基因组学与应用生物学, 2024 年, 第 43 卷, 第 6 期, 第 933-942 页

述评
Review

植物端粒到端粒(T2T)基因组研究进展与展望

宫少达[1] 谢文召[2] 赵如鹏[1] 冯康宁[1] 陈玲玲[1*]
1 广西大学生命科学与技术学院, 南宁, 530004; 2 华中农业大学信息学院, 武汉, 430070
*通信作者, llchen@gxueducn

摘 要 高质量的参考基因组是基因组学研究的基础。目前, 大多数的参考基因组仍然是不完整的。随着长读长测序技术的不断发展, 完成端粒到端粒(telomere-to-telomere, T2T)基因组组装的物种越来越多。T2T 基因组为深入研究着丝粒等复杂区域奠定了基础, 对功能基因的挖掘和重要生物机制的研究具有重要意义。本文概述了植物 T2T 基因组的研究进展, 结合实例介绍了相应的组装策略, 讨论了 T2T 基因组的意义和面临的挑战, 并对未来的发展前景进行了展望。

关键词 T2T 基因组; 测序技术; 基因组组装; 植物基因组

Progress and Prospect of Plant Telomere-to-telomere (T2T) Genome

GONG Shaoda[1] XIE Wenzhao[2] ZHAO Rupeng[1] FENG Kangning[1] CHEN Lingling[1*]
1 College of Life Science and Technology, Guangxi University, Nanning, 530004; 2 College of Informatics, Huazhong Agricultural University, Wuhan, 430070
*Corresponding author, llchen@gxueducn

DOI: 10.13417/j.gab.043.000933

Abstract The high-quality reference genome serves as the foundation for genomics research. Currently, the majority of reference genomes remain incomplete. With the continuous development of long-read sequencing technology, an increasing number of species have been successfully assembled telomere-to-telomere (T2T) genomes. The T2T genome lays the foundation for in-depth study of complex regions such as centromeres, and is of great significance for the excavation of functional genes and the study of important biological mechanisms. This paper provides an overview of the research progress on plant T2T genomes, presents the corresponding assembly strategies with examples, discusses the significance and challenges of T2T genomes, and offers a prospective view of future developments.

Keywords T2T genome; Sequencing technologies; Genome assembly; Plant genome

本世纪初, 被誉为生命科学“登月计划”的人类基因组序列的第一版草图问世, 尽管并不完整, 却在生物医学领域产生了深远的影响, 同时也极大推动了基因组学的发展。随后, 拟南芥(*Arabidopsis thaliana*)、小鼠(*Mus musculus*)、水稻(*Oryza sativa*)、玉米(*Zea mays*)等几百个物种的参考基因组序列草图被相继公布。由于基因组的复杂性, 特别是大量重复序列的存在, 使得参考基因组存在大量的“缺口(gap)”区域。端粒到端粒(telomere-to-telomere, T2T)基因组是利用多种测序策略, 完成一条或多条染色体端粒到端粒无缺口组装的基因组。长期以来, T2T 基因组组装一直是基因组学研究人员的梦想。

近年来, 经过近百名科学家组成的大型团队“T2T 联盟”的共同努力, 完成了最新的人类参考基因组(T2T-CHM13)。该基因组包括了所有 22 条常染色体和 X 染色体的无缺口组装, 标志着人类 T2T 基因

基金项目: 本研究由国家自然科学基金(32270712)和广西科技重大专项(桂科 AA23062085)共同资助。
引用格式: 宫少达, 谢文召, 赵如鹏, 等, 2024. 植物端粒到端粒(T2T)基因组研究进展与展望. 基因组学与应用生物学, 43(6): 933-942. [GONG S D, XIE W Z, ZHAO R P, et al., 2024. Progress and prospect of plant telomere-to-telomere (T2T) genome. Genomics and Applied Biology, 43(6): 933-942.]
通信作者简介: 陈玲玲, 教授, 从事作物生物信息工具开发及数据库的构建, 作物多组学数据整合及挖掘的研究。
收稿日期: 2023-12-26; 接受日期: 2024-02-22

组构建成功（Nurk et al., 2022）。相较于动物基因组，植物基因组的组装通常更具挑战性，因为其往往会经历多倍化事件，并且含有更高比例的重复序列。随着三代测序技术的发展，基因组组装已经进入新时代，特别是高准确性的 PacBio HiFi 和高连续性 ONT Ultra-long 三代测序技术的应用，为破解 T2T 基因组提供了强有力的技术支持。在植物基因组学领域，陆续发表了多个重要的模式物种如拟南芥、水稻等的 T2T 基因组（图 1）（Naish et al., 2021; Song et al., 2021）。迄今为止，已发表的有关植物 T2T 基因组的文章已有几十篇，并且仍在迅速增长（图 2，表 1），T2T 基因组已经成为基因组学研究的重要基础。

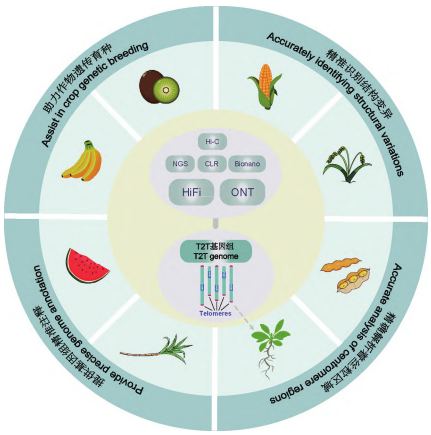

图 1　代表性植物 T2T 基因组及应用

Figure 1　Representative plant T2T genomes and their applications

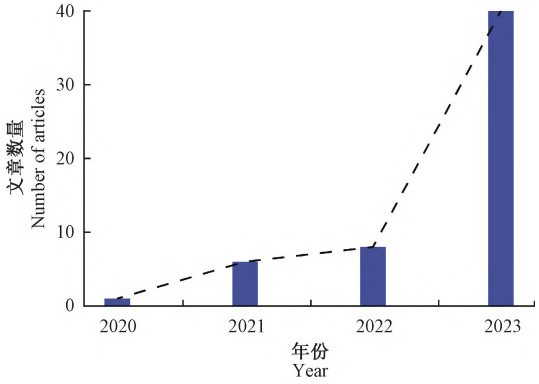

图 2　近年来植物 T2T 基因组文章发表数量

Figure 2　Number of publications on plant T2T genome in recent years

表 1　一些代表性植物 T2T 基因组信息

Table 1　T2T genome information of some representative plants

物种 Species	拉丁名 Latin name	基因组大小/Mb Genome size/Mb	测序策略 Sequencing strategy	文献出处 Reference
水稻	*Oryza sativa*	391.00/395.00	HiFi+CLR+Hi-C	Song et al., 2021
拟南芥	*Arabidopsis thaliana*	133.72	HiFi+ONT+Hi-C	Wang et al., 2022
香蕉	*Musa* spp.	484.00	ONT+Bionano	Belser et al., 2021
澳洲胡桃	*Macadamia integrifolia*	826.00	HiFi+Hi-C	Sharma et al., 2022
西瓜	*Citrullus lanatus*	369.32	HiFi+ONT+Hi-C	Deng et al., 2022
猕猴桃	*Actinidia latifolia*	608.00/640.00	HiFi+ONT+Hi-C	Han et al., 2023
白菜	*Brassica rapa*	424.59	ONT+HiC	Zhang et al., 2023b
草莓	*Fragaria vesca*	220.80	HiFi+ONT+Hi-C	Zhou et al., 2023
蔗茅	*Erianthus rufpilus*	756.60	HiFi+HiC	Wang et al., 2023b
辣根	*Armoracia rusticana*	610.05	HiFi+ONT+Hi-C	Shen et al., 2023
柠檬	*Citrus limon*	633.00	HiFi+ONT+Hi-C	Bao et al., 2023
胡萝卜	*Daucus carota*	427.33	HiFi+ONT+Hi-C	Wang et al., 2023c
玉米	*Zea mays*	2 178.00	HiFi+ONT	Chen et al., 2023
水稻（日本晴）	*Oryza sativa*	985.70	HiFi+ONT+Hi-C	Shang et al., 2023
大豆（Williams 82）	*Glycine max*	1 011.00	HiFi+ONT+Hi-C	Wang et al., 2023a
大豆（ZH13）	*Glycine max*	1 007.00	HiFi+ONT+Hi-C	Zhang et al., 2023a

1　基因组测序技术的发展

以 Sanger 法为代表的第一代测序技术具有读长较长、测序准确度极高等优势，使得其至今仍在单基因测序、克隆验证等领域发挥重要作用。然而，其高耗时、高成本、低通量的特点限制了其进一步发展。相比之下，第二代测序技术，通常被称为下一代测序技术（next generation sequencing, NGS），具有通量高、速度快、成本低等显著优势。二代测序技术克服了 Sanger 测序的一些限制，但仍存在读长较短、GC 偏好性以及难以跨越复杂重复区域等缺点，目前被广泛应用于基因组校正等辅助组装。

第三代测序技术也被称为长读长测序，它的出现使植物基因组的测序组装进入了一个新的阶段。Pacific Biosciences（PacBio）公司的 SMRT（single molecule real-time）测序技术和 Oxford Nanopore Technologies（ONT）公司的纳米孔测序技术是目前主流的两种测序技术。SMRT 测序技术可以产生两种不同的

reads，CLR 模式是为了尽可能地提高读长的长度而研发的，能够测序长度大于 30 kb 的 DNA 片段，但错误率较高；CCS 模式下的 HiFi reads 的读长较短但是准确度能够达到 99.9%。纳米孔测序技术的读长理论上不受设备限制，这使得它能够产生具有超长读长的 ONT ultral-long reads，最高读长可以达到 100~200 kb。HiFi 测序保障了基因组的高质量碱基，而 ONT 测序能够穿越一些较长、复杂的区域，这两者的优势互为补充，使它们在 T2T 基因组组装领域变得不可或缺。

测序 reads 使用组装软件组装到 contig 水平后，需要使用一些辅助方法来得到更连续的染色体水平的组装。相较于传统的遗传图谱、物理图谱、Bionano 等方法，现在应用更广泛的是 Hi-C（high-throughput chromosome conformation capture）测序技术。Hi-C 技术能够捕获全基因组染色质之间的互作信息，进而构建全基因组范围的高分辨率染色质互作图谱。它常被用于与 HiFi、ONT 数据结合生成 T2T 基因组；除此之外，Hi-C 在杂合二倍体、多倍体的组装中也表现良好。

2 基因组组装软件的介绍

随着测序技术的进步，各种组装软件也迅速发展。其中，基于三代测序数据的组装软件包括 Canu（Koren et al., 2017）、NECAT（Chen et al., 2021）、FALCON（Chin et al., 2016）、Flye（Kolmogorov et al., 2019）等都得到了广泛应用。表现较好的组装软件 Canu 将这一步骤划分为三步：校正、修剪和组装。Canu 在校正阶段可以提高碱基准确性，修剪阶段则能够去除一些冗余 reads，保留高质量序列，并最终完成 contig 的组装。随后，Canu 开始支持 HiFi 数据的读取，极大地提升了组装速度和质量。Nextdenovo是一款基于超长 ONT 数据开发的组装工具，已经成为 ONT 数据组装的常用工具，它能够生成高度连续的 contig，能够跨越基因组中大部分难以组装的区域。李恒团队提出了一种全新的针对 PacBio HiFi 数据的单倍型组装算法 Hifiasm（Cheng et al., 2021），使得在组装的过程中能够无损地保留单倍型信息，同时也提升了对基因组高重复和复杂区域的解析能力。最新版本的 Hifiasm 还能够利用超长的 ONT 数据支持 T2T 基因组的组装。Rautiainen 等（2023）通过集成 HiFi 和超长 ONT 数据，开发了用于 T2T 组装的混合基因组组装管道（Verkko），展示了其组装人类 T2T 基因组的能力。

3 T2T 基因组的构建

3.1 常规基因组组装策略

常规的基因组组装过程可以分为以下步骤：（1）基因组特征评估：基于 k-mer 的原理，利用高准确性的二代数据预估目标物种基因组的大小、杂合度、重复序列的含量等特征。基因组越小、杂合度越低、重复序列含量越少，则组装难度越低。（2）对测序 reads 进行校正，提高碱基的准确性。（3）将 reads 拼接成 contig。（4）利用遗传图谱、Bionano 光学图谱、Hi-C 等构建染色体水平的组装。

3.2 构建 T2T 基因组

3.2.1 T2T 基因组组装策略

T2T 基因组的组装策略大致可以分为 3 种（图 3）：（1）将 HiFi 数据组装成 contig，借助高质量的近缘物种基因组或 Hi-C 数据将 contig 提升至染色体水平，生成带有缺口的基因组。同时，对 ONT 数据进行组装校正，然后用 ONT 组装填补 HiFi 组装的缺口。这种方法生成的基因组碱基质量较高，适用于基因组较小的简单物种，西瓜（*Citrullus lanatus*）、草莓（*Fragaria vesca*）、桑树（*Morus notabilis*）等物种的 T2T 组装都采用这种策略（Deng et al., 2022；Ma et al., 2023；Zhou et al., 2023）。在某些物种中，只通过 HiFi 数据就可以完成一条或者多条染色体的 T2T 组装，如蔗茅（*Erianthus rufpilus*）、猕猴桃（*Actinidia latifolia*）（Han et al., 2023；Wang et al., 2023b）。（2）将 ONT 数据组装成 contig，然后进行碱基质量的校正、染色体的挂载，最后用 HiFi 组装补缺口。由于 ONT 的 reads 更长，能够跨越基因组中的高度重复区域，从而生成更为连续的序列。所以对于一些基因组较大、较为复杂的物种，通常需要使用 ONT 的组装结果作为骨架。利用这种方法，研究人员成功完成了大尺寸的玉米（*Zea mays*）T2T 组装（Chen et al., 2023）。（3）Verkko 和 Hifiasm 等软件目前支持同时将 HiFi 数据和 ONT 数据作为输入，以生成 T2T 级别的组装。这种混合组装策略已成功应用于水稻（日本晴）和大豆（*Glycine max*）的基因组组装中（Shang et al., 2023；Zhang et al., 2023a）。另外，还可以通过直接将 reads 比对到基因组的缺口区域来填补缺口，然后再对缺口区域进行碱基的校正。

总之，T2T 基因组的组装策略不是固定的，要通过物种的特异性、测序数据的类型以及不同软件

的组装效果来选择合适的组装策略。同时，在完全没有人工干预的情况下完成 T2T 的组装仍然具有挑战性，因此需要进行手动的检查来完成复杂基因组区域的组装，如端粒和着丝粒区域。

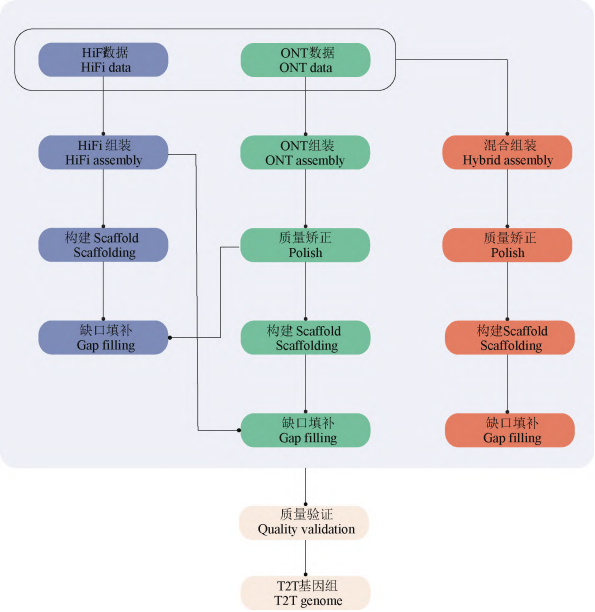

图 3　T2T 基因组组装策略

Figure 3　The assembly strategy of T2T genome

3.2.2　端粒与着丝粒的鉴定

在生物中端粒结构比较保守，一般分成几个部分：串联重复的端粒区域、亚端粒区域和中间的连接部分。在植物基因组中，通常使用七碱基端粒重复序列(5′端的 CCCTAAA 或 3′端的 TTTAGGG)与基因组进行比对来确定端粒区域。

着丝粒在植物细胞分裂过程中发挥着保障染色体正确分离的重要作用。由于着丝粒由高度重复的序列和少量的基因构成（Nagaki et al., 2004），因此可通过重复序列密度和基因密度进行着丝粒位置的预测。西瓜 T2T 组装中，通过对连续串联重复序列的注释，预测出了着丝粒位置（Deng et al., 2022）。猕猴桃 T2T 组装中，通过将串联重复序列、基因密度和 Hi-C 交互信息结合来确定着丝粒位置（Han et al., 2023）。更多的物种基于着丝粒区域的序列特征完成了着丝粒位置的预测工作，如香蕉(*Musa* spp.)、木薯(*Manihot esculenta*)、葡萄(*Vitis vinifera*)等(Huang et al., 2023; Shi et al., 2023; Xu et al., 2023)。近期发表的 quartet 工具中的 CentroMiner 功能可将基因组与重复序列注释文件作为输入，直接输出着丝粒位置的预测结果(Lin et al., 2023)。除此之外，还可以通过将 CENH3(着丝粒特异性组蛋白变体)结合 ChIP-seq 测序与基因组进行比对，完成对着丝粒位置的更精确定位。拟南芥、大麦(*Hordeum vulgare*)等物种使用 ChIP-seq 等多种技术完成了着丝粒位置的准确预测（Naish et al., 2021; Navrátilová et al., 2022）。

3.3　代表性植物 T2T 基因组

3.3.1　水稻

水稻是世界各地的主要粮食作物，也是植物基因组学和育种的重要模型系统。2005 年，国际水稻基因组测序计划(International Rice Genome Sequencing Project)发布了全球首个水稻基因组草图(International Rice Genome Sequencing Project and Sasaki, 2005)，使具有数千年历史的水稻相关研究正式跨入基因组学时代。近年来，一些高质量的水稻基因组陆续被报道，为水稻功能基因组学工作提供了重要的遗传资源。

Song 等(2021)率先报道了两种主要亚洲栽培稻 MH63 和 ZS97 的无缺口参考基因组序列的组装和分析。该研究使用多个软件，包括 Canu、FALCON、MECAT2 等，对 HiFi 和 CLR 的测序数据进行联合组装。随后使用其他组装程序的结果来填补 Canu 组装结果中的缺口。生成的基因组大小分别为 391.56 Mb 和 395.77 Mb，有 7 条和 10 条染色体达到了 T2T 水平。这是植物中首次报道的无缺口参考基因组。同年研究人员使用 Hifiasm 软件进行 contig 组装和缺口填补，成功完成了 MH63 的无缺口基因组的组装(Li et al., 2021)。利用 HIFi+ONT 数据，Zhang 等(2022)在 2022 年 6 月发表了 4 个杂交水稻骨干亲本的 T2T 基因组，这标志着水稻杂种优势与杂交育种研究已进入 T2T 时代。Shang 等(2023)发表了水稻(日本晴)所有染色体 T2T 完整无缺口的组装结果。该研究以日本晴为材料，利用超长 ONT 和 Pacbio HiFi 测序数据，使用 Hifiasm 进行初步组装。随后，使用 3D-DNA 将 contig 进行挂载（Dudchenko et al., 2017）。最后，通过局部组装和手工校正的方式填补了基因组中的 7 个缺口区域和 1 个端粒区域。该基因组中的每条染色体均由一条完整连续的序列组成，同时检测到了 12 条染色体两端的端粒区域。

3.3.2　拟南芥

拟南芥作为被广泛应用的模式植物，其基因组序列极大地加快了植物分子生物学研究。拟南芥 Col-0 基因组序列于 20 世纪初发布（Arabidopsis Genome Initiative, 2000），但在之后的 20 多年里，仍然存在大

量未填补的缺口。2021 年以来，研究学者们接连发布了 3 篇高质量的有关拟南芥 T2T 基因组的文章。

Wang 等(2022)使用 NextDenovo 对 ONT 数据进行组装，得到了 14 条 contig (N50 = 15. 39 Mb)。随后，使用 Hi-C 数对 contig 序列进行排序，获得了 5 条存在 7 个缺口的 scaffolds。整个 ONT-Hi-C 组装被替换为基于 BAC 锚定的 HiFi 序列组装，最终获得了仅有 2 个缺口的高质量 Col-0 基因组组装(命名为 Col-XJTU)。这也是第一篇高质量的有关拟南芥的 T2T 基因组报道。2021 年 11 月，Naish 等(2021)使用 Flye 对 ONT reads 进行组装，随后使用 ragtag 软件(Alonge et al，2022)将 ONT 组装与与 TAIR10 参考基因组对齐，最终通过人工校正等方式，获得了完整的拟南芥 Col-0 的基因组序列，并将其命名为 Col-CEN。焦雨铃团队利用 Nextdenovo 对 ONT 数据进行组装，并使用 HiFi 的初始组装结果填补缺口，最终得到了拟南芥 Col-PEK T2T 基因组的组装(Hou et al.，2022)。Col-PEK 为目前最为完整的拟南芥基因组组装，完成了 1 号、3 号、5 号染色体 T2T 的完整组装，仅 2 号和 4 号染色体的多拷贝核仁组织区(nucleolar organizing region，NOR)尚不完整。

3. 3. 3　大豆

大豆是世界上最重要的作物之一，是人类和牲畜不可或缺的食用油和蛋白质来源。2010 年，大豆品种 Williams 82 的基因组被破译 (Schmutz et al.，2010)，成为首个大豆参考基因组。除了第一个参考基因组之外，最近已经应用长读长测序技术发布了数十种栽培大豆种质的从头基因组组装，然而这些基因组仍存在数千个缺口。

宋庆鑫团队完成了大豆品种(Williams 82)首次 T2T 基因组的组装 (Wang et al.，2023a)。研究团队首先使用 Hifiasm 将 HiFi reads 进行初始组装。随后，使用 3D-DNA 软件生成带有缺口的染色体级别的基因组。同时使用 minimap 和 miniasm 软件将 ONT reads 组装成 contig (Li，2016；Li，2018)。最后利用校正后的 ONT 组装填补染色体上的缺口，获得了 T2T 级别的参考基因组 Wm82-NJAU。中国大豆栽培品种“中黄 13(ZH 13)”因其高产和抗逆性而在中国广泛使用，代表了亚洲大豆的遗传标记。Zhang 等(2023a)近期完成了第一个 ZH13-T2T 的组装，使用 Hifiasm 将 ONT 和 HiFi 数据进行混合组装生成初始的组装结果，进一步利用 ALL-HIC、3D-DNA软件将 contig 序列挂载到染色体上 (Dudchenko et al.，2017；Zhang et al.，2019)，利用 HiFi reads 和 ONT reads 填补染色体上的缺口区域。为了进一步提升基因组质量，使用多款软件包括 Meryl、Winnowmap2、Racon、Merfin 等进行了校正(Vaser et al.，2017；Rhie et al.，2020；Jain et al.，2022)。最终成功完成了 ZH13-T2T 的组装，为进一步研究大豆的功能基因组提供了坚实的数据基础。

4　基因组质量评估

随着长读长组装技术的迅速发展，基因组的质量和完整性有了显著提升，对于 T2T 基因组的组装结果进行多方面评估变得尤为关键。基因组评估可分为连续性、完整性和准确性三个方面，本研究总结了用于评估基因组质量的常见指标和工具。

4. 1　N50

N50 是评价基因组连续性的指标。得到 contig 级别的组装之后，根据 contig 的长度从大到小进行排序，然后逐步开始累加，当加和长度超过总长一半时，加入的序列长度即为 N50 长度。N50 越长，代表基因组组装的连续性越好。

4. 2　BUSCO 完整性评估

Busco (Benchmarking Universal Single-Copy Orthologs)是一种用于评估基因组完整性的工具(Simão et al.，2015)。Busco 首先建立了一个包含多个物种的基准数据库，其中包括已知的单拷贝基因。这些单拷贝基因在不同物种中都有相似的功能，并且相对保守，因此被用作评估基因组完整性的指标。其结果中包括单拷贝比对基因、多拷贝比对基因、部分比对基因和丢失基因，比对成功的基因越多，代表基因组完整度越好。Busco 的优势在于其基于普遍存在的基因集，使其能够适用于不同物种，为基因组学研究提供了一个广泛应用的评估工具。

4. 3　测序数据的比对率与覆盖深度

数据比对率是指成功将原始的测序数据比对到组装结果的比例，它反映了组装结果与原始测序数据的一致性。覆盖深度指某一基因组区域内测序覆盖的深度。通过将原始数据比对回基因组，并手动检查 reads 覆盖深度分布，可以判断基因组中是否存在局部的组装错误。

4. 4　基于 k-mer 的评估

k-mer 提供了一种无需参考基因组即可用于评

估组装质量的途径。Merqury 是一种较新的用于评估基因组质量的工具，其仅利用 k-mer 生成组装评估指标(Rhie et al., 2020)。Merqury 通过将未组装的高精度测序 reads 的 k-mer 集与基因组组装进行比较，可揭示组装中的拷贝数错误，并准确计算一致性质量(QV)和 k-mer 完整性。当亲本 k-mer 可用时，Merqury 还能够检测分型的准确性和单倍型的完整性。与传统的组装指标相比，Merqury 提供了更全面的组装质量评估，在报道植物 T2T 基因组中，这一方法已得到广泛应用。以拟南芥 Col-XJTU 为例，该组装的 5 条染色体的 QV 值均超过 60 (Wang et al., 2022)。与之前的版本相比，西瓜基因组的 QV 值显著提升，从 35.22 增至 76.97，证明了基因组的可靠性和碱基的高准确性(Deng et al., 2022)。

4.5 LAI 值

LTR 组装指数(LTR Assembly Index，LAI)是用完整 LTR-RTs 转座子在所有 LTR-RTs 的占比来评估基因组组装连贯性的一个指数。LAI 独立于基因组大小、LTR-RT 含量以及基因完整性评估指标(如 BUSCO)等。根据基因组的质量，LAI 值一般可以分为三个类别：Draft 级别(0≤LAI<10)，Reference 级别(10≤LAI<20)，Gold 级别(20≤LAI)。蔗茅 T2T 的 LAI 值是 20.87 (Wang et al., 2023b)，达到了黄金标准，表明重复区域的高度完整性。

4.6 SNVs 评估碱基准确性

单核苷酸变异(single nucleotide variations，SNVs)被用于评估碱基的准确性。可以使用 bwa 软件将二代数据比对到参考基因组，随后使用 GATK 等工具进行 SNP calling 并进行过滤 (McKenna et al., 2010)，最终统计纯合和杂合 SNVs 的数量。纯合 SNVs 代表着可能的错误碱基，其占整个基因组长度的比例即为碱基的错误率。基因组的质量越高，碱基的错误率越小。

5 T2T 基因组的意义与应用

5.1 正确地识别结构变异

结构变异(structural variations，SVs)是个体与个体基因组序列中存在的差异，长度一般大于50 bp，包括倒位、插入、缺失等变异类型(Roses et al., 2016)，它们在个体间的存在对于理解基因组的功能和个体间的遗传差异至关重要。通过 T2T 基因组，可以更加准确地识别结构变异，尤其是那些涉及到染色体着丝粒或者端粒区域的复杂变异。Zhang 等(2022)基于对 4 个杂交水稻骨干亲本以及已发表的其余6 个水稻的基因组，与日本晴基因组进行比对，鉴定出不同水稻亲本之间的结构变异。在日本晴和 10 个籼稻基因组之间鉴定出422 858~526 481 个 InDels 和 56 817 个非冗余的 SVs，包括 52 943 个 PAVs。Song 等(2021)在比较了两个无缺口水稻组装 MH63 和 ZS97 每条染色体上的 PAV 分布后，发现了 11 号染色体长臂末端附近存在大量结构变异。在 MH63 中独特地检测到了一个扩张区域(MH-E)和一个插入区域 MH-I，与 ZS97 的相应区域相比，MH63 包含了更多抗性相关基因。

5.2 挖掘新的基因

T2T 基因组为新基因的鉴定提供了重要资源。通过对新基因的研究，能够更深入地了解植物的生物学过程，如生长发育、抗逆性等，为植物生物学的基础研究提供重要信息。在 T2T 香蕉基因组中，研究者新发现了 1 700 个基因，这些新基因主要是串联重复基因，多以基因簇的形式存在于染色体重组区域。这些基因簇包含一些重要的基因家族，如萜类合酶基因家族，它们负责萜类化合物的合成，而这些化合物在植物风味和环境的适应性中发挥着关键作用 (Belser et al., 2021)。

5.3 解析着丝粒等复杂基因组区域

着丝粒等复杂基因组区域是植物基因组组装的难点，T2T 基因组为着丝粒等复杂区域的研究提供了强大的工具，促使我们更全面、更深入地理解这些复杂区域的结构、功能和调控机制，推动对植物基因组结构和进化的理解迈向一个新的阶段。通过鉴定拟南芥着丝粒上的 CEN180 序列，研究者发现不同染色体上 CEN180 序列存在显著差异。同时，同一条染色体内的 CEN180 序列呈现均质化趋势，但是逆转录转座子 ATHILA 的入侵造成了 CEN180 序列的多样化，进而影响了染色体内 CEN180 序列的均质化过程，二者共同推动了其着丝粒结构和功能的进化过程(Naish et al., 2021)。基于从头预测，研究人员鉴定到大豆 13 号染色体着丝粒特异的未发表过的重复序列 GmCent-3，相比于已知的 GmCent-1 和 GmCent-2，GmCent-3 序列间变异更小，这意味着 13 号染色体的着丝粒形成可能晚于其他染色体(Wang et al., 2023a)。

5.4 深入研究重复序列

重复序列在基因组进化中发挥着重要作用。例

如，转座子(TE)参与塑造基因表达模式和基因组结构的形成。对于长末端重复反转录转座子(LTR-RT)来说尤其如此，它是植物基因组中最常见的重复，它的增殖可能导致基因组扩张。高质量的基因组为系统地研究重复序列提供了良好的资源。研究人员通过对高质量的水稻 T2T 基因组 MH63 的深入分析，发现籼稻基因组比粳稻基因组具有更多的转座子(TE)和片段重复(segmental duplications，SD)(Li et al.，2021)。SD 产生大量的重复基因，并通过剂量效应或新/亚功能化影响植物性状。研究发现 TE 的插入不仅影响了重复基因的表达，还加速了这些重复基因的进化，这揭示了 TE 和 SD 对水稻基因组进化的协同作用。玉米 Mo17 的 T2T 基因组鉴定到了约 88.37%的重复序列，包括 75.52%的逆转录转座子和 9.78%的 DNA 转座子(Chen et al.，2023)。同时，Mo17 基因组富含卫星序列，其中 Knob180、CentC 和 TR-1 重复是最丰富的卫星序列，占已确定卫星长度的 90.32%。另外，研究人员还发现玉米基因组内存在主要由 TAG 三核苷酸重复组成的超长简单序列富集区，其长度可达 1.56 Mb，包含近 30 万个 TAG 拷贝。

6 T2T 基因组的挑战

6.1 高重复序列

在植物基因组中，重复序列的比例远高于其他生物，如病毒、细菌和动物，因而成为植物基因组中的重要组成部分。在黄杨(*Buxus sinica*)和水青树(*Tetracentron sinense*)基因组中，重复序列的含量分别占据 76.4%和 78.5%(Chanderbali et al.，2022)。而在蕨类植物基因组中，重复序列的含量更是高达 85.2%，其中 LTR-RTs 约占 67.0%(Marchant et al.，2022)。这些富含重复序列的区域通常涉及到许多关键的遗传功能区域，包括端粒、着丝粒、多拷贝基因以及非重组和高度异染色质的染色体，精确组装它们成为 T2T 基因组的一个难题。在以二代测序为基础的组装中，重复序列的组装通常不够完整。相较于二代测序技术，三代长读长测序能够越过重复序列区域，显著改善组装的连续性以及重复序列组装的完整性和准确性。然而，对于碱基准确度较低的 ONT 测序而言，测序错误给准确区分重复序列中的微小变异和测序错误带来了挑战。一方面，低于测序错误率的低频遗传变异可能会被错误地视为测序错误；另一方面，如果不进行测序错误的纠正，测序错误可能被错误地解释为遗传变异(Kong et al.，2023)。

6.2 高杂合度

由于远缘杂交和自交不亲和性，许多植物基因组具有较高的杂合度。在组装高杂合基因组时，存在无法准确识别和合并高杂合区域的问题，通常会导致这些区域被组装为两个独立的 contig，使得组装基因组的大小超过实际单倍型基因组的大小。在这种情况下，可以采取去除冗余的方法来尽可能地保留一套单倍型。例如，研究人员使用 HaploMerger2 软件分离了荔枝(*Litchi chinensis* Sonn)的两套单倍型(Hu et al.，2022)。随着 Hifiasm 等可以进行分型的软件开发出来，可以直接生成两套单倍型的组装结果，这对于高度杂合物种的组装具有突破性的意义。然而，要同时实现单倍型分离和 T2T 的组装，还是一项十分具有挑战性的任务。

6.3 高倍性

多倍体，定义为拥有三组或更多组同源染色体，在陆生植物中十分常见。两种普遍认可的多倍体形式是同源多倍体和异源多倍体。同源多倍体起源于单个物种的全基因组复制事件，异源多倍体则是两个或更多不同物种之间的杂交形成的。对于异源多倍体来说，区分源自不同祖先物种的亚基因组相对容易，因为它们在漫长的进化历史中保持了较大的遗传变异，如花生(*Arachis hypogaea*)、艾蒿(*Artemisia argyi*)(Zhuang et al.，2019；Miao et al.，2022)。对于同源多倍体来说，由于同源染色体之间的高度相似性，不同单倍型之间几乎相同的同源序列无法准确区分，这可能会产生许多折叠的 contig，为同源多倍体的组装带来了巨大的挑战。

6.4 超大基因组

在已经测序的植物基因组中，基因组大小通常可以跨越几个数量级，从十几 Mb 到几十 Gb 不等。超大基因组的组装面临的主要问题之一是数据处理的复杂性。由于基因组的庞大尺寸，传统的组装算法可能在处理这些超大基因组时效率低下。另一方面，超大基因组的组装也受到数据生成和存储的挑战，生成足够的长读长数据需要昂贵的测序成本，而存储和处理庞大的数据集也需要更大的计算资源。

7 讨论与展望

植物基因组一直是基因组学关注的重点，近年

来 T2T 基因组已经成为植物基因组领域的新趋势。T2T 基因组在识别结构变异、挖掘新基因、探索着丝粒区域、研究重复序列等方面具有重要的意义。通过与其他组学数据如转录组、蛋白组、代谢组等的联合分析，可以深入挖掘物种内的有效遗传信息，从而解决重要的生物学问题。

随着测序技术和组装算法的进步，二倍体的 T2T 基因组组装已经取得了一定进展。同时，在复杂基因的组装中也出现了很多成功的案例，比如具有高重复序列含量的玉米和异源四倍体辣根（*Armoracia rusticana*）（Chen et al., 2023；Shen et al., 2023）。然而，对于具有高重复序列含量、高杂合、高倍性、大尺寸的复杂物种，组装过程中仍然存在极大的挑战性。例如，在同源多倍体的组装中，在多条同源染色体的某个区域的序列高度相似，这个区域有可能只生成一套单倍型组装，并且难以通过 Hi-C 数据等信息准确地将这些 contig 分配给特定的单倍型。此外，超大基因组通常含有大量的重复序列，会显著降低组装的连续性，这对测序数据的长度提出了更高的要求。即使是在简单基因组中，不通过手动干预完成全部染色体的 T2T 组装仍然十分困难。

PacBio 公司近期推出的 Revio 测序平台采用全新的 SMRT Cell 芯片，将 HiFi 数据通量增加了 15 倍，从而显著降低了 HiFi 测序成本。ONT 最新芯片 R10.4.1 在保证读长的前提下，可以将碱基的准确性提升至 99%。这将为未来更高质量的 T2T 基因组的组装奠定基础。毫无疑问，随着测序技术和组装软件的不断改进，获取高质量 T2T 基因组的时间和费用将逐渐降低，使更多的物种能够获得高质量的基因组序列及其他组学数据信息。除了构建高质量基因组，基因组注释包括重复序列注释、基因结构注释和基因功能注释等，也是基因组分析的重要内容之一，但目前基因组精准注释仍然存在很大问题。后续需要进一步开发基因组精准注释及下游分析工具，并加强组学领域之间以及遗传育种学的交叉研究和整合应用，为重要植物优良性状的遗传改良提供更宝贵的资源。

作者贡献

宫少达负责文章的构思以及论文的撰写；谢文召、赵如鹏、冯康宁参与文献整理和论文修改；陈玲玲是本文的指导者和负责人，指导论文框架设计及论文写作与修改。全体作者都已阅读并同意最终的文本。

参考文献

ALONGE M, LEBEIGLE L, KIRSCHE M, et al., 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol., 23(1): 258.

ARABIDOPSIS GENOME INITIATIVE, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408(6814): 796-815.

BAO Y X, ZENG Z Y, YAO W, et al., 2023. A gap-free and haplotype-resolved lemon genome provides insights into flavor synthesis and huanglongbing (HLB) tolerance. Hortic. Res., 10(4): uhad020.

BELSER C, BAURENS F C, NOEL B, et al., 2021. Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. Commun. Biol., 4(1): 1047.

CHANDERBALI A S, JIN L L, XU Q J, et al., 2022. *Buxus* and *Tetracentron* genomes help resolve eudicot genome history. Nat. Commun., 13(1): 643.

CHEN J, WANG Z J, TAN K W, et al., 2023. A complete telomere-to-telomere assembly of the maize genome. Nat. Genet., 55(7): 1221-1231.

CHEN Y, NIE F, XIE S Q, et al., 2021. Efficient assembly of nanopore reads *via* highly accurate and intact error correction. Nat. Commun., 12(1): 60.

CHENG H Y, CONCEPCION G T, FENG X W, et al., 2021. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat. Methods, 18(2): 170-175.

CHIN C S, PELUSO P, SEDLAZECK F J, et al., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nat. Methods, 13(12): 1050-1054.

DENG Y, LIU S C, ZHANG Y L, et al., 2022. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene disco-very and breeding. Mol. Plant, 15(8): 1268-1284.

DUDCHENKO O, BATRA S S, OMER A D, et al., 2017. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science, 356(6333): 92-95.

HAN X, ZHANG Y L, ZHANG Q, et al., 2023. 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, 16(2): 452-470.

HOU X R, WANG D P, CHENG Z K, et al., 2022. A near-complete assembly of an *Arabidopsis thaliana* genome. Mol. Plant, 15(8): 1247-1250.

HU G B, FENG J T, XIANG X, et al., 2022. Two divergent hap-

lotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. Nat. Genet., 54(1): 73-83.

HUANG H R, LIU X, ARSHAD R, et al., 2023. Telomere-to-telomere haplotype-resolved reference genome reveals subgenome divergence and disease resistance in triploid Cavendish banana. Hortic. Res., 10(9): uhad153.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT, SASAKI T, 2005. The map-based sequence of the rice genome. Nature, 436(7052): 793-800.

JAIN C, RHIE A, HANSEN N F, et al., 2022. Long-read mapping to repetitive reference sequences using Winnowmap2. Nat. Methods, 19(6): 705-710.

KOLMOGOROV M, YUAN J, LIN Y, et al., 2019. Assembly of long, error-prone reads using repeat graphs. Nat. Biotechnol., 37(5): 540-546.

KONG W, WANG Y, ZHANG S, et al., 2023. Recent advances in assembly of complex plant genomes. Genom. Proteom. Bioinform., 21(3): 427-439.

KOREN S, WALENZ B P, BERLIN K, et al., 2017. Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. Genome Res., 27(5): 722-736.

LI H, 2016. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics, 32(14): 2103-2110.

LI H, 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18): 3094-3100.

LI K, JIANG W K, HUI Y Y, et al., 2021. Gapless indica rice genome reveals synergistic contributions of active transposable elements and segmental duplications to rice genome evolution. Mol. Plant, 14(10): 1745-1756.

LIN Y Z, YE C, LI X Z, et al., 2023. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centrome-ric repeat identification. Hortic. Res., 10(8): uhad127.

MA B, WANG H H, LIU J C, et al., 2023. The gap-free genome of mulberry elucidates the architecture and evolution of polycentric chromosomes. Hortic. Res., 10(7): uhad111.

MARCHANT D B, CHEN G, CAI S G, et al., 2022. Dynamic genome evolution in a model fern. Nat. Plants, 8(9): 1038-1051.

MCKENNA A, HANNA M, BANKS E, et al., 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res., 20(9): 1297-1303.

MIAO Y H, LUO D D, ZHAO T T, et al., 2022. Genome sequencing reveals chromosome fusion and extensive expansion of genes related to secondary metabolism in *Artemisia argyi*. Plant Biotechnol. J., 20(10): 1902-1915.

NAGAKI K, CHENG Z K, OUYANG S, et al., 2004. Sequencing of a rice centromere uncovers active genes. Nat. Genet., 36(2): 138-145.

NAISH M, ALONGE M, WLODZIMIERZ P, et al., 2021. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. Science, 374(6569): eabi7489.

NAVRÁTILOVÁ P, TOEGELOVÁ H, TULPOVÁ Z, et al., 2022. Prospects of telomere-to-telomere assembly in barley: analysis of sequence gaps in the MorexV3 reference genome. Plant Biotechnol. J., 20(7): 1373-1386.

NURK S, KOREN S, RHIE A, et al., 2022. The complete sequence of a human genome. Science, 376(6588): 44-53.

RAUTIAINEN M, NURK S, WALENZ B P, et al., 2023. Telomere-to-telomere assembly of diploid chromosomes with Verkko. Nat. Biotechnol., 41: 1474-1482.

RHIE A, WALENZ B P, KOREN S, et al., 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol., 21(1): 245.

ROSES A D, AKKARI P A, CHIBA-FALEK O, et al., 2016. Structural variants can be more informative for disease diagnostics, prognostics and translation than current SNP mapping and exon sequencing. Expert Opin. Drug Metab. Toxicol., 12(2): 135-147.

SCHMUTZ J, CANNON S B, SCHLUETER J, et al., 2010. Genome sequence of the palaeopolyploid soybean. Nature, 463(7278): 178-183.

SHANG L G, HE W C, WANG T Y, et al., 2023. A complete assembly of the rice Nipponbare reference genome. Mol. Plant, 16(8): 1232-1236.

SHARMA P, MASOULEH A K, TOPP B, et al., 2022. *De novo* chromosome level assembly of a plant genome from long read sequence data. Plant J., 109(3): 727-736.

SHEN F, XU S X, SHEN Q, et al., 2023. The allotetraploid horseradish genome provides insights into subgenome diversification and formation of critical traits. Nat. Commun., 14(1): 4102.

SHI X Y, CAO S, WANG X, et al., 2023. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. Hortic. Res., 10(5): uhad061.

SIMÃO F A, WATERHOUSE R M, IOANNIDIS P, et al., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19): 3210-3212.

SONG J M, XIE W Z, WANG S, et al., 2021. Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant, 14(10): 1757-1767.

VASER R, SOVI Ć I, NAGARAJAN N, et al., 2017. Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res., 27(5): 737-746.

WANG B, YANG X F, JIA Y Y, et al., 2022. High-quality *Arabidopsis thaliana* genome assembly with nanopore and HiFi long reads. Genomics Proteomics Bioinformatics, 20(1): 4-13.

WANG L F, ZHANG M Z, LI M N, et al., 2023a. A telomere-to-telomere gap-free assembly of soybean genome. Mol. Plant, 16(11): 1711-1714.

WANG T Y, WANG B Y, HUA X T, et al., 2023b. A complete gap-free diploid genome in *Saccharum* complex and the genomic footprints of evolution in the highly polyploid *Saccharum* genus. Nat. Plants, 9(4): 554-571.

WANG Y H, LIU P Z, LIU H, et al., 2023c. Telomere-to-telomere carrot (*Daucus carota*) genome assembly reveals carotenoid characteristics. Hortic. Res., 10(7): uhad103.

XU X D, ZHAO R P, XIAO L, et al., 2023. Telomere-to-telomere assembly of cassava genome reveals the evolution of cassava and divergence of allelic expression. Hortic. Res., 10(11): uhad200.

ZHANG C, XIE L, YU H, et al., 2023a. The T2T genome assembly of soybean cultivar ZH13 and its epigenetic landscapes. Mol. Plant, 16(11): 1715-1718.

ZHANG L, LIANG J L, CHEN H X, et al., 2023b. A near-complete genome assembly of *Brassica rapa* provides new insights into the evolution of centromeres. Plant Biotechnol. J., 21(5): 1022-1032.

ZHANG X T, ZHANG S C, ZHAO Q, et al., 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat. Plants, 5(8): 833-845.

ZHANG Y L, FU J, WANG K, et al., 2022. The telomere-to-telomere gap-free genome of four rice parents reveals SV and PAV patterns in hybrid rice breeding. Plant Biotechnol. J., 20(9): 1642-1644.

ZHOU Y H, XIONG J S, SHU Z Q, et al., 2023. The telomere-to-telomere genome of *Fragaria vesca* reveals the genomic evolution of *Fragaria* and the origin of cultivated octoploid strawberry. Hortic. Res., 10(4): uhad027.

ZHUANG W J, CHEN H, YANG M, et al., 2019. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. Nat. Genet., 51(5): 865-876.

(责任副主编 王海峰)
(责任副主编 罗继景)