



图形泛基因组促进作物遗传改良

罗铮, 杨宁*

作物遗传改良全国重点实验室, 湖北洪山实验室, 华中农业大学植物科学技术学院, 武汉 430070

* 联系人, E-mail: ningy@mail.hzau.edu.cn

收稿日期: 2025-08-12; 接受日期: 2025-12-01; 网络版发表日期: 2026-02-12

国家重点研发计划(批准号: 2023YFF1000401)资助

摘要 随着测序技术与基因组组装方法的突破, 图形泛基因组逐渐成为解析作物遗传多样性和结构变异的核心工具. 传统线性参考基因组的单一性限制了其对物种内广泛变异的表征能力, 而图形泛基因组通过整合多基因型序列, 以图结构的形式直观呈现核心序列与变异的关系, 为重要性状解析、驯化进化研究和杂种优势利用提供了全新视角. 本文系统综述了图形泛基因组在作物遗传改良中的应用进展, 包括重要性状遗传基础挖掘、驯化适应性进化印记解析、杂种优势基因组学机制探索, 以及基因组辅助育种技术创新. 同时, 深入探讨了作物复杂基因组(如多倍体、高重复序列)中SVs鉴定的技术挑战与性能评估瓶颈, 并提出未来需通过算法优化、多组学整合及野生种质资源利用等策略, 推动精准育种和种质资源创新.

关键词 图形泛基因组, 结构变异, 作物遗传改良

随着测序和基因组组装技术的发展, 高质量参考基因组在作物基因注释、结构变异分析和遗传研究中发挥了重要作用, 但传统参考基因组的单一性也暴露出一定局限性. 传统基因组分析主要依赖单一线性参考基因组, 这种线性参考往往基于少数代表性品种或参考株系构建, 其内在的单一性导致对物种内广泛遗传多样性的捕获能力有限. 不同个体之间不仅存在SNP和小型插入缺失变异, 更有大片段插入、缺失、倒位、重复序列及拷贝数变异等复杂结构变异, 这些变异在调控基因表达、影响性状形成和适应性进化中扮演关键角色. 然而, 线性参考基因组无法全面呈现这些非共线性变异信息, 导致对复杂性状如产量、品质、抗逆性和杂种优势的遗传基础解析存在系统性偏差. 以作物为例, 不同的亚种之间基因组差异明显, 水稻参

考基因组主要基于粳稻(*Oryza sativa* subsp. *japonica*)亚种^[1], 而籼稻(*Oryza sativa* subsp. *indica*)亚种的特异性变异难以被充分捕获. 玉米的温热带材料基因组之间的差异也很大^[2]. 此外, 线性参考在多倍体作物或高重复序列基因组中的局限性更加显著, 同源染色体间高度相似的区域容易导致短读长比对错误或伪变异识别, 使得复杂性状的调控网络难以准确描绘. 这种偏差不仅影响变异检测的全面性, 也可能导致下游功能注释、全基因组关联分析(GWAS)和基因组选择(GS)策略的可靠性下降. 因此, 为全面理解作物性状形成的遗传机制, 突破线性参考局限、构建能够整合多基因型变异的图形泛基因组成为当前基因组学研究的核心需求, 它能够在单一框架下呈现核心基因与可变基因之间的多样化连接关系, 从而更准确地解析复杂性状的

引用格式: 罗铮, 杨宁. 图形泛基因组促进作物遗传改良. 中国科学: 生命科学, 2026, 56: 1156–1164

Luo Z, Yang N. Graph pan-genome facilitates crop genetic improvement (in Chinese). Sci Sin Vitae, 2026, 56: 1156–1164, doi: [10.1360/SSV-2025-0256](https://doi.org/10.1360/SSV-2025-0256)

遗传基础, 为精准育种提供坚实的基因组学依据。

泛基因组的概念最早由Tettelin等人^[3]在2005年提出, 用于描述物种内所有基因的集合, 包括所有个体共有的核心基因和部分个体特有的可变基因。近年来, 随着测序技术和研究方法的进步, 越来越多的图形泛基因组(graph pangenome)出现。图形泛基因组通过整合多个个体的基因组信息(包括核心、非核心及特有序列), 以不同个体间的变异作为节点, 构建节点-边的图结构, 能够直观展示物种内所有遗传变异的关系^[4]。这种方法通过图形数据结构将多个基因型序列整合到同一框架中, 能够更高效地捕获变异信息并表示复杂基因组的遗传多样性。当前的图形泛基因组构建主要流程为: 首先利用基因组或长读长测序数据构建包含广泛结构变异的图谱, 其次通过短读长数据鉴定群体中的结构变异, 最后进行下游分析。

图形泛基因组与结构变异为作物遗传改良提供了新的见解, 但作物基因组结构复杂, 存在大量重复序列, 且一些野生近缘种杂合率高, 导致结构变异鉴定困难。本文将对图形泛基因组在育种上应用、结构变异鉴定方法作简要综述和展望。

1 图形泛基因组在育种上应用

图形泛基因组与结构变异的深入研究为作物遗传改良提供了全新的视角。通过整合多物种或种群的基因组信息, 图形泛基因组突破了单一参考基因组的局限性, 系统性解析了物种的遗传多样性, 而结构变异(structural variations, SVs)作为基因组变异的重要形式, 在调控表型多样性中发挥了关键作用。近年来, 不同作物的图形泛基因组陆续发布, 涵盖了水稻^[5,6]、玉

米^[7-9]、小麦^[10]、大麦^[11,12]、大豆^[13-15]、黍^[16]、马铃薯^[17-19]等粮食作物和棉花^[20,21]、油菜^[22]、番茄^[23]、甘蓝^[24]、茄^[25]、花生^[26]、葡萄^[27,28]、西瓜^[29]、鹰嘴豆^[30]、珍珠粟^[31]等经济作物(图1)。以下从多个维度综述其在作物育种中的应用进展(表1)。

1.1 重要性状遗传基础的深度解析

传统基因组分析受限于线性参考的偏差, 难以全面解析复杂性状的遗传机制。图形泛基因组通过整合多物种或多样本的遗传变异, 系统性挖掘SVs在关键农艺性状中的作用。例如, 亚洲栽培稻(*Oryza sativa*)的泛基因组分析表明, 野生稻(*O. rufipogon*)中抗病基因类似物的丰度和多样性显著高于栽培种, 为抗稻瘟病和白叶枯病育种提供了新靶点^[6]。类似地, 马铃薯(*Solanum tuberosum*)泛基因组中高杂合性与野生种渗入片段关联的SVs, 揭示了抗晚疫病基因的进化路径^[19]。在产量与品质性状方面, 大豆(*Glycine max*)通过SV关联分析定位到调控种子蛋白含量(SPC)和油分(OIL)的关键基因*GmMQT*, 其编码的甲基转移酶在栽培种中因SV导致拷贝数扩增, 显著提升油分代谢通路活性^[14]。此外, 番茄(*Solanum lycopersicum*)的图形泛基因组结合多组学数据, 证明SVs通过调控基因表达网络(如果实可溶性固形物相关基因)驱动品质改良^[23]。上述研究揭示了SVs在重要性状中的关键作用, 而进一步解析作物驯化过程中的基因组动态变化, 将有助于挖掘适应性进化的遗传基础。

1.2 驯化与适应性进化的基因组印记

作物驯化过程中的基因组重塑常伴随SVs的动态变化。图形泛基因组通过跨物种/生态型比较, 解析了

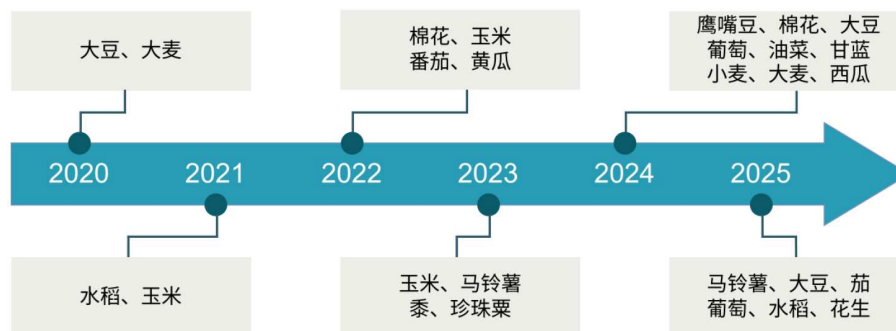


图1 作物图形泛基因组发表时间线

Figure 1 Timeline of publications on crop graph pan-genome

表 1 结构变异对作物表型的影响

Table 1 Impact of structural variations on crop phenotypes

作物	变异类型	相关基因/区域	主要发现	参考文献
水稻	PAV	<i>sh4</i>	<i>sh4</i> 缺失致落粒性丧失	[32]
玉米	SV	<i>ZmACO2</i>	50 kb缺失调控乙烯合成, 提高籽粒灌浆	[9]
大豆	CNV	<i>GmMQT</i>	导致甲基转移酶拷贝扩增, 提升油分含量	[14]
小麦	INV	Chr3A 5.8 Mb	倒位结构来源于欧洲品种渗入, 增强抗条锈病	[33]
番茄	PAV	<i>Solyc09g014280</i>	改变番茄对热胁迫的耐受性	[34]
珍珠粟	CNV	<i>PMF000024.1</i>	转录因子扩增激活ER应激响应, 提高存活率	[31]
西瓜	PAV	<i>CIFH1</i>	拷贝数与果肉颜色强调整正相关	[35]
马铃薯	SV	<i>Rpi-blb1</i>	高杂合性和野生种渗入片段关联抗晚疫病基因	[36]

人工选择对基因组结构的塑造。例如, 亚洲栽培稻的SV分布支持“单次驯化”假说, 其与野生稻的分化事件(如粳稻与籼稻亚种间855122个SNP和13853个PAV)可追溯至祖先群体的遗传瓶颈, 其中*sh4*基因的存在-缺失变异(presence-absence variation, PAV)(1.2 kb缺失)导致种子落粒性丧失^[6]。小麦(*Triticum aestivum*)泛基因组显示, 现代品种中SV的积累(如3A染色体上5.8 Mb倒位)与20世纪欧洲种质渗入密切相关, 显著提升抗锈病能力^[10]。在耐逆性进化方面, 珍珠粟(*Pennisetum glaucum*)的SV分析发现, *PgRWP-RK*转录因子家族成员在耐热种质中扩增(平均拷贝数增加2.5倍), 并通过调控内质网(ER)应激通路基因(如*BiP*和*PDI*)的快速响应, 显著提升高温下的存活率^[31]。甘蓝(*Brassica oleracea*)的SV分析揭示, *BoKANI*基因启动子区的甲基化修饰(由SV诱导)通过延迟开花时间(平均延迟7天), 使其适应高纬度地区的短日照环境^[22]。这些发现不仅阐明了作物适应性进化的分子机制, 也为利用野生种质资源提供了理论依据。

1.3 杂种优势的基因组学基础与利用

杂种优势(heterosis)的遗传本质长期存在争议, 而SVs的互补效应为其提供了新的解释。玉米(*Zea mays*)研究表明, 双亲间SV的数量(如F1杂交种中平均含1200个互补SV)与杂种优势强度(如单株产量提升25%)呈正相关, 其中*ZmACO2*基因的SV(50 kb缺失)通过抑制乙烯合成, 优化籽粒灌浆^[9]。番茄(*Solanum lycopersicum*)图形泛基因组揭示, SVs通过调控等位异质性(如*Solyc09g014280*基因座的PAV)增强杂种优势, 其贡献度在整合SV后遗传力解析提升24%^[23]。此外, 大麦(*Hordeum vulgare*)中SV驱动的淀粉酶基因家族(如

*Amy1*和*Amy2*)扩增(拷贝数增加3倍), 在人工选择下形成麦芽品质优势(酶活性提升40%), 显著优化啤酒工业特性^[12]。

1.4 基因组辅助育种的技术革新

图形泛基因组为精准育种提供了多维度遗传标记。葡萄(*Vitis vinifera*)中整合SV的GWAS识别出50.7%新QTL(如*VvMYB114*的PAV调控花色苷合成), 遗传力约提高22.8%^[27]。番茄图形泛基因组通过覆盖更多SV, 将可溶性固形物含量(SSC)的遗传力解析从0.33提升至0.41^[27]。在分子设计育种中, SVs可作为编辑靶点, 例如西瓜(*Citrullus*)超级泛基因组鉴定的野生种*C. amarus*抗枯萎病位点(如*Fom-2*基因的PAV), 通过渗入育种使栽培种抗病率从20%提升至85%^[29]。此外, 鹰嘴豆(*Cicer arietinum*)整合7个野生种的泛基因组, 系统性挖掘出抗逆基因(如*CaDREB1A*的CNV)和产量相关基因(如*CaCYCD3;1*的SV), 为标记辅助选择提供资源库^[30]。

2 结构变异鉴定方法

由于长读长测序的成本仍然高于短读长测序, 因此当前的结构变异鉴定方法分为两种, 一是基于全基因组或长读长测序数据比对, 在图形泛基因组中鉴定结构变异; 二是将短读长测序数据比对到图形泛基因组, 以鉴定群体水平的结构变异。

2.1 图形泛基因组中结构变异的鉴定

图形泛基因组中结构变异的鉴定主要包括两种策略: 一种是通过将长读长序列直接比对到参考基因组; 另一种是先进进行基因组组装, 再将组装后的基因组与

参考基因组进行比对(图2)。

通过长读长测序技术(如PacBio或Oxford Nanopore)获得的高质量长序列, 能够有效地覆盖复杂的基因组区域, 并进行比对。与传统的短读长测序相比, 长读长测序具有较长的测序片段, 可以跨越复杂的结构变异区域, 直接对这些变异进行捕获。将这些长读长序列与参考基因组进行比对, 能够精准地识别出大范围的结构变异, 如大规模插入、缺失、倒位和拷贝数变异(CNVs)。这种方法特别适用于那些存在大量复杂变异的基因组, 如作物基因组, 这些基因组中往往包含多个重复区域和大规模变异, 短读长技术难以全面覆盖。

另一种常用的全基因组比对方法是先对基因组进行组装, 再将组装得到的基因组与参考基因组进行比对。这种方法适用于高质量基因组的构建, 尤其是在没有参考基因组的情况下或基因组复杂度较高的物种中。在组装完成后, 得到的基因组序列与现有的参考基因组进行比对, 通过比对可以识别出序列中的差异, 包括结构变异。这一方法的优势在于, 它不仅能鉴定出插入和缺失等结构变异, 还能够通过组装过程发现一些未知的变异位点, 尤其是那些传统比对方法难以识别的区域。

在现代作物基因组学研究中, 这两种方法往往是结合使用的。例如在棉花泛基因组中^[21], 本课题组结合了以上两种鉴定结构变异的方法, 将长读长的测序数据通过minimap2^[37]比对到参考基因组, 再通过cu-

teSV^[38], svim^[39], sniffles^[40]等工作鉴定结构变异; 将组装好的基因组与参考基因组比对, 使用Minigraph-Cactus鉴定结构变异, 最后通过合并得到的结构变异用于下游分析, 破解了二倍体和四倍体棉花纤维品质形成的遗传调控共性和分歧。

2.2 群体水平结构变异的鉴定

尽管长读长测序能精准鉴定SVs, 但其高成本限制了群体规模。因此, 基于短读长的群体水平SV鉴定方法仍然是主流。群体水平结构变异的鉴定主要依赖于将短读长的测序数据比对到已有的结构变异图谱上。这种方法的核心在于通过短读长测序技术(如Illumina测序)对大量样本进行高通量数据采集, 然后将这些数据与已知的结构变异图谱进行比对, 从而识别和定位群体中的结构变异。这种策略的一个显著优点是, 相较于长读长测序, 短读长测序技术成本更低, 数据生成速度更快, 且更易于获得大量样本的数据。此外, 使用短读长数据进行结构变异鉴定, 不需要进行基因组组装, 相对减少了计算资源的需求, 这使得短读长测序成为群体水平结构变异研究中的主要技术。

在群体水平结构变异的鉴定过程中, 通常会使用已知的结构变异图谱, 这些图谱包含了不同群体或物种中已识别的变异信息。通过将样本的短读长测序数据与这些图谱比对, 可以识别每个样本中的插入、缺失、倒位等结构变异, 并进一步分析其在群体中的分布情况。这种方法使得研究者能够在大规模群体中快

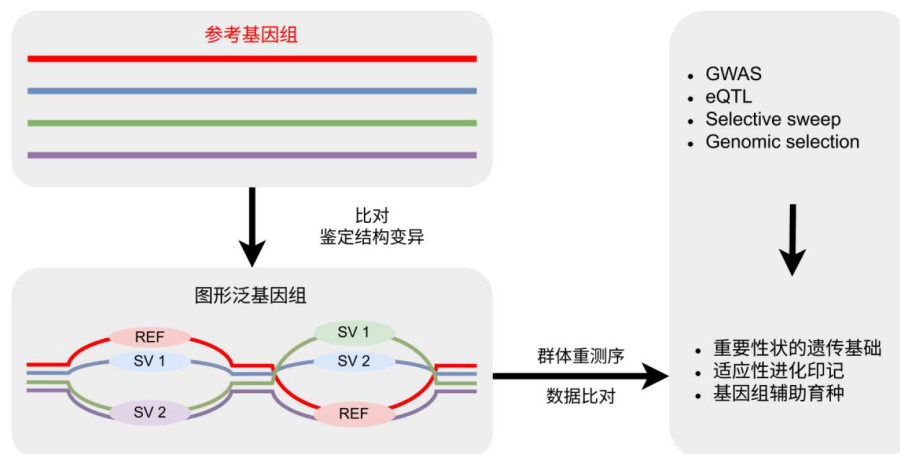


图2 结构变异的鉴定方法及应用

Figure 2 Detection methods and applications of structural variation

速鉴定和分析结构变异, 从而为基因功能研究和作物改良提供重要的信息。

VG^[4], Pangenie^[41], BayesTyper^[42]和Paragraph^[43]都是用于结构变异鉴定的工具, 广泛应用于基因组学研究中, 特别是在群体水平结构变异的识别和分析方面。每个工具都有其独特的功能和适用范围。

2.2.1 基于reads的比对

VG是一个基于图形的方法, 专为结构变异分析设计。VG通过构建变异图来捕捉基因组中的多样性, 而不是传统的线性参考基因组。它能够将短读长的测序数据比对到变异图谱上, 识别群体中的各种结构变异, 包括插入、缺失、倒位等。VG支持高效的结构变异检测, 并且能够处理复杂的基因组变异, 适用于大规模群体数据分析。该工具的一个优势是它能够在图形化表示下捕捉基因组中的所有变异, 特别是对于有多个变异等位基因的区域, VG提供了更为准确的结果。

Paragraph是一个用于从短读长测序数据中推断群体水平结构变异的工具。它通过将多个样本的测序数据进行整合, 结合基因组拼接和结构变异信息, 识别群体中的结构变异。与其他工具相比, Paragraph能够通过高效的算法在计算上节省资源, 并提供精确的变异检测结果。Paragraph支持基因组拼接, 能够对样本中的复杂变异进行高效识别, 特别适用于大规模群体的结构变异研究。

2.2.2 基于K-mer的比对

Pangenie是通过构建包含所有个体变异的泛基因组, 并利用该泛基因组比对样本中的变异, 能够有效地捕捉群体内的遗传多样性。Pangenie与传统的参考基因组比对方法不同, 它能够考虑到群体中的所有变异, 为每个个体提供更加全面的结构变异信息。Pangenie可以处理高通量的短读长数据, 适用于大规模群体结构变异研究, 尤其是复杂遗传变异的识别。但其仅能输出包含特异K-mer的变异, 不适用于结构复杂、重复序列比例高的基因组。

BayesTyper是一个基于贝叶斯统计模型的结构变异检测工具, 专注于通过群体基因组数据识别结构变异。BayesTyper通过构建一个概率模型, 将每个样本的变异与群体内的所有可能变异进行比较, 从而准确地检测出结构变异。

3 复杂基因组结构变异鉴定的探索

3.1 复杂基因组结构变异鉴定的挑战

近年来, 图形泛基因组技术逐渐成为揭示复杂基因组中结构变异的重要工具。其标准流程包括三个关键步骤: 第一步, 对代表性品种进行高质量基因组组装, 获得参考序列; 第二步, 通过全基因组比对鉴定这些基因组之间的结构变异, 构建结构变异图谱; 第三步, 利用这一结构变异图谱, 将群体水平的短读长测序数据与其比对, 从而高效识别和注释结构变异。

尽管这一流程在理论上适用于多种物种, 但实际上, 大多数现有的分析工具(如VG, Paragraph, Pangenie)主要基于人类基因组数据进行开发和优化, 而作物基因组的复杂性(如多倍体、高重复序列)导致这些工具在假阳性率和计算效率上面临显著挑战。

首先, 作物基因组的规模远超人类基因组(如小麦基因组为人类的五倍), 且重复序列占比极高(玉米和棉花分别为85%和70%)。这些重复区域使短读长数据的比对难以区分真实变异与伪变异, 导致假阳性率升高。其次, 多倍体或异源多倍体作物(如小麦、油菜)的亚基因组间存在高度相似性, 进一步加剧了变异检测的复杂性。此外, 作物结构变异的规模和类型更为复杂(如全基因组加倍、转座子插入和大片段倒位), 而现有工具针对人类基因组中小范围变异的优化策略难以适配此类大规模变异, 导致计算效率低下。

相比之下, 人类基因组重复序列仅占50%左右, 现有工具在此类基因组中运行高效。然而, 作物基因组的特殊性要求算法在数据结构、计算模型和参考基因组构建方式上进行革新, 以应对复杂变异的精准检测需求。

为了解决这些问题, 研究者需要开发更适合作物基因组的工具和算法。一些针对作物基因组的专用工具已经开始涌现, 能够更好地处理复杂变异, 同时提高对重复序列的识别能力。然而, 全面适应作物基因组特性的分析工具仍是一个重要的发展方向。这不仅要求算法层面的改进, 也需要在数据结构、计算效率和参考基因组构建方式上进行革新。

3.2 结构变异鉴定性能的评估

结构变异的鉴定性能评估是确保研究结果可靠性和准确性的重要环节。然而, 目前大多数发表的研

究文章中对鉴定到的结构变异缺乏系统性的评估,这在一定程度上影响了研究结论的可信度和普适性.对于性能评估的指标,研究者往往引用机器学习领域的经典评价体系,即精确度(precision)和召回率(recall),但此前的一些文章更倾向于使用准确度(accuracy)作为主要评价指标.这三者虽然相关,但在实际应用中各有侧重:精确度是指在所有被鉴定为阳性的结构变异中,真正阳性结构变异的比率,反映了方法的准确性;召回率是指在所有真实存在的结构变异中,被成功鉴定的比率,能衡量方法的灵敏度;准确度则是指正确鉴定结果(包括正类和负类)在总体鉴定结果中的比率.虽然准确度直观易懂,但结构变异鉴定中的负类(没有变异的区域)通常远多于正类(变异区域),导致准确度容易受到数据分布的偏倚,不能充分反映方法在结构变异鉴定中的性能.因此,精确度和召回率的结合使用能够更全面地评估工具在结构变异鉴定中的性能.

作物基因组的特殊性对结构变异鉴定提出了独特的挑战,尤其是杂合片段的识别.现代作物栽培种大多是自交系,其基因组以高度纯合为主,因此很多研究在评估中将重点放在纯合位点的鉴定性能上.然而,作物的野生近缘种中普遍存在大量的杂合片段,这些杂合结构变异可能具有重要的适应性或进化意义.在当前发表的文章中,没有针对杂合位点的结构变异鉴定性能评估,而这种忽视可能对研究结果的全面性和普适性产生负面影响.鉴于此,需要开发或优化评估方法,专门针对杂合片段的识别精度和召回率进行系统评估,以提高现有方法对复杂基因型的适配性.

评估基准的选取和数据来源也对结果产生重要影响.例如,基于合成基因组或模拟数据的评估,虽然能够提供可控的真值参考,但可能无法完全代表真实生物基因组的复杂性.相比之下,基于实验验证的真实数据评估更加可靠.因此,未来的研究需要在真实数据和模拟数据之间找到平衡,同时引入更多高质量、复杂多样的评估基准,以提高评估结果的全面性和可信度.结构变异鉴定性能的评估在方法开发和应用中具有关键作用,研究者需要在评估指标的选择、杂合片段的识别以及基准数据的选取上进行系统优化,以应对作物基因组中结构变异鉴定的独特挑战,从而为准确、全面地解析复杂基因组结构变异提供更有力的支持.

4 图形泛基因组的局限与技术挑战

尽管图形泛基因组在揭示作物遗传多样性与解析复杂性状方面展现出巨大潜力,但在实际构建与应用过程中仍面临多层次的挑战.这些问题主要集中在构建复杂度、计算效率、标准化、结构变异准确识别、多倍体基因组处理、工具兼容性、性能评估体系以及多组学整合分析等方面.

4.1 构建复杂度与计算资源消耗

图形泛基因组的构建通常需要整合数十至上百个基因组版本,对于植物基因组,构建出的图结构节点数可能超过1亿,边数超过10亿,单次构建过程需消耗数TB级存储与上千CPU核心小时.这种高计算资源消耗限制了图形泛基因组在群体水平上的广泛应用.当前主流工具如VG, PGGB, minigraph在精度与速度间存在权衡.例如, VG的精确度较高但计算成本极大,而minigraph可快速构建大型图形泛基因组,但对复杂重排区域的敏感性不足.未来的发展方向包括: (i) 采用压缩图结构(succinct graph)以减少存储负担; (ii) 基于并行化与GPU加速的算法优化; (iii) 利用AI算法进行变异模式识别与数据降维,提高构建效率.

4.2 标准化与兼容性问题

目前图形泛基因组缺乏统一的标准化数据格式和评估流程.不同软件采用不同的图描述语言,如GFA, GFAv2, VG, ODGI等,导致下游分析工具之间兼容性差.例如,同一数据集在VG与ODGI中生成的节点编号与坐标体系不同,影响注释与GWAS(全基因组关联分析)结果的一致性.此外,图形泛基因组与线性参考的坐标映射存在困难.传统基因功能注释与变异标注依赖线性坐标体系,而图结构为非线性拓扑结构,映射复杂.

4.3 多倍体与复杂基因组的特殊挑战

许多重要作物(如小麦、油菜、甘薯、棉花)为多倍体物种,基因组中存在大量同源或近同源序列,这对图结构构建与结构变异检测提出更高要求.例如,小麦的A, B, D三个亚基因组间高度相似,导致在图构建中节点过度聚合或误连通,从而产生“伪变异”.在油菜中,不同亚基因组间存在广泛的倒位和易位,传统图

结构难以区分同源序列与真实变异。

4.4 多组学整合与功能验证的不足

目前, 图形泛基因组研究仍主要集中于结构层面的变异鉴定, 对于这些变异的功能效应缺乏系统性的验证。尽管结构变异能够解释基因缺失、拷贝数变化或大片段重排, 但其在基因表达调控、表观修饰、代谢通路乃至复杂性状形成中的作用尚未充分解析。现有研究往往停留在单一组学层面, 缺少多组学数据的深度整合, 也缺乏高分辨率的细胞类型特异性分析和实验功能验证。这导致人们对结构变异如何影响表型的理解仍然片面, 限制了泛基因组在精准育种和复杂性状改良中的应用。未来, 需要通过整合转录组、表观组、三维基因组、代谢组及单细胞组学等多维数据, 构建功能化的泛基因组框架, 并结合靶向基因编辑技术进行验证, 从而系统揭示结构变异的生物学功能, 提升其在作物遗传改良中的应用价值。

4.5 AI与自动化分析的潜在突破

当前图形泛基因组构建与分析高度依赖手工或半自动化流程, 涉及序列比对、图结构构建、变异整合及注释的计算量大且复杂, 而人工智能在作物泛基因组领域的应用仍十分有限。现有深度学习或图神经网络方法在作物多倍体、高重复序列基因组上的优化不足, 自动化功能注释、变异优先级判定和多组学数据整合分析仍缺乏成熟方案。未来, AI有潜力通过深度学习和图神经网络实现自动化SV识别、节点关系推断及图压缩, 通过序列模式学习辅助功能注释, 并结合多组学信息预测变异与表型关联, 从而实现从手工分析到智能化、系统化泛基因组解析的转变, 为作物遗传改良提供高效、精准的决策支持工具。

5 展望

随着图形泛基因组研究的深入, 未来在作物遗传改良中的应用前景十分广阔, 但仍面临诸多技术瓶颈

和挑战。首先, 图形泛基因组的构建与分析计算资源消耗大、效率低, 尤其是针对多倍体、高重复序列作物基因组时, 现有算法和数据结构难以高效处理大规模结构变异和复杂图网络。其次, 泛基因组缺乏统一的标准化文件格式和坐标体系, 导致不同工具兼容性差, 下游分析和结果整合困难。此外, 目前缺少成熟的评估体系, 无法全面衡量结构变异鉴定的精度、召回率及功能效应可靠性; 多组学整合和功能验证手段尚不完善, AI及自动化分析的应用仍处于起步阶段, 无法充分发挥智能化数据挖掘和预测潜力。

针对上述局限性, 未来研究可以从以下方向推进: (i) 算法与数据结构优化: 开发高效的图构建与压缩算法, 优化多倍体和高重复序列基因组的处理流程, 降低计算资源消耗, 同时提高SV识别精度; (ii) 标准化与互操作性: 建立统一的图形泛基因组格式和坐标映射体系, 实现与传统线性参考基因组及下游分析工具的兼容; (iii) 多组学功能解析: 整合转录组、表观组、三维基因组、代谢组及单细胞组学数据, 构建功能化泛基因组, 实现结构变异与表型、细胞类型特异性效应的系统关联; (iv) AI驱动的自动化分析: 利用深度学习、图神经网络及序列模式识别模型实现智能化SV检测、图压缩与功能注释, 构建自动化分析管线, 加速数据处理和新变异发现; (v) 精准育种应用: 结合高通量表型测定和基因编辑技术(如CRISPR/Cas), 将功能化图形泛基因组成果转化为标记辅助选择和性状改良的实践工具, 实现产量、品质及抗逆性多维度提升; (vi) 大规模种群与进化研究: 在全球生态区范围内开展大规模泛基因组分析, 挖掘适应性变异, 指导野生种质资源利用和新品种设计。

综上, 未来图形泛基因组的发展需要跨学科技术融合与标准化体系建设, 从算法优化、多组学整合、AI驱动分析到精准育种实践形成闭环。通过这些策略, 图形泛基因组有望在解析作物遗传多样性、挖掘功能变异和加速育种创新中发挥核心作用, 为农业可持续发展提供强有力的技术支撑。

参考文献

- 1 International Rice Genome Sequencing Project, Sasaki T. The map-based sequence of the rice genome. *Nature*, 2005, 436: 793–800
- 2 Yang N, Liu J, Gao Q, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement.

- [Nat Genet](#), 2019, 51: 1052–1059
- 3 Medini D, Donati C, Tettelin H, et al. The microbial pan-genome. [Curr Opin Genet Dev](#), 2005, 15: 589–594
 - 4 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
 - 5 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
 - 6 Guo D, Li Y, Lu H, et al. A pangenome reference of wild and cultivated rice. [Nature](#), 2025, 642: 662–671
 - 7 Hufford M B, Seetharam A S, Woodhouse M R, et al. *De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. [Science](#), 2021, 373: 655–662
 - 8 Gui S, Wei W, Jiang C, et al. A pan-*Zea* genome map for enhancing maize improvement. [Genome Biol](#), 2022, 23: 178
 - 9 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
 - 10 Jiao C, Xie X, Hao C, et al. Pan-genome bridges wheat structural variations with habitat and breeding. [Nature](#), 2025, 637: 384–393
 - 11 Jayakodi M, Padmarasu S, Haberer G, et al. The barley pan-genome reveals the hidden legacy of mutation breeding. [Nature](#), 2020, 588: 284–289
 - 12 Jayakodi M, Lu Q, Pidon H, et al. Structural variation in the pangenome of wild and domesticated barley. [Nature](#), 2024, 636: 654–662
 - 13 Liu Y, Du H, Li P, et al. Pan-genome of wild and cultivated soybeans. [Cell](#), 2020, 182: 162–176.e13
 - 14 Zhang C, Shao Z, Kong Y, et al. High-quality genome of a modern soybean cultivar and resequencing of 547 accessions provide insights into the role of structural variation. [Nat Genet](#), 2024, 56: 2247–2258
 - 15 Yano R, Li F, Hiraga S, et al. The genomic landscape of gene-level structural variations in Japanese and global soybean *Glycine max* cultivars. [Nat Genet](#), 2025, 57: 973–985
 - 16 Chen J, Liu Y, Liu M, et al. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. [Nat Genet](#), 2023, 55: 2243–2254
 - 17 Tang D, Jia Y, Zhang J, et al. Genome evolution and diversity of wild and cultivated potatoes. [Nature](#), 2022, 606: 535–541
 - 18 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
 - 19 Sun H, Tusso S, Dent C I, et al. The phased pan-genome of tetraploid European potato. [Nature](#), 2025, 642: 389–397
 - 20 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
 - 21 Li J, Liu Z, You C, et al. Convergence and divergence of diploid and tetraploid cotton genomes. [Nat Genet](#), 2024, 56: 2562–2573
 - 22 Zhang Y, Yang Z, He Y, et al. Structural variation reshapes population gene expression and trait variation in 2,105 *Brassica napus* accessions. [Nat Genet](#), 2024, 56: 2538–2550
 - 23 Zhou Y, Zhang Z, Bao Z, et al. Graph pangenome captures missing heritability and empowers tomato breeding. [Nature](#), 2022, 606: 527–534
 - 24 Li X, Wang Y, Cai C, et al. Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in *Brassica oleracea*. [Nat Genet](#), 2024, 56: 517–529
 - 25 Benoit M, Jenike K M, Satterlee J W, et al. Solanum pan-genetics reveals paralogues as contingencies in crop engineering. [Nature](#), 2025, 640: 135–145
 - 26 Zhao K, Xue H, Li G, et al. Pangenome analysis reveals structural variation associated with seed size and weight traits in peanut. [Nat Genet](#), 2025, 57: 1250–1261
 - 27 Liu Z, Wang N, Su Y, et al. Grapevine pangenome facilitates trait genetics and genomic breeding. [Nat Genet](#), 2024, 56: 2804–2814
 - 28 Guo L, Wang X, Ayhan D H, et al. Super pangenome of *Vitis* empowers identification of downy mildew resistance genes for grapevine improvement. [Nat Genet](#), 2025, 57: 741–753
 - 29 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
 - 30 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
 - 31 Yan H, Sun M, Zhang Z, et al. Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. [Nat Genet](#), 2023,

- 55: 507–518
- 32 Kou Y, Liao Y, Toivainen T, et al. Evolutionary genomics of structural variation in Asian rice (*Oryza sativa*) domestication. *Mol Biol Evol*, 2020, 37: 3507–3524
- 33 Chen S, Hegarty J, Shen T, et al. Stripe rust resistance gene *Yr34* (synonym *Yr48*) is located within a distal translocation of *Triticum monococcum* chromosome 5A^mL into common wheat. *Theor Appl Genet*, 2021, 134: 2197–2211
- 34 Balyan S, Rao S, Jha S, et al. Characterization of novel regulators for heat stress tolerance in tomato from Indian sub-continent. *Plant Biotechnol J*, 2020, 18: 2118–2132
- 35 Sun H, Zhang J, Liao S, et al. Population-level super-pangenome reveals genome evolution and empowers precision breeding in watermelon. *bioRxiv*, 2025, 666869
- 36 Rakosy-Tican E, Thieme R, König J, et al. Introgression of two broad-spectrum late blight resistance genes, *Rpi-Blb1* and *Rpi-Blb3*, from *Solanum bulbocastanum* into plus race-specific *R* genes into potato pre-breeding lines. *Front Plant Sci*, 2020, 11: 699
- 37 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 2018, 34: 3094–3100
- 38 Jiang T, Liu Y, Jiang Y, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol*, 2020, 21: 189
- 39 Heller D, Vingron M, Birol I. SVIM: structural variant identification using mapped long reads. *Bioinformatics*, 2019, 35: 2907–2915
- 40 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
- 41 Ebler J, Ebert P, Clarke W E, et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet*, 2022, 54: 518–525
- 42 Sibbesen J A, Maretty L, Krogh A. Accurate genotyping across variant classes and lengths using variant graphs. *Nat Genet*, 2018, 50: 1054–1059
- 43 Chen S, Krusche P, Dolzhenko E, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol*, 2019, 20: 291

Graph pan-genome facilitates crop genetic improvement

LUO Zheng & YANG Ning*

College of Plant Science and Technology, Huazhong Agricultural University, Hubei Hongshan Laboratory, National Key Laboratory of Crop Genetic Improvement, Wuhan 430070, China

With breakthroughs in sequencing technology and genome assembly methods, the graph pan-genome has gradually become a core tool for analyzing crop genetic diversity and structural variations. The singularity of traditional linear reference genomes limits their ability to represent the extensive variations within species. In contrast, the graph pan-genome incorporates sequences from multiple genotypes into a single graph structure that captures both conserved regions and genetic variants. This paradigm enables a more comprehensive analysis of agronomic trait architecture, domestication and adaptation processes, and the genomic basis of heterosis. This article systematically reviews the application progress of the graph pan-genome in crop genetic improvement, including the exploration of the genetic basis of important traits, the analysis of domestication and adaptive evolution imprints, the investigation of genomic mechanisms underlying heterosis, and innovations in genomics-assisted breeding technologies. It delves into the technical challenges and performance evaluation bottlenecks in identifying structural variations in complex crop genomes, such as polyploids and genomes with high repeat sequences, and proposes future strategies, such as algorithm optimization, multi-omics integration, and the utilization of wild germplasm resources to advance precision breeding and innovation in germplasm resources.

graph pan-genome, structural variation, crop genetic improvement

doi: [10.1360/SSV-2025-0256](https://doi.org/10.1360/SSV-2025-0256)