

复杂基因组测序技术研究进展

高胜寒, 禹海英, 吴双阳, 王森, 耿佳宁, 骆迎峰, 胡松年

中国科学院北京基因组研究所, 中国科学院基因组科学与信息重点实验室, 北京 100101

摘要: 复杂基因组指的是无法使用常规测序和组装手段直接解析的一类基因组, 通常指包含高比例重复序列、高杂合度、极端 GC 含量、存在难消除异源 DNA 污染的基因组。为了解决复杂基因组的测序和组装问题, 需要分别从基因组测序实验方法、测序技术平台、组装算法与策略 3 个方面进行深入研究。本文详细介绍了复杂基因组测序组装相关的现有技术与方法, 并结合复杂基因组经典实例介绍了复杂基因组测序的技术解决途径和发展历程, 可为制订合适的复杂基因组测序策略提供参考。

关键词: 复杂基因组; 基因组组装; 基因组测序技术

Advances of sequencing and assembling technologies for complex genomes

Shenghan Gao, Haiying Yu, Shuangyang Wu, Sen Wang, Jianing Geng,
Yingfeng Luo, Songnian Hu

CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences,
Beijing 100101, China

Abstract: Complex genomes are noted to be extremely difficult to sequence or assemble by using ordinary methods. Complex genomes are typically characterized as being highly repetitive, highly heterozygous, extremely GC biased, or naturally contaminated, i.e., contaminations which cannot be removed before sequencing. To solve these problems with sequencing and assembling complex genomes, three major techniques include: (1) DNA extraction experiments, (2) Sequencing technologies and platforms, and (3) Algorithms and strategies for assembling. In this review, we summarize these state-of-the-art technologies and strategies used in these directions. We also review the representative projects of complex genome sequencing and address the development of these technologies and strategies for solving the challenges when sequencing or assembling complex genomes.

Keywords: complex genome; genome assembly; genome sequencing technologies

收稿日期: 2018-09-10; 修回日期: 2018-10-29

作者简介: 高胜寒, 博士, 助理研究员, 研究方向: 基因组结构与稳定性。E-mail: gaoshh@big.ac.cn

通讯作者: 胡松年, 博士, 研究员, 博士生导师, 研究方向: 复杂基因组与药物基因组。E-mail: husn@big.ac.cn

DOI: 10.16288/j.ycz.18-255

网络出版时间: 2018/11/6 17:30:31

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20181106.1730.005.html>

基因组是所有生命遗传物质的集合, 为生命行使生物学功能提供指导, 基因组中的碱基序列信息记录着生命进化的历史。因而, 基因组序列的完整解析可极大促进基因功能研究, 更为物种相互作用和基因组比较等生命科学研究提供基础信息。大多数生物的基因组均由 A、T、G、C 4 种碱基组成, 其组合顺序和总长度各不相同, 如何快速和低成本地获取基因组序列一直是基因组学领域的重心。由于测序技术或测序仪器的内在缺陷, 测序读长仍小于基因组长度, 所以除少数基因组较小的 DNA 病毒外, 绝大多数基因组仍无法通过一次测序直接获得全部的序列信息, 需要通过高覆盖度测序和序列组装获得完整的基因组信息。而复杂基因组指的是无法使用常规测序和组装手段直接解析的一类基因组, 通常是指包含高比例重复序列、高杂合度、存在难以消除的异源 DNA 污染的基因组。本文从复杂基因组的特点和来源入手, 结合实例介绍了复杂基因组测序的技术解决途径和发展历程, 为制订合适的复杂基因组测序策略提供参考。

1 复杂基因组概念

复杂基因组是根据重复序列比例和杂合度高低来定义的。通常杂合率大于 0.8%、重复序列比例大于 60% 就称为复杂基因组。基因组的复杂度主要来源于下面几个方面。

1.1 重复序列

复杂基因组组装一直是一个难题, 很大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偏小于实际的基因组大小。重复序列是基因组中重复出现的序列。重复序列在各物种中的比例从病毒(小于 1%)、细菌(啤酒酵母: 3.4%)到真核生物(人: 47%; 玉米: 77%)逐步升高。在对 44 种植物和 68 种脊椎动物分析其全基因组重复水平和基因组大小关系时发现, 重复序列与脊椎动物的基因组大小关联性更高, 植物基因组的重复序列通常比脊椎动物的更高(图 1)^[1]。

根据重复序列结构、位置及功能方面的差别, 可分为散在重复序列(interspersed repeat)、串联重复

序列(tandem repeat)和片段重复序列(segmental duplication)。散在重复序列比较均匀地分布在基因组中, 包含长散在重复(long interspersed nuclear elements, LINE)、短散在重复(short interspersed nuclear elements, SINE)、类反转录病毒转座子(long terminal repeat-retrotransposon, LTR-RT)和 DNA 转座子(DNA transposon)。其中, LTR-RT 是植物中分布最为广泛的一类转座子, 是基因组重复区域的主要成分, 如橡胶(*Hevea brasiliensis*)基因组中 71.2% 为重复序列, 其中 LTR-RT 占主要部分, 它们的大规模复制插入是橡胶基因组明显大于其他近缘物种如木薯(*Manihot esculenta*)、杨树(*Populus trichocarpa*)、蓖麻(*Ricinus communis*)等的主要原因^[2]; 墨西哥蝶螈(*Ambystoma mexicanum*)基因组的重复序列为 65.6%, LTR-RT 是主要成分, 且几乎都分布在 Contig 序列的末端, 给组装带来巨大挑战^[3]。

串联重复是由 1~500 个碱基的重复单元构成, 一般在基因组中重复几十到几百万次, 包含简单重复(simple sequence repeat)和卫星 DNA。如人类基因组的着丝粒周边区域以及染色体近端短臂含有卫星 DNA 和串联重复序列; 一些遗传调控区域序列如核小体结合单元、甲基化位点等都与串联重复有关; 产生茶叶风味的次生代谢产物合成酶基因在基因组上发生拷贝数扩增主要是由串联重复产生^[4]。

1.2 杂合度

杂合度对基因组组装产生很大影响, 以二倍体基因组为例, 通常只组装出一套染色体, 对于杂合度高的区域, 会将两条染色单体都组装出来, 从而造成组装的基因组偏大于实际的基因组大小。对秀丽线虫(*Caenorhabditis elegans*)模拟不同杂合度的数据进行组装, 当杂合度升高时, 各组装软件的 Contig N50 指标都明显下降(图 2)^[5]。相比于动物基因组而言, 植物基因组更加复杂, 很多植物因远源杂交、自交不亲和等因素, 具有基因组杂合高、倍性高等特征, 加上基因组本身比较大, 这些都增加了基因组组装的难度。如自交不亲和的茶树(*Camellia sinensis*)基因组由于种间频繁杂交导致杂合度高达 2.8%^[4]; 异源多倍体的陆地棉(*Gossypium hirsutum*)^[6]、油菜(*Brassica napus*)^[7]等物种需要借助二代和三代测序结合进行组装。

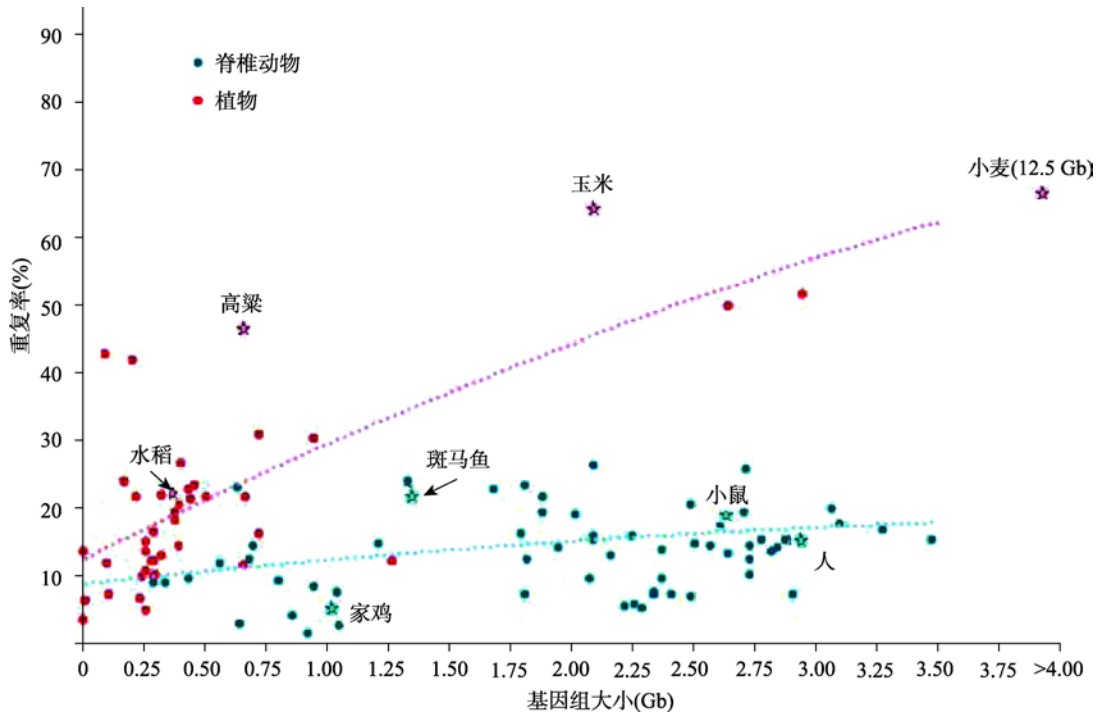


图 1 基因组大小和重复序列的相关性

Fig. 1 The correlation between repetitive sequences and genome size

根据文献[1]修改绘制。

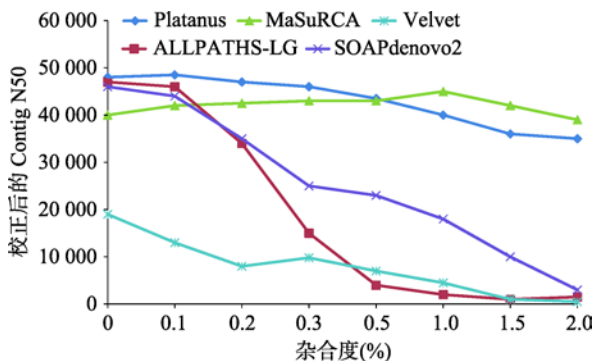


图 2 杂合度对不同算法组装指标的影响

Fig. 2 The impact of heterozygosity on genome assembly using different algorithms

根据文献[5]修改绘制。

1.3 极端 GC 含量

尽管二代测序(next-generation sequencing, NGS)具有很大优势,但是极端碱基组成一直是造成 NGS 数据组装具有挑战性的因素之一。在 PCR 扩增、桥式簇扩增以及测序等 NGS 数据产生过程中(主要是 Illumina 测序平台),由于 GC 偏好性使得基因组中低 GC 或高 GC 区域的测序读长(Reads)覆盖度不均

一。因为 Reads 覆盖度是许多组装软件的关键参数,这种极端的 GC 区域会导致基因组组装碎片化,使其完整性降低^[8]。如恶性疟原虫(*Plasmodium falciparum*)的基因组平均 GC 含量低于 25%,导致许多低 GC 区域的 Reads 覆盖度很少甚至没有 Reads 覆盖^[9];脐形紫菜(*Porphyra umbilicalis*)基因组中 GC 含量高达 65.8%,利用二代测序几乎无法进行组装^[10]。

1.4 基因组污染

基因组中存在污染也是造成其复杂性的一个因素。基因组污染一方面可能来源于 DNA 提取或扩增过程,如 DNA 提取试剂盒、化学试剂和实验室环境中的杂菌很容易造成污染;另一方面可能来源于物种间相互作用/共生/共栖的生活环境。对于藻类来说,共生微生物较多,如紫菜基因组,即使经过抗生素处理,提取出的 DNA 中依然有 50%的测序数据来自共生微生物和其他真核生物污染^[10]。最近,水熊虫(*Hypsibius dujardini*)基因组就因为污染序列事件成为各方争论的热点,最终 Koutsovoulos 等^[11]根据测序 Reads 的覆盖度和 GC 含量不均一性证明了

Boothby 等^[12]发表的版本组装结果中存在大量的 (30%) 来自细菌的污染序列。

2 高复杂基因组测序技术解决途径

2.1 针对高复杂基因组测序的实验方法

复杂基因组测序领域一直是基因组学的重要关注点。在第二代高通量测序之前, 测序费用高且通量低, 对获取高重复序列物种的全基因组序列难度太大或费用太高的物种, 科研人员主要尝试提高“有用”基因组区域的比例, 如利用杂交退火法提高重复序列较少的基因区域的相对丰度^[13], 采用甲基化碱基致死大肠杆菌突变体提高甲基化程度较低基因区及其邻近调控区在质粒克隆中的比例^[14], 或通过外显子芯片杂交获取近缘物种基因区域^[15]。

实验手段“简化”复杂基因组是目前一个重要的研究方向, 成功应用案例包括构建单倍体品系^[16]、染色体分离^[17, 18]以及低识别位点限制性内切酶完全酶切^[19]等。但以上方法都有其技术局限性: 并不是每个物种都可以获得稳定的单倍体品系; 染色体分离技术要求染色体完整且各条染色体具有不同长度或标记信息, 目标样品较少, 而微量样品扩增可能引入偏差; 酶切法无法有效分离不同来源的同源染色体。

2.2 现有高通量测序技术的特点

目前, 使用最多的高通量测序手段分为两种主要类型: (1) 短读长高通量测序, 主要包括 Illumina 的 HiSeq 和 10X Genomics 测序平台; (2) 长读长单分子测序, 主要包括 PacBio 的 SMRT 平台和 Oxford Nanopore Technologies 的 MinION 平台。此外, 还包括一系列辅助分子标记测序系统, 如 BioNano 的 Saphyr 光学酶切图谱系统等, 可用于辅助复杂基因组的组装, 并对组装的准确性进行评估。

从头测序组装(*de novo assembly*)复杂基因组的关键之一是获得较长的读长序列, 以降低重复序列或高相似基因组片段对基因组组装的影响。在第二代高通量测序仪问世之初, Illumina/Solexa 读长不足 35 bp, 基本难以实现对基因组进行有效的组装, 对于复杂基因组更是“束手无策”。为了克服这一困难,

当时发展起来的酶切“延伸”^[20]和局部组装^[21]技术对基因组的组装具有明显的促进作用。经过多年发展, Illumina 的 HiSeq 平台目前的常见读长为 2×100 bp (Illumina HiSeq 2000) 至 2×250 bp (Illumina MiSeq), 虽然较早期有很大提升, 但仍然难以满足复杂基因组组装的需求。尽管 Illumina 测序平台读长较短, 但与长读长技术相比, 其序列的准确性具有明显的优势。因此, 目前针对复杂基因组组装, 需要依靠长读长数据与二代 Illumina 短读长数据相结合的方法来实现, 以充分利用两者之间的互补优势。

Pacific Biosciences (PacBio) 公司于 2010 年发布的基于单分子实时技术(single molecule real time, SMRT)的测序仪 PacBio RS, 目前已经更新到 RSII 和 Sequel 版本。PacBio 测序反应是在 SMRT Cell 反应管中进行, 每个测序芯片(Cell)都有一个厚度为 100 nm 的金属小芯片, 其上面固定着大约 15 万个零模波导孔(zero-mode wave guide, ZMW)。ZMW 是测序技术的核心^[22]。DNA 聚合酶以共价结合的方式锚定在 ZMW 底部, 用来结合单链 DNA 分子模板。PacBio 测序得到的序列是真实的单分子 DNA 序列, 且其读长较长, 典型情况下可达到平均 20~40 kb, 与 Illumina 的最长 250 bp 相比具有明显的优势, 但测序的准确度相对较低, 平均准确度约为 80% 左右。PacBio 产生的数据更适用于复杂基因组的组装, 但需要先进行复杂的校正工作, 才能达到组装要求。目前, PacBio 是用于复杂基因组组装的主流方法。

Oxford Nanopore Technologies Limited 公司在 2012 年推出第一款基于纳米孔测序技术的测序仪。目前测序平台包括 MinION、GridION X5、PromethION 和 SmidgION。其中 SmidgION 是迄今为止体积最小的测序设备, 可在任何地点与智能手机配套使用。纳米孔测序技术的测序原理是: 在纳米孔两边加上一定的电压, 在电势的作用下, DNA 电泳通过纳米孔, 由于 4 种核苷酸的电离水平和空间结构不同, 通过纳米孔时电流强度不同, 根据电流强度准确判断碱基种类^[23]。纳米孔测序的读长可达数百 kb, 在解决复杂基因组组装时, 与 PacBio 相比具有更大的优势。

10X Genomics 平台本质上是一种改进的二代 Illumina 测序技术, 其核心是一种条码标记(barcoding)

技术, 根据 Barcode 信息组装短 Reads 从而获得跨度为几十 kb 到几百 kb 的连锁读长(linked reads), 进而将基因组组装划分成数万乃至数百万个局部组装, 再将局部组装进一步组装到全基因组。该技术可显著降低复杂度, 获得更完整的组装结果, 因此也十分适用于复杂基因组的组装。

对于植物等复杂物种基因组的组装项目, 现有的二代和三代测序仍然难以准确跨过重复序列区域, 而光学图谱技术的出现可以有效克服这一基因组组装难题^[24, 25]。基因组光学图谱是指利用荧光标记酶切技术在全基因组水平上构建限制性内切酶酶切图谱。BioNano Genomics 公司分别在 2014 年和 2017 年推出了 Irys 分析平台和 Saphyr 分析平台。该平台利用限制性内切酶对 DNA 分子进行酶切, 并利用 DNA 聚合酶和不同荧光标记的核苷酸合成带有荧光标记的核酸链; 再利用微流控装置的毛细管电泳将 DNA 分子线性化; 当 DNA 分子通过纳米孔的时候进行高分辨率荧光成像, 从而生成酶切图谱。利用 BioNano 技术和三代 PacBio/Nanopore 相结合, 可有效进行基因组从头测序组装, 解决复杂基因组的组装难题。

Hi-C (high-throughput chromosome conformation capture) 测序是一种以生物细胞核(动物/植物)为研究对象, 研究染色质之间相互作用的技术。该技术可有效进行染色体构象捕获, 从而获得基因组序列信息及其在基因组中的位置信息^[26]。其处理过程: 首先利用染色质与甲醛等交联剂进行交联反应; 再利用 *Hind*、*Mbo* 等限制性内切酶进行酶切反应而获得粘性末端, 并加入生物素标记; 最后进行解交联反应, 利用带有标记的产物进行建库测序。由于 Hi-C 数据可以准确区分细胞核中的不同染色体, 因此对于基因组组装来说, 该技术和三代测序技术结合可以高效进行 Scaffold 乃至染色体级别基因组的构建^[27]。

2.3 复杂基因组组装难点与解决方案

高重复和高杂合对于基因组组装的影响, 在组装结果中表现为两个相反的特性: 由于在组装过程中会将相似的重复序列组装到一起, 因此重复序列会导致基因组组装大小的收缩(小于预估的实际基因组大小); 对于高杂合来说, 染色体组的杂合序列

之间存在一定的序列差异, 因此在组装的时候会被分别独立组装, 从而导致基因组组装大小的扩张(大于预估的实际基因组大小)。因此, 对于具有高重复和高杂合成分的基因组来说, 对其进行正确组装具有较大的挑战性。

目前常用于基因组组装的两种算法 DBG (*De Bruijn Graph*)^[28]和 OLC (*overlap-layout-consensus*)^[29], 虽然在原理和速度上具有较大的差别, 但其本质都是寻找特定序列的最佳连续匹配, 因此在处理高重复和高杂合时, 都存在上述的弱点。相对而言, 由于 DBG 算法是通过 K-mer 的精确匹配进行组装, 可以区分细微的序列差别, 因此在一定程度上可以区分不同的重复序列; 但对于本身具有较大差异的杂合区段来说, DBG 会将其组装成独立的序列, 因此 DBG 对于高重复组装具有相对的优势, 但对于高杂合表现则不佳。而 OLC 算法在寻找最佳比对时, 允许一定的错配, 因此在一定程度上可将杂合区段合并组装, 但对于重复序列来说, 由于大部分重复序列上的差别小于 OLC 允许的错配, 重复序列可能被错误合并组装在一起, 因此 OLC 算法对于高杂合具有相对的优势。如果将两种算法适当地结合在一起, 则可以在一定程度上解决由于高重复和高杂合引起的组装难题。

无论使用哪一种组装方法, 高重复与高杂合在本质上是无法被完美解决的, 只存在解决这两种问题的相对方法。如前所述, 基因组的组装是通过寻找测序数据之间的最佳比对来实现的, 但重复序列是高度相似的, 杂合区段是存在差异的, 因此总会发生将序列错误合并组装, 将本应合并的序列错误分离。为了尽可能降低发生错误的可能性, 就需要在组装时寻找特异性的最佳比对(unique alignment)。由于比对结果的可信度(得分)与比对的长度成正比, 越长的序列, 得到的比对越长, 得到最佳比对的可能性也越高, 因此就要求用于组装的测序数据尽可能长。对于高重复基因组组装来说, 最理想的情况就是测序数据将高度重复序列完全包含在读长中间, 即形成特异-重复-特异(unique-repeat-unique)序列结构, 方能保证重复序列被放置到正确的位置^[30]; 对于高杂合基因组来说情况类似, 最理想的情况就是将杂合区段完全包含在读长中间, 形成特异-杂合-

特异(unique-hetero-unique)序列结构,才能将杂合区段正确识别出来,避免杂合区段被重复组装的问题^[5]。

然而,在实际的测序和组装过程中,最理想的情况是不易获得的。在测序平台方面,目前用于基因组组装的测序平台主要有以 Illumina HiSeq 为代表的二代平台和以 PacBio 为代表的三代平台。这两种平台具有各自不同的特点,前者测序的精度较高(错误率<1%),但测序较短(100~300 bp),而后者测序片段很长(平均可达 20 kb),但错误率很高(一般>15%)。由于基因组中 STR 序列和 LTR 序列的存在,二代测序平台得到的数据无法将重复序列和高杂合区域跨过,导致组装时出现大量的分支(branch)和环(loop),使组装结果碎片化;而三代测序平台得到的数据虽然有助于跨过这些区域,但因其错误率较高,不仅会引入组装错误,同时在一些极端情况下,由于无法和其他序列正确比对,导致组装无法进行(事实上等同于人为引入了杂合)。此外,尽管长片段测序能够帮助解决一部分高重复和高杂合组装的问题,但并不是全部。根据基因组本身特性的差异,即使在 PacBio 平台上得到长度分布完全一样的测序数据,能够跨过重复或者杂合区域的比例也仅在 30%~60%之间浮动。因此,根据基因组本身的特征对测序策略进行优化是十分必要的过程。

针对于不同测序数据的特点,目前产生了多种不同的组装方法,以利用测序数据的特点来尽可能降低高重复和高杂合对于基因组组装的影响,得到尽可能连续的基因组序列。

在产生较为有效的三代数据组装的算法之前,已经有研究对于仅使用二代数据进行高重复和高杂合基因组组装进行了尝试,结果比传统方法具有十分显著的提升。以 Platanus 为例,其组装策略分为 Contig 生成、Scaffold 构建和空洞填补(Gapfill) 3 个部分(图 3)^[5]。该软件主要基于 DBG 算法,并根据 DBG 算法的特点,重复与杂合成分会在 DBG 中形成“接合”(Junction)与“鼓泡”(Bubble),而测序的错误会形成“断头”(Tip)。进行组装优化的目的就是 will Junction、Bubble 和 Tip 尽可能去除,形成线性图(straight)。在 Contig 构建时,Platanus 采取了 3 种创新的策略:(1) 通过拟合泊松分布,将低频“拐点”之下的 K-mer 全部去除,从而有效减少了 Tips 的数

量;(2) 将 Reads 直接比对定位到 Junction 节点,通过定位的质量确定 Junction 的走向,而不是使用 K-mer 深度,可以解决连续 Junction 的组合问题;(3) 使用多 K-mer 延伸策略,从较小的 K-mer 向较大的 K-mer 进行延伸,既可以在初始构建 DBG 时避免杂合造成较低的 K-mer 深度,又可以有效利用测序数据的长度。通过这一策略,杂合形成的 Bubble 被有效地鉴定并被分离出来(注意不是去除)。而在 Scaffold 阶段,Platanus 将大片段 Mate-pair 文库统一定位到 Contig 和去掉的 Bubble 上,将带有杂合的序列作为整体考虑,从而有效利用构建形成 Scaffold 时所包含的 Contig 两端的连接数。在后续过程中,Platanus 同样会识别 Scaffold 图中可能存在的杂合,将覆盖度较低且内部不含 Bubble 的 Scaffold 分支识别为杂合从而去除,带有 Bubble 的进行保留(默认二倍体基因组中不应存在多重杂合)。最后,在空洞填补阶段,将 Reads 重新定位到组装成的 Scaffold 上,将在空洞(Gap)附近的 Reads 筛选出来,进行局部重新组装,将空洞进行填补。综上所述,Platanus 进行高杂合基因组的核心策略是将含有杂合的序列鉴定出来并进行合并去除,相对于传统的方法,可以有效避免在分支处将序列切断,因此可以得到更加连续的组装结果。

与 Platanus 有所不同,另一个组装软件 ALLPATHS-LG^[31]则更适用于处理高重复基因组。与一般的组装程序不同,ALLPATHS-LG 要求在 DNA 文库构建时同时构建小片段(Fragment)文库和长跨度(Jumping)文库。其中,小片段文库在构建时,要求插入片段长度小于 Reads 读长两倍的文库,例如插入长度为 180 bp 的 2×100 bp Pair-end 文库。在组装时,ALLPATHS-LG 会将小片段文库两端的重叠区域结合起来,形成平均长度近似于插入长度的接合片段(end-overlap fragments)。然后,ALLPATHS-LG 使用大小 K-mer 组合的策略,使用较大 K-mer 将结合片段组装成不含任何分支的独立路径(unipath)片段,通过长跨度文库,将路径片段连接成组装图(assembly graph),并通过覆盖度等信息将包含重复序列的分支“扁平化”(flatten),从而形成包含尽可能少的分支或者环路的线性组装序列。ALLPATHS-LG 的组装方法具有以下优势:(1) 基于 180 bp 的 3'末

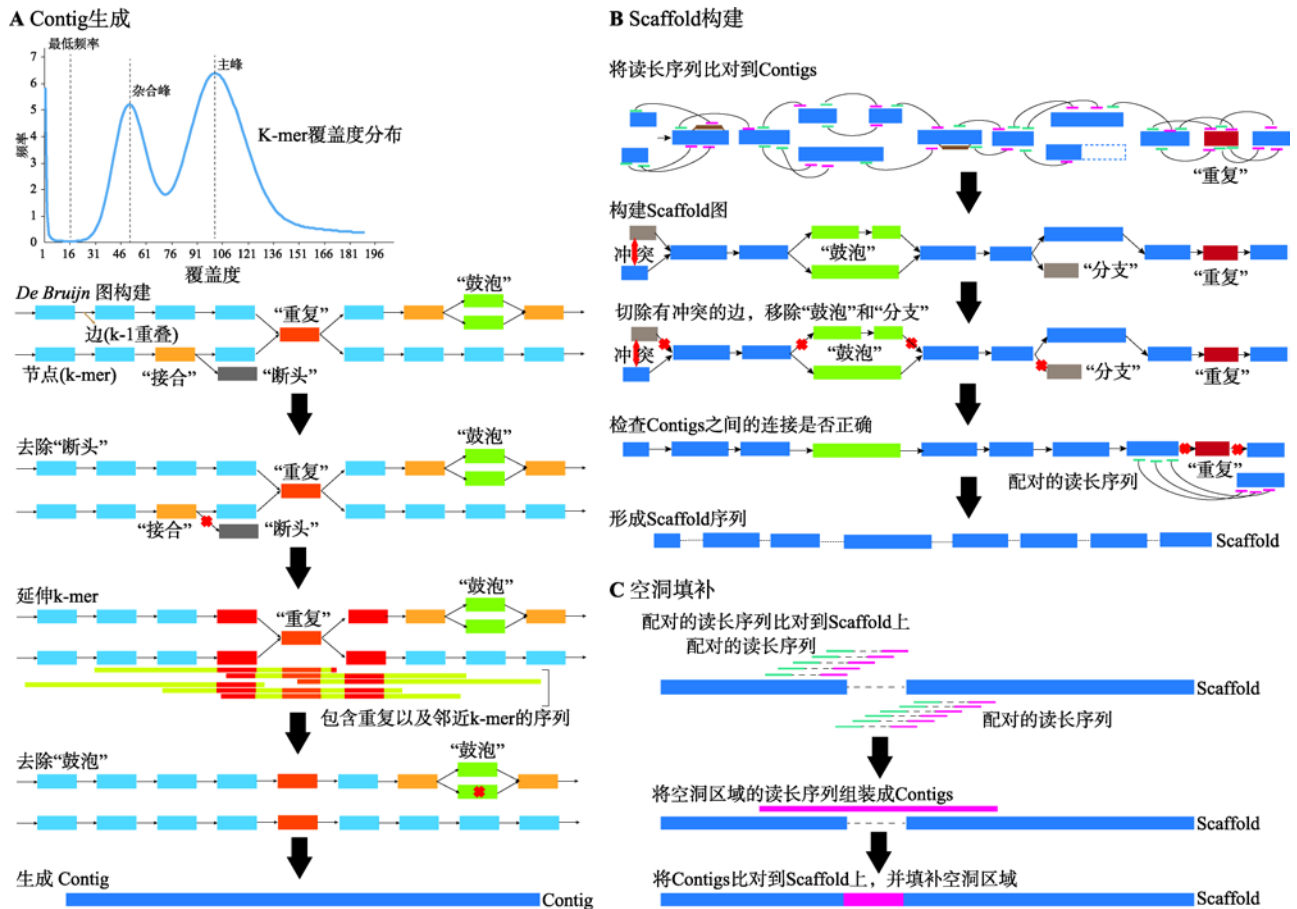


图 3 Platanus 组装策略

Fig. 3 The assembling workflow of Platanus

A: Contig 生成阶段: 首先由原始数据生成 k-mer 分布并拟合泊松分布, 确定最低频率、杂合峰和主峰, 并去除最低频率下的所有 k-mer。然后, 依次通过 De bruijn 图构建、去除“断头”、延伸 k-mer 和去除“鼓泡”4 个步骤, 生成尽可能连续的 Contig 序列。B: Scaffold 构建阶段: 将原始数据比对到生成的 Contig 序列上, 通过配对关系, 首先构建 Scaffold 图, 然后通过切除有冲突的边, 移除“鼓泡”和“分支”, 生成尽可能连续的 Scaffold 序列。C: 空洞填补阶段: 通过比对获取空洞内及其临近的配对序列, 对这些序列进行局部组装, 再通过比对将空洞进行填补。根据文献[5]修改绘制。

端对接片段, 使用较大的 K-mer (默认为 K=96) 进行初始组装, 可以有效地避免重复序列区域产生过多的分支; (2) 在组装前, 使用 24-mer 进行测序数据的矫正, 可以有效降低由于测序错误或者低频 SNP 造成的复杂度; (3) 在处理长跨度文库(包括 Mate-pair 文库和 BAC-end 文库)时, 首先进行嵌合和非环化 Reads 的检查, 消除构建 Scaffold 时的负面影响; (4) 对于由 PCR Bias 或者极端 GC 造成的低覆盖度 ($<10\times$) 的基因组区域, 由于低覆盖度区域的 Reads 之间的重叠长度可能非常短, 因此 ALLPATHS-LG 尝试采用极低的 K-mer (默认为 K=15) 对初始组装中的空洞(可能来自极低覆盖度区域)进行填补。得益于

这些特性, ALLPATHS-LG 可以在仅使用 Illumina 测序数据的情况下, 获得连续性较好的组装结果。以人和小鼠的数据为例, 其 N50 分别可以达到 11.5 Mb 和 7.2 Mb。

在以 PacBio 为代表的三代测序技术逐渐兴起之后, 由于三代测序数据的读长更长, 因此能够有效地解决高重复和高杂合的问题。但受限于三代数据的通量和成本, 在早期, 利用三代数据的方法主要为三代和二代混合组装, 因此也产生了众多的二代、三代数据混合组装的算法, 包括 PBcR^[32]、SPAdes^[33]、DBG2OLC^[34]和 MaSuRCA^[35]等。其中, MaSuRCA 对于高杂合和高重复基因组的混合组装表现较好。

MaSuRCA 是基于 OLC 算法诞生的组装程序, 其最早主要利用二代 Illumina 数据进行组装, 并比较成功地组装了基因组大小约 22 Gb 的火炬松(*Pinus taeda*)为代表的超大植物基因组^[36]。MaSuRCA 在 3.2.0 版本之后开始支持二、三代数据混合组装, 并已经完成了粗山羊草(*Aegilops tauschii*)基因组(古小麦 D 基因组, 4.25 Gb)的组装^[37]。与使用纯二代数据的 SOAPdenovo^[38]和 DenovoMagic 方法相比, MaSuRCA 可以将 N50 提升 30 倍以上(N50 分别为 2.1 kb、16.4 kb 和 486.8 kb)^[35]。MaSuRCA 的混合组装策略如图 4 所示。首先, 对二代 Illumina 数据进行质控, 去掉低质量和接头序列, 并使用 Quorum 算法^[39]进行错误矫正; 然后, 对矫正后的 Reads 进行预组装, 形成称为超级读长(Super-reads)的长片段; 将 Super-reads 比对到三代 PacBio 数据上, 根据比对的顺序和重叠, 将 Super-reads 进一步融合形成更长的巨型读长(Mega-reads); 最后, 使用 Celera Assembler

中的 OLC 算法, 将 Mega-reads 组装成基因组序列。与只使用二代数据进行组装的算法相比, 得益于 PacBio 数据的引导作用, MaSuRCA 可以更好地组装包含重复序列的基因组片段, 同时得益于 OLC 算法允许一定 SNP 差异的特性, MaSuRCA 对于高杂合基因组也有较好的适用性。为了进一步降低高杂合成分对于组装的影响, 在组装最后阶段, MaSuRCA 会使用 MUMMer^[40]将基因组中较长的冗余片段合并去除, 从而有效降低高杂合造成的冗余组装问题。但同时也需要注意, 由于 MaSuRCA 的组装策略是使用三代长序列作为“骨架”引导二代数据组装, 三代数据并不构成序列的主要部分, 因此仍然会受限于 Illumina 测序平台对 GC 含量和扩增偏好性(PCR Bias)的敏感性, 理论上仍会存在基因组组装的覆盖度盲区。此外, 由于 Super-reads 的产生仍然依赖于较短的测序长度, 因此对于较长的重复序列, 如 LTR 等, 仍然会存在较大的困难。

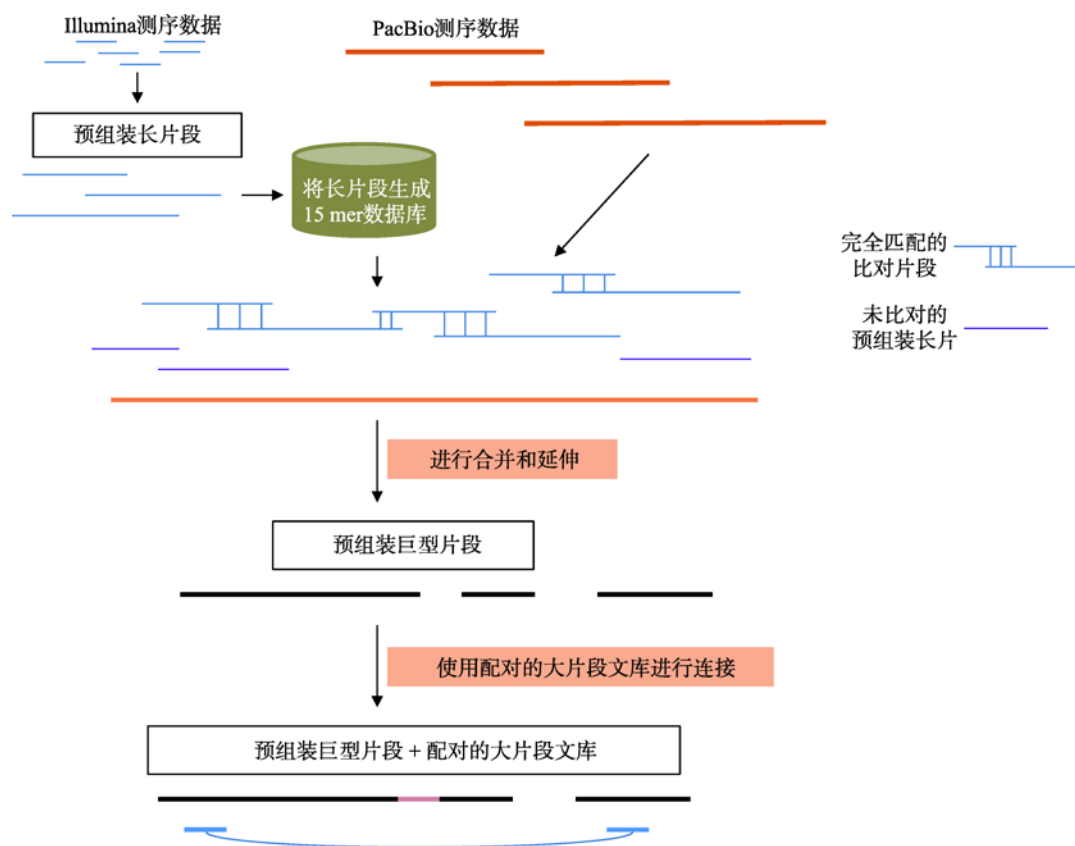


图 4 MaSuRCA 混合组装策略

Fig. 4 The hybrid assembling strategy of MaSuRCA

根据文献^[35]修改绘制。

随着三代测序技术的进步,且三代测序在偏好性方面有二代技术无法比拟的优势,新的策略开始着眼于仅使用三代数据来进行基因组的组装。除了 PacBio 官方 SMRT Pipeline 中的 PBcR 和 Falcon 之外,仅使用三代数据的组装方法还包括 Canu^[30]、miniasm^[41]、MECAT^[42]、SMARTdenovo 和 WTDGB 等。在这些软件中,Canu 具有较好的组装准确性,但是由于其基于完全的 OLC 算法,对于资源消耗十分巨大;而以 miniasm 为代表的加速算法,虽然在准确性上欠佳,但是其资源消耗小,更适用于深度测序或对超大基因组进行组装^[43]。以目前使用较多且组装准确性较高的 Canu 为例,该方法是 CABOG 组装程序和 PBcR 自我矫正程序的延伸版本,其基本组装策略同样为 OLC 算法,但其核心的优势在于整合了高效的比对算法 MHAP^[32]和矫正算法 Falcon^[44],从而在连续性和准确性方面具有较好的平衡性。如图 5 所示,Canu 的组装流程分为 3 个阶段:矫正(correction)、修剪(trim)和组装(assembly)。在矫正阶段,Canu 首先将原始数据使用 MHAP 算法进行比对,根据比对结果将 Reads 进行聚类,然后根据聚类结

果生成一致性(consensus)序列,从而对测序数据进行自我矫正。在修剪阶段,Canu 采用 CABOG 中的重叠修剪(overlap-based trim)方法,将测序数据中不产生重叠的部分切除。最后,Canu 使用矫正与修剪后的 Reads 进行基于 OLC 算法的组装,生成 Contig,从而完成组装。目前,根据 Canu 官方的测试结果,Canu 在完整程度和速度上,均优于 Falcon (纯三代数据组装)和 SPAdes (二代三代混合组装)。但是,对于高杂合的基因组,Canu 目前仍然将差异较大(超过 1.5% 以上的差异)的单倍型(Haplotypes)序列分开组装,且目前尚无较好的方法来继续分辨与处理这些序列。目前,使用 Canu 进行高杂合或多倍基因组组装,建议的解决思路有两个:(1) 尽可能保留单倍型组装,然后再去掉或合并同源区段:该方法通过设置严格的比对错误率,尽可能将杂合部分单独组装,产生多倍于预期基因组的组装,然后通过“Purge Haplotigs”方法^[45]去除杂合或者多倍部分,但有可能去除掉部分基因组重复;(2) 尽可能在组装时合并杂合或多倍部分:该方法通过设置较宽松的的错误率,将杂合或多倍部分合并到一起,但有可能造成组装错

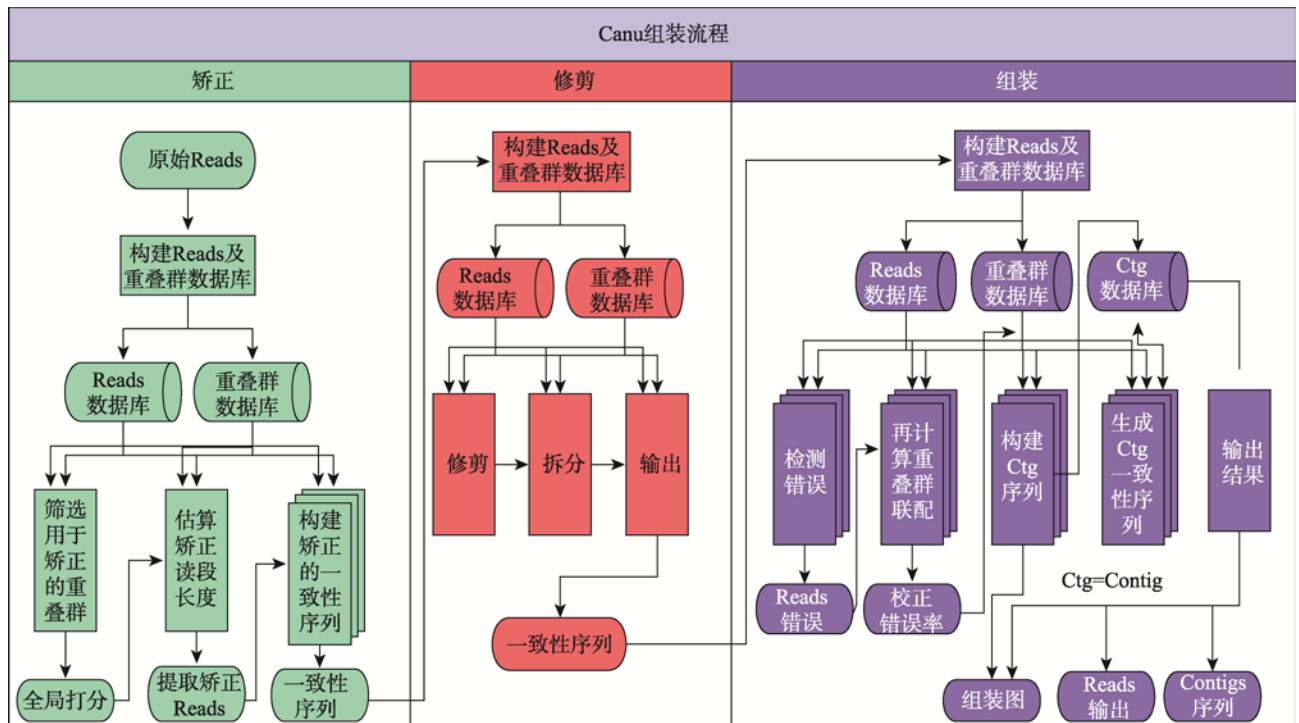


图 5 Canu 组装流程

Fig. 5 The workflow of Canu pipeline

根据文献^[30]修改绘制。

误。同时, Canu 计划在未来与 10X Genomics 或者 Hi-C 等数据结合, 综合解决该问题。

随着以 Oxford Nanopore Technology (ONT) 为代表的纳米孔测序技术的发展, 其测序的读长和通量相对于 PacBio 测序更胜一筹, 因此基于 ONT 数据完成的组装实例也日渐增多。与 PacBio 相比, ONT 在数据上的特点表现为平均读长更长, 但是错误率相对来说较高。目前, 可应用于 PacBio 数据的组装方法, 包括 Canu^[30]、miniasm^[41]和 TULIP 等, 同样适用于 ONT 数据, 但需要对参数进行调整^[46]。以 Canu 为例, 根据软件作者的测试, 在处理 ONT 数据时, 需要将矫正预期的错误率(corrected error rate)适当调高, 比如在组装 Nanopore R7 2D 和 R9 1D 数据时, 建议将预期错误率从 8.5% 调整到 14.4%, 以适应 ONT 数据的特点。

正如前面所述, 解决复杂基因组组装难题的关键, 在于获得尽可能长的测序数据。但受限于现有的 DNA 实验技术与测序技术, 可获得的测序长度是有限的。即便是宣称最长读长可超过 1 Mb 的 ONT 技术, 在实际的复杂基因组组装中, 其一般建库测序方法得到的文库平均长度与 PacBio 相比并无显著优势^[43]。因此, 目前尚不能从根本上解决复杂基因组的精准组装问题。但近年来, 一些特殊的技术如 BioNano、10X Genomics 和 Hi-C, 虽然无法直接获得完整的 DNA 长片段, 但可以在现有测序技术的基础上获得更大的跨度, 在一定程度上解决了复杂基因组组装的难题, 为复杂基因组结构分析提供帮助。

BioNano(以 Irys 版本为例)是利用光学标记构建基因组酶切图谱的技术。在进行复杂基因组组装时, BioNano 技术的作用体现在两方面: (1) 由于 BioNano 读取的分子长度最高可以达到 1 Mb 以上, 平均长度也可以达到 200 kb, 因此在构建 Scaffold 时可以有效跨过较长的重复序列区域, 获得较为连续的 Scaffold 序列; (2) BioNano 可以用于鉴定单倍型基因组上的酶切位点差异, 从而将单倍型基因组序列分离开, 因此有助于将高杂合基因组中的杂合区域完全分离组装。同时, BioNano 读取的光学酶切图谱可以有效纠正组装中的结构错误。对于复杂基因组组装来说, 使用 BioNano 官方的 BioNano Sovle 软件进行辅助组装的基本流程包括: (1) 生成基因组

草图的 BioNano 酶切图谱; (2) 将特定酶切的 BioNano Molecule 进行自身一致性组装, 生成光学图谱(optical map); (3) 将光学图谱和基因组序列的酶切图谱进行共线性比对, 以光学图谱为骨架生成 Scaffold 序列; (4) 如果使用双酶切文库, 可进一步使用 TGH Pipeline 生成双酶切一致的 Scaffold, 实现最佳组装效果。

10X Genomics 提供的技术(以 Chromium 为例)类似于早期的 BAC 混合池(BAC pooling)技术, 该技术通过油包水微体系, 为 DNA 片段加入条码标记(barcode)并测序, 之后可通过条码标记将原本属于同一 DNA 片段的测序数据进行聚类。10X Genomics 技术可以有效降低组装的复杂度, 通过局部的 Contig 组装和不同条码标记聚类之间的重叠, 可以有效地构建出连续的基因组组装。与 BioNano 技术类似, 10X Genomics 技术对复杂基因组组装的作用也体现在两个方面: (1) 目前其条码标记序列的平均长度可达到 50 kb 左右, 理论上可以跨过大部分的转座子重复, 且限制在局部片段上, 其复杂度相对于全基因组来说要低很多, 因此该技术可以有效减少由于重复序列造成的组装碎片化问题; (2) 由于 10X Genomics 技术可以在 50 kb 的分辨率上对单倍型差异通过条码标记进行分离, 因此同样可以对高杂合基因组中的每一个单倍型基因组进行独立组装, 在一定程度上解决杂合造成的组装难题。目前, 已经有多个利用 10X Genomics 技术和其官方提供的 Supernova 组装方法完成的基因组组装实例, 其 Scaffold N50 基本都可以达到 Mb 级别^[47, 48]。

Hi-C 技术是通过固定染色体在三维空间上的交联, 对染色体的空间结构进行概率性的推断。由于相邻序列之间发生交联的概率要大于相距较远的序列, 因此, 同一染色体内的空间交联要高于不同染色体之间的交联, 因此利用 Hi-C 技术, 可以在染色体级别的长跨度上对染色体进行组装和分离。同时, 由于单倍型基因组分属于不同的染色体组, 因此利用 Hi-C 技术, 同样可以实现对异源二倍体基因组中的单倍型基因组进行分离。对于多倍体超大基因组(如六倍体小麦基因组), Hi-C 同样可以将多倍体基因组中不同来源的基因组成分的分离, 从而在一定程度上解决多倍体超大基因组的组装难题。相对于早期利用遗传图谱进行染色体构建的方法, Hi-C 技术

优势主要表现在：(1) 不需要构建大量的 F_1 群体，只需个体即可完成；(2) 不需要对亲本进行纯化，就可以对单倍型基因组进行分离，因此适用于不易纯化的高杂合基因组组装。以具有代表性的使用 Juicer 方法组装的埃及伊蚊(*Aedes aegypti*)基因组^[49]为例，使用 Juicer 进行辅助组装的步骤包括：生成与 Hi-C 文库对应的全基因组电子酶切图谱；将 Hi-C 文库比对定位到基因组上，筛选酶切位点附近的测序数据，并将完全重复的位点去除；(3) 构建基于 Hi-C 文库的全基因组交联图(contact map)；(4) 根据交联图中 Contig 之间的频率，将属于同一个拓扑结构域(TAD)的序列进行聚类，并对 Contig 进行重排；(5) 构建出 Hi-C Scaffolds，并通过多次迭代纠正组装的错误，将无法正确放置的序列碎片剔除。

综上所述，利用 BioNano、10X Genomics 或 Hi-C 技术可以有效弥补现有测序技术在测序长度上的不足，辅助构建出更加连续完整的基因组序列。同时，从 3 种技术对于杂合处理的趋势来看，进行多倍型基因组完整组装是复杂基因组组装未来发展的趋势。

3 复杂基因组测序经典案例

3.1 梨基因组

梨(*Pyrus bretschneideri*)由于其自交不亲和、排斥近交等特点使其基因组杂合度达到 1%，是高度杂合的物种。因此，不适合采用全基因组鸟枪法(whole genome shotgun, WGS)测序组装策略。研究人员在梨基因组测序中采用新一代 Illumina 测序平台并结合 BAC-to-BAC 策略，其中 BAC-to-BAC 策略是成功的关键^[50]。该方法首先构建插入片段长度为 80~180 kb 的 BAC 文库，文库构建按照 Agilent 公司标准流程，在文库构建过程中加入了不同接头，从而实现将 2208 个样品混合进行 1 个测序泳道(lane)的测序，最终得到 38 304 个 BAC 的测序结果，相当于 10 倍基因组的长度。同时，进行了全基因组的 Pair-end 测序(插入片段长度 180~800 bp)和 Mate-pair 测序(插入片段长度大于 2000 bp)。在组装方面，首先利用 SOAPdenovo 软件和 Pair-end 的数据进行单个 BAC 的测序组装，然后利用 SSPACE 软件^[51]和全基因组

的 Mate-pair 数据进行 Scaffolds 构建。利用几轮的相互叠加的测序验证将单个 BAC 的组装结果进行混合组装，同时将 Scaffold 末端 3 kb 的序列和 Scaffold 进行比对，相同序列进行组装，同时去除冗余序列，最终得到可信的 Scaffolds，进一步利用 SSPACE 软件、Mate-pair 的全基因组测序数据和 Pair-end 的测序数据将 Scaffold 组装成 Super-scaffold。最后利用 NUCmer^[52]和 BLAST^[53]将 BAC 的组装结果和 Scaffold 进行组装。梨基因组研究在没有物理图谱辅助的情况下，完成了高杂合、高重复序列的二倍体果树基因组组装，积累的组装经验对于其他高度杂合的基因组研究具有很好的借鉴价值。

3.2 太平洋牡蛎基因组

太平洋牡蛎(*Crassostrea gigas*)与大多数海洋无脊椎动物类似，具有高度杂合的基因组，其杂合度高达 2.3%，基于常规的策略，几乎不可能对其进行从头组装。研究人员采用 Fosmid 克隆混合池(Fosmid pooling)、辅助短序列以及逐级组装策略，比较成功地解决了牡蛎基因组高杂合造成的组装难题^[54]。首先将 145 170 个 Fosmid 克隆混合成 1613 个混合池进行测序，每个混合池测序数据量达到 60×覆盖度。将得到的序列按照每个混合池单独组装，得到 Contig 序列，这些序列通过 Contig 之间的重叠序列进行混合组装得到 Super-contig。利用 LASTZ 和测序程度信息对 Super-contig 进行自我全基因组比对，移除组装过程中产生的冗余序列，将得到的序列进一步用 Pair-end 数据矫正组装得到最终的 Scaffold 序列。最后利用短序列对 Scaffold 序列进行矫正，最终得到的牡蛎基因组大约为 559 Mb，总共约有 28 000 个基因。牡蛎基因组的这种 Fosmid 混合池结合短序列的测序和组装方法，为研究人员更好地破译具有高度杂合性和高度多态性的基因组开辟了一条新途径。

3.3 硬橡胶树杜仲基因组

杜仲(*Eucommia ulmoides*)基因组杂合率约为 0.9%~1%，重复序列在 66%以上，属于高杂合、高重复的复杂基因组。在杜仲基因组测序策略上，采用了全基因组鸟枪法和第二代(Illumina HiSeq 2000 和 MiSeq)、第三代测序技术(PacBio)和 BioNano 光

学图谱技术的有机结合^[55]。利用 Platanus 组装软件将高质量数据组装成 Contigs 和 Scaffolds, 然后用 SSPACE 软件, 利用 Pair-end 和 Mate-pair 数据将 Scaffolds 组装成更长的 Super-scaffolds。使用 PBJelly^[56]和 PacBio 数据完成上述 Scaffold 的补洞工作。最后利用 BioNano 数据完成 Scaffold 的定位。在组装过程中, 采用了适用于杂合物种的软件 Platanus, 同时利用 PBJelly 软件进一步通过三代数据完善组装结果。

3.4 六倍体小麦基因组

六倍体小麦(*Triticum aestivum*)基因组庞大, 是典型的异源多倍体基因组, 由 3 套相似而又不同的基因组整合形成一个极为复杂的六倍体基因组。其中 3 套不同的基因组分别是 A、B、D 基因组, 每个基因组含有 7 条染色体, 共有 21 条染色体组成, 并且每套基因组上都有一套同源性比较高的相关基因, 而且这些基因在每个同源染色体上的排列顺序也都不相同, 这就使基因组序列的组装工作变的非常复杂。由于其高倍性, 加之高度重复序列, 给小麦基因组的研究带来巨大的挑战, 对六倍体小麦的基因组研究更是延续了 13 年之久。对于这种基因组体量大、重复序列含量高的复杂基因组, 研究者采取“分而治之”的组装策略: 首先利用 NRGene 的 DeNovoMagic2 软件对小麦基因组进行组装, 然后通过 BAC 克隆文库, 对分属于 3 套不同基因组的序列进行分离。然后, 在分离组装的基础上, 结合三维基因组 Hi-C 技术和高密度遗传图谱(population sequencing, POPSEQ)对基因组序列进行染色体挂载, 重新修正了基因组, 大小为 15.4~15.8 Gb。在组装过程中, 研究者还参照已经发表的乌拉尔图小麦(*Triticum urartu*) (A 基因组)、粗山羊草(*Aegilops tauschii*) (D 基因组)、拟斯卑尔脱山羊草(*Aegilops speltoides*) (B 基因组)和硬粒小麦(*Triticum durum*) (B 基因组)基因组。最终, 组装得到 1601 条 Scaffold 序列, 总大小为 14.5 Gb, 是迄今为止完成度最高、质量值最好的小麦基因组序列。同时, 为了验证组装的准确性, 研究者还构建了 BioNano 光学图谱、辐射杂交物理图谱等, 用于检查和修正基因组 Scaffolds 中存在的组装错误^[57]。综上所述, 对于高倍体基因组, 对其中的单倍型基因组进行分离是完

成该类型复杂基因组组装的关键。

3.5 野生番茄基因组更新

野生番茄(*Solanum pennellii*)基因组大小约为 1.2 Gb, 属于较大的复杂基因组。之前已有研究报道过基于二代数据组装的野生番茄基因组, 但受限于技术水平, 其组装的 Scaffold N50 为 1.7 Mb, 组装大小为 942 Mb, 但序列中存在较多的空洞, 其 Contig N50 仅为 2.18 kb^[58]。随着 Nanopore 技术的进步, 近期有研究团队对野生番茄基因组进行了基于纯 Nanopore 测序数据的更新。Nanopore 单分子实时测序技术, 其一般建库方法的平均读长为 10 kb, 超长建库方法平均读长可达到 100 kb, 因此可以用来测序、组装 Gb 级别的高质量基因组, 尤其是可以用于高杂合、高重复等复杂基因组。在该研究中, 通过 Nanopore 测序得到了约 110.96 Gb 的有效数据, 并使用 Canu 和 SMARTdenovo 两个软件联合进行组装, 最终得到的更新版本基因组组装大小为 915 Mb, Contig 数目为 889, Contig N50 达到了 2.52 Mb, 与之前发表的 2.18 kb 相比有了质的飞跃^[59]。该研究充分表明了使用 Nanopore 进行复杂大型基因组进行组装的可行性。同时, 该研究还对各种 Nanopore 数据组装方法的效果进行了对比, 具有一定的参考价值。

4 本课题组研究领域及成果

本课题组主要从事基因组结构和功能解析, 从病毒、细菌和原生生物等简单生物到高等生物(橡胶、木薯等)皆有涉及。本课题组根据物种特点结合当时测序技术分别制定测序策略, 以获得较好的基因组组装结果。在二代测序技术发展早期, 细菌等基因组主要采用一代测序技术, 该技术具有准确度高、测序读长长等优点, 是简单基因组测序的优先选择, 如本课题组已测序完成的副血链球菌(*Streptococcus parasanguinis*)^[60]等。但是由于一代测序技术的花费高、通量低, 耗时长等缺点, 导致该技术在复杂基因组上的应用具有很大的局限性。

随着二代测序技术的应用以及相应软件的开发, 大量物种采用该技术进行测序。本课题组开展了多个物种的测序与分析工作, 包括枣椰树(*Phoenix*

dactylifera)^[61]、牛带绦虫(*Taenia saginata*)^[62]、橡胶树(*Hevea brasiliensis*)^[2]、茶薪菇(*Agrocybe chaxingu*)等。

以橡胶树基因组的测序与组装策略为例(图 6), 主要过程包括: (1) 构建不同片段插入长度的文库, 包括 Pair-end 和 Mate-pair 文库; (2) 利用二代组装软件将测序数据搭建成为 Scaffolds; (3) 使用 cDNA 进一步延伸 Scaffolds; (4) 筛选到 5912 个 BAC 克隆, 将这些 BAC 分到 143 个 384 板中, 构建 BAC Pooling 文库, 利用 Illumina 测序平台获得数据, 基于序列质量, 将其中 124 个 384 板的数据单独组装; (5) 基于 BAC 组装数据, 模拟 10 kb、20 kb、30 kb、40 kb、50 kb、70 kb 和 100 kb 插入长度片段的测序数据; (6) 将原有 Scaffolds 从空洞区域断开, 拆分成 Contigs, 然后利用模拟数据将 Contigs 重新搭建成为 Scaffolds。橡胶树基因组采用 BAC 数据模拟成不同插入长度片段文库测序数据来辅助基因组组装, 这与之前已发表的牡蛎基因组^[54]所采用 Fosmid 混合池辅助组装的策略不同。在牡蛎基因组测序与组装中, 每个 Fosmid 混合池数据单独组装并修正组装

数据; 然后, 利用 OLC 方法将所有混合池组装互连接成 Super-contigs; 最后, 利用二代 Mate-pair 大片段数据将 Super-contigs 进一步组装成 Scaffold, 并对空洞区域进行填补。由于在橡胶树基因组中 BAC 测序数据大约为 4×, 理论上只覆盖基因组 98% 的区域, 并且只有 78% 的转录本比对到 BAC 数据上, 因此本实验室采用不同的方法处理 BAC 数据, 包括 BAC-Scaffold 与 WGS-Scaffold 合并, BAC-Scaffold 比对到 WGS-Scaffold 上以延长 WGS-Scaffold 以及 BAC 数据模拟同插入长度片段文库测序数据等。综合比较这些策略, BAC 数据模拟是最优策略。该策略的优点主要体现在: (1) 降低 Mate-pair 文库的测序量; (2) 模拟产生更大的 Mate-pair 片段, 提高 Scaffold 完整性; (3) 相比 Mate-pair 文库, 通过 BAC 克隆得到的数据可避开环化失败、PCR 重复扩增导致的偏好性, 产生更多有效数据。

枣椰树基因组采用 454、SOLiD 和 3730XL 测序平台相结合的策略, 具体过程包括: (1) 利用 454 数据, 使用 Newbler 程序构建 Contigs; (2) 使用

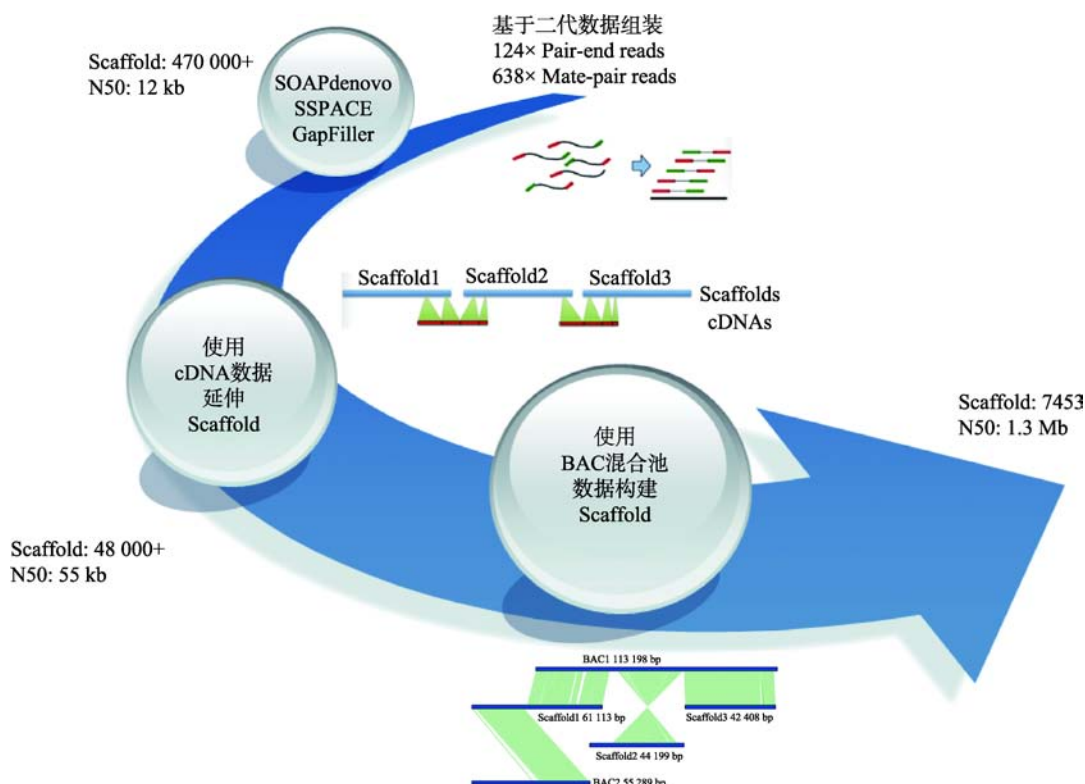


图 6 橡胶树基因组的组装流程

Fig. 6 The workflow of para rubber tree genome assembly

根据文献[2]修改绘制。

SOLiD 的不同插入片段长度的 LMP (long mate-pair) 文库搭建 Scaffolds; (3) 利用高质量测序数据填补 Scaffolds 中的空洞区域; (4) 采用 BAC 末端测序数据进一步延伸 Scaffolds。为了降低高重复引起的复杂度问题, 本实验室采用 BAC 混合池测序来辅助枣椰树基因组的组装(图 7)。

二代测序数据读长较短, 快速获得高重复高杂合基因组的高质量序列图谱仍面临着极大困难, 而

三代测序数据读长较长, 可更好地处理重复和杂合问题。本实验室已将 PacBio 等三代测序技术作为主流策略, 现已利用 PacBio 技术对水稻(*Oryza sativa* ssp. indica cv. 93-11)进行测序, 并使用 BioNano 光学图谱和 Hi-C 三维基因组技术, 对基因组进行染色体级别的 Scaffold 组装。在此基础上, 进一步循环使用 Hi-C 和 BioNano 光学图谱数据, 对组装中存在的错误进行多轮校正, 优化提升组装质量(图 8)。

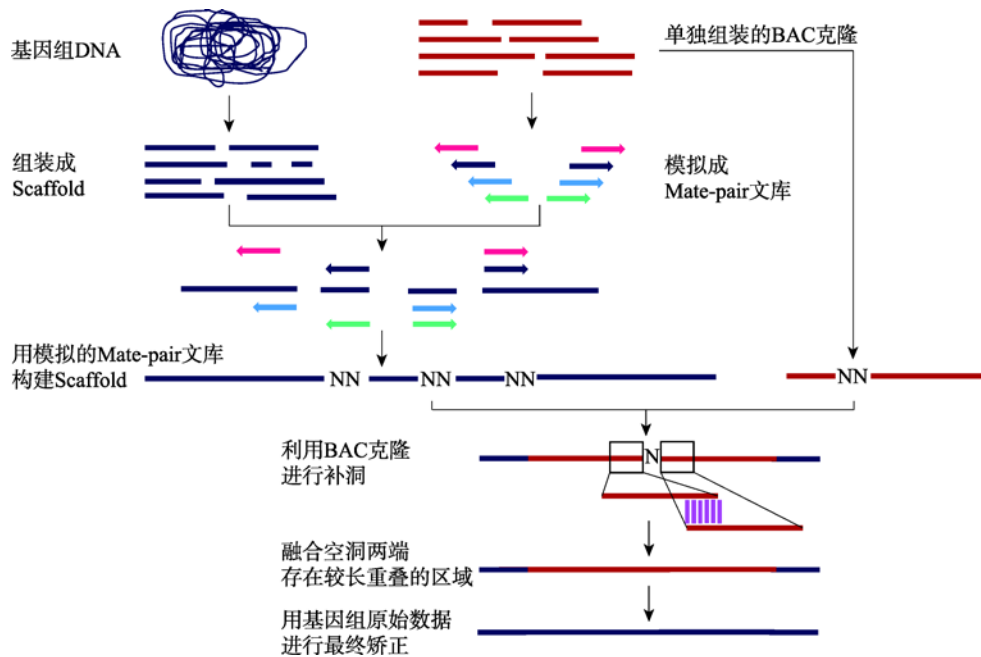


图 7 BAC 混合池辅助基因组组装流程

Fig. 7 The workflow of BAC pooling assisted genome assembly

根据文献[61]修改绘制。

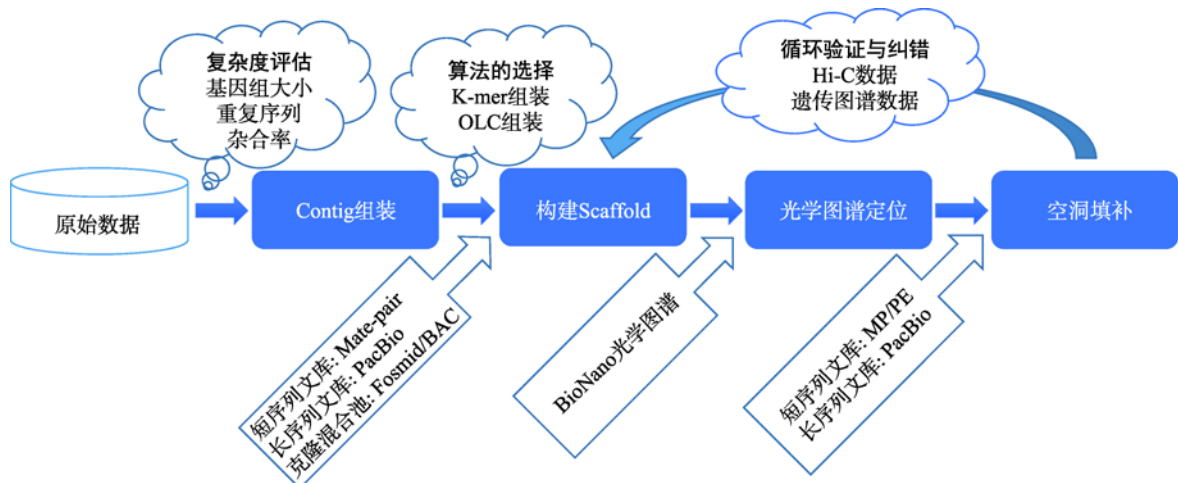


图 8 使用 Hi-C 和 BioNano 技术进行辅助优化组装流程

Fig. 8 The workflow of genome assembly improvement using Hi-C and BioNano data

综合考虑当前的测序技术特点,针对如何挑选进行基因组测序的物种或样品,我们建议应充分考虑基因组测序难度,相近物种间或品种间可能存在的异质性,优先选择容易测序的,如杂合度较低、重复比例较低、易获得高质量基因组 DNA、易获得单倍体材料的物种或品种;同时也要考虑到同一物种不同组织间也可能存在差异,优先选择方便提取高质量基因组 DNA,且外源 DNA 污染可能性小的组织。而如何进行复杂基因组测序,总原则是先短序列评估,再进行长短结合深度测序,具体可参考如下流程:(1) 首先采用二代测序数据进行基因组复杂度评估,估计基因组大小,杂合度和重复序列比例,完成 GC 含量分布图,比较公共微生物基因组数据库,评估是否存在基因组污染,可对多个近缘物种或品种平行测序评估,选择复杂度最低的物种或品种进行深度测序,并测试能否提取高质量长片段基因组 DNA;(2) 考虑到测序费用仍然较高,进行深度测序前,可以使用已完成全测序的基因组大小和重复序列比例最接近的近缘基因组序列,参考前期评估的杂合度,进行基于 10X Genomics 或 PacBio/Nanopore 测序的技术的数据模拟,评估方案效果和费用;(3) 进行高通量测序,获得二代和三代测序数据,三代序列质量矫正和组装,整合 BioNano 光学图谱,参考 Hi-C 结果和遗传图谱评估和反馈优化组装结果,使用长短序列进行组装结果矫正,最终获得高质量的高杂合基因组序列图谱。

5 结语与展望

总结近 10 年来基因组测序技术领域的发展脉络,不难看出新老技术不断融合与发展的力量:BAC 混合池测序实质是 BAC-to-BAC 测序策略与二代片段测序技术的结合,PacBio 和 Nanopore 测序虽然是基于实时测序的新技术,测序组装策略也采用直接对全基因组 DNA 进行测序,但其数据处理的算法则可以追溯到 Sanger 测序时代。此外,二代和三代测序技术流程基本已经成熟,组装成功与否和质量的好坏已不再受制于测序技术本身,而是取决于是否能够获得高质量长片段基因组 DNA。尤其对于植物而言,如何降低多样化次生代谢物质的影响,以

及如何从包含共生微生物的藻类中提取高质量 DNA,仍然充满挑战。而目前最有效的高质量基因组 DNA 提取流程的源头就是 Sanger 测序时代的 BAC 文库构建技术。从这个层面上看,充分认识和尊重基因组学技术的继承性和发展连续性特点弥足珍贵。

随着技术的发展,基因组测序难易的概念也随之发展和变化,早期认为难以测序的复杂基因组将成为容易测序的“简单”基因组。当然,基因组完成图的标准也将随之不断提升,即使是通常认为的高质量基因组完成图,如人类基因组和水稻基因组,实际上仍然存在着大量未知区域。未来,使用最新技术更新重要物种基因组,解析未知区域的序列信息也将是未来数年基因组研究领域的重点之一。

毫无疑问,随着测序技术的进步,获取高质量基因组序列的时间和费用成本将变得越来越低,更多物种都将有高质量的基因组序列以及其它多组学数据信息。因此,基因组学专家的工作重点应逐渐转为利用比较基因组学技术解决生物学问题,或通过整合多组学数据进行信息挖掘,指导和推动以生物学问题为中心的科学研究。

参考文献(References):

- [1] Jiao WB, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. *Curr Opin Plant Biol*, 2017, 36: 64–70. [DOI]
- [2] Tang C, Yang M, Fang Y, Luo Y, Gao S, Xiao X, An Z, Zhou B, Zhang B, Tan X, Yeang HY, Qin Y, Yang J, Lin Q, Mei H, Montoro P, Long X, Qi J, Hua Y, He Z, Sun M, Li W, Zeng X, Cheng H, Liu Y, Yang J, Tian W, Zhuang N, Zeng R, Li D, He P, Li Z, Zou Z, Li S, Li C, Wang J, Wei D, Lai CQ, Luo W, Yu J, Hu S, Huang H. The rubber tree genome reveals new insights into rubber production and species adaptation. *Nat Plants*, 2016, 2(6): 16073. [DOI]
- [3] Nowoshilow S, Schloissnig S, Fei JF, Dahl A, Pang AWC, Pippel M, Winkler S, Hastie AR, Young G, Roscito JG, Falcon F, Knapp D, Powell S, Cruz A, Cao H, Habermann B, Hiller M, Tanaka EM, Myers EW. The axolotl genome and the evolution of key tissue formation regulators. *Nature*, 2018, 554(7690): 50–55. [DOI]
- [4] Wei C, Yang H, Wang S, Zhao J, Liu C, Gao L, Xia E, Lu Y, Tai Y, She G, Sun J, Cao H, Tong W, Gao Q, Li Y, Deng

- W, Jiang X, Wang W, Chen Q, Zhang S, Li H, Wu J, Wang P, Li P, Shi C, Zheng F, Jian J, Huang B, Shan D, Shi M, Fang C, Yue Y, Li F, Li D, Wei S, Han B, Jiang C, Yin Y, Xia T, Zhang Z, Bennetzen JL, Zhao S, Wan X. 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(18): E4151–E4158. [DOI]
- [5] Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, Kohara Y, Fujiyama A, Hayashi T, Itoh T. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res*, 2014, 24(8): 1384–1395. [DOI]
- [6] Li F, Fan G, Lu C, Xiao G, Zou C, Kohel RJ, Ma Z, Shang H, Ma X, Wu J, Liang X, Huang G, Percy RG, Liu K, Yang W, Chen W, Du X, Shi C, Yuan Y, Ye W, Liu X, Zhang X, Liu W, Wei H, Wei S, Huang G, Zhang X, Zhu S, Zhang H, Sun F, Wang X, Liang J, Wang J, He Q, Huang L, Wang J, Cui J, Song G, Wang K, Xu X, Yu JZ, Zhu Y, Yu S. Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat Biotechnol*, 2015, 33(5): 524–530. [DOI]
- [7] Sun F, Fan G, Hu Q, Zhou Y, Guan M, Tong C, Li J, Du D, Qi C, Jiang L, Liu W, Huang S, Chen W, Yu J, Mei D, Meng J, Zeng P, Shi J, Liu K, Wang X, Wang X, Long Y, Liang X, Hu Z, Huang G, Dong C, Zhang H, Li J, Zhang Y, Li L, Shi C, Wang J, Lee SM, Guan C, Xu X, Liu S, Liu X, Chalhoub B, Hua W, Wang H. The high-quality genome of *Brassica napus* cultivar ‘ZS11’ reveals the introgression history in semi-winter morphotype. *Plant J*, 2017, 92(3): 452–468. [DOI]
- [8] Chen YC, Liu T, Yu CH, Chiang TY, Hwang CC. Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS One*, 2013, 8(4): e62856. [DOI]
- [9] Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*, 2009, 6(4): 291–295. [DOI]
- [10] Brawley SH, Blouin NA, Ficko-Blean E, Wheeler GL, Lohr M, Goodson HV, Jenkins JW, Blaby-Haas CE, Helliwell KE, Chan CX, Marriage TN, Bhattacharya D, Klein AS, Badis Y, Brodie J, Cao Y, Collen J, Dittami SM, Gachon CMM, Green BR, Karpowicz SJ, Kim JW, Kudahl UJ, Lin S, Michel G, Mittag M, Olson B, Pangilinan JL, Peng Y, Qiu H, Shu S, Singer JT, Smith AG, Sprecher BN, Wagner V, Wang W, Wang ZY, Yan J, Yarish C, Zauner-Riek S, Zhuang Y, Zou Y, Lindquist EA, Grimwood J, Barry KW, Rokhsar DS, Schmutz J, Stiller JW, Grossman AR, Prochnik SE. Insights into the red algae and eukaryotic evolution from the genome of *Porphyra umbilicalis* (Bangiophyceae, Rhodophyta). *Proc Natl Acad Sci USA*, 2017, 114(31): E6361–E6370. [DOI]
- [11] Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc Natl Acad Sci USA*, 2016, 113(18): 5053–5058. [DOI]
- [12] Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Nishimura EO, Tintori SC, Li Q, Jones CD, Yandell M, Messina DN, Glasscock J, Goldstein B. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci USA*, 2015, 112(52): 15976–15981. [DOI]
- [13] Whitelaw CA, Barbazuk WB, Perteza G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, SanMiguel P, Lakey N, Bedell J, Yuan Y, Budiman MA, Resnick A, Van Aken S, Utterback T, Riedmuller S, Williams M, Feldblyum T, Schubert K, Beachy R, Fraser CM, Quackenbush J. Enrichment of gene-coding sequences in maize by genome filtration. *Science*, 2003, 302(5653): 2118–2120. [DOI]
- [14] Palmer LE, Rabinowicz PD, O’Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR. Maize genome sequencing by methylation filtration. *Science*, 2003, 302(5653): 2115–2117. [DOI]
- [15] George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res*, 2011, 21(10): 1686–1694. [DOI]
- [16] Potato Genome Sequencing C, 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, Mejia 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, Sonderkaer M, Iovene M, Torres GA, Jiang J, Veilleux RE, Bachem CW, de Boer J, Borm T, Kloosterman B, van Eck H, Datema E, Hekkert B, Goverse A, van Ham RC, Visser RG. Genome sequence and analysis of the tuber crop potato. *Nature*, 2011, 475(7355): 189–195. [DOI]
- [17] Yang H, Chen X, Wong WH. Completely phased genome sequencing through chromosome sorting. *Proc Natl Acad Sci USA*, 2011, 108(1): 12–17. [DOI]
- [18] Fan HC, Wang J, Potanina A, Quake SR. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol*, 2011, 29(1): 51–57. [DOI]
- [19] Young AL, Abaan HO, Zerbino D, Mullikin JC, Birney E, Margulies EH. A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome Res*, 2010, 20(2): 249–256. [DOI]
- [20] Sorber K, Chiu C, Webster D, Dimon M, Ruby JG, Hekele A, DeRisi JL. The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS One*, 2008, 3(10): e3495. [DOI]
- [21] Hiatt JB, Patwardhan RP, Turner EH, Lee C, Shendure J. Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods*, 2010, 7(2): 119–122. [DOI]
- [22] Rhoads A, Au KF. PacBio sequencing and its applications. *Genom Prot Bioinf*, 2015, 13(5): 278–289. [DOI]
- [23] Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol*, 2016, 17(1): 239. [DOI]
- [24] Staňková H, Hastie AR, Chan S, Vrána J, Tulpová Z, Kubaláková M, Visendi P, Hayashi S, Luo M, Batley J, Edwards D, Doležel J, Šimková H. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol J*, 2016, 14(7): 1523–1531. [DOI]
- [25] Chen P, Jing X, Liao B, Zhu Y, Xu J, Liu R, Zhao Y, Li X. BioNano genome map resource for *Oryza sativa* ssp. japonica and indica and its application in rice genome sequence correction and gap filling. *Mol Plant*, 2017, 10(6): 895–898. [DOI]
- [26] Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, 2012, 58(3): 268–276. [DOI]
- [27] Paulsen J, Liyakat Ali TM, Collas P. Computational 3D genome modeling using Chrom3D. *Nat Protoc*, 2018, 13(5): 1137–1152. [DOI]
- [28] Zerbino DR, Birney E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res*, 2008, 18(5): 821–829. [DOI]
- [29] Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, Johnson J, Li K, Mobarry C, Sutton G. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, 2008, 24(24): 2818–2824. [DOI]
- [30] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, 2017, 27(5): 722–736. [DOI]
- [31] Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA*, 2011, 108(4): 1513–1518. [DOI]
- [32] Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*, 2015, 33(6): 623–630. [DOI]
- [33] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 2012, 19(5): 455–477. [DOI]
- [34] Ye C, Hill CM, Wu S, Ruan J, Ma ZS. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci Rep*, 2016, 6: 31900. [DOI]
- [35] Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. *Bioinformatics*, 2013, 29(21): 2669–2677. [DOI]
- [36] Zimin A, Stevens KA, Crepeau MW, Holtz-Morris A, Koriabine M, Marcais G, Puiu D, Roberts M, Wegrzyn JL,

- de Jong PJ, Neale DB, Salzberg SL, Yorke JA, Langley CH. Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics*, 2014, 196(3): 875–890. [DOI]
- [37] Zimin AV, Puiu D, Luo MC, Zhu T, Koren S, Marcais G, Yorke JA, Dvorak J, Salzberg SL. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res*, 2017, 27(5): 787–792. [DOI]
- [38] Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, 2012, 1(1): 18. [DOI]
- [39] Marcais G, Yorke JA, Zimin A. QuorUM: an error corrector for illumina reads. *PLoS One*, 2015, 10(6): e0130821. [DOI]
- [40] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. *Genome Biol*, 2004, 5(2): R12. [DOI]
- [41] Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. *Bioinformatics*, 2016, 32(14): 2103–2110. [DOI]
- [42] Xiao CL, Chen Y, Xie SQ, Chen KN, Wang Y, Han Y, Luo F, Xie Z. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat Methods*, 2017, 14(11): 1072–1074. [DOI]
- [43] Schmidt MH-W, Vogel A, Denton A, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Mass J, Pfaff C, Schurr U, Chetelat R, Maumus F, Aury J-M, Fernie AR, Zamir D, Bolger AM, Usadel B. Reconstructing the gigabase plant genome of *Solanum pennellii* using nanopore sequencing. *bioRxiv*, 2017, doi: 10.1101/129148. [DOI]
- [44] Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*, 2016, 13(12): 1050–1054. [DOI]
- [45] Roach MJ, Schmidt SA, Borneman AR. Purge haplotigs: synteny reduction for third-gen diploid genome assemblies. *bioRxiv*, 2018, doi: 10.1101/286252. [DOI]
- [46] de Lannoy CV, de Ridder D, Risse J. A sequencer coming of age: *de novo* genome assembly using MinION reads. *bioRxiv*, 2017, doi: 10.1101/142711. [DOI]
- [47] Jones SJ, Haulena M, Taylor GA, Chan S, Bilobram S, Warren RL, Hammond SA, Mungall KL, Choo C, Kirk H, Pandoh P, Ally A, Dhalla N, Tam AKY, Troussard A, Paulino D, Coope RJN, Mungall AJ, Moore R, Zhao Y, Birol I, Ma Y, Marra M, Jones SJM. The genome of the northern sea otter (*Enhydra lutris kenyoni*). *Genes (Basel)*, 2017, 8(12): genes8120379. [DOI]
- [48] Zhang GQ, Liu KW, Li Z, Lohaus R, Hsiao YY, Niu SC, Wang JY, Lin YC, Xu Q, Chen LJ, Yoshida K, Fujiwara S, Wang ZW, Zhang YQ, Mitsuda N, Wang M, Liu GH, Pecoraro L, Huang HX, Xiao XJ, Lin M, Wu XY, Wu WL, Chen YY, Chang SB, Sakamoto S, Ohme-Takagi M, Yagi M, Zeng SJ, Shen CY, Yeh CM, Luo YB, Tsai WC, Van de Peer Y, Liu ZJ. The Apostasia genome and the evolution of orchids. *Nature*, 2017, 549(7672): 379–383. [DOI]
- [49] Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, Aiden EL. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, 2017, 356(6333): 92–95. [DOI]
- [50] Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, Khan MA, Tao S, Korban SS, Wang H, Chen NJ, Nishio T, Xu X, Cong L, Qi K, Huang X, Wang Y, Zhao X, Wu J, Deng C, Gou C, Zhou W, Yin H, Qin G, Sha Y, Tao Y, Chen H, Yang Y, Song Y, Zhan D, Wang J, Li L, Dai M, Gu C, Wang Y, Shi D, Wang X, Zhang H, Zeng L, Zheng D, Wang C, Chen M, Wang G, Xie L, Sovero V, Sha S, Huang W, Zhang S, Zhang M, Sun J, Xu L, Li Y, Liu X, Li Q, Shen J, Wang J, Paull RE, Bennetzen JL, Wang J, Zhang S. The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res*, 2013, 23(2): 396–408. [DOI]
- [51] Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 2011, 27(4): 578–579. [DOI]
- [52] Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*, 2002, 30(11): 2478–2483. [DOI]
- [53] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 1990, 215(3): 403–410. [DOI]
- [54] Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, Xiong Z, Que H, Xie Y, Holland PW, Paps J, Zhu Y, Wu F, Chen Y, Wang J, Peng C, Meng J, Yang L, Liu J, Wen B, Zhang N, Huang Z, Zhu Q, Feng Y, Mount A, Hedgecock D, Xu Z, Liu Y, Domazet-Loso T,

- Du Y, Sun X, Zhang S, Liu B, Cheng P, Jiang X, Li J, Fan D, Wang W, Fu W, Wang T, Wang B, Zhang J, Peng Z, Li Y, Li N, Wang J, Chen M, He Y, Tan F, Song X, Zheng Q, Huang R, Yang H, Du X, Chen L, Yang M, Gaffney PM, Wang S, Luo L, She Z, Ming Y, Huang W, Zhang S, Huang B, Zhang Y, Qu T, Ni P, Miao G, Wang J, Wang Q, Steinberg CE, Wang H, Li N, Qian L, Zhang G, Li Y, Yang H, Liu X, Wang J, Yin Y, Wang J. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, 2012, 490(7418): 49–54. [DOI]
- [55] Wuyun TN, Wang L, Liu H, Wang X, Zhang L, Bennetzen JL, Li T, Yang L, Liu P, Du L, Wang L, Huang M, Qing J, Zhu L, Bao W, Li H, Du Q, Zhu J, Yang H, Yang S, Liu H, Yue H, Hu J, Yu G, Tian Y, Liang F, Hu J, Wang D, Gao R, Li D, Du H. The hardy rubber tree genome provides insights into the evolution of polyisoprene biosynthesis. *Mol Plant*, 2018, 11(3): 429–442. [DOI]
- [56] English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, Gibbs RA. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One*, 2012, 7(11): e47768. [DOI]
- [57] Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Pozniak CJ, Stein N, Choulet F, Distelfeld A, Eversole K, Poland J, Rogers J, Ronen G, Sharpe AG, Pozniak C, Ronen G, Stein N, Barad O, Baruch K, Choulet F, Keeble-Gagnère G, Mascher M, Sharpe AG, Ben-Zvi G, Josselin A-A, Stein N, Mascher M, Himmelbach A, Choulet F, Keeble-Gagnère G, Mascher M, Rogers J, Balfourier F, Gutierrez-Gonzalez J, Hayden M, Josselin A-A, Koh C, Muehlbauer G, Pasam RK, Paux E, Pozniak CJ, Rigault P, Sharpe AG, Tibbits J, Tiwari V, Choulet F, Keeble-Gagnère G, Mascher M, Josselin A-A, Rogers J, Spannagl M, Choulet F, Lang D, Gundlach H, Haberer G, Keeble-Gagnère G, Mayer KFX, Ormanbekova D, Paux E, Prade V, Šimková H, Wicker T, Choulet F, Spannagl M, Swarbreck D, Rimbart H, Felder M, Guilhot N, Gundlach H, Haberer G, Kaithakottil G, Keilwagen J, Lang D, Leroy P, Lux T, Mayer KFX, Twardziok S, Venturini L, Appels R, Rimbart H, Choulet F, Juhász A, Keeble-Gagnère G, Choulet F, Spannagl M, Lang D, Abrouk M, Haberer G, Keeble-Gagnère G, Mayer KFX, Wicker T, Choulet F, Wicker T, Gundlach H, Lang D, Spannagl M, Lang D, Spannagl M, Appels R, Fischer I, Uauy C, Borrill P, Ramirez-Gonzalez RH, Appels R, Arnaud D, Chalabi S, Chalhoub B, Choulet F, Cory A, Datla R, Davey MW, Hayden M, Jacobs J, Lang D, Robinson SJ, Spannagl M, Steuernagel B, Tibbits J, Tiwari V, van Ex F, Wulff BBH, Pozniak CJ, Robinson SJ, Sharpe AG, Cory A, Benhamed M, Paux E, Bendahmane A, Concia L, Latrasse D, Rogers J, Jacobs J, Alaux M, Appels R, Bartoš J, Bellec A, Berges H, Doležel J, Feuillet C, Frenkel Z, Gill B, Korol A, Letellier T, Olsen O-A, Šimková H, Singh K, Valárik M, van der Vossen E, Vautrin S, Weining S, Korol A, Frenkel Z, Fahima T, Glikson V, Raats D, Rogers J, Tiwari V, Gill B, Paux E, Poland J, Doležel J, Číhalíková J, Šimková H, Toegelová H, Vrána J, Sourdille P, Darrier B, Appels R, Spannagl M, Lang D, Fischer I, Ormanbekova D, Prade V, Barabaschi D, Cattivelli L, Hernandez P, Galvez S, Budak H, Steuernagel B, Jones JDG, Witek K, Wulff BBH, Yu G, Small I, Melonek J, Zhou R, Juhász A, Belova T, Appels R, Olsen O-A, Kanyuka K, King R, Nilsen K, Walkowiak S, Pozniak CJ, Cuthbert R, Datla R, Knox R, Wiebe K, Xiang D, Rohde A, Golds T, Doležel J, Čížková J, Tibbits J, Budak H, Akpinar BA, Biyiklioglu S, Muehlbauer G, Poland J, Gao L, Gutierrez-Gonzalez J, N'Daiye A, Doležel J, Šimková H, Číhalíková J, Kubaláková M, Šafář J, Vrána J, Berges H, Bellec A, Vautrin S, Alaux M, Alfama F, Adam-Blondon A-F, Flores R, Guerche C, Letellier T, Loaec M, Quesneville H, Pozniak CJ, Sharpe AG, Walkowiak S, Budak H, Condie J, Ens J, Koh C, Maclachlan R, Tan Y, Wicker T, Choulet F, Paux E, Alberti A, Aury J-M, Balfourier F, Barbe V, Couloux A, Cruaud C, Labadie K, Mangenot S, Wincker P, Gill B, Kaur G, Luo M, Sehgal S, Singh K, Chhuneja P, Gupta OP, Jindal S, Kaur P, Malik P, Sharma P, Yadav B, Singh NK, Khurana J, Chaudhary C, Khurana P, Kumar V, Mahato A, Mathur S, Sevanthi A, Sharma N, Tomar RS, Rogers J, Jacobs J, Alaux M, Bellec A, Berges H, Doležel J, Feuillet C, Frenkel Z, Gill B, Korol A, van der Vossen E, Vautrin S, Gill B, Kaur G, Luo M, Sehgal S, Bartoš J, Holušová K, Plíhal O, Clark MD, Heavens D, Kettleborough G, Wright J, Valárik M, Abrouk M, Balcárková B, Holušová K, Hu Y, Luo M, Salina E, Ravin N, Skryabin K, Beletsky A, Kadnikov V, Mardanov A, Nesterov M, Rakitin A, Sergeeva E, Handa H, Kanamori H, Katagiri S, Kobayashi F, Nasuda S, Tanaka T, Wu J, Appels R, Hayden M, Keeble-Gagnère G, Rigault P, Tibbits J, Olsen O-A, Belova T, Cattonaro F, Jiumeng M, Kugler K, Mayer KFX, Pfeifer M, Sandve S, Xun X, Zhan B, Šimková H, Abrouk M, Batley J, Bayer PE, Edwards D, Hayashi S, Toegelová H, Tulpová Z, Visendi P, Weining S, Cui L, Du X, Feng K,

- Nie X, Tong W, Wang L, Borrill P, Gundlach H, Galvez S, Kaithakottil G, Lang D, Lux T, Mascher M, Ormanbekova D, Prade V, Ramirez-Gonzalez RH, Spannagl M, Stein N, Uauy C, Venturini L, Stein N, Appels R, Eversole K, Rogers J, Borrill P, Cattivelli L, Choulet F, Hernandez P, Kanyuka K, Lang D, Mascher M, Nilsen K, Paux E, Pozniak CJ, Ramirez-Gonzalez RH, Šimková H, Small I, Spannagl M, Swarbreck D, Uauy C. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 2018, 361(6403): science.aar7191. [DOI]
- [58] Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G, Fich EA, Conte M, Keller H, Schneeberger K, Schwacke R, Ofner I, Vrebalov J, Xu Y, Osorio S, Aflitos SA, Schijlen E, Jiménez-Gómez JM, Ryngajllo M, Kimura S, Kumar R, Koenig D, Headland LR, Maloof JN, Sinha N, van Ham RCHJ, Lankhorst RK, Mao L, Vogel A, Arsova B, Panstruga R, Fei Z, Rose JKC, Zamir D, Carrari F, Giovannoni JJ, Weigel D, Usadel B, Fernie AR. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nature Genetics*, 2014, 46: 1034. [DOI]
- [59] Schmidt MH, Vogel A, Denton AK, Istace B, Wormit A, van de Geest H, Bolger ME, Alseekh S, Mass J, Pfaff C, Schurr U, Chetelat R, Maumus F, Aury JM, Koren S, Fernie AR, Zamir D, Bolger AM, Usadel B. *De novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell*, 2017, 29(10): 2336–2348. [DOI]
- [60] Geng J, Chiu CH, Tang P, Chen Y, Shieh HR, Hu S, Chen YY. Complete genome and transcriptomes of *Streptococcus parasanguinis* FW213: phylogenic relations and potential virulence mechanisms. *PLoS One*, 2012, 7(4): e34769. [DOI]
- [61] Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, Yu X, Liu J, Pan L, Zhang T, Yin Y, Xin C, Wu H, Zhang G, Ba Abdullah MM, Huang D, Fang Y, Alnakhli YO, Jia S, Yin A, Alhuzimi EM, Alsaihati BA, Al-Owayyed SA, Zhao D, Zhang S, Al-Otaibi NA, Sun G, Majrashi MA, Li F, Tala, Wang J, Yun Q, Alnassar NA, Wang L, Yang M, Al-Jelaify RF, Liu K, Gao S, Chen K, Alkhaldi SR, Liu G, Zhang M, Guo H, Yu J. Genome sequence of the date palm *Phoenix dactylifera* L. *Nat Commun*, 2013, 4: 2274. [DOI]
- [62] Wang S, Wang S, Luo Y, Xiao L, Luo X, Gao S, Dou Y, Zhang H, Guo A, Meng Q, Hou J, Zhang B, Zhang S, Yang M, Meng X, Mei H, Li H, He Z, Zhu X, Tan X, Zhu XQ, Yu J, Cai J, Zhu G, Hu S, Cai X. Comparative genomics reveals adaptive evolution of Asian tapeworm in switching to a new intermediate host. *Nat Commun*, 2016, 7: 12845. [DOI]

(责任编辑: 赵方庆)