中国科学: 生命科学 2025 年 第 55 卷 第 9 期: 1793 ~ 1811
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

评 述

植物超泛基因组研究现状与展望

黄小琴[1], 胡丽松[1,2,3], 范睿[1,2,3], 张兴坦[4*], 郝朝运[1,2,3*]

1. 中国热带农业科学院香料饮料研究所, 万宁 571533
2. 农业农村部香辛饮料作物遗传资源利用重点实验室, 万宁 571533
3. 海南省热带香辛饮料作物遗传改良与品质调控重点实验室, 万宁 571533
4. 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; haochy79@163.com

收稿日期: 2024-12-03; 接受日期: 2025-03-18; 网络版发表日期: 2025-08-05
国家自然科学基金(批准号: 32460741)、海南省院士工作站专项(批准号: YSPTZX202154)和中国热带农业科学院基本科研业务费专项(批准号: 1630142023002)资助

摘要 近年来, 高通量测序技术的迅猛发展极大地推动了植物全基因组测序进程. 然而, 单一参考基因组无法全面覆盖物种的全部遗传信息, 限制了基因组学的深入研究. 泛基因组(pan-genome)通过整合多个代表性个体的基因组数据, 有效克服了单一参考基因组的局限性, 能够更全面展示物种的基因多样性, 包括基因结构和变异, 为基因组学研究开辟了新方向. 然而, 泛基因组更多关注单一物种内的基因多样性, 为了捕捉植物界丰富的遗传多样性, 泛基因组进一步扩展为超泛基因组(super-pangenome). 植物超泛基因组从物种层面扩展至属或更高分类单元, 涵盖栽培种及其二级、三级基因库中的野生近缘种. 通过整合多元物种的种质资源, 不仅构建了丰富的遗传变异图谱, 还能捕捉到稀有基因和独特变异, 为植物改良提供了新的潜力与方向. 这将有助于通过分子标记辅助选择或基因编辑技术, 将野生近缘种有价值的性状转移到优质品种, 为植物品种改良提供宝贵基础, 推动农业生产中更高效、更具适应性的品种创新. 本文回顾了植物基因组和泛基因组研究的发展历程, 总结了目前已发表的植物超泛基因组及其应用, 并探讨了植物超泛基因组未来的前景和面临的挑战.

关键词 植物超泛基因组, 植物泛基因组, 基因组组装, 植物基因组学, 测序技术

随着测序通量和准确率的提高以及测序成本的降低, “万种植物基因组计划(The Plant 10,000 Genomes Project, 10KP)”[1]、“地球生物基因组计划(Earth Bio-Genome Project, EBP)”[2]等全球范围内的大型基因组测序项目相继启动. 随着大量植物基因组数据的积累, 在对同一物种多个基因组进行比较分析时, 发现不同个体间存在大量存在-缺失变异(presence-absence variations, PAVs)[3]. 这揭示了单一参考基因组在代表性上的局限性, 仅能反映单个样本的遗传信息, 无法涵盖物种内所有个体的基因序列变异, 难以全面反映物种的遗传多样性. 随着学者开始将物种内更多样本纳入研究范畴, 单一参考基因组逐渐转向泛基因组构建[4]. 泛基因组通过整合单个物种的所有遗传信息, 更全面地展示物种内存在的多样性区域[3]. 近年来, 植物泛基因组研究在功能基因挖掘、基因组进化以及物种起源与驯化等方面取得了一系列重要进展.

引用格式: 黄小琴, 胡丽松, 范睿, 等. 植物超泛基因组研究现状与展望. 中国科学: 生命科学, 2025, 55: 1793–1811
Huang X Q, Hu L S, Fan R, et al. Current status and prospects of super-pangenome studies in plants (in Chinese). Sci Sin Vitae, 2025, 55: 1793–1811, doi: 10.1360/SSV-2024-0337

为了拓宽泛基因组的研究范围, 引入了“超泛基因组”的概念, 通过整合属级甚至更高分类单元中多样化的野生近缘种和栽培品种来捕获更广泛的遗传多样性. 野生近缘种(crop wild relatives, CWR)拥有更广泛的基因资源库, 包含许多改变农业生产力和植物抗性的关键特性. 这些特性包括抗病虫害、耐旱性、耐盐性等, 这对提高植物的适应性和生产力至关重要. 为了更好地利用和展示CWRs的遗传多样性, 通过构建“超泛基因组”将泛基因组从物种水平扩展到属级甚至更高分类单元水平. 超泛基因组有助于挖掘CWRs中隐藏的遗传变异, 并探索物种基因组中非核心的基因部分, 这对理解植物基因组的进化和适应具有重要意义[5].

鉴于此, 本综述系统总结了植物基因组学从单一基因组到泛基因组, 再到超泛基因组的发展历程, 讨论了植物超泛基因组的样本选择、构建方式和分析工具. 通过梳理植物超泛基因组的主要研究成果与前沿进展, 为全面理解超泛基因组在物种基因组进化、植物驯化及表型多样性中的重要价值提供新视角. 同时, 本综述还探讨了其潜在应用方向与发展前景, 为未来植物超泛基因组学研究的创新和突破提供见解与启示.

1 植物基因组的发展

1.1 植物基因组的发展历程

1975年, 由Sanger和Coulson[6]提出双脱氧链终止法(Sanger法). 1977年, Maxam与Gilbert[7]提出化学降解进行测序的方法. 同年, Sanger使用双脱氧链终止法测定了第一个DNA基因组序列, 即全长约5375个核苷酸的噬菌体phiX-174[8]. 一代测序技术就此诞生, 从此测序技术快速发展, Sanger测序法迅速成为DNA测序的主要方法. Sanger测序主要基于重叠布局共识(overlap-layout-consensus, OLC)组装算法[9], 读长通常为500~1000 bp. 拟南芥(*Arabidopsis thaliana*)作为最早用于基础功能研究的模式生物, 其第一个基因组于2000年发表, 这标志着植物基因组学研究的开始[10]. 2002年发表了第一个单子叶禾本科植物水稻(*Oryza sativa*)基因组[11]. 第一代测序技术虽然具有序列读长长和准确率高等优势, 但其测序速度慢、费用高、通量低等缺陷使其不能满足大规模测序的需求, 导致植物基因组研究进展缓慢.

Illumina Solexa, Roche 454和ABI SOLiD等第二代测序技术(next-generation sequencing, NGS)的出现促进了基于德布鲁因图(de Bruijn graph, DBG)组装算法的发展[12]. DBG适合处理较短的reads, 常用组装软件有ABySS, ALLPATHS-LG, SOAPdenovo, Velvet等[13]. 二代测序技术显著提高了测序速度和数据产出量, 大大地降低了测序成本, 改变了测序的规模化进程, 推动了植物全基因组测序变革性的爆发式增长[12]. 例如, Roche 454技术被用于可可(*Theobroma cacao*)基因组测序, 葡萄(*Vitis vinifera*)和苹果(*Malus domestica*)结合Sanger和Roche 454技术, 棉花结合Illumina Solexa和Roche 454技术, 而香蕉则结合Sanger, Illumina Solexa和Roche 454技术[14,15]. Roche 454平台读长约400 bp左右, Solexa(Illumina)和SOLiD平台的读长则在100 bp左右或更短[9].

复杂基因组存在大量重复序列, 短读长序列片段难以准确拼接. 为了解决这些问题, 新的测序技术不断涌现. 2009年, PacBio公司的单分子实时测序技术(single molecule real-time, SMRT)和Oxford Nanopore Technologies公司的纳米孔测序技术(nanopore sequencing)相继推出[16,17]. 三代测序技术在测序过程中不需要PCR扩增, 其读长可达到二代测序的100倍, 大大降低了基因组拼接的难度. 复活草(*Oropetium thomaeum*)[18]和拟南芥[19]是最早通过PacBio SMRT数据独立组装的植物基因组. 拟南芥基因组达到染色体水平, 而245 Mb的复活草基因组组装contig N50达到2.4 Mb, 这种连续性在短读长技术中是无法实现的[20]. 随后苹果(*M. domestica*)三倍体“Hanfu”[21]、番木瓜(*Carica papaya*)[22]、荔枝(*Litchi chinensis*)[23]等高质量基因组成功组装. 三代测序技术推动了组装算法和软件的发展, 如Canu[24], Falcon/Falcon-Unzip[25]和HGAP[26]等工具能够更好地组装长片段并校正错误. 这些软件显著提高了基因组组装的精度和效率, 特别是在处理复杂或大型基因组时[27]. 2019年, PacBio公司推出了基于环形一致性测序(circular consensus sequencing, CCS)模式的HiFi(high fidelity)测序技术, 其能够生成长读长(10~20 kb)且高精度(准确率>99%)的读长序列[28]. 高通量染色体构象捕获技术(high-throughput chromosome conformation capture, Hi-C)[29]、Bionano Genomics光学图谱[30]、10x Genomics linked read[31]测序等辅助基因组组装测序技术的出现极大地帮助实现

染色体水平的组装.

测序技术与组装算法的发展，加快了植物基因组组装进程. 从2000年到2020年，公布了782种植物的1144个基因组. 2021至2023年期间，已有1031种植物的2373个基因组，其中793个为新测序物种[32]. 过去三年，植物基因组测序数量迅速增长，尤其是高质量基因组序列的可用性，极大地推动了功能基因组学和群体遗传学等植物学科的发展[33].

1.2　基因组到泛基因组的转变

2005年，Tettelin等人[34]首次在细菌中提出微生物泛基因组的概念. 随着越来越多植物基因组被组装出来，人们发现单一参考基因组不能反映一个物种的基因多样性，这导致植物泛基因组概念的产生(图1). 2007年，Morgante等人[35]提出将泛基因组引入到植物中，在玉米基因组中描述了可变基因组在遗传多样性和性状差异中的重要作用，并探讨其组成、起源和功能. 随着二代测序技术的应用和普及，2014年通过从头组装策略构建了7株代表性野生大豆的植物泛基因组，鉴定出大量与抗逆性、抗病性、开花期、种子成分、器官大小和生物量等重要农艺性状相关的基因与变异. 例如, 野生大豆和栽培大豆的开花期与调控开花基因中的单核苷酸多态性(single nucleotide polymorphism，SNP)及插入缺失(InDel)变异有关. 该研究在植物研究领域中开启了泛基因组研究历程[36]. 短读长的二代测序难以检测大尺度的结构变异(>50 bp), 随着三代长读长测序技术的出现，2020年在大豆中构建了第一个基于三代测序数据的植物图形泛基因组，以中国大豆品种“中黄13”(ZH13)为参考在28个大豆基因组中识别出723862个PAV、27531个拷贝数变异(copy number variation，CNV)、21886个易位事件和3120个倒位事件，该研究促进了图形泛基因组的发展[37]. 迄今已在水稻[38]、玉米[39]、番茄[40]、黄瓜[41]等植物中开展了泛基因组测序研究.

1.3　泛基因组到超泛基因组的转变

迄今为止，传统的植物泛基因组研究主要集中于单一物种内的农家种和栽培种等种质资源，尽管番茄[40]、大豆[37]等研究涵盖了少量的野生近缘种. 然而，整体而言，传统的泛基因组研究未充分涵盖更多

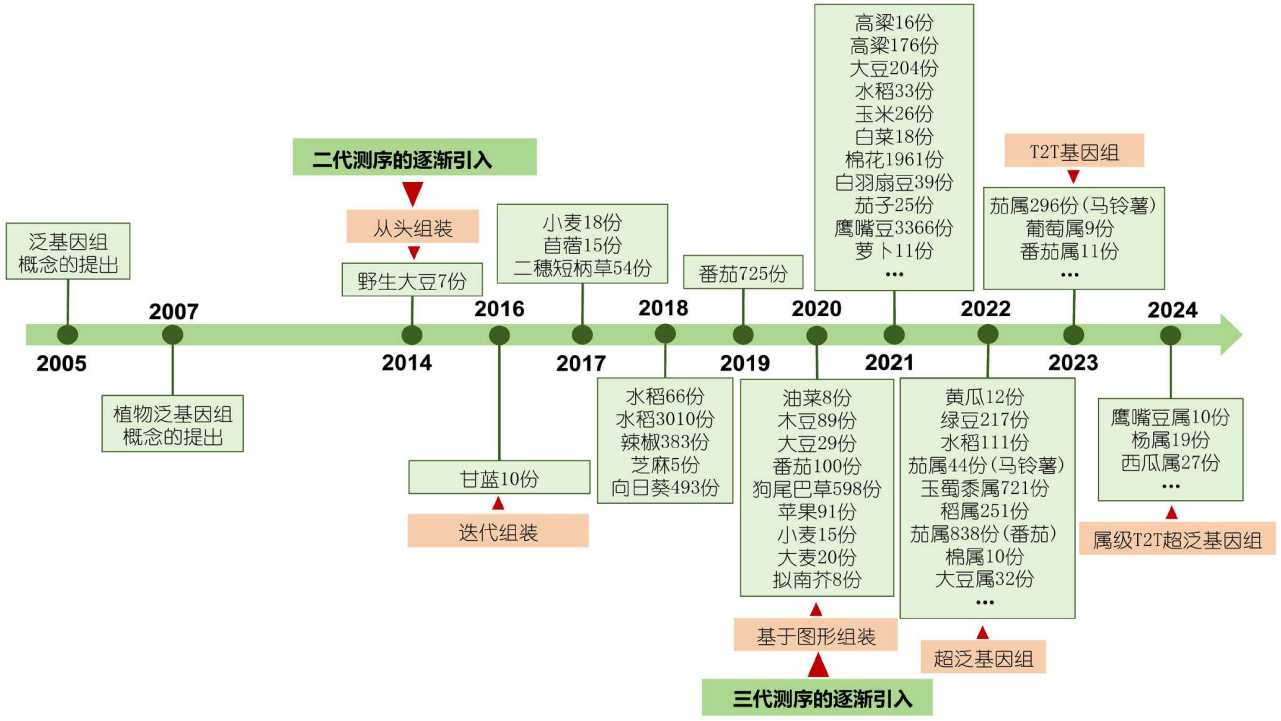

图 1　植物泛基因组发展史
Figure 1　History of plant pan-genome development.

野生近缘种, 在种质资源代表性方面仍显不足, 无法全面反映属级甚至更高分类单元中物种的遗传变异. 2020年, Khan等人[42]提出了超泛基因组(super-pangenome)的概念, 并指出了在属级水平的基因组多样性. 超泛基因组是泛基因组的延伸, 通过整合属级不同物种(如野生近缘种、地方种、改良种等)的基因组数据, 构建出包含属级水平的完整基因组序列变异库. 2022年, 大豆[43]、水稻[44]和玉米[45]的植物超泛基因组相继发布, 将植物超泛基因组推向了高潮. 2024年, Zhang等人[46]发表了西瓜属7个种28份种质的T2T(telomere-to-telomere genome)高质量基因组图谱, 成功构建首个属级的T2T水平超泛基因组, 为植物基因组学研究带来新的突破. 目前, 许多植物已构建了超泛基因组, 包括水稻[44]、玉米[45]、番茄[47]、葡萄[48]、鹰嘴豆[49]、杨树[50]等(表1). 然而, 整体而言, 植物超泛基因组研究仍处于早期阶段, 目前应用研究主要集中在水稻、玉米等主要粮食作物, 而次要作物及园艺植物(如果树、花卉、药用植物)较少; 在样本层面, 样本规模一般较小, 尚缺乏统一的选择标准; 在技术层面, 现有手段难以解析复杂基因组, 分析算法和工具仍在不断优化.

CWRs具有高度的遗传多样性和强大的环境适应能力, 与栽培种相比, 其在应对干旱、盐害等非生物胁迫方面表现出更强的抗逆性. 超泛基因组的构建能够有效引入CWRs, 在识别丰富的结构变异(structural variant, SV)方面具有显著的优势. 在番茄研究过程中, Alonge等人[40]通过对100个栽培番茄和野生番茄种质进行ONT长读长测序, 建立了番茄全基因组结构变异(pan-SV)图谱, 共捕获到了238490个SV. 与该pan-SV图谱相比, 在11个番茄种质的超泛基因组中, 鉴定到的224447个SV中有180314个是其独有的, 其中大多数是由于纳入了CWRs而被捕获到[47]. 超泛基因组可为植物遗传改良提供更加丰富的基因资源, 特别是与重要农艺性状(如抗病性、开花时间和器官大小等)相关的基因. 此外, 超泛基因组还可以作为进化研究的优秀资源, 帮助准确检测物种分化时间, 并提供不同进化事件的历史估算[42].

2 植物超泛基因组的构建

2.1 超泛基因组的定义与概念

超泛基因组涵盖了属级或更高分类单元多个物种非冗余的基因组序列. 根据组成, 超泛基因组包括两部分: 所有属级或更高分类单元基因组中都存在的核心基因组(core genome)和仅在部分基因组中特有的非核心基因组(variable genome)[64]. 核心基因组的基因通常维持个体的生存和基本功能, 非核心基因组的基因通常参与生物和非生物胁迫(如抗病性)的响应[65]. 其中, 非核心可变基因组中基因来源主要有基因组复制(whole genome duplication, WGD)、局部串联重复、转座子(transposable element, TE)介导的插入、基因从头生成(如突变)、片段复制、近缘物种渗入以及不同物种间水平基因转移[66]. 非核心基因组反映了个体之间的差异, 可划分为附属基因和特有基因, 前者指基因存在于两个或多个个体中, 后者指基因仅存在于一个个体中[67]. 根据开放程度, 超泛基因组有两种类型: 开放型超泛基因组(open super-pangenomes)和闭合型超泛基因组(closed super-pangenomes). 随着测序样本数量的增加, 开放型超泛基因组大小无法预测, 其需要更多样本才能使超泛基因组和核心基因组的大小达到平台期. 而闭合型超泛基因组和核心基因组的大小趋于稳定且可预测, 几乎可以获得属内物种的所有基因序列信息, 例如水稻[44]、玉米[45]、番茄[47]、葡萄[48]、鹰嘴豆[49]、杨树[50]、西瓜[46]等植物的超泛基因组都是闭合型.

2.2 超泛基因组的样本选择

样本特性和样本数量对超泛基因组研究的分析效率与完整性至关重要. 选择的种质既是为了生物学兴趣, 也是为了研究的多样性和全面性, 从而更全面地了解超泛基因组的结构和功能. 为了确保所选样本能够全面捕捉到物种的遗传多样性且具有代表性, 科学合理的样本选择是构建超泛基因组的关键. 可以通过低深度测序进行初步遗传多样性评估, 以高效且经济的方式获取多样性信息; 结合系统发育关系, 选择具有代表性的样本; 最后, 结合植物形态特征、农艺性状、抗性等数据, 进一步筛选与特定性状或环境条件相关的样本, 以确保充分覆盖物种的遗传变异范围并能够反映不同遗传背景下的多样性. 为了捕捉属内的最大多样性, 应选择具有不同形态、表型和地理起源的种质[68].

利用属内不同物种的种质, 可以深入探讨属内物种间的起源和进化等生物学问题. 其次, 野生种与栽

表 1　植物超泛基因组研究汇总表

Table 1　Summary of plant super-pangenome studies

物种	属名	种质数量	种质种类	参考文献
大豆 (*Glycine max*)	大豆属 (*Glycine*)	32个	多年生二倍体: *G. falcata*, *G. stenophita*, *G. cyrtoloba*, *G. syndetika*, *G. tomentella* D3; 异源多倍体: *G. dolichocarpa*; 栽培种: *G. max*; 野生种: *G. soja*	Zhuang等人[43]
水稻 (*Oryza sativa*)	稻属 (*Oryza*)	251个	栽培稻: *O. sativa*; 野生稻: *O. rufipogon*; 非洲栽培稻: *O. glaberrima*; 短舌野生稻: *O. barthii*	Shang等人[44]
		23个	野生稻: *O. alta*, *O. australiensis*, *O. brachyantha*, *O. eichingeri*, *O. glumaepatula*, *O. grandiglumis*, *O. latifolia*, *O. malampuzhaensis*, *O. minuta*, *O. officinalis*, *O. punctata*, *O. rhizomatis*, *O. meyeriana*, *O. rufipogon*; 栽培稻: *O. sativa*, *O. glaberrima*	Long等人[51]
棉花 (*Gossypium herbaceum*)	棉属 (*Gossypium*)	22个	栽培种: *G. arboreum*, *G. herbaceum*; 野生种: *G. herbaceum* wild, *G. anomalum*, *G. sturtianum*, *G. raimondii*, *G. stocksii*, *G. longicalyx*, *G. australe*, *G. rotundifolium*	Wang等人[52]
番茄 (*Solanum lycopersicum*)	茄属 (*Solanum*)	725个	栽培种: *S. lycopersicum*; 野生种: *S. pimpinellifolium*, *S. cheesmaniae*, *S. galapagense*	Gao等人[53]
		838个	栽培种: *S. lycopersicum*; 野生种: *S. pimpinellifolium*, *S. cheesmaniae*, *S. galapagense*; 706个番茄种质的二代Illumina短读长数据	Zhou等人[54]
		11个	野生种: *S. habrochaites*, *S. chilense*, *S. peruvianum*, *S. corneliomulleri*, *S. neorickii*, *S. chmielewskii*, *S. pimpinellifolium*, *S. galapagense*, *S. lycopersicoides*; 栽培种: *S. lycopersicum*	Li等人[47]
葡萄 (*Vitis vinifera*)	葡萄属 (*Vitis*)	9个	北美野生种: *V. acerifolia*, *V. aestivalis*, *V. arizonica*, *V. berlandieri*, *V. girdiana*, *V. monticola*, *V. mustangensis*, *V. riparia*, *V. rupestris*	Cochetel等人[48]
		18个	栽培种: *V. vinifera*; 野生种: *V. retordii*, *V. arizonica*, *V. labrusca*	Liu等人[55]
鹰嘴豆 (*Cicer arietinum*)	鹰嘴豆属 (*Cicer*)	10个	野生种: *C. reticulatum*, *C. echinospermum*, *C. bijugum*, *C. judaicum*, *C. pinnatifidum*, *C. yamashitae*, *C. chorassanicum*, *C. cuneatum*; 栽培种: *C. arietinum*	Khan等人[49]
马铃薯 (*Solanum tuberosum*)	茄属 (*Solanum*)	44个	栽培种/地方种: *S. tuberosum*; 栽培种的祖先: *S. candolleanum*; 野生种: *S. andreanum*, *S. boliviense*, *S. brevicaule*, *S. buesii*, *S. bulbocastanum*, *S. burkartii*, *S. cajamarquense*, *S. chacoense*, *S. chomatophilum*, *S. commersonii*, *S. jamesii*, *S. lignicaule*, *S. morelliforme*, *S. multi-interruptum*, *S. neorossii*, *S. paucissectum*, *S. pinnatisectum*, *S. piurae*, *S. sogarandinum*, *S. vernei*; 不结薯的马铃薯姊妹类群Etuberosum: *S. palustre*, *S. etuberosum*	Tang等人[56]
		296个	栽培种/地方种: *S. phureja*, *S. juzepczukii*, *S. goniocalyx*, *S. stenotomum*, *S. curtilobum*, *S. chaucha*, *S. ajanhuiri*, *S. andigena*, *S. tuberosum*; 野生种: *S. abancayense*, *S. achacachense*, *S. acroglossum*, *S. acroscopicum*, *S. albornozii*, *S. ambosinum*, *S. andreanum*, *S. avilesii*, *S. berthaultii*, *S. blanco*, *S. boliviense*, *S. brevicaule*, *S. bukasovii*, *S. bulbocastanum*, *S. cajamarquense*, *S. canasense*, *S. cardiophyllum*, *S. chacoense*, *S. chomatophilum*, *S. commersonii*, *S. gourlayi*, *S. hondelmannii*, *S. hypacrarthrum*, *S. incamayoense*, *S. infundibuliforme*, *S. jamesii*, *S. kurtzianum*, *S. laxissimum*, *S. leptophyes*, *S. limbaniense*, *S. marinasense*, *S. medians*, *S. megistacrolobum*, *S. megistracrolobum*, *S. microdontum*, *S. multidissectum*, *S. multiinterruptum*, *S. okadae*, *S. pampasense*, *S. paucissectum*, *S. pinnatisectum*, *S. polyadenium*, *S. raphanifolium*, *S. sogarandinum*, *S. sparsipilum*, *S. spegazzinii*, *S. stenophyllidium*, *S. tarijense*, *S. vernei*, *S. verrucosum*, *S. violaceimarmoratum*	Bozan等人[57]
玉米 (*Zea mays*)	玉蜀黍属 (*Zea*)	721个	栽培种: *Z. may*; 野生近缘种: *Z. nicaraguensis*, *Z. luxurians*, *Z. diploperennis*, *Z. perennis*	Gui等人[45]

(表1续)

物种	属名	种质数量	种质种类	参考文献
猕猴桃 (*Actinidia chinensis*)	猕猴桃属 (*Actinidia*)	15个	净果: *A. arguta*, *A. ploygama*; 斑果: *A. chinensis*, *A. eriantha*, *A.hemsleyana*, *A. latifolia*, *A. rufa*, *A. zhejiangensis*	Yu等人[58]
西瓜 (*Citrullus lanatus*)	西瓜属 (*Citrullus*)	547个	地方种/栽培种: *C. lanatus*; 野生种: *C. mucosospermus*, *C. amarus*, *C. colocynthis*	Wu等人[59]
		28个	地方种/栽培种: *C. lanatus*; 野生种: *C. amarus*, *C. mucosospermus*, *C. colocynthis*, *C. ecirrhosus*, *C. naudinianus*, *C. rehmii*	Zhang等人[46]
杨树(*Populus*)	杨属 (*Populus*)	19个	野生种: *P. adenopoda*, *P. alba*, *P. davidiana*, *P. deltoids* v2.1, *P. euphratica*, *P. ilicifolia*, *P. koreana*, *P. lasiocarpa*, *P. pruinosa*, *P. pseudoglauca*, *P. qiongdaoensis*, *P. rotundifolia*, *P. simonii*, *P. szechuanica*, *P. tremula*, *P. trichocarpa* v4.1, *P. wuana*, *P. yunnanensis*	Shi等人[50]
白蜡树(*Fraxinus*)	梣属 (*Fraxinus*)	39个	*F. albicans*, *F. americana*, *F. angustifolia*, *F. anomala*, *F. apertisquamifera*, *F. baroniana*, *F. chinensis*, *F. cuspidata*, *F. dipetala*, *F. excelsior*, *F. floribunda*, *F. goodingii*, *F. greggii*, *F. griffithii*, *F. hupehensis*, *F. latifolia*, *F. mandschurica*, *F. mandshurica*, *F. nigra*, *F. ornus*, *F. paxiana*, *F. pennsylvanica*, *F. platypoda*, *F. quadrangulata*, *F. sieboldiana*, *F. sogdiana*, *F.* sp. 1973-6204, *F.* sp. D2006-0159, *F. velutina*, *F. xanthoxyloides*	Liu等人[60]
石斛兰 (*Dendrobiu nobile*)	石斛属 (*Dendrobium*)	17个	*D. aphyllum*, *D. Chao Praya Smile*, *D. chrysotoxum*, *D. crocatum*, *D. crumenatum*, *D. discolor*, *D. ellipsophyllum*, *D. formosum*, *D. hercoglossum*, *D. jenkinsii*, *D. leonis*, *D. lindleyi*, *D. nobile*, *D. officinale*, *D. secundum*, *D. smilliae*, *D. tetragonum*	Li等人[61]
杜鹃花 (*Rhododendron*)	杜鹃花属 (*Rhododendron*)	9个	*R. simsii*, *R. championae*, *R. micranthum*, *R. molle*, *R. mucronulatum*, *R. neriiflorum*, *R. nivale*, *R. ovatum*, *R. redowskianam*	Xia等人[62]
苹果 (*Malus domestica*)	苹果属 (*Malus*)	13个	栽培种: *M. domestica*; 野生种: *M. sylvestris*, *M. sieversii*, *M. orientalis*, *M. asiatica*	Wang等人[63]

培种的结合, 有助于挖掘重要的功能基因, 从而科学指导驯化和育种研究. 此外, 不同地理环境种质的收集, 可以研究属内物种的适应性进化, 识别环境适应基因, 并探讨物种入侵等问题. 以水稻超泛基因组为例, Shang等人[44]选取了来源于44个国家的251份水稻种质, 包括202份亚洲栽培稻(*O. sativa*, Os)、28份亚洲野生稻(*O. rufipogon*, Or)、11份非洲栽培稻(*O. glaberrima*, Og)和10份短舌野生稻(*O. barthii*, Ob)样本. Og, Ob和Or样本因其地理多样性而被收集. Os样本中的22个优良现代水稻品种因其高产、抗病性、氮素利用效率和其他关键农艺性状而被收集. 其余180份Os样本保留了50526份水稻品种的大部分遗传变异和表型变异. 简而言之, 选择样本的意图是确保研究的代表性, 涵盖广泛的遗传和表型多样性, 以便更全面地理解植物的驯化和改良过程.

根据物种特性、生物学背景和研究目标来进行样本数量选择. 样本选择过多会导致高成本和资源浪费, 而样本选择过少则可能无法获得全面的超泛基因组信息. 对于遗传多样性较高的物种, 需要更多样本以充分覆盖其遗传变异; 而对于遗传多样性较小的物种, 较少样本即可代表其基因组多样性. 当揭示基因型与特定表型性状(如抗旱性、抗病性等)的关系时, 需选择表型差异明显的样本, 避免无关样本干扰. 若构建属内丰富的基因组变异图谱, 则需要更多样本以确保对超泛基因组变异的全面覆盖. 最后, 可通过分析超泛基因组和核心基因组的饱和曲线来确定最优样本数量.

2.3 超泛基因组的构建方式

现阶段, 虽然超泛基因组在概念上拓展了泛基因组的研究范围, 但其构建方式和技术流程仍与泛基因组基本一致, 尚未针对超泛基因组需求进行专用软件工具的研发和优化. 目前构建泛基因组的方法有三种

(图2): 从头组装(*de novo* assembly)、迭代组装(iterative assembly)和基于图形组装(graph-based assembly)[66,69]. 从头组装是从测序数据中对多个样本进行从头组装和注释, 然后进行相互比较来构建泛基因组. 该策略需要高质量和高覆盖度的测序数据, 成本高、耗时长, 适用于小规模样本, 同时需要较大的计算资源, 但该策略不依赖参考基因组且能检测到更多的结构变异[70]. 拟南芥[71]、野生大豆[36]、绿豆[72]、水稻[73]等都是采用从头组装策略. 迭代组装是基于物种的单个参考基因组, 然后逐个将多个样本的测序数据比对到参考基因组. 对未比对上的数据单独组装, 非冗余序列合并到参考基因组上, 通过不断迭代, 逐步扩展和完善泛基因组. 该策略避免了组装和注释从而节省了计算资源, 成本低且速度快, 适用于大规模样本, 但该策略对于结构变异的检测效果不佳[70]. 甘蓝[74]、小麦[75]、番茄[40]、棉花[76]等都是采用迭代组装策略来构建. 图形泛基因组是将所有的样本之间的变异整合到泛基因组图谱中, 使用图形结构来表示基因组及其变异. 与线性基因组相比, 图形泛基因突破了传统线性基因组的存储形式, 可以更直观地展示物种的遗传信息和序列结构变异信息[70]. 该策略适用于大规模样本, 可以检测到更全面的变异信息, 但需要消耗更多的数据存储和计算资源[77]. 大豆[37]、黄瓜[41]、苹果[78]等都是基于图形组装来构建泛基因组.

超泛基因组拥有属内多个物种丰富的遗传资源库, 可以对属内遗传多样性进行全面的解析. 与泛基因组一致, 超泛基因组采用同样的三种构建策略. 截至目前, 西瓜[46]、鹰嘴豆[49]、水稻[44]、葡萄[48]、番茄[47]采用图形组装策略, 杨树[50]采用从头组装, 而马铃薯[57]、玉米[45]则采用迭代组装.

2.4　超泛基因组的分析工具

超泛基因组构建及分析过程主要包括高质量基因组组装与注释、核心与可变基因识别、构建超泛基因组、超泛基因组注释、遗传变异检测(如SV和PAV)以及可视化与下游分析. 构建超泛基因组的过程中涉及多种分析工具的协同(图3).

目前超泛基因组的构建主要基于图形构建策略. 图形超泛基因组构建方法可以分为三种类型: 基于参考基因组以及变异信息的构建方式、基于参考基因组直接构建方式、无参考基因组的构建方式[79]. 第一种构建方法最简单且高效, 依赖高质量基因组和上游比

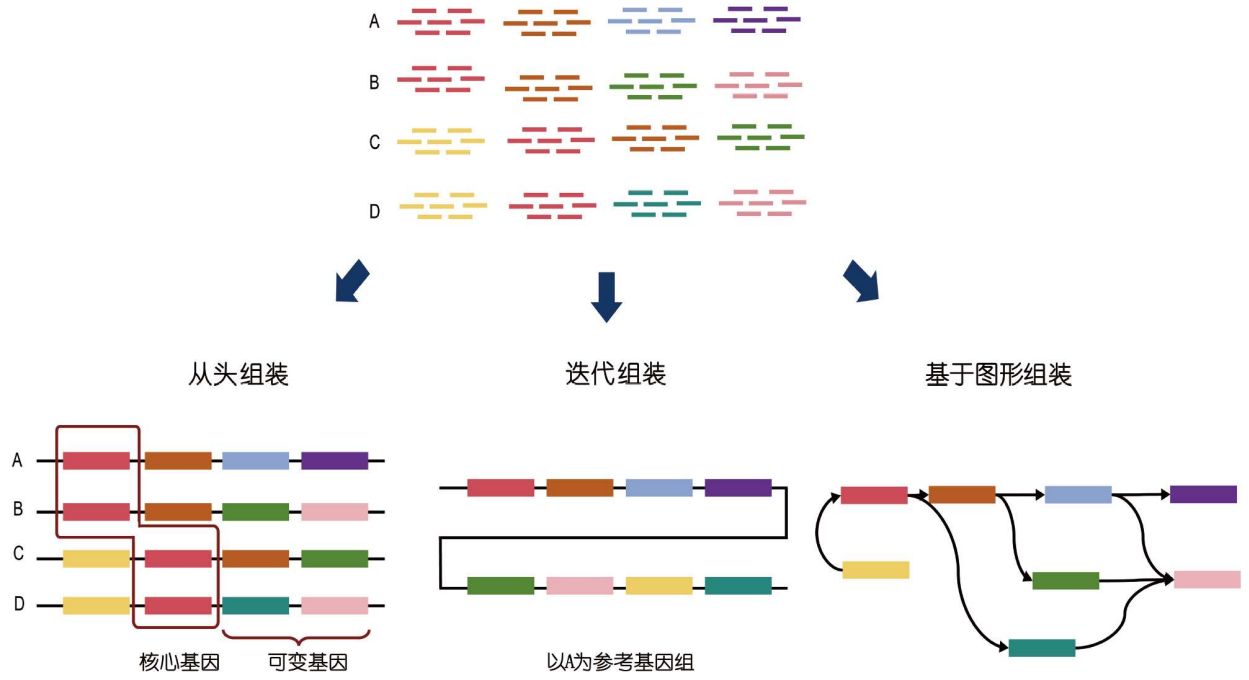

图 2　泛基因组三种构建方式
Figure 2　Three methods of pan-genome construction

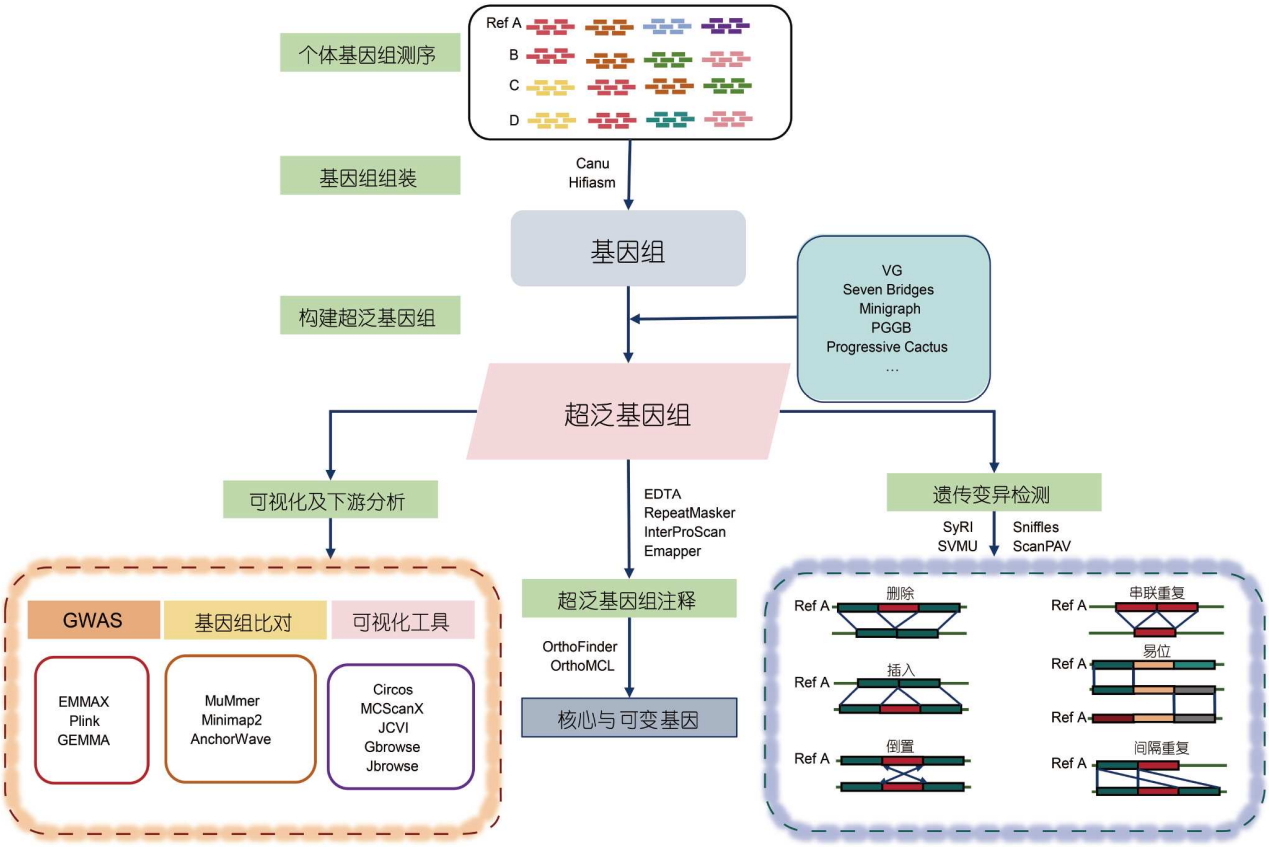

图 3 超泛基因组研究工作流程
Figure 3 Workflow of super-pangenome research

对步骤，但无法准确描述嵌套或更复杂的变异以及较长的插入序列[77]. 利用MuMmer[80], Minimap2[81], AnchorWave[82]等常用的全基因组比对工具将其他基因组与参考基因组进行比对，然后通过SyRI[83], SVMU[84]等工具识别变异并生成变异文件(如VCF). 小型遗传变异(<50 bp)和大型遗传变异(>50 bp)信息常储存在VCF格式文件中[85]. 使用variation graph toolkit (vg)[86]整合变异和参考基因组，生成图形化超泛基因组. vg计算速度较快，能够整合SNP, InDel和SV，依赖于上游比对步骤，可适配线性参考基因组的下游分析，但需要大量内存和计算资源[79]. 该方法已经被运用于番茄[47]、水稻[44]、鹰嘴豆[49]等物种的超泛基因组构建. 第二种构建方法保留了线形参考基因组的坐标和注释，可适配线性参考基因组的下游分析(如GWAS)及可视化(如Gbrowse)[77,87]. 然而，该方法存在参考基因组偏倚，可能忽略非参考基因组特有的遗传变异，且不同参考基因组或添加顺序会导致超泛基因组图谱不同[77]. 首先确定参考基因组，再将所有目标基因组输入构建软件(如Minigraph[88]和Minigraph-Cactus[89])，软件自动生成图形超泛基因组，无需独立进行变异分析. Minigraph计算速度快，能捕捉大于100 bp的SV，但无法捕捉SNP. Minigraph-Cactus结合了Minigraph高效图谱构建和Cactus多序列比对优势，扩展了变异范围，支持小范围变异(如SNPs和小型InDels)和大范围变异(SVs)，但在大型基因组和样本数量多的情况下计算资源需求较高[87]. 目前主要用于牛[90]等动物的超泛基因组构建. 第三种构建方法消除了参考基因组的偏倚，任何基因组都可作为参考. 然而，对于大型植物基因组，该方法需要大量的计算资源和时间，且需要多次调试最佳参数[77]. 该构建方法主要利用Progressive Cactus[91]和PanGenome Graph Builder(PGGB)[92]等工具，捕获不同基因组间的差异. PGGB集成了wfmash,

seqwish和smoothxg工具，分别进行全基因组比对、图谱的构建及结构优化处理．通过综合处理的方法，PGGB确保了超泛基因组图谱能够充分捕捉基因组的多样性，减少参考基因组偏差的潜在影响[87]. PGGB计算需求较高，特别是在处理具有大量重复序列的植物基因组时，在一定程度上限制了其广泛应用．Progressive Cactus适合多物种比较且比对质量较高，但计算资源需求大[87]．葡萄[48]超泛基因组构建采用了该方法．随着新工具和方法的不断涌现，基于图形的超泛基因组构建将成为未来的重要参考标准.

3　植物超泛基因组的应用

3.1　功能基因挖掘

构建完整的参考基因组并解析农艺性状的遗传基础对植物遗传改良至关重要[93]．超泛基因组蕴含着丰富的进化信息和大量有价值的基因资源，为功能基因挖掘提供了应用前景．Long等人[51]在稻属中共鉴定出101723个基因家族，核心基因家族达9.66%(9834个). 与栽培稻相比，野生稻存在63881个未发现的新基因家族．西瓜属超泛基因组大小约为单个西瓜基因组的1.5倍，与已报道的栽培种基因组T2T-G42相比，增加了399.2 Mb序列和11225个基因家族[46]．大豆属(涵盖一年生和多年生物种)的超级泛基因组共注释了109827个非冗余的蛋白质编码基因，其中约70%在一年生栽培大豆中未被检测到[43]．野生近缘种中丰富的新基因与结构变异为基因功能研究和育种应用提供了宝贵的遗传资源，使得可以有针对性地将在驯化过程中丧失的关键基因有效地整合到栽培优良品系中．在西瓜属中，通过QTL定位从野生种和栽培种杂交选育的重组自交系“PKR6”中鉴定出与抗枯萎病生理小种1 (Fon race 1)相关的Qfon1.1，该QTL位于Chr01的5 cM区域. 通过将PKR6的QTL区域基因组序列与易感品系G42进行比对，成功将QTL缩小至364 kb，进而从野生种中精确定位到特有的抗性基因[46].

随着超泛基因组数据的可用性不断提高，可以利用这些数据进行功能基因挖掘．Lv等人[94]基于超泛基因组数据，构建了一个涵盖属内多个核心种质资源的高质量水稻超泛着丝粒图谱，探究了不同染色体和亚群中着丝粒的多样性及其进化过程．此外，通过水稻超泛基因组的SV分析，在着丝粒区域鉴定出正向调控水稻分蘖数的*OsMAB*．植物野生种与栽培种之间存在显著影响农艺性状的SV，超泛基因组能够系统地解析这些关键序列的差异，为开发分子标记奠定基础. 超泛基因组可作为一个强大的基因分型平台，通过整合SV与SNP等数据，开展基于结构变异的全基因组关联分析(SV-GWAS)、基于单核苷酸多态性的全基因组关联分析(SNP-GWAS)等标记-性状的关联研究，在野生种质资源中精准定位到与重要农艺性状相关的遗传信号. Wei等人[95]从水稻超泛基因组的SNP中鉴定出与耐盐相关的eQTL，结合全基因组关联研究(genome-wide association study, GWAS)，快速定位到一个主要的耐盐性位点*qSTS5*．借助超泛基因组数据，能够快速识别与农艺性状相关的新基因，为基因功能挖掘提供了新的思路和方向.

3.2　探索属内物种驯化过程

植物驯化是一个复杂的进化过程．通过改变植物的形态和生理特征，使其更具适应性[96,97]．在驯化过程中，瓶颈效应和连续的人工选择导致栽培种的遗传多样性大幅减少．野生种丰富的遗传多样性蕴含了宝贵的育种材料，有益的野生种渗入能够促进植物遗传改良[47]．因此，通过整合野生种的基因组信息来构建属内的超泛基因组为植物驯化提供更好的视角[98].

在植物驯化过程中，基因选择性丢失和新基因获得等信息通常难以被准确鉴定．基于超泛基因组，可以鉴定属内每个个体的遗传变异，揭示属级分类中基因丢失和获得模式，帮助识别受选择的基因和区域. 部分SVs在物种演化和驯化过程中经历了选择性保留或去除. 通过比较野生稻和栽培稻的基因组，发现了多个与驯化相关的基因，如*SHAT1*，*PROG1*等，这些基因在亚洲和非洲水稻中独立发生了变异，导致不同的驯化性状[44]．野生番茄和栽培番茄之间存在244 bp的缺失，该缺失发生在编码细胞色素P450蛋白基因(*Sgal12g015720*)的第一个外显子中，*Sgal12g015720*基因主要影响番茄的侧枝数量和果实数量，这段244 bp的缺失可能在番茄驯化过程中发生，并对栽培番茄的性状产生重要影响[47]．在西瓜驯化过程中，伴随着多个与含糖量增加、果肉变红相关的基因簇扩张，大量抗病功能相关的基因簇丢失．Zhang等人[46]在西瓜超泛基因组中发现，苦味丧失、糖分积累和果肉颜色增加等关键性状与功能基因的SVs相关．在苦味丧失方

面, *ClBt*在野生种*C. colocynthis*和*C. amarus*中分别存在6和18 bp缺失, 影响苦味葫芦素E(CuE)的合成. 在糖分积累方面, *TST2*的拷贝数变异影响糖分含量, 栽培西瓜果实中表达较低, 而野生种几乎无表达. 在果肉颜色方面, *ClPHT4*;*2*的CDS在野生种*C. colocynthis*和*C. amarus*中分别存在6和12 bp缺失, 其表达量上调促进胡萝卜素在栽培西瓜中积累. 同时, 超泛基因组还可以帮助揭示属内不同物种在驯化过程中面对胁迫环境时的适应性机制. 例如, 亚洲和非洲水稻在适应淹没胁迫时采用不同基因策略. 亚洲水稻通过*Sub1A*基因的"淹没静止策略"来抗淹没, 而非洲水稻中该基因发生缺失. 相反, *SNORKEL1*/*2*和*ACE1*基因在两者中均存在, 并通过促进节间伸长来适应淹没胁迫[44].

核苷酸结合富含亮氨酸重复序列受体(nucleotide-binding domain leucine-rich repeat, NBS-LRR/NLR)是植物细胞内重要的免疫受体, 感知病原体的攻击并启动免疫反应. NLR基因常以成对或串联簇的形式排列. 通过在属级水平或更高分类单元构建超泛NLRome (super pan-NLRome)数据集, 可以有效解决GWAS和图位克隆在鉴定特定病害相关NLR基因研究中的效率低下问题[99]. 通过分析不同个体间NLR的差异, 捕获野生近缘种等非驯化种质中的NLR多样性, 有助于衡量超泛NLRome的复杂性. 基于24份野生种和20份栽培种的优质二倍体马铃薯种质构建了超泛NLRome, 共获得57683个NLRs, 并揭示了驯化过程中NLR显著扩张现象, 不同种质间NLR基因拷贝数差异显著(范围为478~1976个)[56]. 在2个栽培稻和21个野生稻的超泛基因组研究中, 构建了超泛NLRome, 共鉴定出7048个NLRs. 栽培稻NLRs数量高于野生稻, 出现了扩增现象. 然而, 栽培稻的抗病性和多样性低于野生稻, 可能在驯化和人工选择过程中丢失了一些抗性基因, 或特定抗性基因被固定和扩增所致[51]. 超泛NLRome的广泛应用, 使得在属级或更高分类单元水平上对不同物种个体间特定基因家族的比较成为可能. 总之, 超泛基因组有助于理解植物驯化过程, 并为精准育种和从头驯化提供新信息.

3.3 揭示属内物种的进化史

利用超泛基因组对不同物种基因组序列进行比较分析, 探讨物种间不同时间尺度上的进化历史, 揭示生物演化脉络[100]. 通过获取属内不同物种间的遗传变异(如SNP, SV等)深入探索植物的起源、进化、驯化过程以及基因流动. 倒位、CNV等结构变异以及TE转座活动是基因组进化的重要驱动力, 不仅在生物环境适应中扮演重要角色, 还通过影响基因流动、基因剂量以及基因组结构来塑造生物的进化路径[101].

染色体结构改变可能阻碍物种间的基因流动, 进而影响杂交后代的可育性, 并减少种间基因重组, 从而导致生殖隔离, 最终推动物种分化[102]. 以西瓜属为例, Wu等人[59]发现野生种*C. colocynthis*与近缘野生种(*C. amarus*, *C. mucosospermus*)及栽培种*C. lanatus*三个物种之间存在显著的染色体重排现象, 并推测*C. colocynthis*可能保留了祖先染色体核型. 通过系统发育分析, 将西瓜属物种的分化时间追溯至约4.54~2.41百万年前(Mya), 推测染色体结构变异可能是该属物种形成的主要驱动力. 为深入解析西瓜属的进化历程, Zhang等人[46]通过构建西瓜7个种的T2T水平超泛基因组, 在栽培种*C. lanatus*中识别出362个SV, 其中33个来自野生种*C. mucosospermus*, 68个来自亚种*C. lanatus* subsp. *cordophanus*, 另有200个为二者共有. 这表明, 除了*C. lanatus* subsp. *cordophanus*, 栽培西瓜谱系中还包含其他野生祖先物种的遗传成分.

转座子在SV形成中具有主导作用[50]. 在杨属中发现TEs驱动的SVs不仅通过改变编码序列和顺式调控元件影响基因表达, 还会引发局部染色质结构和表观遗传标记(如DNA甲基化)的变化, 进而对基因功能产生多维度调控. 其中Gypsy和Copia超家族的TEs与基因表达呈正相关, 而与Helitrons呈负相关[50]. 植物基因组中, TEs通过快速扩增、消除和转位等动态过程驱动基因组进化[47,103]. TEs的增减变化是植物基因组大小变化的主要来源, 使得植物基因组的复杂性远高于脊椎动物[104]. 例如, 在茄属[47]、杨属[50]和稻属[44]等超泛基因组中TEs含量较高, 物种间基因组大小的差异主要由TEs的扩张或收缩决定, I型反转录转座子占主导地位. 此外, 在马铃薯超泛基因组研究中发现, 野生马铃薯的基因数量显著少于地方种和栽培种, 表明驯化与育种过程伴随基因含量的动态调整. 同时, TEs含量在进化分支间存在显著差异, 进一步表明TEs在马铃薯属内基因组进化中起着重要作用[57].

基因组复制事件或多倍化事件影响着基因组大小、基因数目和基因功能[105]. 杨属超泛基因组研究揭示核心基因主要源自WGD衍生的重复基因, 这些基因

在属级水平上更保守. Shi等人[50]通过结合转录组、DNA甲基化和染色质可及性数据, 解析了杨属的进化轨迹, 发现WGD衍生的重复基因更易被保留. 其保留机制与高表达水平、低组织特异性、邻近区域低甲基化及高染色质可及性密切相关, 且这些基因主要富集于基础代谢和发育调控通路. 此外, WGD重复基因的序列进化和功能分化受到甲基化修饰的调控. 超泛基因组的多物种比较, 进一步表明WGD衍生的遗传多样性为植物适应性进化和物种多样化提供了重要的遗传基础.

3.4 属内遗传变异检测

相较于单一的线性参考基因组, 超泛基因组可以提供物种内和物种间不同类型的遗传变异, 如SNP, InDel, CNV, PAV等. 植物基因组中存在从SNP到大规模的SV[106]. 通过全基因组比较, 可以识别出单核苷酸变异(single nucleotide variant, SNV)、小于50 bp的插入缺失和大于50 bp的SV[107]. Wang等人[108]结合水稻超泛基因组对10548份水稻群体进行群体水平最大的自然变异分型, 构建出水稻超级变异图谱RSPVM, 包含了54378986个SNP和11119947个InDel, 其中84%的SNP和92%的InDel属于稀有变异.

与SNPs相比, SV对适应性进化、功能基因变异和物种多样化影响更大[101]. SV是植物进化和驯化的主要驱动力[109]. Shi等人[50]发现, 杨属SV热点区域富含防御反应、次生代谢物合成和信号转导相关基因, 这在适应生物和非生物胁迫中起重要作用. SVs是个体间的遗传差异, 包括缺失、插入、CNV、倒位和易位[110]. 这些变异可以导致基因丢失、基因复制和新基因的产生, 从而引起物种的表型变异[111]. 例如, 在水稻超泛基因组中发现与籽粒重相关的SVs通过影响其附近基因(*HGW*, *OsNaPRT1*等)的表达来决定性状[44]. SVs影响玉米超泛基因组中的*Zm00001d023299*基因在干旱胁迫下的表达[45]. 不同杨属物种(Clade-Ⅰ和Clade-Ⅱ)研究发现1号染色体上约104 kb的倒位区域与叶缘形态差异相关, 该区域内的关键基因*CUC2*是调控叶缘锯齿形成的主要因素. Clade-Ⅱ物种中*CUC2*启动子区域180 bp的插入导致其表达水平显著高于Clade-Ⅰ, 使Clade-Ⅱ叶缘产生更多锯齿[50]. 在猕猴桃属中发现*AtPIR*同源基因*Ach22g02880DH*外显子有55 bp插入, *WER*同源基因*Ach19g03580DH*内含子有114 bp缺失, *CPC*同源基因*Ach13g13590DH*和*GL3*同源基因*Ach25g04680DH*的启动子上分别存在218和205 bp插入, 这些SVs影响着猕猴桃毛状体发育[58].

3.5 遗传改良与适应性育种

基于基因组学辅助育种的方法已成功应用于多种作物, 其中泛基因组技术有效地描绘了物种内的遗传变异[70]. 在水稻[38]和黄瓜[41]等作物泛基因组中, 已成功挖掘出与复杂性状相关的关键基因[54]. 传统泛基因组难以有效解析从远缘杂交物种或更高分类群(如属、科)基因库中渗入的基因或遗传变异. 相比于泛基因组, 超泛基因组更加注重更高分类单元种间遗传变异的挖掘, 能更精准地解析驯化过程中渗入的序列片段. 在多倍体植物研究中, 超泛基因组能够在属级水平解析倍性变化、大范围基因组重排、亚基因组起源、演化路径及其功能分化, 还能识别多倍化过程亚基因组中与抗病性、耐旱性和产量等性状密切相关的遗传变异和关键基因. 超泛基因组极大地拓展了遗传改良可利用的基因池, 帮助育种者更有针对性地育种改良, 从而培育出更优质、更具适应性的品种.

超泛基因组研究揭示了物种间及物种内的遗传多样性. 回交可以将野生种质中的目标基因引入栽培种, 从而增加栽培种的遗传多样性. 然而, 大规模的倒位会发生重组抑制, 在回交育种过程中导致连锁累赘, 即引入不利的等位基因[112,113]. 超泛基因组有助于识别出不同种质间的大规模倒位, 帮助选择不含倒位片段的供体亲本. 以番茄为例, *S. pennellii*番茄不含3号染色体上7.1 Mb的倒位片段, 可作为理想的供体亲本. 通过回交将该片段中有利的基因引入优良品种中, 达到改良番茄品种的目的[47].

自交不亲和及自交衰退是妨碍马铃薯育种进程的两大障碍[114]. 马铃薯基因组中大量杂合有害突变的积累阻碍着优良自交系的发展. 为了推动马铃薯基因组设计育种体系的构建, Wu等人[115]开发了茄科的“进化透镜”方法. 该方法利用192份二倍体马铃薯的SNP位点数据, 其中44份样本来自马铃薯超泛基因组数据[56], 并结合95份茄科(涵盖87个物种)和5份旋花科(涵盖5个物种)基因组的进化保守位点, 成功鉴定出有害突变[115]. 在该研究中, 超泛基因组数据为构建更高效的育种模型和筛选优良基因提供了数据支撑[115]. 利用野生种质资源挖掘优异基因在遗传改良中具有重要意

义. Eggers和Ma[116,117]在野生二倍体马铃薯*S. chacoense*克隆了自交不亲和抑制基因*Sli*, 该基因编码F-box蛋白, 并通过与多种类型S-核糖核酸酶(S-RNase)互作来克服自交不亲和, 该研究为用二倍体杂交种子替代四倍体薯块繁殖提供了可能性. 通过将栽培西瓜*C. lanatus*与野生种*C. amarus*和*C. mucosospermus*进行种间杂交, 选育出具有多种病害抗性基因的"PKR6"重组自交系. 栽培西瓜与野生种的远缘杂交, 不仅通过引入驯化过程丢失的抗病基因来增强其抗病性, 还增加了新的遗传多样性, 对长期可持续育种和改良至关重要[46]. 高效的育种模型能有效缩短育种年限、降低育种成本、提高育种效率. Wei等人[118]构建了水稻数量性状核苷酸(quantitative trait nucleotide, QTN)综合图谱, 开发了基因组导航系统RiceNavi并在水稻主栽品种"HHZ"中成功应用. "HHZ"借助RiceNavi的选配指导和路线优化, 获得株型紧凑、生育期短、有香味的改良型"HHZ". 此外, 大规模的基因组数据基于机器学习(machine learning, ML)能有效开发分子标记和应用于全基因组选择育种模型[55], 如智慧育种平台Smart Breeding Platform[119]和Oryza CLIMtools[120]等. 这些应用充分体现出未来超泛基因组数据在遗传改良和智能育种方面的巨大潜力.

3.6 超泛基因组数据库的构建

随着植物测序数据的爆发式增长, 这些蕴含巨大潜力的数据推动了物种数据库的构建. 植物超泛基因组的综合性数据库平台是一个强有力的工具, 为基因组学、功能基因组学和系统生物学等基础研究提供了宝贵资源, 也可为未来分子育种和农业生产提供重要支持[70]. 截至目前, 已构建了一些植物的超泛基因组综合性数据库, 如稻属[44]、杨属[50]、西瓜属[46]. RiceSuperPIRdb数据库提供了稻属的SVs、基因注释、TE注释、超泛基因组图谱等信息, 同时集成了BLAST等多种分析工具. 通过整合稻属的基因组资源并呈现可视化数据集, 方便挖掘与农艺性状相关的关键遗传变异位点(如SNPs, CNVs和PAVs), 进一步推动水稻功能基因组学研究[44]. WaGMDB数据库整合了西瓜属的基因组、转录组、代谢组等多组学数据, 并整合EMS突变体信息及图形超泛基因组. 同时, 该数据库集成JBrowse, BLAST, Primer design, Batch Query等多种分析工具. 基于花粉EMS诱变技术, 构建了G42西瓜EMS突变体库. 该数据库有利于继续挖掘基因资源、推动西瓜种质创新[46]. PSIR为杨属超泛基因组数据库, 可以访问泛基因类型、功能注释及结构变异, 支持在物种内和物种间搜索直系同源和旁系同源基因, 同时可进行多组学的资源访问和下载, 为杨属遗传解析和分子育种改良提供重要的遗传资源和参考信息[50].

4 展望

4.1 更高分类水平的进阶

植物基因组庞大且复杂, 未来可以整合更高分类水平的基因组数据来探索植物进化. 目前, 原核生物的泛基因组已突破物种甚至达到门的界限, Maistrenko等人[121]使用来自10个原核门的7104个高质量基因组构建了门级水平的超泛基因组[66]. 由于原核生物基因组较小且为单倍体, 因此该研究在计算上是可行的. 目前植物的超泛基因组研究范围达到属级甚至更高层次. Huang等人[122]利用12个新组装基因组和已发表的6个基因组构建了柑橘亚科的超泛基因组图谱, 揭示了柑橘亚科的起源与演化历程, 并阐明了*PH4*在柑橘中对柠檬酸积累具有核心作用. Ma等人[123]构建了来自11个竹亚科不同属代表性物种的亚科级超泛基因组, 包括2种草本竹子和9种木本竹子, 涵盖了从二倍体(草本)到四倍体和六倍体(木本)的谱系. 该研究发现木本竹子亚基因组的核型稳定性及优势进化, 揭示了多倍体化对基因组演化和物种多样性的影响. 虽然超泛基因组在不同物种及更高分类层次的比较中具有巨大潜力, 但面临的技术挑战也不容忽视. 不同物种在染色体数目、倍性、基因组大小、重复序列分布以及序列差异(如SNP, SV, WGD事件)等方面存在显著差异, 这些差异导致基因组比对率偏低、基因数目统计不一致等问题. 高杂合性、高重复性、异源多倍体、大型基因组(>10 Gb)的组装面临巨大挑战. 为了克服这些困难, 需要开发新的工具或优化现有的组装、注释和比对工具, 进行基于图形超泛基因组的比对算法创新, 建立多倍体专用比对流程, 采用先进的测序技术等手段. 长读长测序技术(如PacBio和ONT)可生成10~100 kb, 甚至达到1 Mb的读长, 在组装过程可以结合其他技术进行解析. 例如, 四倍体马铃薯栽培种"Otava"的单倍型基因组重建采用了高精度的长读长测序、717个二倍体

花粉基因组的单细胞测序和Hi-C数据[124]. 十二倍体甘蔗现代栽培种“R570”基因组的组装综合运用了PacBio HiFi长读长测序、Illumina短读长测序、Bionano光学图谱、遗传图谱和Hi-C数据等技术[125].

通过与相关领域的植物分类学家进行合作, 可以有效地增加样本的多样性和代表性. 同时, 整合已有的基因组数据和样本, 将有助于提升对多样性研究的全面性和深度. 随着测序成本下降和计算能力提升, 未来植物超泛基因组研究有望扩展到更大的分类单元, 超越属级和亚科层次, 进入科甚至目的水平[66]. 未来, 在更广泛的绿色植物王国中构建超泛基因组将是一个重要的研究方向, 从整个分类学上提高我们对植物基因组的认识.

4.2 高分辨率超泛基因组的构建

对于已经具有参考基因组的植物物种, 基因组研究已经转向生成更高质量的基因组, 比如端粒到端粒T2T基因组、单倍型分型基因组(haplotype-resolved genome)等. 目前, 准确性高、连续性高和完整性高的T2T基因组已成为高质量基因组组装的新标准, 广泛应用于植物基因组研究[126]. 单倍型超泛基因组(haplotype super-pangenome)和端粒到端粒超泛基因组(T2T super-pangenome)代表了未来植物基因组研究和应用的前沿方向.

单倍型超泛基因组通过高分辨率的单倍型组装, 在等位基因特异性变异和复杂基因区域的解析上有独特优势. T2T超泛基因组则通过解决基因组中高度重复序列和复杂结构区域, 实现了基因组的完整连续组装, 有助于解析端粒、着丝粒等复杂结构的变异特征和进化模式. 目前, 在西瓜属中构建了首个属级T2T水平的超泛基因组[46]. 在葡萄属中通过对9份北美野生葡萄进行单倍型分离, 构建了18个单倍型参考基因组, 完成了葡萄属超泛基因组的构建[48]. 这些工作为植物基因组学研究提供了新的视角和工具. 尽管PacBio HiFi和ONT长读长测序技术取得了进展, 但对于许多物种来说, 构建单倍型基因组和无缺口的T2T基因组仍然困难. 植物基因组具有高杂合性、高重复性、高倍性的特性以及染色体端粒和着丝粒区域的复杂性, 给单倍型分离和T2T基因组构建带来了极大的技术挑战. 随着测序技术的进步, 未来更高效、更高精度的测序方法将不断被研发出来, 这将为高分辨率超泛基因组的构建提供更加可靠的数据基础及遗传变异检测.

4.3 多组学整合的创新应用

高通量测序技术的进步为多组学研究铺平了道路. 应用基因组学、转录组学、蛋白质组学、代谢组学和表观遗传学等多组学生物数据将有助于揭示植物生长发育和适应环境的全局机制[127]. 随着NGS技术产生的大规模测序数据, 多组学技术已开始应用于超泛基因组研究. 杨树超泛基因组研究利用转录组、重亚硫酸盐测序和转座酶可及染色质测序(assay for transposase-accessible chromatin using sequencing, ATAC-seq)技术, 探讨了杨属多个物种间的基因表达模式、表观遗传特征(如DNA甲基化)以及调控结构(染色质可及性区域), 并通过多组学整合深入了解了杨属的遗传调控机制[50]. 多组学整合不仅涵盖多个维度的数据类型, 还强调各类数据之间的交互, 能够全面反映植物在不同环境条件下的生长发育、代谢过程和应对机制, 提供更加全面和精准的植物生物学图谱. 未来, 可充分利用前沿的多组学数据进行系统性分析, 识别并验证关键候选基因, 揭示植物发育及其应对压力的分子机制, 从而为植物遗传改良提供新的策略.

4.4 数据算法与工具的开发

首先, 处理和分析属内甚至更高分类单元多样性高且复杂而庞大的数据集需要强大的计算资源和高效算法. vg[86], Minigraph[88]等工具提供了完整的图形超泛基因组构建与分析流程, 支持快速生成和后续分析. 但随着基因组数据量急剧增加, 传统数据处理方法难以应对“海量数据”. 为了实现高效的超泛基因组构建, 必须开发新的算法和工具, 以支持数据存储、分析和可视化. 图形超泛基因组能够有效整合遗传变异, 但其存储和可视化仍是主要技术难题[66]. 常见线性超泛基因组可视化工具主要包括Gbrowse[128], Jbrowse[129]等. 图形超泛基因组常见的宏观可视化工具有Bandage[130], GfaViz[131], Sequence Tube Map[132]等, 微观可视化工具有vg viz[85], ODGI[133]等[87]. 尽管目前有多种可视化工具能直观地展示基因组信息, 但在处理大规模数据集、多样化变异、复杂度高和基因组庞大的物种难以可视化. 其次, 当前超泛基因组分析往往缺乏统一标准, 样本数量、测序深度、构建策略和序列注释方法各不相同, 导致结果的可比性和可重复性受到

影响. SV基因分型在识别嵌套变异、大规模倒位或易位等复杂变异类型时具有挑战性. 尽管短读长测序的基因分型方法(如Graphtyper2[134], BayesTyper[135]等)已有应用, 但在复杂变异的解析上存在局限. 采用高精度的长读长测序基因分型方法, 如Sniffles[136], cuteSV[137]等, 能更有效识别这些变异, 通过更长的读长跨越重复区域, 减少比对错误, 从而提高准确性. 因此, 结合长读长测序技术是提升结构变异基因分型准确性的关键. 图形超泛基因组能够存储各种遗传变异, 但如何有效整合和利用这些数据仍然是一个挑战. 此外, 针对复杂植物基因组的高质量组装和注释算法缺乏, 也限制了超泛基因组研究的发展.

为了克服这些困难, 需要不断探索新的技术和策略, 如利用长读长测序技术和改进算法, 提高基因组组装的质量和效率. 其次, 需要加强对数据处理工具的开发, 包括优化存储方式和建立标准格式. 此外, 机器学习和深度学习等先进算法可用于数据挖掘和结构变异识别, 解析基因型与性状间复杂关系, 帮助从大规模、噪声较大的数据中预测重要农艺性状. 例如, 基于机器学习, 利用葡萄超泛基因组对466份重测序数据进行了SNP, InDel和SVs等复杂遗传变异的检测, 并构建了预测模型. 根据模型评分, 实现对早期个体农艺性状的预测和选择, 充分展示机器学习模型在植物性状预测中的可行性[55]. 这些新兴技术有助于实现对超泛基因组和多组学数据的高效整合与分析, 从而推动对植物基因组学的深入理解.

参考文献

1 Cheng S, Melkonian M, Smith S A, et al. 10KP: a phylodiverse genome sequencing plan. GigaScience, 2018, 7: 1–9
2 Lewin H A, Robinson G E, Kress W J, et al. Earth BioGenome Project: sequencing life for the future of life. Proc Natl Acad Sci USA, 2018, 115: 4325–4333
3 Liu Y, Tian Z. From one linear genome to a graph-based pan-genome: a new era for genomics. Sci China Life Sci, 2020, 63: 1938–1941
4 Hübner S. Are we there yet? Driving the road to evolutionary graph-pangenomics. Curr Opin Plant Biol, 2022, 66: 102195
5 Bohra A, Kilian B, Sivasankar S, et al. Reap the crop wild relatives for breeding future crops. Trends Biotechnol, 2022, 40: 412–431
6 Sanger F, Coulson A R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol, 1975, 94: 441–448
7 Maxam A M, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci USA, 1977, 74: 560–564
8 Sanger F, Air G M, Barrell B G, et al. Nucleotide sequence of bacteriophage φX174 DNA. Nature, 1977, 265: 687–695
9 Miller J R, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics, 2010, 95: 315–327
10 Kaul S, Koo H, Jenkins J, et al. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 2000, 408: 796–815
11 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
12 Michael T P, VanBuren R. Building near-complete plant genomes. Curr Opin Plant Biol, 2020, 54: 26–33
13 Sohn J, Nam J W. The present and future of *de novo* whole-genome assembly. Brief Bioinform, 2016, 19: 23–40
14 Edwards D, Batley J. Plant genome sequencing: Applications for crop improvement. Plant Biotechnol J, 2010, 8: 2–9
15 Imelfort M, Edwards D. *De novo* sequencing of plant genomes using second-generation technologies. Brief BioInf, 2009, 10: 609–618
16 Eid J, Fehr A, Gray J, et al. Real-time DNA sequencing from single polymerase molecules. Science, 2009, 323: 133–138
17 Clarke J, Wu H C, Jayasinghe L, et al. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotech, 2009, 4: 265–270
18 VanBuren R, Bryant D, Edger P P, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature, 2015, 527: 508–511
19 Berlin K, Koren S, Chin C S, et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol, 2015, 33: 623–630
20 Jiao W B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol, 2017, 36: 64–70
21 Zhang L, Hu J, Han X, et al. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nat Commun, 2019, 10: 1494

22 Yue J, VanBuren R, Liu J, et al. SunUp and Sunset genomes revealed impact of particle bombardment mediated transformation and domestication history in papaya. Nat Genet, 2022, 54: 715–724

23 Hu G, Feng J, Xiang X, et al. Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. Nat Genet, 2022, 54: 73–83

24 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

25 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

26 Chin C S, Alexander D H, Marks P, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods, 2013, 10: 563–569

27 Hu T, Chitnis N, Monos D, et al. Next-generation sequencing technologies: an overview. Hum Immunol, 2021, 82: 801–811

28 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

29 Lieberman-Aiden E, van Berkum N L, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 2009, 326: 289–293

30 Lam E T, Hastie A, Lin C, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol, 2012, 30: 771–776

31 Zhang L, Zhou X, Weng Z, et al. Assessment of human diploid genome assembly with 10x Linked-Reads data. GigaScience, 2019, 8: giz141

32 Xie L, Gong X, Yang K, et al. Technology-enabled great leap in deciphering plant genomes. Nat Plants, 2024, 10: 551–566

33 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

34 Tettelin H, Masignani V, Cieslewicz M J, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci USA, 2005, 102: 13950–13955

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

36 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

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

38 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

39 Wang B, Hou M, Shi J, et al. *De novo* genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. Nat Genet, 2023, 55: 312–323

40 Alonge M, Wang X, Benoit M, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell, 2020, 182: 145–161.e23

41 Li H, Wang S, Chai S, et al. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. Nat Commun, 2022, 13: 682

42 Khan A W, Garg V, Roorkiwal M, et al. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends Plant Sci, 2020, 25: 148–158

43 Zhuang Y, Wang X, Li X, et al. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. Nat Plants, 2022, 8: 233–244

44 Shang L, Li X, He H, et al. A super pan-genomic landscape of rice. Cell Res, 2022, 32: 878–896

45 Gui S, Wei W, Jiang C, et al. A pan-*Zea* genome map for enhancing maize improvement. Genome Biol, 2022, 23: 178

46 Zhang Y, Zhao M, Tan J, et al. Telomere-to-telomere *Citrullus* super-pangenome provides direction for watermelon breeding. Nat Genet, 2024, 56: 1750–1761

47 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

48 Cochetel N, Minio A, Guarracino A, et al. A super-pangenome of the North American wild grape species. Genome Biol, 2023, 24: 290

49 Khan A W, Garg V, Sun S, et al. Cicer super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement

in chickpea. Nat Genet, 2024, 56: 1225–1234
50 Shi T, Zhang X, Hou Y, et al. The super-pangenome of *Populus* unveils genomic facets for its adaptation and diversification in widespread forest trees. Mol Plant, 2024, 17: 725–746
51 Long W, He Q, Wang Y, et al. Genome evolution and diversity of wild and cultivated rice species. Nat Commun, 2024, 15: 9994
52 Wang M, Li J, Qi Z, et al. Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. Nat Genet, 2022, 54: 1959–1971
53 Gao L, Gonda I, Sun H, et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet, 2019, 51: 1044–1051
54 Zhou Y, Zhang Z, Bao Z, et al. Graph pangenome captures missing heritability and empowers tomato breeding. Nature, 2022, 606: 527–534
55 Liu Z, Wang N, Su Y, et al. Grapevine pangenome facilitates trait genetics and genomic breeding. Nat Genet, 2024, 56: 2804–2814
56 Tang D, Jia Y, Zhang J, et al. Genome evolution and diversity of wild and cultivated potatoes. Nature, 2022, 606: 535–541
57 Bozan I, Achakkagari S R, Anglin N L, et al. Pangenome analyses reveal impact of transposable elements and ploidy on the evolution of potato species. Proc Natl Acad Sci USA, 2023, 120: e2211117120
58 Yu X, Qu M, Wu P, et al. Super pan-genome reveals extensive genomic variations associated with phenotypic divergence in Actinidia. Mol Hortic, 2025, 5: 4
59 Wu S, Sun H, Gao L, et al. A *Citrullus* genus super-pangenome reveals extensive variations in wild and cultivated watermelons and sheds light on watermelon evolution and domestication. Plant Biotechnol J, 2023, 21: 1926–1928
60 Liu J N, Yan L, Chai Z, et al. Pan-genome analyses of 11 *Fraxinus* species provide insights into salt adaptation in ash trees. Plant Commun, 2025, 6: 101137
61 Li Y, Zhang B, Zhang S, et al. Pangeneric genome analyses reveal the evolution and diversity of the orchid genus *Dendrobium*. Nat Plants, 2025, 11: 421–437
62 Xia X M, Du H L, Hu X D, et al. Genomic insights into adaptive evolution of the species-rich cosmopolitan plant genus *Rhododendron*. Cell Rep, 2024, 43: 114745
63 Wang T, Duan S, Xu C, et al. Pan-genome analysis of 13 *Malus* accessions reveals structural and sequence variations associated with fruit traits. Nat Commun, 2023, 14: 7377
64 Golicz A A, Bayer P E, Bhalla P L, et al. Pangenomics comes of age: from bacteria to plant and animal applications. Trends Genet, 2020, 36: 132–145
65 Tranchant-Dubreuil C, Rouard M, Sabot F. Plant pangenome: impacts on phenotypes and evolution. Annu Plant Rev, 2019, 2
66 Bayer P E, Golicz A A, Scheben A, et al. Plant pan-genomes are the new reference. Nat Plants, 2020, 6: 914–920
67 Chavan S, Karla U. Concept of pan-genomics in crop improvement. Agric Food E-Newsletter, 2022, 12: 156–169
68 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
69 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 [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]
70 Li W, Liu J, Zhang H, et al. Plant pan-genomics: recent advances, new challenges, and roads ahead. J Genet Genomics, 2022, 49: 833–846
71 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
72 Liu C, Wang Y, Peng J, et al. High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. Plant Commun, 2022, 3: 100352
73 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
74 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
75 Montenegro J D, Golicz A A, Bayer P E, et al. The pangenome of hexaploid bread wheat. Plant J, 2017, 90: 1007–1013
76 Li J, Yuan D, Wang P, et al. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. Genome Biol, 2021, 22: 119

77 Hu H, Li R, Zhao J, et al. Technological development and advances for constructing and analyzing plant pangenomes. Genome Biol Evol, 2024, 16: evae081

78 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

79 Wang S, Qian Y Q, Zhao R P, et al. Graph-based pan-genomes: increased opportunities in plant genomics. J Exp Bot, 2023, 74: 24–39

80 Marçais G, Delcher A L, Phillippy A M, et al. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol, 2018, 14: e1005944

81 Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 2018, 34: 3094–3100

82 Song B, Marco-Sola S, Moreto M, et al. AnchorWave: sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. Proc Natl Acad Sci USA, 2022, 119: e2113075119

83 Goel M, Sun H, Jiao W B, et al. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol, 2019, 20: 277

84 Chakraborty M, Emerson J J, Macdonald S J, et al. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun, 2019, 10: 4872

85 Hickey G, Heller D, Monlong J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol, 2020, 21: 35

86 Garrison E, Sirén J, Novak A M, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol, 2018, 36: 875–879

87 Hu H, Wang J, Nie S, et al. Plant pangenomics, current practice and future direction. Agr Commun, 2024, 2: 100039

88 Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. Genome Biol, 2020, 21: 265

89 Hickey G, Monlong J, Ebler J, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. Nat Biotechnol, 2024, 42: 663–673

90 Leonard A S, Crysnanto D, Mapel X M, et al. Graph construction method impacts variation representation and analyses in a bovine super-pangenome. Genome Biol, 2023, 24: 124

91 Armstrong J, Hickey G, Diekhans M, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature, 2020, 587: 246–251

92 Garrison E, Guarracino A, Heumos S, et al. Building pangenome graphs. Nat Methods, 2024, 21: 2008–2012

93 Shang L, He W, Wang T, et al. A complete assembly of the rice Nipponbare reference genome. Mol Plant, 2023, 16: 1232–1236

94 Lv Y, Liu C, Li X, et al. A centromere map based on super pan-genome highlights the structure and function of rice centromeres. J Integr Plant Biol, 2024, 66: 196–207

95 Wei H, Wang X, Zhang Z, et al. Uncovering key salt-tolerant regulators through a combined eQTL and GWAS analysis using the super pan-genome in rice. Natl Sci Rev, 2024, 11: nwae043

96 Diamond J. Evolution, consequences and future of plant and animal domestication. Nature, 2002, 418: 700–707

97 Hancock J F. Contributions of domesticated plant studies to our understanding of plant evolution. Ann Bot, 2005, 96: 953–963

98 Raza A, Bohra A, Garg V, et al. Back to wild relatives for future breeding through super-pangenome. Mol Plant, 2023, 16: 1363–1365

99 Jayakodi M, Shim H, Mascher M. What are we learning from plant pangenomes? Annu Rev Plant Biol, 2025, 76: 663–686

100 Kersey P J. Plant genome sequences: past, present, future. Curr Opin Plant Biol, 2019, 48: 1–8

101 Wellenreuther M, Mérot C, Berdan E, et al. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. Mol Ecol, 2019, 28: 1203–1209

102 Baack E, Melo M C, Rieseberg L H, et al. The origins of reproductive isolation in plants. New Phytol, 2015, 207: 968–984

103 Chen J, Lu L, Benjamin J, et al. Tracking the origin of two genetic components associated with transposable element bursts in domesticated rice. Nat Commun, 2019, 10: 641

104 Kress W J, Soltis D E, Kersey P J, et al. Green plant genomes: what we know in an era of rapidly expanding opportunities. Proc Natl Acad Sci USA, 2022, 119: e2115640118

105 Ren R, Wang H, Guo C, et al. Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. Mol Plant, 2018, 11: 414–428

106 Saxena R K, Edwards D, Varshney R K. Structural variations in plant genomes. Brief Funct Genomics, 2014, 13: 296–307

107 Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. Hum Genome Var, 2024, 11: 18
108 Wang T, He W, Li X, et al. A rice variation map derived from 10548 rice accessions reveals the importance of rare variants. Nucleic Acids Res, 2023, 51: 10924–10933
109 Allaby R. Clonal crops show structural variation role in domestication. Nat Plants, 2019, 5: 915–916
110 Feulner P G D, De Kayne R. Genome evolution, structural rearrangements and speciation. J Evol Biol, 2017, 30: 1488–1490
111 Yuan Y, Bayer P E, Batley J, et al. Current status of structural variation studies in plants. Plant Biotechnol J, 2021, 19: 2153–2163
112 Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends Ecol Evol, 2018, 33: 427–440
113 Huang K, Rieseberg L H. Frequency, origins, and evolutionary role of chromosomal inversions in plants. Front Plant Sci, 2020, 11: 296
114 Jansky S H, Spooner D M. The evolution of potato breeding. Plant Breed Rev, 2018, 41: 169–214
115 Wu Y, Li D, Hu Y, et al. Phylogenomic discovery of deleterious mutations facilitates hybrid potato breeding. Cell, 2023, 186: 2313–2328.e15
116 Eggers E J, van der Burgt A, van Heusden S A W, et al. Neofunctionalisation of the Sli gene leads to self-compatibility and facilitates precision breeding in potato. Nat Commun, 2021, 12: 4141
117 Ma L, Zhang C, Zhang B, et al. A nonS-locus F-box gene breaks self-incompatibility in diploid potatoes. Nat Commun, 2021, 12: 4142
118 Wei X, Qiu J, Yong K, et al. A quantitative genomics map of rice provides genetic insights and guides breeding. Nat Genet, 2021, 53: 243–253
119 Li H, Li X, Zhang P, et al. Smart breeding platform: a web-based tool for high-throughput population genetics, phenomics, and genomic selection. Mol Plant, 2024, 17: 677–681
120 Ferrero-Serrano Á, Chakravorty D, Kirven K J, et al. Oryza CLIMtools: a genome–environment association resource reveals adaptive roles for heterotrimeric G proteins in the regulation of rice agronomic traits. Plant Commun, 2024, 5: 100813
121 Maistrenko O M, Mende D R, Luetge M, et al. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. ISME J, 2020, 14: 1247–1259
122 Huang Y, He J, Xu Y, et al. Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. Nat Genet, 2023, 55: 1964–1975
123 Ma P F, Liu Y L, Guo C, et al. Genome assemblies of 11 bamboo species highlight diversification induced by dynamic subgenome dominance. Nat Genet, 2024, 56: 710–720
124 Sun H, Jiao W B, Krause K, et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. Nat Genet, 2022, 54: 342–348
125 Healey A L, Garsmeur O, Lovell J T, et al. The complex polyploid genome architecture of sugarcane. Nature, 2024, 628: 804–810
126 Garg V, Bohra A, Mascher M, et al. Unlocking plant genetics with telomere-to-telomere genome assemblies. Nat Genet, 2024, 56: 1788–1799
127 Yang Y, Saand M A, Huang L, et al. Applications of multi-omics technologies for crop improvement. Front Plant Sci, 2021, 12: 563953
128 Donlin M J. Using the Generic Genome Browser(GBrowse). Curr Protoc Bioinf, 2009, 28: 9.
129 Buels R, Yao E, Diesh C M, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol, 2016, 17: 1–2
130 Wick R R, Schultz M B, Zobel J, et al. Bandage: interactive visualization of *de novo* genome assemblies. Bioinformatics, 2015, 31: 3350–3352
131 Gonnella G, Niehus N, Kurtz S, et al. GfaViz: flexible and interactive visualization of GFA sequence graphs. Bioinformatics, 2019, 35: 2853–2855
132 Beyer W, Novak A M, Hickey G, et al. Sequence tube maps: making graph genomes intuitive to commuters. Bioinformatics, 2019, 35: 5318–5320
133 Guarracino A, Heumos S, Nahnsen S, et al. ODGI: understanding pangenome graphs. Bioinformatics, 2022, 38: 3319–3326
134 Eggertsson H P, Kristmundsdottir S, Beyter D, et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nat Commun, 2019, 10: 5402
135 Sibbesen J A, Maretty L, Krogh A. Accurate genotyping across variant classes and lengths using variant graphs. Nat Genet, 2018, 50: 1054–1059
136 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
137 Jiang T, Liu Y, Jiang Y, et al. Long-read-based human genomic structural variation detection with cuteSV. Genome Biol, 2020, 21: 189

Current status and prospects of super-pangenome studies in plants

HUANG XiaoQin[1], HU LiSong[1,2,3], FAN Rui[1,2,3], ZHANG XingTan[4*] & HAO ChaoYun[1,2,3*]

[1] *Spice and Beverage Research Institute, Chinese Academy of Tropical Agricultural Sciences, Wanning 571533, China*
[2] *Key Laboratory of Genetic Resources Utilization of Spice and Beverage Crops, Ministry of Agriculture and Rural Affairs, Wanning 571533, China*
[3] *Hainan Provincial Key Laboratory of Genetic Improvement and Quality Regulation for Tropical Spice and Beverage Crops, Wanning 571533, China*
[4] *Shenzhen Key Laboratory of Agricultural Genomics, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China*
* *Corresponding authors, E-mail: zhangxingtan@caas.cn; haochy79@163.com*

In recent years, the rapid development of high-throughput sequencing technology has greatly promoted the process of plant whole genome sequencing. However, a single reference genome cannot completely cover the entire genetic information of a species, thus limiting the in-depth study of genomics. By integrating the genomic data of multiple representative individuals, pan-genome effectively overcomes the limitations of a single reference genome, and it can more comprehensively display the genetic diversity of a species including gene structure and variation, opening up a new direction for genomics research. However, pan-genome focuses more on the genetic diversity within a single species. To capture the rich genetic diversity of the plant kingdom, the pan-genome is further expanded to super-pangenome. The plant super-pangenome extends from the species level to genus or higher taxon, covering cultivated species and their wild relatives in the secondary and tertiary gene pools. By integrating the germplasm resources of diverse species, not only a rich genetic variation map is constructed, rare genes and unique variations are also captured, providing new potential and direction for plant improvement. This will facilitate the transfer of valuable traits from wild relatives to high-quality varieties through marker-assisted selection or gene editing, providing a valuable basis for plant variety improvement and promoting more efficient and adaptive variety innovation in agricultural production. In this article, we review the development of plant genome and pan-genome research, summarize the published plant super-pangenome and its applications, and discuss the future prospects and challenges of plant super-pangenome research.

plant super-pangenome, plant pan-genome, genome assembly, plant genomics, sequencing technology

doi: 10.1360/SSV-2024-0337