]。此外，近年来一些其他辅助基因组拼接的技术也陆续出现，如高通量染色体构象捕获技术(Hi-C)，不仅可以构建染色体交互矩阵，还能在 Scaffold 的聚类、排序、定向中起到重要作用[36]；而 BioNano 光学图谱技术通过生成单 DNA 分子全基因组限制性内切酶图谱，可以增加基因组 Scaffold 的长度，对已拼接的基因组进行纠错以及检测大片段结构变异等[37]。结合这些新兴技术，可使基因组组装达到染色体水平，且准确性更高，提供更多的变异信息，能够逐步解决复杂基因组拼接的需求，优化基因组组装。

2　已测序植物物种基因组大小分布

基因组的多倍化以及重复序列的积累是植物基因组扩张的主要机制。多倍化对基因组大小有直接影响，即使在没有杂交的情况下，也会增强遗传多样性和基因组动态性，为基因的新功能化和亚功能化提供契机[38]。植物中，重复序列是基因组的重要组成成分，已测序植物基因组中重复序列平均含量约占拼接序列的 50%，大蒜(*Allium sativum* L.)中其含量甚至高达 91.3%[39](附表 1[1])。基因组上重复序列的动态变化可在多个方面影响植物基因组进化，包括基因组的结构与组成，表观遗传效应，以及更微观的调控基因表达等[40]。

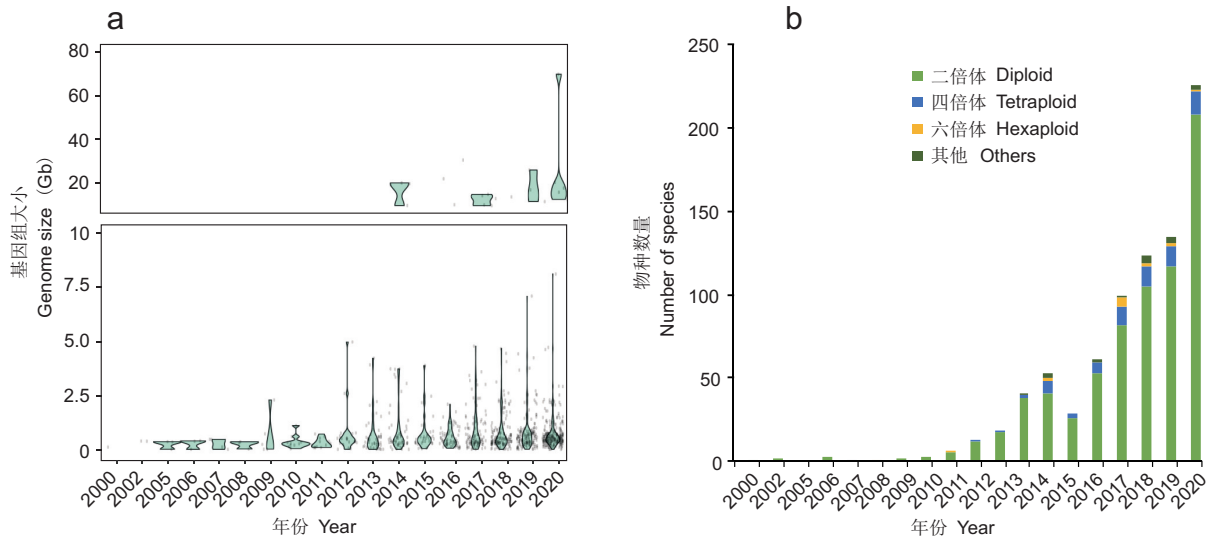
对 2000－2020 年测序的植物基因组拼接大小进行统计，发现植物基因组大小跨度很大，从极端性红藻(*Galdieria phlegrea*)的 11.4 Mb[41]到百合科重楼属七叶一枝花(*Paris polyphylla* var. *yunnanensis*)的 82.6 Gb[42]，相差了 3 个数量级(图 2：a)。历年测序植物物种基因组拼接大小基本分布在 700 Mb － 1.5 Gb 左右。基因组在 10 Gb 左右或者更大的植物主要是小麦(*Triticum aestivum* L.)[43, 44]和松(*Pinus lambertiana* Dougl.)、杉(*Picea glauca* (Moench) Voss)[45-47]等林木。2010 年之前，除了玉米(*Zea mays* L.)的拼接基因组超过 2 Gb(2.4 Gb)[48]之外，其余都小于 730 Mb，反映了当时测序与拼接技术的局限性。直到 2012 年后，相对复杂的植物基因组才陆续被研究，测序基因组的大小随着年份变化呈增加趋势，大量重复序列超过 60%的基因组也陆续发表，这是测序技术发展、拼接算法提升等原因共同促成的结果(图 2：a，附表 1[1])。植物多倍化对其进化具有重要意义，通过遗传物质的加倍可以丰富物种的多样性，增加物种对环境的适应性，为植物育种提供重要材料和资源。二倍体物种的基因组相对简单，对拼接算法等要求相对较低。在已测序的植物物种中，近 80%都是二倍体(图 2：b)。近年来，多倍体植物的研究也得到了快速发展，尤其是四倍体(图 2：b)。如，2018 年研究人员对野生马铃薯 M6(*Solanum tuberosum* L.)进行了测序组装，获得了农艺性状向栽培马铃薯渗入的重要资源[49]；2020 年科学家对 5 种异源多倍体棉花(*Gossypium*)进行了基因组进化分析，发现他们通过亚基因组转座子交换而呈现多样化[50]；此外，像油菜(*Brassica napus* L.)[51, 52]、烟草(*Nicotiana tabacum* L.)[53]等作为重要的模式生物，其基因组也被多次测序并广为研究。

3　植物基因组测序物种分类和分布

3.1　测序植物分类统计

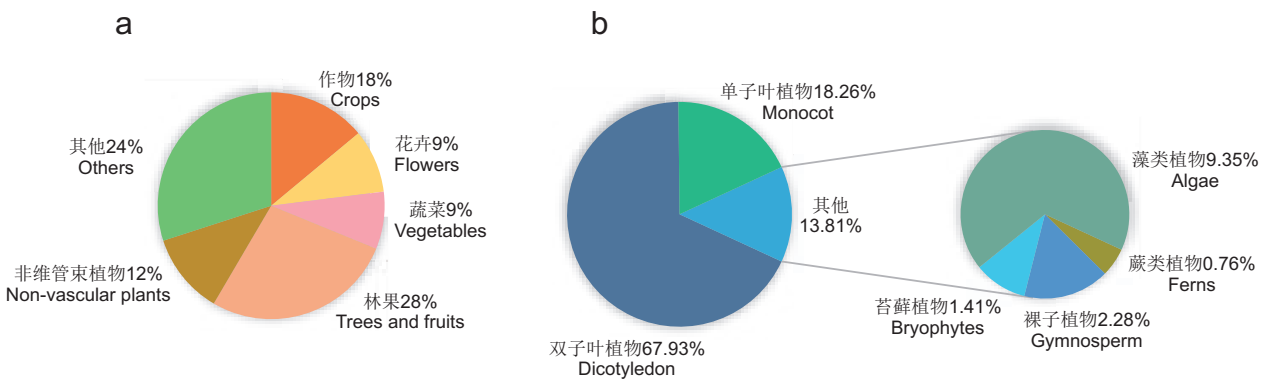
根据植物的用途不同，大致可分为大田作物、花卉、蔬菜、林果、非维管束等类型。统计近 20 年的测序物种，林木果树所占比例最高，为 28%，大田作物为 18%，花卉 9%，蔬菜 9%，藻类等非维管束植物占 12%(图 3：a)。可以看出，尽管作物基因组研究的相对较早，但林木果树的研究发展更加快速。对测序植物类型进行统计分析，发现被子植物占比最高，达 86%以上，其中双子叶植物是单子叶植物的近 4 倍，这是因为双子叶植物中作物、水果、种子油料等物种的重要性程度更高[14]；测序最少的是蕨类植物，不到 1%；此外，裸子植

1)如需查阅附表内容请登录《植物科学学报》网站(http：//www.plantscience.cn)查看本期文章。



a：植物基因组拼接大小分布；b：所有植物测序基因组倍性情况。
a: Distribution of assembly size from sequenced genomes in plants; b: Chromosome ploidy of sequenced plant genomes.

图 2 历年测序植物物种基因组大小和倍性统计
Fig. 2 Statistics on genome size and ploidy of sequenced plant species over the years



a：根据不同功能用途划分；b：根据不同类型进行划分。
a: Classification by different functional groups; b: Classification by different taxonomy types.

图 3 已测序植物物种类型统计
Fig. 3 Statistics of sequenced plant species

物、藻类植物和苔藓植物分别为 2.28%、9.35%和 1.41%(图 3：b)。

3.2 测序植物物种科目分布

在已测序的植物物种中，研究数量排前 10 的科是禾本科(Poaceae)、十字花科(Brassicaceae)、豆科(Fabaceae)、蔷薇科(Rosaceae)、茄科(Solanaceae)、桃金娘科(Myrtaceae)、木犀科(Oleaceae)、锦葵科(Malvaceae)、唇形科(Lamiaceae)、葫芦科(Cucurbitaceae)、菊科(Asteraceae)、杨柳科(Salicaceae)，表 1 总结了这 10 个科的主要代表物种。作为重要模式物种，这些植物的基因组测序是革命性的里程碑事件，大大加速了相关研究进程并拓宽了研究方向。例如，2014 年研究人员结合高密度小麦 SNP 遗传图谱和二代测序技术绘制了六倍体普通小麦品种‘中国春’的基因组草图[43]，对小麦亚基因组和小麦近缘二倍体、四倍体进行比较基因组学研究发现，亚基因组间序列相似性高，结构保守，基因丢失较少。但由于小麦谱系的分化，在整个基因组中仍出现了较多基因新增、丢失和复制的证据。此外，将每条染色体分离出来分别测序，化整为零的思路大大避免了同源染色体之间的相互混淆，为今后复杂多倍体基因组研究提供了借鉴[43]。此前，Guo 等[54]对广泛存在的稻田杂草—稗草(*Echinochloa crus-galli* (L.) P. Beauv.)进行了测序，揭示了稗草与水稻竞争的分子机制，并为水稻育种了

表 1　文章数量排名前 10 的测序植物基因组科名及代表物种

Table 1　The names and representative species of the top 10 plant families by number of articles

科 Family	文章数量 Number of articles	代表物种 Representative species
禾本科 Poaceae	111	水稻(*Oryza sativa* L.)、小麦(*Triticum aestivum* L.)、大麦(*Hordeum vulgare* L.)、玉米(*Zea mays* L.)、高粱(*Sorghum bicolor* (L.) Moench)、甘蔗(*Saccharum officinarum* L.)、稗草(*Echinochloa crus-galli* (L.) P. Beauv.)、竹子(*Bambusoideae*)、茭白(*Zizania latifolia* (Griseb.) Stapf)
十字花 Brassicaceae	70	拟南芥(*Arabidopsis thaliana* (L.) Heynh.)、油菜(*Brassica napus* L.)、芥菜(*Brassica juncea* (L.) Czern. & Coss.)
豆科 Fabaceae	53	大豆(*Glycine max* (L.) Merr.)、花生(*Arachis hypogaea* L.)、紫花苜蓿(*Medicago sativa* L.)
蔷薇科 Rosaceae	40	蔷薇(*Rosa* sp.)、草莓(*Fragaria* × *ananassa* Duch.)、苹果(*Malus pumila* Mill.)、杏(*Armeniaca vulgaris* Lam.)、桃(*Amygdalus persica* L.)、梨(*Pyrus* spp.)、月季(*Rosa chinensis* Jacq.)
茄科 Solanaceae	40	马铃薯(*Solanum tuberosum* L.)、烟草(*Nicotiana tabacum* L.)、番茄(*Solanum lycopersicum* L.)、辣椒(*Capsicum annuum* L.)
桃金娘科 Myrtaceae	27	巨桉(*Eucalyptus grandis* Hill.)、白桉(*Eucalyptus alba* Reinw.)、蓝桉(*Eucalyptus globulus* Labill.)、稀花桉(*Eucalyptus pauciflora* Sieb.ex Spreng)
木犀科 Oleaceae	26	油橄榄(*Olea europaea* ssp.*africana*)、桂花(*Osmanthus fragrans* (Thunb.) Lour.)、白蜡树(*Fraxinus chinensis* Roxb.)
锦葵科 Malvaceae	22	棉花(*Gossypium* spp.)、可可(*Theobroma cacao* L.)、榴莲(*Durio zibethinus* Murr.)
唇形科 Lamiaceae	17	薰衣草(*Lavandula angustifolia* Mill.)、薄荷(*Mentha haplocalyx* Briq.)、一串红(*Salvia splendens* Ker-Gawler)
葫芦科 Cucurbitaceae	17	南瓜(*Cucurbita moschata* Duchesne)、冬瓜(*Benincasa hispida* (Thunb.) Cogn.)、西瓜(*Citrullus lanatus* (Thunb.) Matsum. et Nakai)、葫芦(*Lagenaria siceraria* (Molina) Standl.)、丝瓜(*Luffa cylindrica* (L.) Roem.)、苦瓜(*Momordica charantia* L.)、罗汉果(*Siraitia grosvenorii* Swingle)
菊科 Asteraceae	17	菊花(*Dendranthema morifolium* (Ramat.)Tzvel.)、向日葵(*Helianthus annuus* L.)、莴苣(*Lactuca sativa* L.var. *angustanaIrish.*)
杨柳科 Salicaceae	17	胡杨(*Populus euphratica* Oliv.)、细叶杨(*Populus deltoids* cv. Zhonghua hongye)、旱柳(*Salix matsudana* Koidz.)

提供重要的基因遗传资源。2019 年，研究人员利用基因组领域前沿的组装、光学图谱等技术，组装了连续性提高近几十倍的棉花(*Gossypium hirsutum* L.、*Gossypium barbadense* L. var. *acuminatum* (Roxb.) Mast.)参考基因组，有助于科学家解析陆地棉和海岛棉在产量、适应性和品质方面的差异原因，并为培育优良棉花新品种奠定了坚实的基础[55]。研究较多的禾本科、十字花科和豆科植物包含了许多大田作物和园艺植物，这些基因组序列将在植物分子设计、品种创制以及设计育种等方面发挥重要作用，产生巨大的社会和经济效益。

3.3　不同用途植物基因组的分布

测序刚兴起时，对进行测序的植物一般都有特定的要求，如研究群体规模大、模式植物、基因组较小、倍性低、重复序列含量低、低杂合度等[12]。因此，最初的测序物种主要集中于模式植物和简单藻类，主要目的在于获得其全基因组序列以推进植物生物学研究。然而模式植物虽然有助于理解基本的生物机制，却不能完全反映其他物种的特异性和复杂的遗传机制，加上基因型与环境互作的影响，仅用模式植物进行研究存在很多的不确定性[17]。因此，近二十年来，研究者对越来越多的植物物种甚至是一个物种的多个品种进行了基因组构建。这些测序物种集中在对人类社会生活产生重要功能的植物，如作物、林果、花卉、蔬菜等(图 4)。从年度分布来看，对作物的研究开始最早(2002 年水稻)[21]，紧接着是非维管束(2004 年温泉红藻

Cyanidioschyzon merolae)[56]、林果(2006 年毛果杨 *Populus trichocarpa* Torr. & Gray)[57]、蔬菜(2009 年黄瓜 *Cucumis sativus* L.)[58]，而花卉相关研究在 2013 年才开始有报道[59]。测序对象从最初的实用性植物逐渐向观赏性植物转变，反映出基因组测序的普适性和重要性，已经跟人类生活密不可分。作物和林果的研究趋势大体相同，在数量上林果相关研究反而要超过作物。据此可以推测，未来基因组测序的植物物种还是会集中在具有重要经济价值的粮食、饲料、纤维和油料作物等类型中，以持续为社会发展创造更大的效益。

3.4 测序物种的系统关系和分布

基于 Angiosperm Phylogeny Group Ⅳ(APG Ⅳ)和 1000 个植物转录组的系统发生推论[60]，我们构建了目前已发表的 843 个基因组的系统发生树并绘制了无根树(图 5，附图 1[1])。结果显示，测序植物分布在 69 个目(Order)，其中包含物种数量最多的是禾本目(Poales)，含有 113 个物种，其次是十字花目(Brassicales)、唇形目(Lamiales)和蔷薇目(Rosales)，分别含有 74、72 和 58 个物种。包含物种最少的 21 个目仅分别含有 1 个已测序物种，如团藻目(Volvocales)、褐指藻目(Phaeodactylales)、葫芦藓目(Funariales)、金鱼藻目(Ceratophyllales)等。无根树(附图 1)分析结果显示，已测序物种分为被子植物门(Angiospermae)(双子叶植物纲(Dicotyledons)，单子叶植物纲(Monocotyledons)、裸子植物门(Gymnospermae)、蕨类植物门(Pteridophyta)、石松门(Lycopodiophyta)、苔藓植物门(Bryophyta)、轮藻门(Charophyta)、绿藻门(Chlorophyta)、蓝藻门(Cyanophyta)以及红藻门(Rhodophyta)。双子叶植物纲的物种被测序的最多，特别是蔷薇亚纲分支已有超过 259 个物种被测序，说明这一类物种受到的关注更多。裸子植物门下分 5 纲，其中银杏纲(Ginkgopsida)、松柏纲(Pinopsida)和买麻藤纲(Gnetopsida)的物种测序较多，结合裸子植物在建材、燃料等领域的重要作用，推测未来会有更多苏铁纲(Cycadopsida)和红豆杉纲(Taxopsida)的物种被测序。绿藻门是藻类植物中最大的一门，分为绿藻纲(Chlorophyceae)和轮藻纲(Charophyceae)，分别有 48 和 23 个物种被测序。藻类植物富含蛋白质，可作为食品或动物饲料，此外，绿藻还可用作藻类生理生化研究的材料，宇宙航行的供氧体，以及水体指示生物等。可以预测，绿藻门、蓝藻门、轮藻门、红藻门等藻类物种，在测序数量和类型上将不断扩增和丰富。

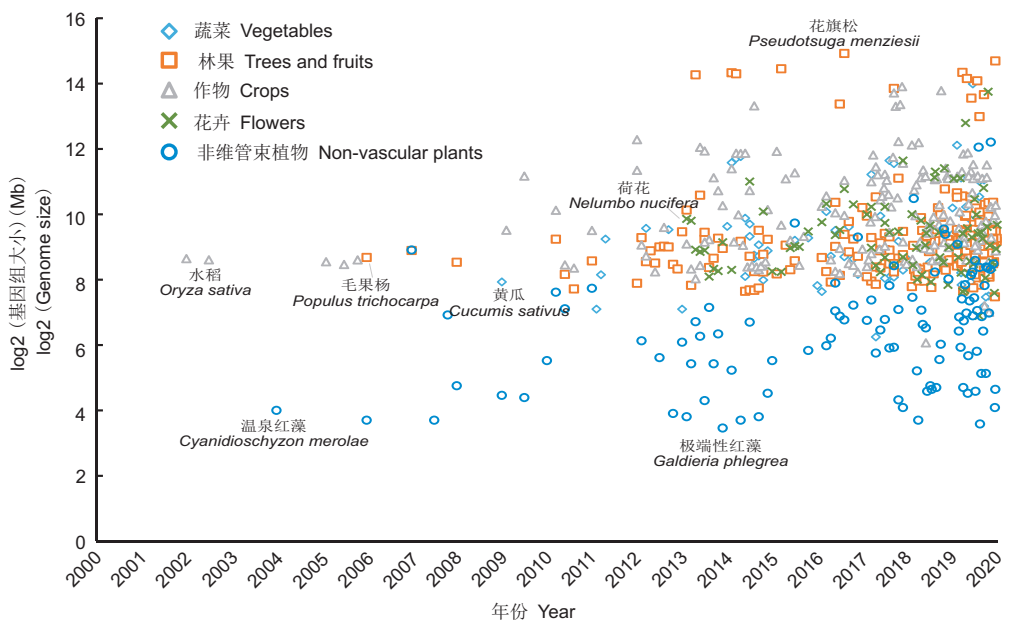


图 4 植物功能大类的测序进展

Fig. 4 Sequencing progress of plant functional groups

1) 如需查阅附图内容请登录《植物科学学报》网站(http://www.plantscience.cn)查看本期文章。

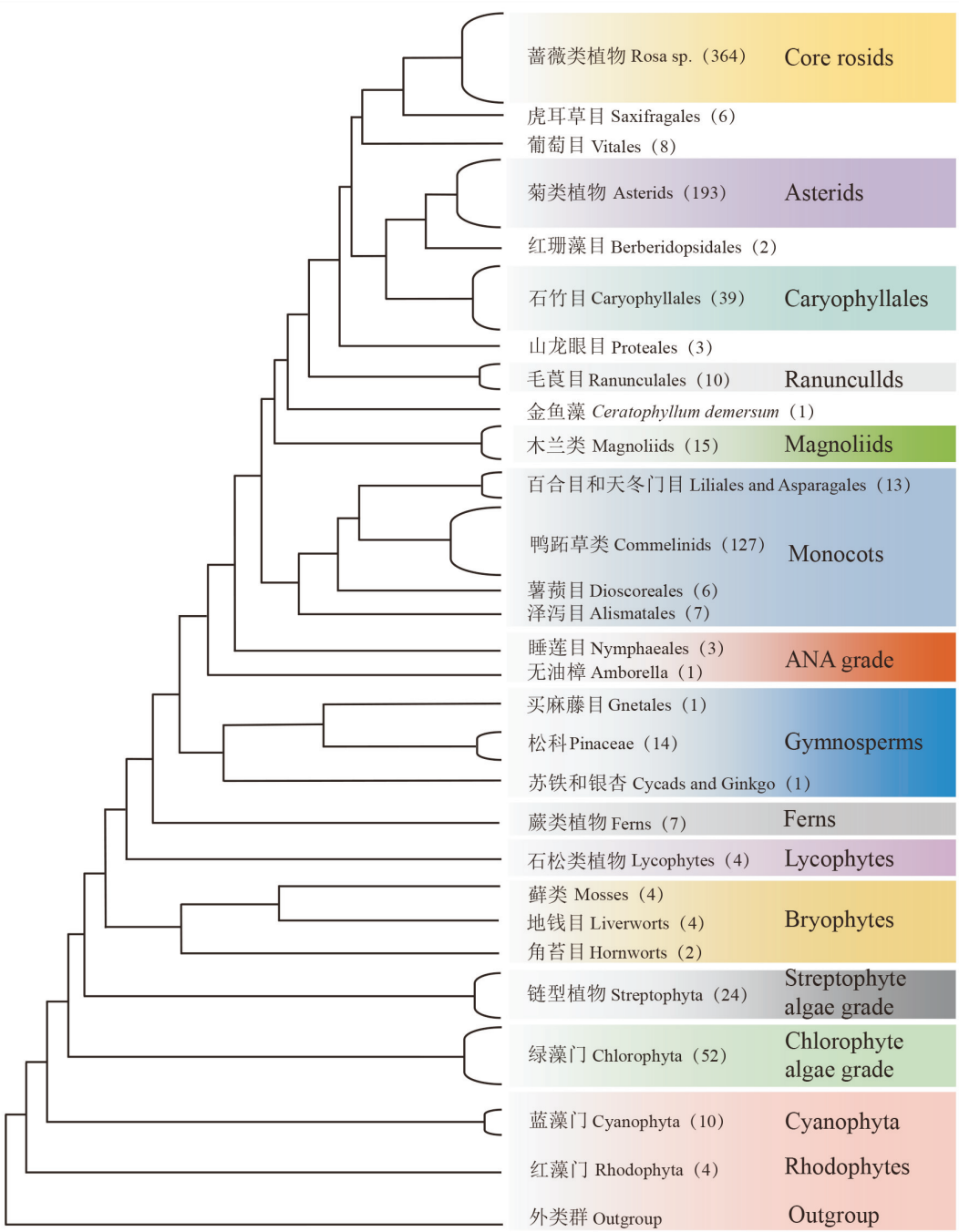


图 5　植物测序物种的系统发生关系

Fig. 5　Phylogenic tree of sequenced plant species

4　讨论

从桑格测序技术到全基因组鸟枪测序法（whole genome shotgun sequencing），从第二代测序技术到如今广泛使用的第三代测序技术，技术的革新使得基因组测序成本快速下降，通量不断提升，读序不断增长，植物基因组的组装质量与 20 年前相比有了质的飞跃[20]。2018 年 10 月 PacBio 公司发布高保真单分子长读长测序（HiFi）模式后，其在基因组组装中的应用便如雨后春笋，势不可挡。HiFi 可兼顾长读长和高准确度，对复杂基因组的组装特别有效，还可完成高质量的单倍型基因组的组装，这种特性使其在基因组结构变异检测和从头组装领域具有巨大的应用价值[61]，是未来基因

组拼接的重要选择。2020年5月，科学家利用Hi-Fi测序等技术实现了同源四倍体紫花苜蓿(*Medicago sativa* L.)基因组的组装[31]。同年10月，利用PacBio长读长与限制性位点相关DNA测序相结合的方法，研究人员构建了异源四倍体草坪植物染色体级别的基因组[62]。可以看到，科学家们在利用HiFi技术组装高度重复、复杂、异源多倍体和同源多倍体基因组等方面积累了丰富经验，为进一步探索植物基因组打下了坚实的基础。

随着越来越多的植物基因组被测序，复杂基因组的研究成了下一个热点。高倍性的植物物种往往具有大量重复序列和巨大的基因组，如何正确拼接这些基因组序列是亟待解决的问题。目前已有出现多种高通量作图技术，包括BioNano[63]、高通量染色体构象捕获技术(high-throughput/resolution chromosome conformation capture，Hi-C)[64]及10x Genomics[65]等，可以有效避免传统作图方法费时费力的缺点[66]。例如，Hi-C和光学图谱等技术的应用使拟南芥基因组的完整性得到了显著提升[67]。除了应用于二倍体植物，Hi-C技术还常被用来处理多倍体基因组，如Monat等人[68]利用该技术检测到小麦基因组中的大规模染色体重排。此外，异源四倍体画眉(*Eragrostis tef* (Zucc.) Trotter)[69]、八倍体甘蔗(*Saccharum officinarum* L.)[70]、异源四倍体花生(*Arachis hypogaea* L.)[71]等都利用Hi-C进行了高质量的组装。另一种基因组数据获取技术10x Genomics，虽然目前仅在为数不多的植物中应用，如二倍体辣椒(*Capsicum annuum* var. *grossum*)[72]和茄子(*Solanum melongena* L.)[66]，但该方法具有细胞通量高、建库周期短、成本低等优点[73]，预计未来在泛基因组研究和基因组辅助组装方面有很大的应用潜力。这些新生技术为复杂基因组的研究创造了条件，多倍体和杂合性等问题将不再会是挑战，有助于获得大量单个物种从头组装的基因组，为研究者提供更为广泛的遗传多样性资源。

自2000年和2002年报道拟南芥和水稻全基因组测序工作后，植物基因组测序工作的顺利进行涌现出了海量数据，加之生物信息学技术的普及使用，对生物学研究的推动起到了举足轻重的重要作用。基于现有趋势分析，我们推测植物基因组测序相关文章数量仍会逐年递增，更多的植物物种将被测序，但长久来看热度将逐渐降低。值得注意的是，对热带植物的测序起始于2008[74]，截至2020年底共有波罗蜜(*Artocarpu heterophyllus* Lam.)[75]、野生香蕉(*Musa schizocarpa*)[76]和非洲栽培稻(*Oryza glaberrima*)[77]等在内的81种热带植物被测序，且近4年(2017-2020)测序的物种数就已达到了这12年测序物种总数的一半以上(55.56%)。可以预见，随着热带植物在食用、饲用、能源等方面的功能被逐步挖掘，将会有更多的热带植物被测序。

长久以来，特定物种的参考基因组被认为是研究序列差异的“标准参照”，但科学家逐渐意识到，单个参考基因组并不能完全代表一个物种的全部遗传多样性，基于此，泛基因组研究成了近几年的主流趋势。目前已经报道了多种重要作物的泛基因组，包括水稻[78-80]，大豆[81]，番茄(*Solanum lycopersicum* L.)[82]，向日葵(*Helianthus annuus* L.)[83]和小麦[84, 85]。泛基因组的构建需要对物种内足够多的个体进行测序，通过序列比较发现一系列结构变异和基因拷贝数变异，有助于科研人员对物种多样性的理解，并促进破译核心基因组和非核心基因组的重要功能。值得一提的是，目前所构建的泛基因组都是种内的，在未来有可能构建属间或者种间的泛基因组。

植物与其他生物间的协同进化对生物演化具有重要意义，包括影响生物多样性、环境适应性以及群落稳定性等。得益于测序技术的发展，未来可以进一步研究作物与杂草的互作，植物与微生物的互作，植物与昆虫的互作，植物与动物的互作等。例如，利用榕树及其授粉小蜂基因组，可为研究榕属的气生根、雌雄异株和共生系统中的协同进化提供新的角度[86]。此外，结合生物信息学及比较基因组学等手段，可以进一步探索不同物种之间的基因转移、起源与进化，以及更复杂的遗传机制。

尽管植物基因组的研究已经取得了长足的发展，但要实现设计植物还有很长的路要走。由于测序技术、拼接算法、生物信息学工具的不断发展，测序目标植物将不再受成本或复杂度的限制，参考基因组质量也必将逐渐趋向“完美”，一系列的难点也将会被逐步克服，研究者可以专注于生物学问题本身的思考而不必担心技术保障，有利于在全球变暖、未来粮食需求、燃料需求等时代背景下更好

地探索植物生物学的未解之谜。

参考文献：

[1] *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*[J]. *Nature*, 2000, 408(6814): 796-815.

[2] Hamilton JP, Robin C. Advances in plant genome sequencing[J]. *Plant J*, 2012, 70(1): 177-190.

[3] Neale DB, Martínez PJ, Delatorre AR, Montanari S, Wei XX. Novel insights into tree biology and genome evolution as revealed through genomics[J]. *Annu Rev Plant Biol*, 2017, 68: 457-483.

[4] Isobe S, Shirasawa K, Hirakawa H. Challenges to genome sequence dissection in sweetpotato[J]. *Breed Sci*, 2017, 67(1): 35-40.

[5] Thorsten L, Sophien K, Khaoula B. CRISPR crops: plant genome editing toward disease resistance[J]. *Annu Rev Phytopatholy*, 2018, 56(1): 479-512.

[6] Negrao S, Oliveira MM, Jena KK, Mackill D. Integration of genomic tools to assist breeding in the *japonica* subspecies of rice[J]. *Mol Breed*, 2008, 22: 159-168.

[7] Xu X, Pan SK, Cheng SF, Zhang B, Mu DS, *et al*. Genome sequence and analysis of the tuber crop potato[J]. *Nature*, 2011, 475(7355): 189-195.

[8] Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K. Crop genome sequencing: lessons and rationales[J]. *Trends Plant Sci*, 2011, 16(2): 77-88.

[9] Albert VA, Barbazuk WB, de Pamphilis CW, Der JP, Leebens-Mack J, *et al*. The *Amborella* genome and the evolution of flowering plants[J]. *Science*, 2013, 342(6165): 1241089.

[10] Soundararajan P, Won SY, Kim JS. Insight on rosaceae family with genome sequencing and functional genomics perspective[J]. *Biomed Res Int*, 2019: 7519687.

[11] Ahmad R, Anjum MA, Balal RM. From markers to genome based breeding in horticultural crops: an overview[J]. *Phyton-Int J Exp Bo*, 2020, 89(2): 183-204.

[12] Michael TP, Jackson S. The first 50 plant genomes[J]. *Plant Genome*, 2013, 6(2): 1-7.

[13] Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, *et al*. DNA sequencing at 40: past, present and future[J]. *Nature*, 2017, 550(7676): 345-535.

[14] Belser C, Istace B, Denis E, Dubarry M, Baurens FC, *et al*. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps[J]. *Nat Plants*, 2018, 4(11): 879-887.

[15] Chen F, Chen JH, Wang ZJ, Zhang JW, Li XJ, *et al*. Genomics: cracking the mysteries of walnuts[J]. *J Genet*, 2019, 98(2): 1-3.

[16] Singh B, Salaria N, Thakur K, Kukreja S, Gautam S, *et al*. Functional genomic approaches to improve crop plant heat stress tolerance[J]. *F1000Res*, 2019, 8: 1721.

[17] Song SH, Tian DM, Zhang Z, Hu SN, Yu J. Rice genomics: over the past two decades and into the future[J]. *Genom Proteomics Bioinformatics*, 2018, 16(6): 397-404.

[18] Isobe S, Shirasawa K, Hirakawa H. Current status in whole genome sequencing and analysis of *Ipomoea* spp.[J]. *Plant Cell Rep*, 2019, 38(11): 1365-1371.

[19] Isobe S, Shirasawa K, Hirakawa H. Advances of whole genome sequencing in strawberry with NGS technologies[J]. *Hort J*, 2020, 89(2): 108-114.

[20] Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, *et al*. Plant genome sequencing- applications for crop improvement[J]. *Curr Opin Biotechnol*, 2014, 26: 31-37.

[21] Yu J, Hu SN, Wang J, Wong GKS, Li SG, *et al*. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)[J]. *Science*, 2002, 296(5565): 79-92.

[22] Derelle E, Ferraz C, Rombauts S, Rouzé P, Worden AZ, *et al*. Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features[J]. *Proc Natl Acad Sci USA*, 2006, 103(31): 11647-11652.

[23] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors[J]. *Proc Natl Acad Sci USA*, 1977, 74(12): 5463-5467.

[24] Ronaghi M, Mathias U, Pål N. A sequencing method based on real-time pyrophosphate[J]. *Science*, 1998, 281(5375): 363-365.

[25] Avni R, Nave M, Barad O, Baruch K, Twardziok SO, *et al*. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication[J]. *Science*, 2017, 357(6364): 93-97.

[26] Mitros T, Session AM, James BT, Wu GHA, Belaffif MB, *et al*. Genome biology of the paleotetraploid perennial biomass crop *Miscanthus*[J]. *Nat Commun*, 2020, 5442(11): 1-11.

[27] Gui S, Peng J, Wang XL, Wu ZH, Cao R, *et al*. Improving *Nelumbo nucifera* genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements[J]. *Plant J*, 2018, 94(4): 721-734.

[28] Eid J, Fehr A, Gray J, Luong K, Lyle J, *et al*. Real-time DNA sequencing from single polymerase molecules[J]. *Science*, 2009, 323(5910): 133-138.

[29] Hon T, Mars K, Young G, Tsai YC, Karalius JW, *et al*. Highly accurate long-read HiFi sequencing data for five complex genomes[J]. *Sci Data*, 2020, 7(1): 1-11.

[30] Zhou Q, Tang D, Huang W, Yang ZM, Zhang Y, *et al*. Haplotype-resolved genome analyses of a heterozygous diploid potato[J]. *Nat Genet*, 2020, 52(10): 1018-1023.

[31] Chen HT, Zeng Y, Yang YZ, Huang LL, Tang BL, *et al*. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa[J]. *Nat Commun*, 2020, 11(1): 2494.

[32] Sun XP, Jiao C, Schwaninger HD, Chao CT, Ma YM, *et al*. Phased diploid genome assemblies and pan-ge-

nomes provide insights into the genetic history of apple domestication[J]. *Nat Genet*, 2020, 52(12): 1423–1432.

[33] Ma DN, Dong SS, Zhang SC, Wei XQ, Xie QJ, *et al*. Chromosome-level reference genome assembly provides insights into aroma biosynthesis in passion fruit (*Passiflora edulis*)[J]. *Mol Ecol Resour*, 2020, 21(3): 955–968.

[34] Leggett RM, Clark MD. A world of opportunities with Nanopore sequencing[J]. *J Exp Bot*, 2017, 68(20): 5419–5429.

[35] Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, *et al*. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome[J]. *Nat Biotechnol*, 2019, 37(10): 1155–1162.

[36] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds[J]. *Science*, 2017, 356(6333): 92–95.

[37] Bocklandt S, Hastie A, Cao H. Bionano genome mapping: high-throughput, ultra-long molecule genome analysis system for precision genome assembly and haploid-resolved structural variation discovery[J]. *Adv Exp Med Biol*, 2019, 1129: 97–118.

[38] Van de PY, Mizrachi E, Marchal K. The evolutionary significance of polyploidy[J]. *Nat Rev*, 2017, 18: 411–424.

[39] Sun X, Zhu S, Li N, Cheng Y, Zhao J, *et al*. A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and Allicin biosynthesis[J]. *Mol Plant*, 2020, 13(9): 1328–1339.

[40] Dodsworth S, Chase MW, Kelly LJ, Leich IJ, Macas J, *et al*. Genomic repeat abundances contain phylogenetic signal[J]. *Syst Biol*, 2015, 64(1): 112–126.

[41] Qiu H, Price DC, Weber APM, Reeb V, Yang EC, *et al*. Adaptation through horizontal gene transfer in the cryptoendolithic red alga *Galdieria phlegrea*[J]. *Curr Biol*, 2013, 23(19): 865–866.

[42] Gao XY, Zhang X, Chen W, Li J, Yang WJ, *et al*. Transcriptome analysis of *Paris polyphylla* var. *yunnanensis* illuminates the biosynthesis and accumulation of steroidal saponins in rhizomes and leaves[J]. *Phytochemistry*, 2020, 178(7): 112460.

[43] The International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome[J]. *Science*, 2014, 345(6194): 1251788.

[44] Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*[J]. *GigaScience*, 2017, 6(11): gix097.

[45] Stevens KA, Wegrzyn JL, Zimin A, Puiu D, Crepeau M, *et al*. Sequence of the sugar pine megagenome[J]. *Genet*, 2016, 204(4): 1613–1626.

[46] Warren RL, Keeling CL, Yuen MMS, Raymond A, Taylor GA, *et al*. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism[J]. *Plant J*, 2015, 83(2): 189–212.

[47] Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, *et al*. The Norway spruce genome sequence and conifer genome evolution[J]. *Nature*, 2013, 497(7451): 579–584.

[48] Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, *et al*. The B73 maize genome: complexity, diversity, and dynamics[J]. *Science*, 2009, 326(5956): 1112–1115.

[49] Leisner CP, Hamilton JP, Crisovan E, Manrique-Carpintero NC, Marand AP, *et al*. Genome sequence of M6, a diploid inbred clone of the high-glycoalkaloid-producing tuber-bearing potato species *Solanum chacoense*, reveals residual heterozygosity[J]. *Plant J*, 2018, 94(3): 562–570.

[50] Chen ZJ, Sreedasyam A, Ando A, Song QX, de Santiago LM, *et al*. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement[J]. *Nature Genet*, 2020, 52(5): 525–533.

[51] Sun FM, Fan GY, Hu Q, Zhou YM, Guan M, *et al*. The high-quality genome of *Brassica napus* cultivar 'ZS11' reveals the introgression history in semi-winter morphotype[J]. *Plant J*, 2017, 92(3): 452–468.

[52] Bayer PE, Hurgobin B, Golicz AA, Chan CKK, Yuan YX, *et al*. Assembly and comparison of two closely related *Brassica napus* genomes[J]. *Plant Biotechnol J*, 2017, 15(2): 1602–1610.

[53] Sierro N, Battey JND, Ouadi S, Bakaher N, Bovet L, *et al*. The tobacco genome sequence and its comparison with those of tomato and potato[J]. *Nat Commun*, 2014, 5: 3833.

[54] Guo LB, Qiu J, Ye CY, Jin GL, Mao LF, *et al*. *Echinochloa crus-galli* genome analysis provides insight into its adaptation and invasiveness as a weed[J]. *Nat Commun*, 2017, 8(1): 1031.

[55] Hu Y, Chen JD, Fang L, Zhang ZY, Ma W, *et al*. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton[J]. *Nature Genet*, 2019, 51(4): 739–748.

[56] Matsuzaki M, Misumi O, Shin T, Maruyama S, Takahara M, *et al*. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D[J]. *Nature*, 2004, 428(6983): 653–657.

[57] Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, *et al*. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)[J]. *Science*, 2006, 313(5793): 1596–1604.

[58] Huang S, Li RQ, Zhang ZH, Li L, Gu XF, *et al*. The genome of the cucumber, *Cucumis sativus* L.[J]. *Nature Genet*, 2009, 41(12): 1275–1281.

[59] Ming R, VanBuren R, Liu YL, Yang M, Han YP, *et al*. Ge-

nome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.) [J]. *Genome Biol*, 2013, 14(5): R41.

[60] One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants[J]. *Nature*, 2019, 574(7780): 679-685.

[61] Lang DD, Zhang SL, Ren PP, Liang F, Sun ZY, *et al*. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacbio Sequel Ⅱ system and ultralong reads of Oxford Nanopore[J]. *GigaScience*, 2020, 9(12): 123.

[62] Zhou CX, Olukolu B, Gemenet DC, Wu S, Gruneberg W, *et al*. Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outbred mapping populations[J]. *Nature Genet*, 2020, 52(11): 1256-1264.

[63] Jiao YP, Peluso P, Shi JH, Liang T, Stitzer MC, *et al*. Improved maize reference genome with single-molecule technologies[J]. *Nature*, 2017, 546(7659): 524-527.

[64] Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes[J]. *Methods*, 2012, 58(3): 268-276.

[65] Phillippy AM. New advances in sequence assembly[J]. *Genome Res*, 2017, 27(5): xi-xiii.

[66] Wei QZ, Wang JL, Wang WH, Hu TH, Hu HJ, Bao CG. A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant[J]. *Hortic Res*, 2020, 7(1): 153.

[67] Xie T, Zheng JF, Liu S, Peng C, Zhou YM, *et al*. *De novo* plant genome assembly based on chromatin interactions: a case study of *Arabidopsis thaliana*[J]. *Mol Plant*, 2015, 8(3): 489-492.

[68] Zhang XT, Zhang SC, Zhao Q, Ming R, Tang HB. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data[J]. *Nat Plants*, 2019, 5(8): 833-845.

[69] Vanburen R, Wai CM, Wang XW, Pardo J, Yocca AE, *et al*. Exceptional subgenome stability and functional divergence in the allotetraploid Ethiopian cereal teff[J]. *Nat Commun*, 2020, 11(1): 884.

[70] Zhang JS, Zhang XT, Tang HB, Zhang Q, Hua XT, *et al*. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. [J]. *Nature Genet*, 2018, 50(11): 1565-1573.

[71] Bertioli DJ, Jenkins J, Clevenger J, Dudchenko O, Gao DY, *et al*. The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*[J]. *Nature Genet*, 2019, 51(5): 877-884.

[72] Hulsekemp AM, Maheshwari S, Stoffel K, Hill TA, Jaffe D, *et al*. Reference quality assembly of the 3. 5-Gb genome of *Capsicum annuum* from a single linked-read library[J]. *Hortic Res*, 2018, 5: 4.

[73] Gao CX, Zhang MN, Chen L. The Comparison of two single-cell sequencing platforms: BD Rhapsody and 10x genomics chromium[J]. *Curr Genomics*, 2020, 21(8): 602-609.

[74] Ming R, Hou SB, Feng Y, Yu QY, Dionne-Laporte A, *et al*. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)[J]. *Nature*, 2008, 452(7190): 991-996.

[75] Sahu SK, Liu M, Yssel A, Kariba R, Muthemba S, *et al*. Draft genomes of two artocarpus plants, Jackfruit (*A. heterophyllus*) and Breadfruit (*A. altilis*) [J]. *Genes*, 2020, 11(1): 27.

[76] Dhont A, Denoeud F, Aury JM, Baurens FC, Carreel F, *et al*. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants[J]. *Nature*, 2012, 488(7410): 213-217.

[77] Monat C, Pera B, Ndjiondjop MN, Sow M, Tranchant-Dubreuil C, *et al*. *De novo* assemblies of three *Oryza glaberrima* accessions provide first insights about pan-genome of African rices[J]. *Genome Biol Evol*, 2017, 9(1): 1-6.

[78] Qin P, Lu HW, Du HL, Wang H, Chen WL, *et al*. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations[J]. *Cell*, 2021, 184(13): 3542-3558.

[79] Zhao Q, Feng Q, Lu HY, Li Y, Wang AH, *et al*. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice[J]. *Nat Genet*, 2018, 50(2): 278-284.

[80] Wang WS, Mauleon R, Hu ZQ, Chebotarov D, Tai SS, *et al*. Genomic variation in 3, 010 diverse accessions of Asian cultivated rice[J]. *Nature*, 2018, 557(7703): 43-49.

[81] Liu YC, Du HL, Li PC, Shen YT, Peng H, *et al*. Pan-genome of wild and cultivated soybeans[J]. *Cell*, 2020, 182(1): 162-176.

[82] Gao L, Gonda I, Sun HH, Ma QY, Bao K, *et al*. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor[J]. *Nat Genet*, 2019, 51(6): 1044-1051.

[83] Hübner S, Bercovich, Todesco M, Mandel JR, Odenheimer J, *et al*. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance [J]. *Nat Plants*, 2019, 5(1): 54-62.

[84] Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee HT, *et al*. The pangenome of hexaploid bread wheat[J]. *Plant J*, 2017, 90(5): 1007-1013.

[85] Walkowiak S, Gao LL, Monat, C, Haberer G, Kassa MT, *et al*. Multiple wheat genomes reveal global variation in modern breeding[J]. *Nature*, 2020, 588(7837): 277-283.

[86] Zhang X, Wang G, Zhang S, Chen S, Wang Y, *et al*. Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution[J]. *Cell*, 2020, 183(4): 875-889.

（责任编辑：李惠英）