中国科学: 生命科学 2024 年 第 54 卷 第 2 期: 233 ~ 246
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
评 述 青年科学家专栏

植物复杂基因组与泛基因组研究现状与展望

王英豪[1], 余嘉鑫[2], 唐海宝[1], 张兴坦[2*]

1. 福建农林大学生命科学学院, 福州 350000;
2. 中国农业科学院深圳农业基因组研究所, 深圳 518124
* 联系人, E-mail: zhangxingtan@caas.cn

收稿日期: 2023-06-28; 接受日期: 2023-10-12; 网络版发表日期: 2023-12-11
国家自然科学基金(批准号: 32222019)资助

摘要 基因组是指一个生物体内遗传物质的总和, 是生物学研究的关键之一. 自2000年拟南芥基因组被测序发表以来, 已有超过800个植物基因组相继被破解, 极大促进了植物分子生物学、遗传学等领域的发展. 即便如此, 植物基因组学研究仍然面临一系列挑战, 包括高杂合、高重复度、高倍性等复杂基因组的组装和泛基因组的构建等. 本文从植物基因组学的发展概况、基因组测序技术、组装算法等三个方面, 全面展示了植物基因组的快速发展. 其中, 介绍了简单基因组装和复杂基因组装的相关策略, 总结了"端粒到端粒"(telomere-to-telomere或称T2T)的组装和泛基因组构建方法以及其重要性. 最后, 对未来植物基因组的发展进行了展望, 认为随着技术的不断进步, 基因组解析技术和方法将会更加完善, 为植物基因组的深入研究提供更多支持. 本文为植物T2T、复杂基因组组装和泛基因组的构建方法研究提供了参考依据.

关键词 植物基因组学, 测序技术, 组装算法, 植物复杂基因组, 植物泛基因组

植物基因组研究不仅是对植物遗传信息的解码, 它涉及生态、进化、农业、分子育种等多个领域, 为人们提供了前所未有的理解与应用机会. 然而, 在遗传学和基因组学的研究历程中, 植物复杂基因组的解读与分析始终是一个颇具挑战性的领域. 与其他生物体相比, 植物拥有相对较大且结构复杂的基因组. 随着测序技术、组装算法的飞速发展, 研究者正逐步揭示植物基因的奥秘. 在此背景下, 本文综述了植物复杂基因组研究的进展, 介绍了简单基因组装和复杂基因组装的相关策略, 总结了"端粒到端粒"(telomere-to-telomere或称T2T)的组装和泛基因组(pan-genome)构建方法以及其重要性.

1 植物基因组发展概况

1.1 组学技术发展意义

基因组是指一个生物体内遗传物质的总和, 其中蕴含着生命体丰富的遗传信息[1], 是生物学研究中至关重要的基础. 2000年12月拟南芥[2]基因组在*Nature*期刊上的发表正式叩开了植物基因组研究的大门. 近20多年来随着测序技术、组装算法、遗传学、生物信息学的飞速进步[3], 截至2022年12月底已有825个植物

引用格式: 王英豪, 余嘉鑫, 唐海宝, 等. 植物复杂基因组与泛基因组研究现状与展望. 中国科学: 生命科学, 2024, 54: 233–246
Wang Y H, Yu J X, Tang H B, et al. Research status and prospect of plant complex genomes and pan-genomes (in Chinese). Sci Sin Vitae, 2024, 54: 233–246, doi: 10.1360/SSV-2023-0068

物种(统计数据来自于NCBI以及GWH)基因组相继发表(图1). 近年来, 植物基因组迎来了快速发展期. 在2000~2010年期间受限于测序技术与成本以及组装算法, 只发表了26个植物物种的参考基因组, 但近三年已有超过500个植物物种的基因组被发表[4,5]. 这其中以被子植物居多约占比88%, 并主要在禾本科、十字花科、豆科这三大科, 大小从64 Mb到31 Gb跨度三个数量级[6~8]. 参考基因组的公布极大地推动了对植物中丰富的基因遗传资源的挖掘, 对关键性状候选基因的筛选, 以及植物基础生物学、遗传学、功能基因组学、植物育种学等相关领域的快速发展.

1.2 测序技术发展概况

推动植物基因组快速发展的主要动力之一是测序技术的巨大进步. 测序技术的进步主要体现在读长提升、测序质量提高、通量提升, 同时带来测序成本的显著下降. 起初在拟南芥[2]、水稻[9]、玉米[10]等最先公布的植物基因组中所采用的技术为Sanger的双脱氧终止法[11], 被称为第一代测序方法. 该方法准确度高, 但耗时长、成本高、通量低, 因此当时的基因组项目往往需要消耗大量的时间以及科研成本. 二代测序技术(next-generation sequencing, NGS)[12]极大地缓解了这一困境, 二代测序技术因其具有高通量、成本低、准确度高等特点, 直到目前在基因组分析中都占据着十分重要的地位[13~15]. 但二代测序存在读长短的弊端, 因此在面对如高重复基因组等复杂组装时仅使用二代测序会引入较多的错误.

随着测序技术的飞速进步, 三代测序正式踏入了基因组学. 三代测序主要分为两类: 第一类是美国Pacific Bioscience公司研发的单分子实时测序技术, 其CCS模式产生的HiFi读数准确度可以达到99%, 长度可以达到15 kb[16,17]. 第二类是Oxford Nanopore Technologies公司的纳米孔测序技术, 在测序精度上略低于HiFi reads(读长, 指的是测序仪单次测序所得到的碱基序列), 但长度可以达到10~100 kb. 目前三代测序是基因组组装中所用的主流测序手段, 如马铃薯[18]、野苹果[19]、西番莲[20]等. 长读长测序技术极大地提高了组装的连续性, 平均contig N50水平(contig N50是一种用于衡量基因组组装质量的指标, 它是指所有contig的长度从小到大排序后, 加起来达到基因组总长度的50%

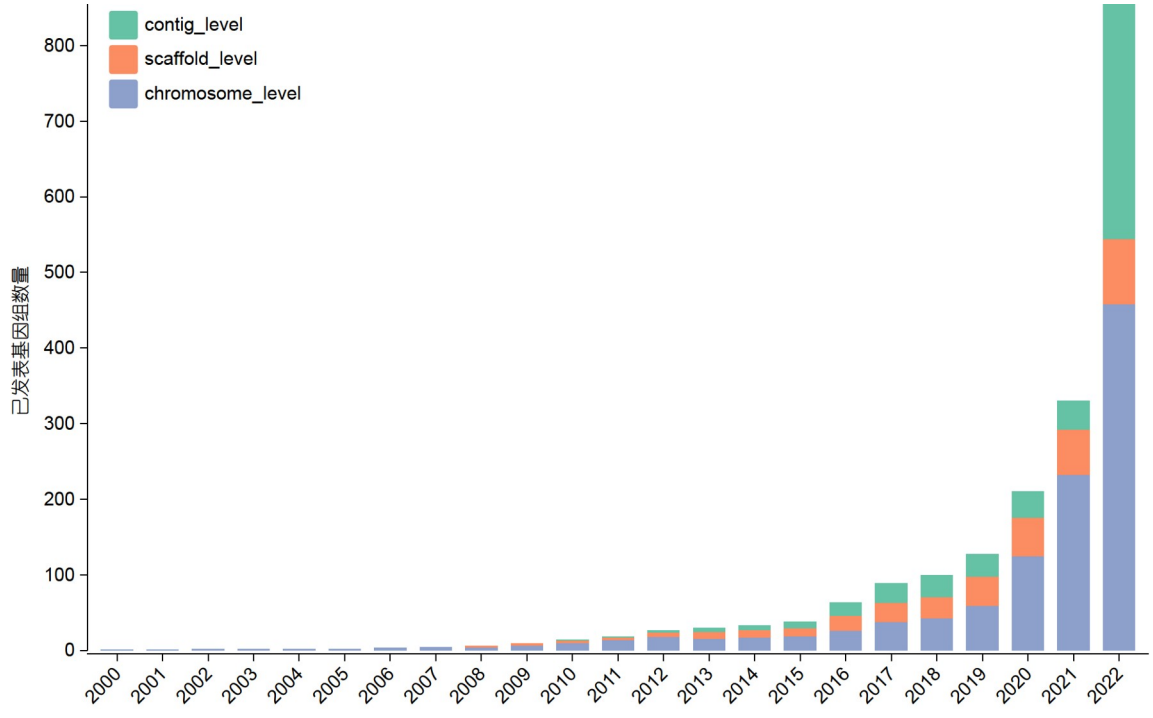


图 1 2000~2022年发表基因组情况(数据来自于NCBI以及GWH)

Figure 1 Genomes published from 2000 to 2022 (data from NCBI and GWH)

对应contig的长度)从2010年(99.5±48.1) kb增加到了2020年的(3395.2±735.4) kb[5]. 与此同时, Hi-C[21]、BioNano光学图谱[22]以及Pore-C[23]等技术的推出, 可以呈现染色体内部互作信号, 将线性的测序数据提升到立体空间水平, 从而应用于染色体水平辅助组装. 由此可见, 测序技术是组装基因组的根基, 测序技术的发展使得测序成本日趋下降, 读长持续升高, 测序质量逐步提升, 极大地推动着植物基因组学的发展.

1.3　组装算法发展概况

一个高质量完整基因组不仅依托测序技术, 组装算法对其也至关重要. 无论是一代Sanger测序、二代NGS测序乃至三代长读长测序, 相比于基因组的长度仍是有限的. 测序结果犹如一块块拼图碎片而基因组是最后的图案, 组装算法的目的就是将众多的测序片段, 找到其中正确的前后关系并将其拼接起来. 目前主要的算法可以分为三大类.

第一类贪婪算法. 贪婪算法首先选择满足一定要求的读长序列作为contig的种子, 然后寻找和读长序列的两端含有重叠区域的读长序列, 对选作种子的读长序列进行扩展, 直到当前拼接的序列两端无法继续扩展. 采用贪婪算法的软件主要有: SHARCGS[24], SSAKE[25]和VCAKE[26]. 若存在两个及以上的读长序列与当前拼接的序列的某一段含有重叠区域时, 算法无法确定应该选择哪一条读长序列进行扩展, 因此当遇到这种情况时贪婪算法所拼接的contig的长度往往较短, 所以目前组装工具大多基于下述两种方法(图2).

第二类OLC(overlap-layout-consensus)算法. 该方法主要应用于一代测序数据以及三代数据的长读长组装. 其主要原理是通过reads之间两两比较找到相互重叠的部分, 以此找到局部拼接最优解, 随后构建区域图从而将reads拼接至contig以及scaffolds(图2). 采用OLC算法的组装软件有Canu[27], MECAT[28], NECAT[29].

第三类DBG(de Bruijn graph, 德布鲁因图)算法. OLC主要针对读长较长的片段, 对于二代长度为100~150 bp的序列而言, 因其信息过于碎片化并不是十分适用, 因此需要DBG等算法进行优化. DBG算法首先将序列*k*-mer化, 所谓*k*-mer即长度为*k*步长为1的子序列. 根据*k*-mer的交叠关系, 将有*k*−1碱基重叠的*k*-mer连接起来构建德布鲁因图, 消除由测序引起的错误后, 拼接至contig水平. 采用DBG算法的软件有Velvet[30], ABySS[31], SOAPdenovo[32].

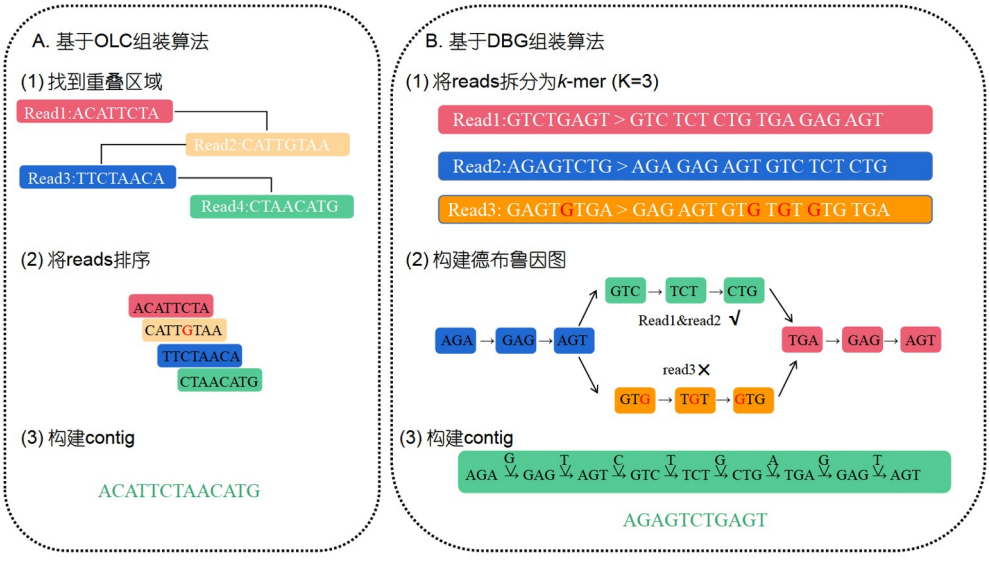


图 2　OLC组装算法以及DBG组装算法流程. A: 基于OLC算法, 通过比对找到片段间的重叠信息后寻找一条质量最高的序列路径, 并获得与路径对应的序列, 即contig; B: 基于DBG算法, 通过*k*-mer化后构建德布鲁因图, 随后寻找最佳路径构建contig

Figure 2　Overview of OLC assembly algorithm and DBG assembly algorithm. A: Based on the OLC algorithm, the highest quality sequence path is identified by finding overlapping information between fragments, and the corresponding sequence, known as a contig, is obtained; B: using the DBG algorithm, the contig is constructed by building a de Bruijn graph after *k*-merization and subsequently searching for the optimal path

由于PacBio CCS和Nanopore在长度上的优势，三代长读长测序已成为应用最广泛的组装测序手段. 目前已有针对HiFi数据的组装软件，比如Hifiasm[33]以及HiCanu[34]，Hifiasm可以在单个机器上多线程运行，在较少的资源消耗下快速完成基因组组装. HiCanu也是组装PacBio CCS数据的常用软件之一，其组装流程可以分为三个阶段: 校正、修整和装配. 由此可见，有效的组装算法将会在已有测序数据的基础上助力基因组装，提高组装连续性，节约组装时间与计算资源，加快复杂基因组的组装.

2 植物复杂基因组组装策略

植物基因组的组装一般可以分成如下步骤(图3). (ⅰ) 基因组特征评估: 在组装前首先需要对待组装的物种进行基因组调查(survey)从而评估基因组大小、杂合度、GC含量、重复序列等重要的基因组特征，这几点决定了物种组装的难度以及成本[35]. (ⅱ) 基因组初步组装: 通过短reads之间的交叠关系构建成无缝隙(gap)的contig(重叠群，通过reads的交叠关系拼接而成的长片段). 随后根据大片段文库以及双端测序调整contig的排序以及方向，将contig进一步组装成更长的片段scaffold. (ⅲ) 使用Hi-C数据或者近缘种信息将contig或scaffold挂载至染色体. (ⅳ) 基因组质量评估: 初步组装好的基因组需要通过BUSCO、HiC热图、近缘种共线性等方式评估组装质量.

随着测序技术和组装算法的不断改进，大部分的简单基因组(基因组大小不超过1 Gb，杂合度小于0.5%，重复序列低于50%，GC含量在35%~65%之间[36])可通过多种测序技术结合组装算法有效解决. 在10年前，多国合作耗费许多人力和时间才完成了马铃薯基因组[37]. 如今，单个团队使用HiFi结合Hi-C图谱构建的染色体水平基因组，可以将contig N50提高500倍以上(从32 kb到17.3 Mb).

在植物基因组中，相当大一部分的基因组属于复杂基因组. 复杂基因组指的是一类无法直接使用常规的测序和组装方法进行解析的基因组，通常包括以下特点: 基因组杂合率大于0.8%、重复序列占比高于60%、GC含量高于65%或低于35%、高倍性以及难以去除异源DNA污染等[38]. 中国农业科学院深圳农业基因组研究所唐蝶和周倩[39]2021年在《生物技术通报》上发表的名为“植物基因组组装技术研究进展”的

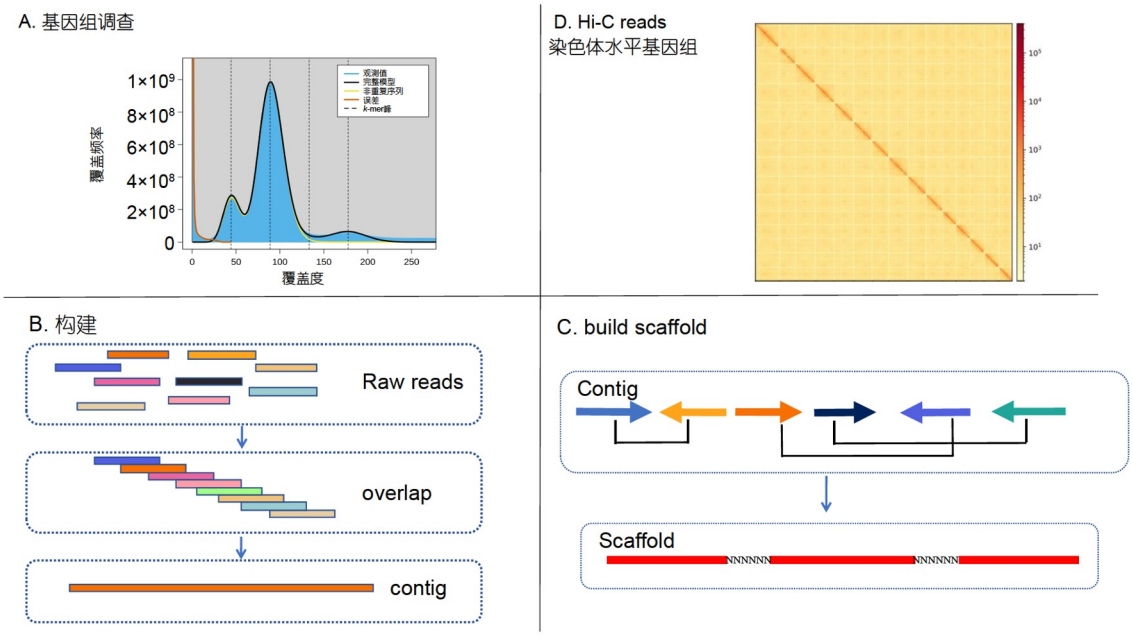


图 3 一般基因组组装流程. A: 基因组调查; B: 构建contig; C: 构建scaffold; D: 挂载至染色体水平

Figure 3 Workflow of genome assembly process: A: An example of genome survey; B: workflow of contig assembly; C: genome scaffolding; D: chromosomal-scale genome assembly based on Hi-C heatmap

综述文章, 已详细地阐述了复杂基因组组装的进展以及策略, 因此本文只做简单介绍.

2.1　长读长测序是解决高重复基因组组装的关键

重复序列一直是基因组装过程中一个难题, 主要原因是由于重复序列含量高并且分布在基因组的不同位置, 往往造成组装的基因组小于实际的基因组大小. 在自然界中重复序列在各物种的比例从病毒(小于1%)、细菌(3%左右)、人(47%)、玉米(77%~85%)逐步升高. 在一个44种植物和68种脊椎动物全基因组重复水平和基因组大小的关系研究中发现, 植物基因组的重复序列明显高于脊椎动物[40]. 但同时重复序列在物种进化和功能调控中扮演着重要的角色.

目前发表了如火炬松(22 Gb, 重复序列82%)[41]、银杏(10 Gb, 重复序列80%)[42]、大蒜(16.9 Gb, 重复序列91.3%)[43]的基因组数据. 二代数据因为其读长较短往往无法跨过重复序列, 因此在overlap构建contig的过程中可能会丢失掉许多片段, 三代数据更可能跨过重复区段, 因此极大地提高了重复序列的区分度, 组装的完整性以及准确性. 例如, 银杏基因组采用了PacBio数据重新组装过后对比二代组装contig N50从48 kb[42]提升至1.58 Mb[44], 由此可见, 高精度的长读长测序将显著改善高重复基因组组装.

2.2　去冗余算法实现高杂合二倍体单套基因组组装

因远缘杂交、自交不亲和以及无性繁殖等特性, 许多植物基因组高度杂合, 含有频繁的等位基因变异, 如自交不亲和茶树基因组杂合度高达2.8%[45]. 这种高度的杂合性导致同源片段的一致性较小, 使得基因组组装时出现许多分支结构, 更容易引起错误, 破坏组装的准确性和连续性.

长读长测序的出现同样驱动了高杂合基因组组装领域的进步. 在通过reads重叠构建初步的组装图后, 长逾10 kb的三代测序reads(表1)能够横跨数个单核苷酸多态性(single nucleotide polymorphisms, SNPs)位点, 为临近的SNPs位点提供相位信息, 进而帮助解开杂合区域的分支结构, 拓展组装局部的连续性, 产生高度连续的contig组装. 然而, 高度杂合的基因组区域会在局部形成复杂的组装图. 尽管上述方法能够大幅提升contig的连续性, 但在拆解复杂的局部组装图时仍会产生错误组装, 将两套不同单倍型的基因序列混在一起, 形成遗传信息的冗杂, 对下游分析造成困扰. 因此初始contig通常需要去除冗余序列, 才能获得准确的单套基因组组装.

目前组装高杂合单套基因组主要有以下几个思路. (ⅰ) 基于全基因组测序深度(reads depth, RD)的策略. 将全基因组测序数据比对到contig组装中, 由于杂合contig存在两份冗余拷贝, 其测序深度通常会呈现为平均测序深度的一半. 利用Purge_haplotigs[46]等软件基于测序深度的原理, 识别并去除冗余的杂合contig, 从而获得有代表性的单套contig水平基因组, 进一步将其挂载到染色体基因组. 目前, 金荞麦[47]基因组已经成功根据该策略组装完成. (ⅱ) 基于全基因组比对(whole genome alignment comparison, WGAC)的方法. 通过全基因组的自身contig比对能够发现具有高度相似性的contig对, 将长度较短的冗余片段删除后, 最终获得单套基因组. 然而, 由于上述两种方法执行全基因组水平的序列比对, 其对算力的要求非常高. (iii) 基于*k*-mer的方法检测基因组中的冗余contig. 该策略首先对contig组装进行*k*-mer计数, 排除高频出现的*k*-mer, 随后将contig之间的共有*k*-mer转换为contig的相似度, 具有高相似性的contig即为冗余contig. 因为该方法不需要广泛的全基因组比对, 因此极大了提高了计算效

表 1　不同测序技术特点

Table 1　Characteristics of different sequencing technologies

	Sanger	NGS	Pacbio CCS	Oxford Nanopore
测序长度	600~1000 bp	100~150 bp	10~20 kb	10~100 kb
错误率	<0.001%	<0.1%	<0.1%	<5%
检测方法	荧光/光学	荧光/光学	荧光/光学	电流
优点	准确度高	通量大, 准确度高	长读长且准确度高	长读长
缺点	通量低, 成本高	读长短	成本高	错误率高, 成本高

率, 并且降低了成本, 在大基因组中具有较好的应用前景[48]. 例如, Khaper(https://github.com/tangerzhang/khaper)成功利用该策略组装了茶树铁观音单倍体[49]基因组.

以上的三种做法主要目的是组装出单倍体基因组(monoploid genome). 单套基因组主要思路是将杂合的区域整合在一起. 虽然目前可以依靠上述策略解析杂合二倍体的单倍体基因组, 但依然面临着大量遗传信息的丢失, 无法获得全面的遗传信息.

2.3 分型算法助力单倍型分型基因组破译

随着基因组学的发展, 在研究中遇到高杂合或多倍体基因组时, 越来越多研究者会选择组装其单倍型分型基因组(haplotype-resolved genome). 单倍型分型基因组是将高度相似的同源染色体序列(即单倍型)拆分出来, 最大程度保留等位序列, 提高了组装的准确性和完整性. 这种方法在一些高杂合和多倍体的植物基因组中如二倍体马铃薯[18]、四倍体紫花苜蓿[50]等的研究中已经得到了广泛的应用. 单倍型分型基因组保有同源染色体之间的遗传差异, 在挖掘优异等位基因以及探索杂种优势等方面具有单套基因组所不能比拟的优势.

根据输出的contig基因组的不同, 单倍型分型基因组的组装方法主要分为以下三类. (ⅰ) 基于一致性序列的基因组分型. 该方法首先从头组装出单倍体基因组, 随后将reads比对到单倍体基因组, 利用杂合的SNPs位点进行定相. HapCUT2[51]及WhatsHap[52]是采用该方法的经典工具. 该方法无法处理大尺度的结构变异. (ⅱ) 混合单倍型基因组组装. 在早期的单倍型分型基因组中, 只能在组装图的分支结构局部进行定相, 缺乏临近分型区块之间的相位信息, 因此会采用单倍体基因组加杂合区域单倍型基因组的方式展示, 例如Falcon-unzip[53]. (iii) 完整的单倍型基因组装, 得益于高精度的HiFi reads, hifiasm[33], HiCanu[34]等工具能够极大地提升分型区块的连续性, 再辅以额外的连接信息, 例如ONT超长读长序列, Hi-C等, 便能够将相邻的分型区块联系起来, 输出兼具完整性和精确性的单倍型分型基因组. 此外, 对于已知亲本的物种, 可以对双亲本分别进行二代测序后, 利用双亲的特异*k*-mer对分型区块添加标记, 区分来自不同亲本的单倍型, 例如Trio binning[54]等.

植物基因组存在复杂的倍性, 其形成原因包括杂交和基因组加倍等. 多倍化有利于增加物种变异, 改良作物品质. 常见的重要作物如小麦、甘蔗、棉花、马铃薯等都是多倍体植物. 多倍体可以分为两类: 一类称为异源多倍体, 由已经产生生殖隔离的不同物种间杂交加倍产生. 因为杂交产生所以每套染色体之间有明显差异, 因此可以采用纯合染色体组装后通过祖先种与近祖先种拆分出亚基因组, 相对组装难度较小. 例如, 四倍体油菜[55]基因组借助二倍体祖先成功分出了两个亚基因组. 另一类称为同源多倍体, 即所有染色体组来源于个体内的染色体组自我复制或同个物种内不同个体间的杂交与多倍化. 这样形成的同源染色体相似度高, 在组装过程中易产生错误的信号, 因此同源多倍体基因组的难度明显大于异源多倍体. Hi-C分型工具结合三代长读长数据为同源多倍体解析提供了巨大的推动力. 例如, 同源多倍体甘蔗基因组这一世界技术难题在2018年被本团队[56]成功破解, 也是全球首个组装到染色体水平的同源多倍体基因组. 其解析方法是首先通过BAC文库以及三代测序克服序列相似性组装出全部contig, 随后使用自主开发的ALLHiC Prune[57]算法有效实现了同源染色体的拆分, 并利用遗传算法解决了多倍体染色体内短序列之间的排序和定向. 随着基因组学研究的不断发展, 目前通过多技术整合成功解析了越来越多同源多倍体物种的基因组, 并且在这些基因组的组装连续性方面也有了明显的提升. 例如, 最近发表的同源四倍体马铃薯[58], 该研究引入了一种基于遗传图谱和遗传计量评分的方法(polyploid graph binning), 并结合HiFi读长序列的分组信息进行单倍型分型.

3 基因组测序技术的最新进展

基因组学发展日新月异, 本文归纳总结了近年间测序技术的最新进展.

3.1 PacBio Revio平台

PacBio是行业里较早成功的商业化单分子测序平台. 其在2019推出的sequel II HiFi reads因其高精度长读长的特点极大地促进了基因组学的发展. 在一段时间内, HiFi reads是市场上唯一在准确度和读长两个维度上保持平衡的技术. 然而, HiFi测序平台存在一个致

命的缺点: 其测序通量较低, 从而导致测序价格高, 无法大规模普及.

2022年10月, PacBio宣布推出Revio长读长测序系统. 该系统在HiFi测序的基础上显著的提高了测序通量. 相比于Sequel II测序平台, Revio单张测序芯片测序通量提高313%, 单次测序量从原来的20 Gb提高至90 Gb. 并且其可同时支持四张独立操作芯片, 四张芯片测序量可达360 Gb, 测序时间由原来的30小时降低到24小时. 结合计算方面的重大进步, Revio可将HiFi数据通量增加15倍. Revio平台的推出将强力助推基因组快速组装、降低高精度长读长测序成本、加快基因组变异检测等.

3.2 高精度长读长ONT测序

截至目前, 牛津纳米孔测序技术(Oxford Nanopore Technology, ONT)是测序读长最长的一种技术, 其超长建库方法可以使read N50达到50~100 kb, 最长读长可达Mb级别. 因此ONT超长读长可以轻松跨越重复区域, 显著提升组装连续性. 然而, 测序准确率在一段时间内限制着Nanopore在植物基因组学的应用. 其上一代R9.4芯片准确率约为92%[59], 相比HiFi reads 99%以上的准确率存在明显差距. 所以截至目前在大多基因组项目中读长更长的Nanopore往往不是首选.

2022年Oxford Nanopore Technologies发布了新R10.4.测序芯片. 该芯片与kit 14试剂盒相结合. 在标准的400碱基/秒(bp/s)的条件下, 使测序质量在单链模式下达到Q20(测序准确度大于99%), 双链模式下达到Q30(测序准确度大于99.9%), 既保证了准确度, 同时实现了高产出. ONT reads有着显著的长度优势, 在保证长度的同时提升精度, 将极大地推动T2T基因组等高连续性组装的进展.

4 植物端粒到端粒T2T基因组进展

4.1 开展T2T基因组学的意义

T2T基因组指的是具有端粒到端粒高质量、高准确性和高连续性的完整基因组. 三代测序技术的发展, 特别是高连续性的ONT超长测序和高准确性的HiFi测序强强联合, 克服了着丝粒和高重复区域的组装难题. 端粒以及着丝粒在生命活动中发挥着重要的作用, 因此进行高标准基因组组装对探索这些区域的功能和结构至关重要. 如探索染色体的起源, 驯化以及鉴定性别关键基因, 解析着丝粒和端粒等复杂结构的变异特征等.

自从2021年6月首个籼稻T2T基因组[60]发布以来, 截至2023年3月已发布了14个植物T2T植物基因组(表2). 如北京大学现代农学院张兴平团队[61]基于西瓜T2T基因组利用重测序数据快速鉴定了2个显性突变体(长果和雄性不育果), 在T2T基因组水平上发现均由典型的EMS诱变碱基突变类型G>A导致的. 中国农业科学院蔬菜花卉研究所分子育种创新团队[62]通过白菜T2T基因组发现, 着丝粒区域主要富集*ALE*和*CRM*类型的LTRs. 通过比较不同白菜亚种间的着丝粒序列, 发现白菜亚种间着丝粒分化显著. 进一步研究发现, 着丝粒与泛着丝粒区域分别富集Copia以及Gypsy类型的LTRs, 而且发现着丝粒区域的LTRs插入时间要显著晚于泛着丝粒区域. 这些结果暗示, 白菜LTRs的插入导致着丝粒区域经历着快速进化.

4.2 T2T基因组构建策略

目前T2T组装方法主要有以下两种. (ⅰ) 使用HiFi数据组装到contig水平, 随后使用Hi-C数据将contig挂载至染色体, 最后使用ONT超长读长填补gap. 该策略适合于基因组小于500 Mb, 且HiFi数据组装后contig数目较小的基因组. 该策略的优势是组装比较简单, 计算资源消耗少, 并且组装后准确性高. (ⅱ) 使用ONT超长读长直接进行组装, 随后使用二代纠错. 这种方法凭借ONT数据长度, 可以得到较大片段的contig, 矫正过后使用HiC挂载至染色体水平, 最后再次使用N50大于100 kb的ONT数据填补gap. 这种做法理论上比较普适但准确性略低.

4.3 高精度基因组组装算法的开发

一般来说, T2T基因组组装采用多种长读长测序技术联用的策略. 美国国立卫生研究院于2022年在*Nature Biotechnology*发表了一篇用于T2T高精度基因组装的新方法——Verkko[74]. 该方法首先通过HiFi reads构建德布鲁因图, 再将ONT reads对齐到图上, 逐步解决循环和缠结区域, 最终借用Canu的consensus板块得到组装结果. 该方法能够在较短的时间内, 完成多种物种真正完整的T2T基因组工具. 在人类HG002基因组测试中, Verkko可以将46条染色体中的20条组装

表 2 已发表植物T2T基因组

Table 2 Published plant T2T genomes

发表时间	物种	基因组大小(Mb)	组装策略
2021	籼稻[60]	397	CCS+Bionano+Hi-C
2021	香蕉[63]	484	ONT+Bionano
2021	澳洲胡桃[64]	826	HiFi+Hi-C
2021	番茄[65]	799	HiFi+ONT+Hi-C
2022	拟南芥[66]	133	HiFi+ONT+Hi-C
2023	猕猴桃[67]	640	HiFi+ONT+Hi-C
2023	草莓[68]	784	HiFi+ONT
2022	水稻[69]	398	HiFi+ONT
2022	西瓜[61]	369	HiFi+ONT+Bionano
2022	大麦[70]	4700	HiFi+Hi-C
2022	苦瓜[71]	286	HiFi+Hi-C
2023	白菜[62]	425	ONT+Hi-C
2023	柠檬[72]	633	HiFi+ONT+Hi-C
2023	桃金娘[73]	450	HiFi+Hi-C

至99.997%的准确率, 并且具有较高的自动化.

2022年3月加利福尼亚大学Anton Bankevich教授团队在*Nature Biotechnology*发表了一款用于高精度基因组组装的软件La Jolla Assembler (LJA)[75]. 该方法基于HiFi reads, 使用Bloom过滤, 随后在不同*k*-mer大小下, 构建多路德布鲁因图进行组装. 该团队通过近期发表的人类T2T基因组的长读长reads进行自动化组装, 以此来展示LJA的效果. 与Hifiasm[33]和HiCanu[34]相比, 结果显示LJA生成了更为连续的组装结果, 其中包括6条没有任何装配错误的完整染色体, 并且整个人类基因组组装中仅存在10个装配错误, 组装错误的数量减少了80%. 由此可见,新的组装工具在长读长测序的加持下, 将会使高精度基因组的组装变得更加简便与快捷.

5 植物泛基因组进展

5.1 开展泛基因组学研究的意义

近年来, 随着多种植物参考基因组的不断公布以及同种不同个体植物基因组的相互比较, 人们逐渐认识到单一个体的参考基因组无法全面代表整个物种的遗传多样性, 会遗漏大量的物种间遗传变异信息. “泛基因组”(Pangenome)的概念在微生物学领域中首次被提出, 并逐渐在各种生物体中得到应用, 指某一物种的全部基因组信息的集合[76]. 随着测序成本下降、技术提升, 2014年玉米[77]、野生大豆[78]、水稻[79]泛基因组的发表正式打开了植物泛基因组的大门. 迄今已在20余种植物开展了泛基因组研究. 植物泛基因组的发展有助于揭示丰富的遗传变异, 挖掘新的功能基因, 解析品种形成的分子基础, 深化对物种遗传多样性的认识.

5.2 泛基因组构建策略

在泛基因组样本选择时, 主要考虑样本数量与样本特性, 应尽可能的选择更全面、具有代表性的样本, 同时包括野生种与近代栽培种等. 这样构建的泛基因组可以提供更丰富的遗传变异信息以及更全面揭示基因组动态变化的过程.

泛基因组的构建方式主要分为三种(图4). (ⅰ) 迭代组装(Iterative): 基于物种的参考基因组, 将其他个体的基因组逐个与参考基因组比对, 将未比对上的reads组装成新的contig, 并添加至原始的参考基因组中, 通过多个物种不断迭代从而构建出一个泛基因组. 该策略具有成本低、构建速度快等优点. 早期的甘蓝[80]、油菜[81]等泛基因组通过该方法构建. (ⅱ) 从头组装(*de novo*): 该策略通过对同一个物种的多个样本进行完整

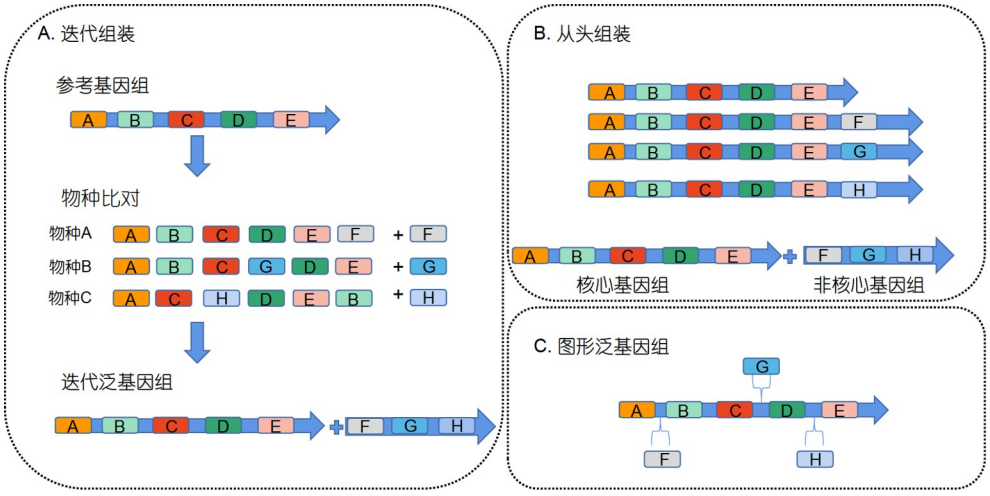


图 4　三种泛基因组构建方法. A: 迭代组装, 通过全基因组比对后, 将未比对到参考基因组的片段添加至参考基因组; B: 从头组装, 从头组装多个个体基因组后, 鉴定核心基因组与非核心基因组; C: 图形泛基因组, 用图的形式标注变异位点

Figure 4　Three pan-genome construction approaches. A: Iterative assembly. After whole genome alignment, fragments that are absent in the reference genome are added to the pan-genome; B: *de novo* assembly. Integration of multiple *de novo* assembled individual genomes leads to identification of core and dispensable genomes; C: graph-based pan-genome. The structural variations across the multiple individuals are represented in a graphical form

的从头组装, 再通过全基因组的比对分析鉴定出核心基因组(存在于所有样本中的基因组)与非核心基因组(存在于一个或多个样本中的基因组), 从而得到物种泛基因组. 该策略虽花费成本较高且需要较多计算资源, 但结果更为准确, 可以全面解析物种内的差异, 是目前主流的构建策略. 水稻[82]、芝麻[83]、大麦[84]、小麦[85]、拟南芥[86]等皆采用该策略构建泛基因组. (iii) 图形泛基因组(graph-based pan-genome): 以上两种做法所构建的泛基因组都为线性结构, 无法有效地储存和描述基因组变异. 图形结构泛基因组基于基因组从头组装, 或迭代组装的基础上, 将基因组及变异位点以图形的结构表示. 该策略除了考虑泛基因组的序列信息, 还能提供变异信息的空间位置. 与线性泛基因组相比, 图形结构泛基因组可以在群体中检测到更全面的变异信息, 与此同时也给基因组储存和计算带来了新的挑战. 目前大豆[87]、水稻[88]、高粱[89]、狼尾草[90]、番茄[91]等构建了图形泛基因组.

6　总结与展望

自拟南芥基因组发表的22年以来, 目前植物组学领域已经破译了800多个基因组, 范围从非维管植物到开花植物, 涵盖了植物进化历程上的主要进化分支. 伴随着测序技术、组装算法的快速发展, 植物复杂基因组、T2T基因组、泛基因组等前沿领域也取得了令人瞩目的进步. 丰富的组学遗传信息加深了人们对植物多样性遗传基础的理解, 并推动优良基因挖掘、植物育种等科学和技术进步.

然而全球约有45万至50万种植物物种[92], 尽管目前公布的基因组数量已呈现指数式的增长, 但目前破译的物种仅代表绿色植物多样性的一小部分. 例如, 在拥有1000多个物种的裸子植物门, 仅有其中13个物种的基因组得到了测序解析[93]. 另一方面, 高复杂的基因组, 如具有高重复性、高杂合性、同源多倍化的物种, 在组装质量上仍有较大的进步空间. 以目前发表的最大的植物基因组——糖松为例, 在高达31 Gb的基因组大小中, contigN50仅为4.5 kb[94], 在连续性上有待提升. 此外, 一些重要的经济作物如现代栽培种甘蔗, 因其为同源多倍体, 并具有复杂的染色体倍性关系, 至今都无法得到有效的破译. 在测序方面, 应当在保证准确性的前提下提升读长, 更准确地反馈染色体内的互作信号; 在组装算法方面, 应当提高组装的准确性和计算效率.

同时, 植物基因组的前沿领域方兴未艾, 仍需进一步探索. 高精度(精确度≥99.9%)、长读长(长度≥100 kb)的测序技术的出现, 使得构建T2T基因组

成为可能，但现今破译的植物T2T基因组大多为简单基因组. 因为部分复杂的着丝粒区域需要大量的手工校准，并且ONT超长等数据中长读长得率并不稳定，仅少数读长能够达到100 kb乃至1 Mb的长度，无法跨越一些片段重复区域及着丝粒区域中长达数Mb的串联重复阵列.

在泛基因组研究中，也存在着一系列挑战. 首先，构建策略不够完善; 其次, 泛基因组包含了大量的遗传信息, 在定位遗传变异, 关联位点的过程需要消耗大量的时间和计算资源; 最后，为了更好地应用泛基因组，规范的泛基因组可视化工具、群体遗传和比较基因组工具也亟需开发.

统而言之，随着粮食安全问题、植物的环境适应性、多态性问题逐渐受到重视，人们开始认识到基因组学是21世纪植物科学发展的核心之一. 植物基因组学研究的深入，催化了更广泛的植物学领域的突破. 现如今，植物基因组学研究正处于快速发展的阶段，测序技术和组装算法的进步能够帮助研究人员克服组装难题. 基因组学带来的洞见可以为植物的生长发育调控、环境适应性、植物系统进化以及资源植物利用等研究提供宝贵的信息，进而在农业和环境科学等领域提供潜在的应用.

参考文献

1 Hamilton J P, Robin Buell C. Advances in plant genome sequencing. Plant J, 2012, 70: 177–190
2 The Arabidopsis Genome Initiative . Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 2000, 408: 796–815
3 Shendure J, Balasubramanian S, Church G M, et al. DNA sequencing at 40: past, present and future. Nature, 2017, 550: 345–353
4 Marks R A, Hotaling S, Frandsen P B, et al. Representation and participation across 20 years of plant genome sequencing. Nat Plants, 2021, 7: 1571–1578
5 Sun Y, Shang L, Zhu Q H, et al. Twenty years of plant genome sequencing: achievements and challenges. Trends Plant Sci, 2022, 27: 391–401
6 Isobe S, Shirasawa K, Hirakawa H. Advances of whole genome sequencing in strawberry with NGS technologies. Hort J, 2020, 89: 108–114
7 Niu S, Li J, Bo W, et al. The Chinese pine genome and methylome unveil key features of conifer evolution. Cell, 2022, 185: 204–217.e14
8 Singh B, Salaria N, Thakur K, et al. Functional genomic approaches to improve crop plant heat stress tolerance. F1000Res, 2019, 8: 1721
9 Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science, 2002, 296: 79–92
10 Schnable P S, Ware D, Fulton R S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science, 2009, 326: 1112–1115
11 Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, 74: 5463–5467
12 Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science, 1998, 281: 363–365
13 Mitros T, Session A M, James B T, et al. Genome biology of the paleotetraploid perennial biomass crop *Miscanthus*. Nat Commun, 2020, 11: 5442
14 Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 2012, 490: 55–60
15 Avni R, Nave M, Barad O, et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science, 2017, 357: 93–97
16 Wenger A M, Peluso P, Rowell W J, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol, 2019, 37: 1155–1162
17 Hon T, Mars K, Young G, et al. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci Data, 2020, 7: 399
18 Zhou Q, Tang D, Huang W, et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. Nat Genet, 2020, 52: 1018–1023
19 Sun X, Jiao C, Schwaninger H, et al. Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. Nat Genet, 2020, 52: 1423–1432
20 Ma D, Dong S, Zhang S, et al. Chromosome-level reference genome assembly provides insights into aroma biosynthesis in passion fruit (*Passiflora edulis*). Mol Ecol Resour, 2021, 21: 955–968
21 Dudchenko O, Batra S S, Omer A D, et al. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science, 2017, 356: 92–95
22 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. In: Suzuki Y, ed. Single Molecule and Single Cell Sequencing. Advances in

Experimental Medicine and Biology. Singapore: Springer, 2019. 97–118
23 Ulahannan N, Pendleton M, Deshpande A, et al. Nanopore sequencing of DNA concatemers reveals higher-order features of chromatin structure. bioRxiv, 2019, 833590
24 Dohm J C, Lottaz C, Borodina T, et al. SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. Genome Res, 2007, 17: 1697–1706
25 Warren R L, Sutton G G, Jones S J M, et al. Assembling millions of short DNA sequences using SSAKE. Bioinformatics, 2007, 23: 500–501
26 Jeck W R, Reinhardt J A, Baltrus D A, et al. Extending assembly of short DNA sequences to handle error. Bioinformatics, 2007, 23: 2942–2944
27 Koren S, Walenz B P, Berlin K, et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res, 2017, 27: 722–736
28 Xiao C L, Chen Y, Xie S Q, et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods, 2017, 14: 1072–1074
29 Chen Y, Nie F, Xie S Q, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun, 2021, 12: 60
30 Namiki T, Hachiya T, Tanaka H, et al. MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. Nucleic Acids Res, 2012, 40: e155
31 Jackman S D, Vandervalk B P, Mohamadi H, et al. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res, 2017, 27: 768–777
32 Xie Y, Wu G, Tang J, et al. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics, 2014, 30: 1660–1666
33 Cheng H, Concepcion G T, Feng X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods, 2021, 18: 170–175
34 Nurk S, Walenz B P, Rhie A, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res, 2020, 30: 1291–1305
35 Yang H M. Genomics (in Chinese). Beijing: Science Press, 2016 [杨焕明. 基因组学. 北京:科学出版社, 2016]
36 Xie L J, Ye C Y, Shen E H. Research progress in genome sequencing (in Chinese). J Plant Sci, 2021, 39: 681–691 [谢玲娟, 叶楚玉, 沈恩惠. 基因组测序研究进展. 植物科学学报, 2021, 39: 681–691]
37 The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. Nature, 2011, 475: 189–195
38 Gao S H, Yu H Y. Research progress of complex genome sequencing technology (in Chinese). Heredity, 2018, 40: 944–963 [高胜寒, 禹海英. 复杂基因组测序技术研究进展. 遗传, 2018, 40: 944–963]
39 Tang D, Zhou Q. Research progress of plant genome assembly technology (in Chinese). Biotechnol Bull, 2021, 37: 1–12 [唐蝶, 周倩. 植物基因组组装技术研究进展. 生物技术通报, 2021, 37: 1–12]
40 Jiao W B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol, 2017, 36: 64–70
41 Neale D B, Wegrzyn J L, Stevens K A, et al. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol, 2014, 15: R59
42 Guan R, Zhao Y, Zhang H, et al. Draft genome of the living fossil *Ginkgo biloba*. GigaScience, 2016, 5: 49
43 Sun X, Zhu S, Li N, et al. A chromosome-level genome assembly of garlic (*Allium sativum*) provides insights into genome evolution and allicin biosynthesis. Mol Plant, 2020, 13: 1328–1339
44 Liu H, Wang X, Wang G, et al. The nearly complete genome of *Ginkgo biloba* illuminates gymnosperm evolution. Nat Plants, 2021, 7: 748–756
45 Wei C, Yang H, Wang S, et al. Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality. Proc Natl Acad Sci USA, 2018, 115: E4151–E4158
46 Carlsen S A, Schmell E, Weigel P H, et al. The effect of the method of isolation on the surface properties of isolated rat hepatocytes. J Biol Chem, 1981, 256: 8058–8062
47 He M, He Y, Zhang K, et al. Comparison of buckwheat genomes reveals the genetic basis of metabolomic divergence and ecotype differentiation. New Phytol, 2022, 235: 1927–1943
48 Sedlazeck F J, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods, 2018, 15: 461–468

49 Zhang X, Chen S, Shi L, et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. Nat Genet, 2021, 53: 1250–1259

50 Chen H, Zeng Y, Yang Y, et al. Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nat Commun, 2020, 11: 2494

51 Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res, 2017, 27: 801–812

52 Patterson M, Marschall T, Pisanti N, et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. J Comput Biol, 2015, 22: 498–509

53 Chin C S, Peluso P, Sedlazeck F J, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods, 2016, 13: 1050–1054

54 Koren S, Rhie A, Walenz B P, et al. *De novo* assembly of haplotype-resolved genomes with Trio binning. Nat Biotechnol, 2018, 36: 1174–1182

55 Chalhoub B, Denoeud F, Liu S, et al. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science, 2014, 345: 950–953

56 Zhang J, Zhang X, Tang H, et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat Genet, 2018, 50: 1565–1573

57 Zhang X, Zhang S, Zhao Q, et al. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants, 2019, 5: 833–845

58 Bao Z, Li C, Li G, et al. Genome architecture and tetrasomic inheritance of autotetraploid potato. Mol Plant, 2022, 15: 1211–1226

59 Sanderson N D, Kapel N, Rodger G, et al. Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. Microb Genomics, 2023, 9

60 Li K, Jiang W, Hui Y, et al. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. Mol Plant, 2021, 14: 1745–1756

61 Deng Y, Liu S, Zhang Y, et al. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. Mol Plant, 2022, 15: 1268–1284

62 Zhang L, Liang J, Chen H, et al. A near-complete genome assembly of *Brassica rapa* provides new insights into the evolution of centromeres. Plant Biotechnol J, 2023, 21: 1022–1032

63 Belser C, Baurens F C, Noel B, et al. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun Biol, 2021, 4: 1047

64 Sharma P, Masouleh A K, Topp B, et al. *De novo* chromosome level assembly of a plant genome from long read sequence data. Plant J, 2022, 109: 727–736

65 Su X, Wang B, Geng X, et al. A high-continuity and annotated tomato reference genome. BMC Genomics, 2021, 22: 898

66 Hou X, Wang D, Cheng Z, et al. A near-complete assembly of an *Arabidopsis thaliana* genome. Mol Plant, 2022, 15: 1247–1250

67 Han X, Zhang Y, Zhang Q, et al. Two haplotype-resolved, gap-free genome assemblies for *Actinidia latifolia* and *Actinidia chinensis* shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. Mol Plant, 2023, 16: 452–470

68 Zhou Y, Xiong J, Shu Z, et al. The telomere-to-telomere genome of *Fragaria vesca* reveals the genomic evolution of *Fragaria* and the origin of cultivated octoploid strawberry. Hortic Res, 2023, 10: uhad027

69 Zhang Y, Fu J, Wang K, et al. The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. Plant Biotechnol J, 2022, 20: 1642–1644

70 Navrátilová P, Toegelová H, Tulpová Z, et al. Prospects of telomere-to-telomere assembly in barley: analysis of sequence gaps in the MorexV3 reference genome. Plant Biotechnol J, 2022, 20: 1373–1386

71 Fu A, Zheng Y, Guo J, et al. Telomere-to-telomere genome assembly of bitter melon (*Momordica charantia* L. var. *abbreviata* Ser.) reveals fruit development, composition and ripening genetic characteristics. Hortic Res, 2023, 10: uhac228

72 Bao Y, Zeng Z, Yao W, et al. A gap-free and haplotype-resolved lemon genome provides insights into flavor synthesis and huanglongbing (HLB) tolerance. Hortic Res, 2023, 10: uhad020

73 Li F, Xu S, Xiao Z, et al. Gap-free genome assembly and comparative analysis reveal the evolution and anthocyanin accumulation mechanism of *Rhodomyrtus tomentosa*. Hortic Res, 2023, 10: uhad005

74 Rautiainen M, Nurk S, Walenz B P, et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol, 2023, 41: 1474–1482

75 Bankevich A, Bzikadze A V, Kolmogorov M, et al. Multiplex de Bruijn graphs enable genome assembly from long, high-fidelity reads. Nat Biotechnol, 2022, 40: 1075–1081

76 Morgante M, Depaoli E, Radovic S. Transposable elements and the plant pan-genomes. Curr Opin Plant Biol, 2007, 10: 149–155

77 Hirsch C N, Foerster J M, Johnson J M, et al. Insights into the maize pan-genome and pan-transcriptome. Plant Cell, 2014, 26: 121–135

78 Li Y, Zhou G, Ma J, et al. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol, 2014, 32: 1045–1052

79 Schatz M C, Maron L G, Stein J C, et al. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. Genome Biol, 2014, 15: 506

80 Golicz A A, Bayer P E, Barker G C, et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun, 2016, 7: 13390

81 Hurgobin B, Golicz A A, Bayer P E, et al. Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid *Brassica napus*. Plant Biotechnol J, 2018, 16: 1265–1274

82 Zhao Q, Feng Q, Lu H, et al. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet, 2018, 50: 278–284

83 Yu J, Golicz A A, Lu K, et al. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. Plant Biotechnol J, 2019, 17: 881–892

84 Gordon S P, Contreras-Moreira B, Woods D P, et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat Commun, 2017, 8: 2184

85 Walkowiak S, Gao L, Monat C, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature, 2020, 588: 277–283

86 Jiao W B, Schneeberger K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. Nat Commun, 2020, 11: 989

87 Liu Y, Du H, Li P, et al. Pan-genome of wild and cultivated soybeans. Cell, 2020, 182: 162–176.e13

88 Qin P, Lu H, Du H, et al. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell, 2021, 184: 3542–3558.e16

89 Tao Y, Luo H, Xu J, et al. Extensive variation within the pan-genome of cultivated and wild sorghum. Nat Plants, 2021, 7: 766–773

90 Yan H, Sun M, Zhang Z, et al. Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. Nat Genet, 2023, 55: 507–518

91 Li N, He Q, Wang J, et al. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet, 2023, 55: 852–860

92 Lughadha E N, Govaerts R, Belyaeva I, et al. Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. Phytotaxa, 2016, 272: 82

93 One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. Nature, 2019, 574: 679–685

94 Stevens K A, Wegrzyn J L, Zimin A, et al. Sequence of the sugar pine megagenome. Genetics, 2016, 204: 1613–1626

Research status and prospect of plant complex genomes and pan-genomes

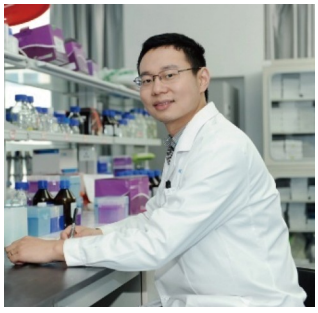
WANG YingHao[1], YU JiaXin[2], TANG HaiBao[1] & ZHANG XingTan[2]

1 College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou 350000, China;
2 Shenzhen Key Laboratory of Agricultural Genomics, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China

The genome refers to the sum of genetic materials within an organism and is one of the key areas of study in biology. Since the publication of the *Arabidopsis* genome sequence in 2000, more than 800 plant genomes have been published, which greatly promotes the development of plant molecular biology, genetics, and other fields. However, plant genomics research still faces a series of challenges, including the assembly of complex genomes with high heterozygosity, high repetitiveness, and high ploidy, as well as the construction of pan-genomes. This article comprehensively demonstrates the rapid development of plant genomics from three aspects: an overview of the development of plant genomics, genome sequencing technologies, and assembly algorithms. This review describes relevant strategies for assembling simple and complex genomes, and summarizes the assembly of "Telomere-to-Telomere" (T2T) and pan-genome construction methods, as well as their importance. Finally, the future development of plant genomics was discussed, and it was suggested that with the continuous advancement of technology, genome analysis techniques and methods will become more sophisticated, providing more support for further research on plant genomics. This review provides an important reference for research on the T2T assembly, assembly of complex genomes, and construction of pan-genomes in plants.

plant genomics, sequencing technology, assembly algorithm, complex plant genomes, plant pan-genomes

doi: 10.1360/SSV-2023-0068



张兴坦, 中国农业科学院深圳农业基因组研究所研究员、博士生导师、国家自然科学基金优秀青年科学基金获得者. 2009年于哈尔滨工业大学本科毕业, 2015年于重庆大学获得植物学博士学位. 长期致力于植物功能基因组学研究, 研究领域包括复杂基因组的组装和分型、生物信息分析技术的开发、基于组学大数据挖掘重要功能基因等. 研究成果以第一或通讯作者(含共同)在*Cell*, *Nature*, *Nature Genetics*, *Nature Plants*, *Nature Communications*等期刊发表.