



刘羽诚, 2016—2020 年就读于中国科学院遗传与发育生物学研究所, 在田志喜课题组攻读博士学位; 2021—2023 年在该课题组开展博士后工作; 2023 年至今任中国科学院遗传与发育生物学研究所副研究员, 从事大豆功能基因组学、比较基因组学、大数据挖掘与数据库开发相关研究。博士期间, 开展大豆泛基因组工作, 完成 26 个大豆种质的高质量参考基因组, 在植物中创造性实践了图泛基因组构建策略, 系统阐释了染色体结构变异在大豆演化/驯化过程中的作用, 为后续泛基因组研究提供了经典的思路和范例。获得“博士后创新人才计划”、“中国科学院稳定支持青年团队”项目资助; 主持国家自然科学基金委青年科学基金项目。博士论文《大豆泛基因组研究》荣获 2023 年中国科学院优秀博士生论文。

大豆泛基因组研究进展

刘羽诚¹, 申妍婷¹, 田志喜^{1,2}

1. 中国科学院遗传与发育生物学研究所, 种子创新重点实验室, 北京 100101
2. 中国科学院大学, 北京 101408

摘要: 人工驯化为农业发展提供了原始驱动力, 也深刻地改变了许多动植物的遗传背景。伴随组学大数据理论和技术体系的发展, 作物基因组研究已迈入泛基因组时代。借助泛基因组的研究思路, 通过多基因组间的比较和整合, 能够评估物种遗传信息上界和下界, 认知物种的遗传多样性全貌。此外, 将泛基因组与染色体大尺度结构变异、群体高通量测序及多层次组学数据相结合, 可以进行更为深入的性状-遗传机制解析。大豆(*Glycine max* (L.) Merr.)是重要的粮油经济作物, 大豆产能关乎国家粮食安全。对大豆遗传背景形成、重要农艺性状关键位点的解析, 是实现更高效的大豆育种改良的前提。本文首先对泛基因组学的核心问题进行了阐述, 解释了从头组装/比对组装、迭代式组装和图基因组等泛基因组研究策略的演变历程和各自特征; 接着对作物泛基因组研究的热点问题进行了概括, 并且以大豆为例详细阐释了包括类群选择、泛基因组构建、数据挖掘等方面在内的泛基因组研究的开展思路, 着重说明染色体结构变异在大豆演化/驯化历程中的贡献及其在农艺性状遗传基础挖掘上的价值; 最后讨论了图泛基因组在数据整合、结构变异计算方面的应用前景。本文对作物泛基因组未来的发展趋势进行了展望, 以期对作物基因组学及数据科学研究提供参考。

关键词: 大豆; 泛基因组; 结构变异; 演化; 驯化

收稿日期: 2023-12-29; 修回日期: 2024-02-09; 网络发布日期: 2024-02-22

基金项目: 国家自然科学基金项目(编号: 32201775, U22A20473)和中国科学院稳定支持青年团队计划(编号: YSBR-078)资助[Supported by the National Natural Science Foundation of China (Nos.32201775, U22A20473) and CAS Project for Young Scientists in Basic Research (No. YSBR-078)]

作者简介: 刘羽诚, 副研究员, 研究方向: 大豆比较基因组学。E-mail: ychliu@genetics.ac.cn

通讯作者: 田志喜, 研究员, 博士生导师, 研究方向: 大豆种质资源基因组演化与分子遗传解析。E-mail: zxtian@genetics.ac.cn

DOI: 10.16288/j.yczs.23-321

Frontiers of soybean pan-genome studies

Yucheng Liu¹, Yanting Shen¹, Zhixi Tian^{1,2}

1. Key Laboratory of Seed Innovation, Institute of Genetics and Development of Biology, Chinese Academy of Sciences, Beijing 100101, China

2. University of Chinese Academy of Sciences, Beijing 101408, China

Abstract: Artificial domestication provided the original motivation to the blooming of agriculture, following with the dramatic change of the genetic background of crops and livestock. According to theory and technology upgradation that contributing to the omics, we appreciate using the pan-genome instead of single reference genome for crop study. By comparison and integration of multiple genomes under the guidance of pan-genome theory, we can estimate the genomic information range of a species, leading to a global understanding of its genetic diversity. Combining pan-genome with large size chromosomal structural variations, high throughput population resequencing, and multi-omics data, we can profoundly study the genetic basis behind species traits we focus on. Soybean is one of the most important commercial crops over the world. It is also essential to our food security. Dissecting the formation of genetic diversity and the causal loci of key agricultural traits of soybean will make the modern soybean breeding more efficiently. In this review, we summarize the core idea of pan-genome and clarified the characteristics of construction strategies of pan-genome such as *de novo*/mapping assembly, iterative assembly and graph-based genome. Then we used the soybean pan-genome work as a case study to introduce the general way to study pan-genome. We highlighted the contribution of structural variation (SV) to the evolution/domestication of soybean and its value in understanding the genetic bases of agronomy traits. By those, we approved the value of graph-based pan-genome for data integration and SV calculation. Future research directions are also discussed for crop genomics and data science.

Keywords: soybean; pan-genome; structural variation; evolution; domestication

近 20 年来基因组学经历了爆发式的发展, 如今已经成为生命科学领域研究的重要范畴。基因组承载着生命体的基本遗传信息, 一个高质量的基因组是展开深度遗传学及分子功能研究的先决条件。然而, 随着基因组学理论体系的延展、测序技术的革新、数据维度和数据需求的不断丰富, 研究者对基因组本身的认知经历了不断的扩充与迭代。将单一的参考基因组作为特定物种或者类群基因组的“标准品”, 其代表性和蕴含的生物多样性始终是有限的。物种内、种系间的差异是解析种群演化和表型特征形成的关键, 不能被忽视。针对这些问题, 研究人员不断探索新的研究方法思路, 这此过程中考虑多个代表性基因组比较与整合的泛基因组学 (pan-genomics) 框架得以建立, 成为现今研究的热点

方向。

作物分子设计育种是解决国家粮食安全问题的的重要手段, 而高质量的作物基因组是遗传学家、育种家认识改造作物的关键基础。作物基因组演化存在诸多特征。一方面, 植物基因组中基因组序列重复、基因组加倍、多倍化等事件更为频繁, 使得植物在染色体水平上积累了更多的结构差异^[1]; 另一方面, 作物驯化改良是一致性和多样化兼有的过程, 尽管品种/品系之间具备高度的可比性, 但单个品种/品系的基因组并不能代表整个作物的遗传背景。因此研究者认识到, 使用单个基因组作为参考开展作物遗传与功能基因组研究, 很可能低估研究对象遗传分化的程度并遗失诸多重要的遗传变异^[2,3]。以上特征表明作物是开展泛基因组研究的良好素材, 而泛

基因组也是深度解析作物基因组多样性、挖掘农艺性状相关位点的重要方法。作为传统基因组形式的补充和扩展,泛基因组现今已成为作物基因组图谱绘制和遗传解析的常用手段^[4,5]。

大豆(*Glycine max*)是我国重要的作物和经济物资,由于需求的激增导致供给不足,国内大豆不得不大量依赖进口。改良种质,培育高产、稳产、高品质、适应不同农田环境的大豆,是提高大豆产量的关键。中国拥有最丰富的大豆遗传资源以及多样的栽植生态区系,采用泛基因组的研究方法,厘清大豆的遗传变异,发掘新的或未被充分使用的遗传位点,结合分子设计育种等手段,对于推进中国大豆品种的选优改良,具有重要意义。

1 泛基因组概述

1.1 泛基因组概念的发展

泛基因组(pan-genome)的词缀“pan”来源于希腊语,意为“全”、“一切”。泛基因组通常意义上是指代一个物种/类群所有基因组,或代表性基因组的总和。在研究的早期,测序技术产出的数据质量有限,测序成本高昂,在许多真核生物中获得单个高质量组装基因组是十分困难的事情。因此,往往用单个或少数高完成度的基因组作为一个物种或是一个类群的代表或参考。而在一些原核生物中,由于基因组规模小,获取基因组相对容易,研究人员通常可以获得同一个类群中多个个体的完整基因组,并且开展多基因组间的整体比对。这类工作最早由 Tettelin 等^[6]于 2005 年在无乳链球菌(*Streptococcus agalactiae*)中开展,是泛基因组研究的雏形。

然而泛基因组的概念推广到更复杂的动植物等真核生物类群并没有那么迅速。首先,通常情况下真核生物基因组相比细菌要大得多,这意味着基因组测序的成本和后续组装消耗的算力、时间资源都很巨大。其次,真核生物基因组更为复杂,多倍体、高重复序列、高杂合度等情况都会增加基因组组装的难度^[7~10]。并且由于基因组成分复杂,有大量非基因区序列、重复序列的存在,使得泛基因组组分评估及基因组差异的鉴定也不易进行^[11]。近几年,随着测序技术的发展,测序成本下降,比较基因组

学手段不断完善,这些问题才逐渐得到解决。从原核生物到真核生物,泛基因组的范畴也从包含全体注释基因扩展到包含所有基因组序列。而伴随组学研究维度的开拓,泛组学概念的应用也从基因组层面延伸到如泛转录组、泛三维基因组等层面^[12,13]。

1.2 泛基因组研究的核心问题

泛基因组研究的核心问题,是对物种/类群基因组完备性或者代表性遗传信息的描述^[14]。与群体遗传学类似,泛基因组的研究对象并非单一个体。然而群体遗传学层面的基因组研究侧重于发掘变异位点及遗传多态性,即个体间的异质性。而个体间的异质性和同质性,即共享与差异的基因组成分,均为泛基因组研究描述的内容。通过泛基因组研究,人们能了解一个物种/类群的完整基因组架构,并借此推断构成这一物种/类群的核心遗传信息(即基因组下界),以及物种/类群的遗传分化程度(即基因组上界)。

此外,泛基因组研究涉及基因组间的比较和整合,其中对不同基因组间染色体结构变异(structural variation, SV)的挖掘和处理也成为研究的重要环节^[15]。相较于单核苷酸多态性(single nucleotide polymorphism, SNP),结构变异的长度不定,变异类型更为复杂,处理难度也更高。同时,结构变异引起的基因组改变更为剧烈,更易引起物种间表型特征的多态性。这类变异在基因组学研究的早期,因为技术和成本的限制,很难作为重要的研究方向,而如今则成为泛基因组研究聚焦的重点之一。对于染色体结构变异的处理,也体现了泛基因组实践策略的不同发展阶段。

2 泛基因组实践策略及研究实例

2.1 从头组装/比对组装基因组

泛基因组构建需要对物种/类群的代表性个体进行仔细筛选,进行基因组测序。获得数据后,最常规的策略是分别对每个个体进行基因组从头组装,将单独组装的基因组数据集作为泛基因组^[16~18];或者将测序数据比对到一个高质量的参考基因组上,并将无法比对的数据分类出来单独进行组装,作为

现有参考基因组的扩展集,形成“参考基因组+额外序列”,即“共有序列+染色体差异序列”的形式^[19~21]。

这类方法在实践层面上最为简单,在泛基因组研究的早期有较多应用,但也存在诸多问题。单独基因组形式的泛基因组通常包含过多冗余的数据量和数据维度。而“参考基因组+额外序列”的方式对于泛基因组的组织并不直观有效。因此研究者需要探索更为高效合理的泛基因组数据组织形式。

2.2 迭代式泛基因组

迭代式基因组是一类经过实践的参考基因组整合方法。该类方法从一个参考基因组(往往是高质量或已被广泛认可的基因组)开始,依次将其他样品的测序读段比对到参考基因组上,并且直接修改当前参考基因组,在恰当的位置添入非冗余的染色体差异序列。参考基因组在这个过程中不断被迭代升级,最终成为一个兼容多基因组状态的线性基因组^[2,22]。这类方法主要在甘蓝中得以实践,获得了 99 Mb 的额外序列,并且绘制了多个体来源整合的染色体变异图谱^[22]。

迭代式泛基因组相较于从头组装的泛基因组整合度高,不引入额外序列,并且类似传统的线性基因组,更易于理解。但实现过程中对于原有基因组的覆盖将不可避免丢失许多单独基因组状态下的特征。因此,迭代式组装尽管减少了信息的冗余,也同时存在大量的信息丢失^[11]。

2.3 基于图论的泛基因组

基因组学的快速发展对泛基因组提出了更全面的数据结构诉求。泛基因组除了提供个体间共享和特异序列信息的记录存储外,还承担着数据的调用、检索、可视化、比对等多种功能。基于图论的基因组(即图基因组)是满足以上需求的有效形式。该方法首先选择一个基因组作为本底,通过读段比对或者染色体共线性比较的方式,获得各个样品相对于参考基因组的变异位置及变异内容。最后依照上述信息,采用图论的方式将参考序列与变异序列以节点方式存储,并且用边代表他们的连接关系^[2,4,11,23]。

尽管图基因组并不像传统线性基因组那样直观,但其最大程度压缩了冗余信息,并且保留了有意义信息。此外图基因组可以灵活地进行数据组合与还原,

保证了组学数据的可读性。对于基因组较大,变异复杂的真核生物,图基因组是更合适的方法,也成为现在的趋势^[24~28]。此外,图基因组更兼容计算机的 I/O 形式,能够更快、更有效地进行基于二代测序数据的比对和结构变异检测。目前,图基因组是泛基因组数据存储、调用、展示等综合性能最佳的形式,越来越多的基因组分析工具开始向该方向发展,如 vg (Variation Graph toolkit)^[26]、GraphTyper2^[25]、Giraffe^[29]、odgi (Optimized Dynamic Genome/Graph Implementation)^[30]、pggb (PanGenome Graph Builder)^[31]等。一些经典的工具,如 HISAT2^[32]也有此方面功能的拓展。图基因组在泛基因组,尤其是植物泛基因组学领域,目前已经有了很多实践,逐渐成为研究的主流方法。

2.4 作物泛基因组研究

2011 年, Gan 等^[33]对拟南芥(*Arabidopsis thaliana*)自然群体材料的基因组比较是植物泛基因组研究的开端。该工作从头组装了 18 个拟南芥的单拷贝序列基因组,通过比较发现了相对参考基因组共有 28.3 Mb 非冗余变异序列,平均每个样品 4.5~7.6 Mb。此后泛基因组研究逐渐在植物中兴起,并且在近 10 年间高速发展。目前许多植物,特别是作物都完成了从单一参考基因组到泛基因组的整合与跨越^[20,22,34~39]。早期植物泛基因组多采用从头组装/比对组装的策略进行构建,部分研究采用了迭代组装方式(表 1)。在近期的研究中,从头组装结合图泛基因组已经成为主流的泛基因组研究策略(表 1)。泛基因组研究在一定程度上揭示了作物物种内或近缘种间的基因组变异规模。对比一些研究结果可以得出,在不同植物类群的泛基因组中,核心基因家族占总基因家族数量的 40%~70%,表明 30%~60% 的基因家族在物种内发生了获得/丢失的变异^[16,17,19~22,40,41]。

泛基因组是深度挖掘农艺性状与基因组变异,尤其是染色体结构变异关联性的有效手段。一方面,对于已知基因或位点,泛基因组能够提供更新、更全面的变异认知。野生大豆(*Glycine soja*)的泛基因组研究比较了大豆开花途径基因的变异,发现 *PHY4*、*E3*、*E4*、*E1*、*FT*、*LFY* 等基因在野生及栽培大豆基因组间均存在蛋白差异,并且 *FT* 在野生大豆中存在一个参考基因组 WM82 中没有的亚型^[17]。这些变异

表 1 植物泛基因组研究实例汇总

Table 1 Case studies of plant pan-genome

类群	发表年份	样品数	测序方式	泛基因组构建策略	参考文献
拟南芥(<i>Arabidopsis thaliana</i>)	2011	18	二代测序	迭代组装+从头组装	[33]
野生大豆(<i>Glycine soja</i>)	2014	7	二代测序	从头组装	[17]
甘蓝(<i>Brassica oleracea</i>)	2016	9	二代测序	迭代组装	[22]
苜蓿(<i>Medicago truncatula</i>)	2017	15	二代测序	从头组装	[76]
二穗短柄草(<i>Brachypodium distachyon</i>)	2017	54	二代测序	从头组装	[16]
水稻(<i>Oryza sativa</i>)	2018	3010	二代测序+三代测序	比对组装	[21]
野生及栽培水稻(<i>O. rufipogon</i> , <i>O. sativa</i>)	2018	66	二代测序	比对组装	[42]
水稻属及亲缘物种(<i>Oryza</i> , <i>Leersia</i>)	2018	13	三代测序+二代测序	从头组装	[18]
辣椒属(<i>Capsicum</i>)	2018	168	二代测序	比对组装	[77]
芝麻(<i>Sesamum indicum</i>)	2018	5	二代测序	比对组装	[78]
番茄及野生亲缘种(<i>Solanum</i> section <i>Lycopersicon</i>)	2019	725	二代测序	比对组装	[19]
向日葵(<i>Helianthus annuus</i>)	2019	287	二代测序	比对组装	[20]
油菜(<i>Brassica napus</i>)	2020	8	三代测序	从头组装	[43]
野生及栽培大豆(<i>Glycine</i> subgenus <i>Soja</i>)	2020	29	三代测序	从头组装+图基因组	[39]
大麦(<i>Hordeum vulgare</i>)	2020	20	二代测序+三代测序	从头组装	[79]
番茄及野生亲缘种(<i>Solanum</i> section <i>Lycopersicon</i>)	2020	14	二代测序+三代测序	比对组装(泛结构变异)	[45]
鹰嘴豆(<i>Cicer arietinum</i>)	2021	3366	二代测序	比对组装	[80]
棉花及亲缘种(<i>Gossypium</i>)	2021	1961	二代测序	比对组装	[81]
野生及栽培高粱(<i>Sorghum bicolor</i>)	2021	13	三代测序	从头组装	[82]
玉米(<i>Zea mays</i>)	2021	26	三代测序	从头组装	[83]
水稻(<i>O. sativa</i>)	2021	33	三代测序	从头组装+图基因组	[34]
野生及栽培萝卜(<i>Raphanus</i>)	2021	11	三代测序	从头组装+图基因组	[84]
黄瓜 (<i>Cucumis sativus</i>)	2022	12	三代测序	从头组装+图基因组	[38]
水稻属(<i>Oryza</i>)	2022	251	三代测序	从头组装+图基因组	[85]
棉花属(<i>Gossypium</i>)	2022	10	三代测序	从头组装+图基因组	[86]
多年生大豆(<i>Glycine</i> subgenus <i>Glycine</i>)	2022	6	三代测序	从头组装	[62]
野生及栽培马铃薯(<i>Solanum</i> section <i>Petota</i>)	2022	44	三代测序	从头组装	[87]
番茄(<i>Solanum lycopersicum</i>)	2022	32	三代测序	从头组装+图基因组	[35]
野生及栽培谷子(<i>Setaria</i>)	2023	110	三代测序	从头组装+图基因组	[40]
茶(<i>Camellia sinensis</i>)	2023	22	三代测序	从头组装+图基因组	[41]
柑橘属(<i>Citrus</i>)	2023	12	三代测序	从头组装+图基因组	[36]
番茄及野生亲缘种(<i>Solanum</i> section <i>Lycopersicon</i>)	2023	13	三代测序	从头组装+图基因组	[85]
玉米(<i>Z. mays</i>)	2023	12	三代测序	从头组装	[88]
野生及栽培黍(<i>Panicum miliaceum</i>)	2023	32	三代测序	从头组装+图基因组	[46]

可能导致了野生和栽培大豆开花特征的分化。66 份野生和栽培水稻的泛基因组研究充分挖掘了 *waxy*、*Hd1* 等位点的多种单倍型, 涉及 SNP 和 Indel 的多种组合, 加深了对水稻品质、花期等复杂农艺性状的理解^[42]。谷子(*Setaria italica*)泛基因组研究表明,

种质间落粒性、籽粒大小差异与染色体结构变异相关。其中, 在其他谷物中被平行选择的 *sh1* 基因, 在谷子中也发生了一个 855 bp 的存在/缺失变异 (presence and absence variation, PAV), 造成基因的获得/缺失, 进而控制落粒性的变化^[40]。这也体现出

sh1 在谷物中功能的保守性和利用改造价值。另一方面, 群体结构变异数据可以用作关联分析, 发挥和 SNP 相当或者互补的效力。Song 等^[43]在油菜 (*Brassica napus*) 泛基因组研究中使用 PAV 数据进行种子重量的全基因组关联分析 (genome wide associated study, GWAS), 其信号区间和使用 SNP 的计算结果重叠, 而其中一个 3.6 kb 的 PAV 位于信号峰值。该变异为转座元件 (transposable element, TE) 插入, 统计 NAM 群体的表型发现该变异的存在/缺失和角果长度和种子重量都显著相关。而该 TE 下游为 *CYP78A9* 基因, 推测变异影响了该基因的表达, 从而造成性状的变化。谷子泛基因组研究中对千粒重、粒宽的 SV-GWAS 分析找到一个控制相关表型的基因及变异位点^[40]。该基因启动子区发生了 366 bp 的 PAV。实验表明, 该序列变异导致基因表达量改变, 相关过表达株系也表现出粒宽的显著下降。水稻中对于产量的分析发现, 使用结构变异进行 GWAS 分析能够检测到比 SNP 更为显著的关联位点, 其中位于 *OsNPY2* 基因上游的一个 1.4 kb 序列存在/缺失与产量表型密切相关^[44]。

3 大豆泛基因组研究

3.1 大豆属泛基因组组成

2014 年野生大豆的泛基因组研究是植物中第一项明确泛基因组概念的工作^[17]。然而其数据质量、全面性和挖掘深度都受到了时代和技术的制约。2020 年一项包含大豆属 *Soja* 亚属的野生、栽培大豆在内, 26 个大豆种质材料基因组、转录组及近 3000 份种质材料重测序的工作则更精准地描绘了大豆的遗传变异图谱, 系统阐述了染色体结构变异在大豆演化/驯化中发挥的作用^[39]。该研究从 2898 份来自世界大豆主要栽植区的种质资源中共检测到约 3 千万个单核苷酸变异位点。根据系统发育关系, 挑选出 26 个代表性的种质, 进行基因组从头组装和泛基因组构建。这 26 个种质按类群划分包括野生、农家种、栽培品种, 按用途划分包括骨干亲本及区域主栽品种等, 从头组装基因组大小在 992.3~1059.8 Mb 之间, 样品序列锚定在染色体上的比率平均为 99.0%, 二代测序比对回自身基因组的比对率平均在 99.4%。

基因组重复序列注释检测到大豆基因组的平均重复序列比例为 54.4%, 蛋白编码基因注释表明大豆泛基因组样品平均注释基因数量为 56,522, BUSCO 检验平均达到 95.6%。以上结果符合大豆基因组的基本特征, 说明基因组组装注释质量达到高水平。

对 26 个大豆从头组装基因组, 连同已经报道的 ZH13 的基因组进行基因家族聚类, 所有基因被分入 57,492 个基因家族, 这与之前野生大豆中报道的数量接近^[17]。对不同品种数量构建的泛基因、核心基因家族数目的抽样统计显示, 泛基因组的数量在 25 个样品时到达了平台期, 意味着该研究的取样对于大豆基因组已具有足够的代表性。将基因家族按样品出现的频数作为划分, 得到大豆的核心基因家族 (频数为 27) 20,623 个, 松弛核心基因家族 (频数为 25、26) 8163 个, 非必需基因家族 (频数为 2~24) 28,679 个, 私有基因家族 (频数为 1) 27 个。由此得出, 大豆泛基因组中核心 (及松弛核心) 基因家族占总基因家族的 50.1%, 非必需及私有家族 (可变家族) 的数量占 49.9%。该结果符合以往研究得出的植物中 30%~60% 的基因家族为可变家族的认知^[16,17,19-22,40,41]。

3.2 大豆属泛基因组变异

泛基因组包含的变异是否能反应物种群体水平的变异, 是值得探讨的问题。以 ZH13 基因组作为参考, 结合 26 个泛基因组样品和已报道的 WM82 及 W05 的基因组数据, 在 29 个大豆基因组上检测到 14,604,953 个 SNP 和 12,716,823 个 Indel (≤ 50 bp)^[39]。该数据与 2898 份重测序的变异数据进行比较, 尽管 SNP 数量比 2898 份重测序要少, 但是二者分布特征相似。以 500 kb 区间为窗口进行全基因组扫描, 过滤 2898 份重测序中次等位基因频率 (minor allele frequency, MAF) < 0.01 的位点后, 其与 29 个基因组中 SNP 数量的皮尔森相关性系数为 0.553。此外 π 、 d_N/d_S 等群体遗传学特征值在 29 个基因组与 2898 份重测序数据间同样具有很高的相关性。这表明泛基因组对于变异的检测具有群体水平的代表性。

大尺度结构变异 (> 50 bp) 采用短序列测序方式往往很难鉴定。通过基因组比对的方式, 以 ZH13 为参考在 28 个大豆基因组中检测到共计 776,399 个结构变异, 其中 723,862 个 PAV、27,531 个拷贝数

变异(copy number variation, CNV)、21,886 个易位事件、3120 个倒位事件^[39]。PAV 的长度主要分布在 1~2 kb, 易位长度主要分布在 10~30 kb, 倒位长度主要分布在 100~200 kb。CNV 的变化倍数主要在 2~3 倍。泛基因组中检测到的 723,862 个 PAV 共计 4.71 Gb 序列长度, 平均每个样品 167.09 Mb, 占基因组大小约 16%。比较每个样品的获得与缺失序列长度之差, 及其与 ZH13 基因组大小之差, 发现二者具有很高的相关性, 说明 PAV 是造成样品间基因组大小差异的主要来源。在大豆中结构变异在基因组重复序列区域显著富集, 其中 78.5% 的 PAV 来自

于 DNA 重复。对番茄(*Solanum lycopersicum*)泛基因组研究发现 84% 的序列缺失与 76% 的序列插入变异与重复序列重合(>100 bp)^[45]。对黍(*Panicum miliaceum*)的泛基因组研究发现 PAV 与 TE 的重合比例在 70% 左右^[46]。这些结果暗示一些植物中序列重复事件可能是结构变异发生的重要驱动力, 进而导致物种内基因组大小的波动。

3.3 大豆属图泛基因组构建

大豆是首个实践了图泛基因组构建的植物, 为后续作物的泛基因组研究开拓了新思路(图 1A)。构

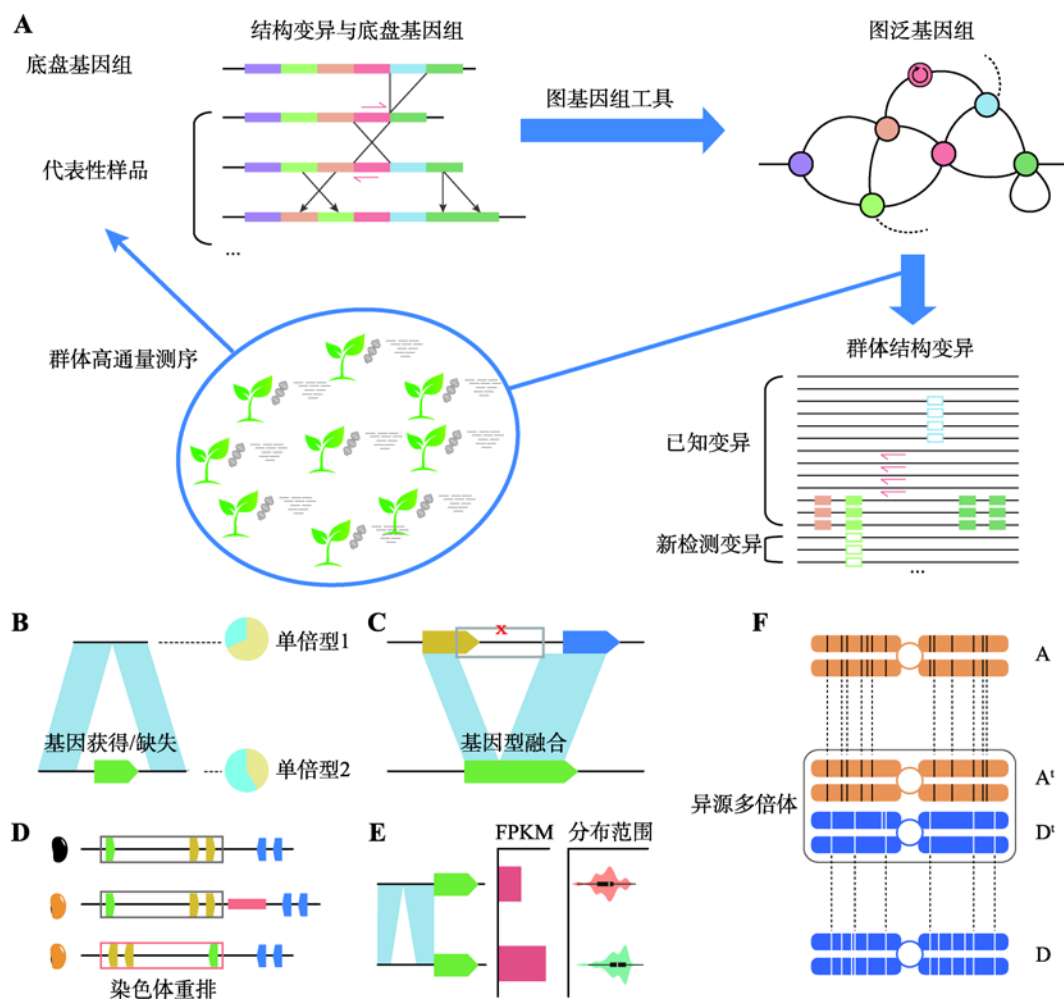


图 1 作物泛基因组研究策略及认知

Fig. 1 Crop pan-genome strategy and knowledge

A: 图泛基因组研究基本流程, 包括群体测序筛选代表性样品、结构变异分析、图泛基因组构建、群体结构变异检测等; B~E: 泛基因组视角下的大豆农艺性状、演化历程遗传机制认知, 包括基因获得/缺失与种皮亮度(B)、基因融合与 *E3* 基因多态性(C)、染色体重排与种皮颜色(D)、结构变异对基因表达调控与种质分布(E); F: 异源多倍体大豆的冗余基因丢失与亚基因组偏好性。

建图泛基因组,需要对结构变异数进行合并和过滤,一方面降低构建图基因组的计算负担,另一方面减少最终图基因组的复杂度和假阳性。在 29 个大豆基因组中检测到 776,399 个结构变异,根据位点和类型进行合并,非冗余结构变异总数随样品增加而增加,最终趋于稳定,得到共计 124,222 个非冗余结构变异位点^[39]。与此同时,共有的结构变异最终收敛到 130 个。野生大豆相较于栽培大豆,私有结构变异所占的比例更大。

此外,研究表明将结构变异中重复序列占总长度 90% 的条目过滤,是有效的数据压缩、降低错误率的策略。Liu 等^[39]采用 vg 工具,以过滤后的结构变异数据为输入,ZH13 基因组为底盘基因组,构建可用于检索和二代数据比对的大豆图泛基因组索引文件。将 2898 个大豆样品重测序数据比对到图泛基因组上,共计检测到 55,402 个结构变异。采用图泛基因组检测结构变异的精确率、召回率及 F-score 分别为 0.94、0.75 和 0.83,表明图泛基因组结合群体二代测序数据是作物中进行大规模结构变异检测的可行方法。图泛基因组流程检测的结构变异 N50 为 659/595 bp(缺失/插入),远高于 GATK 流程的 3/3 bp,说明图泛基因组流程对于大尺度结构变异检测具有很好的效力。相对于 28 个基因组中检测到的变异,在约 3000 份群体水平找到 3584 个新的结构变异,占总变异数的 6.5%,并且这些变异的出现频率较低。野生大豆中检测到的已有和新结构变异的数量均明显高于农家种和栽培大豆。水稻中相似研究检测到的新结构变异占总变异数的 16.4%^[34],但该研究的图泛基因组构建仅针对栽培稻进行。这也侧面反应出作物的野生种可能持有更丰富的变异类型,在作物泛基因组研究中加入野生类群可以很好地提升遗传变异的覆盖度。

3.4 泛基因组助力大豆演化/驯化遗传基础

GWAS 分析是检测与表型关联的遗传变异的有效手段,而群体水平检测的结构变异同样能够辅助农艺性状相关位点的挖掘(图 1B)。大豆种皮亮度是一个重要的性状,以往研究报道其与一种大豆疏水性蛋白(HPS)的积累有关^[47],但具体相关的基因仍未明确。Liu 等^[39]以图泛基因组检测的结构变异为基因型,对种皮亮度表型进行了 GWAS 分析,在 15

号染色体上定位到一个信号区间。其中一个 10 kb 的 PAV 包含了一个编码 HPS 结构域的基因,并造成该基因在品种间的获得/缺失。表型统计发现,存在该 10 kb 序列的样品种皮光亮的比例更高,说明该 PAV 是控制大豆种皮亮度的遗传位点之一。

位于基因区的结构变异可能造成基因开放阅读框(open reading frame, ORF)的改变,进而导致功能的丢失或分化。其中结构变异造成的转录本通读是一种较为特殊的情况,即由于序列丢失导致原本独立转录的基因融合为一个转录本。转录本通读引起的基因融合在基因进化过程中起到重要作用^[48]。依赖大规模的泛基因组数据,不仅能确认已有报道的等位基因,也能鉴定包括融合基因在内的基因新结构(图 1C),例如大豆开花相关的主效基因 *E3*^[49]。自然状态下,*E3* 以复等位基因的形式存在^[50]。26 个从头组装基因组的注释基因与 ZH13 的 *E3* 进行比较,可以找到一个从 *E3* 第 3 个内含子开始的 13.3 kb 缺失。该变异造成了其中一个基因(*SoyZH13_19G210500*)的完全丢失^[39]。RNAseq 数据显示该变异除了导致 *E3* 的最后一个外显子及 *SoyZH13_19G210500* 的缺失外,还造成了 *E3* 和 *SoyZH13_19G210600* 的转录本读通。此外,该变异还造成了 *E3* 在缺失最后一个外显子后获得了一个额外的外显子。PCR 片段测序验证了 *E3* 与 *SoyZH13_19G210600* 的基因融合事件,以及外显子改变事件是真实存在且相互独立的。泛基因组挖掘并验证了 *E3* 基因由结构变异产生的大量多态性,包括基因融合与 ORF 改变等,这可能是塑造大豆区域适应性分化的重要原因。

大豆的许多性状控制遗传位点,由于变异类型复杂、涉及基因多而难以被克隆^[17,51-54]。大规模从头组装的基因组使得这类解析变得可能(图 1D)。大豆种皮颜色相关的 *I* 位点是受驯化位点^[54,55],使大豆种皮从黑色转变为黄色。该位点为一系列异黄酮代谢途径中查尔酮合成酶(*CHS*)基因组成的基因簇,存在同源依赖的基因沉默(homology dependent gene silencing, HDGS)机制,调控 *CHS* 基因的表达^[56-58]。Liu 等^[39]在 29 个大豆基因组中调查种皮颜色的表型以及 *I* 位点,发现 4 个野生大豆和农家种 *SoyL02* 表现为黑色种皮,其余栽培大豆均为黄色种皮。*I* 位点及周边的 SNP 构建系统发育树发现黑或黄种皮的样品各自聚类在一起。结构变异分析表明,

相对于黑种皮类型基因组,一部分黄种皮样品的基因组上存在一个约 100 kb 的倒位以及 *CHS* 序列单元的重复,这与之前的报道相符^[59]。然而另一部分样品中,虽然这个约 100 kb 的倒位变异不存在,仍然表现出黄色种皮。尽管如此,其上有一段约 23 kb 的序列发生了重复,并且插入到其后的 *CHS* 反向重复基因簇中,而这很可能导致了双交换事件并造成周围 *CHS* 单元的假基因化。因此,1 位点周围的染色体变异得到完整的解析,而调控机制有待于进一步探索。

基因表达可能受到基因附近调控区序列变异的影响,进而导致农艺性状的变化。泛基因组结合转录组的研究策略能够深入挖掘由染色体结构变异导致的表达量差异,从而定位农艺性状的候选基因和变异(图 1E)。缺铁萎黄是大豆在石灰土中种植时常见的病症。Lin 等^[60]的研究已定位到若干与铁离子利用效率相关的 QTL 位点,其中一个位于 14 号染色体。该 QTL 中存在一个注释为铁/锌离子调控转运蛋白的基因 *SoyZH13_14G179600*,其 5'启动子区在泛基因组中检测到一个 1.4 kb 的 PAV^[39]。该 PAV 满足转座子 *DNA Mutator* 的序列特征^[61],并且可以将 26 个大豆种质分成两组:未发生序列缺失和发生序列缺失的类型。RNA-seq 数据表明,后者相对前者具有更高的表达量。结合群体基因型数据和样品信息记录发现,1.4 kb 序列缺失的样品主要分布在纬度更高的种植区,而未发生序列缺失样品分布在纬度较低的地理区域。中国不同地理区域的土壤 pH 不同,进而影响铁离子浓度。因此,区域差异可能是造成遗传分化的诱因。

3.5 多年生大豆泛基因组研究

大豆属除了分布于东亚地区的一年生大豆(*Soja* 亚属)之外,还有约 30 个分布于澳大利亚的多年生大豆物种(*Glycine* 亚属)。该类群虽然和栽培大豆分化较大,但是部分物种染色体数目与栽培大豆相同,可能是栽培大豆潜在的遗传改良基因资源库,具有研究价值。2022 年,一项针对 *Glycine* 亚属 6 个物种(5 个二倍体和 1 个四倍体)的泛基因组研究系统地揭示了多年生大豆的基因组演化特征^[62]。二倍体物种基因组大小为 935.6~1373.8 Mb,平均大小 1105 Mb 左右,与 *Soja* 亚属大致接近,而基因组预测的蛋白

质编码基因有 70% 在一年生大豆中缺失。多年生大豆相对栽培大豆而言,整体基因组变异幅度较大,遗传资源应用可能更侧重于定向基因改造或替换而非远源杂交。

以菜豆(*Phaseolus vulgaris*)为参考的比较基因组发现,多年生大豆相对于一年生大豆,基因组重排事件更少,染色体更为稳定。Zhuang 等^[62]研究计算了同源基因家族在一年生、多年生大豆中的 Ka/Ks,发现 52 个家族在两个亚属中发生了净化选择;其中 *PHP*、*D14* 等是与开花、植株发育相关的基因,在两个亚属内计算 Ka/Ks 值低,但是在亚属间计算则具有较高的 Ka/Ks 值,暗示这些基因可能参与了亚属间生活史策略的分化。

物种多倍化后,往往会发生冗余基因的丢失,导致亚基因组的分化,这种分化通常具有偏好性^[63,64]。Zhuang 等^[62]分别比较四倍体多年生大豆 *G. dolichocarpa* 的两套亚基因组(A'A'D'D'),发现多倍化前后两套对应基因组间染色体序列重排少相对保守,而多倍化后的基因组上发生了不同程度的基因丢失;在 *G. dolichocarpa* 中,D'基因组上丢失了 4019 个基因,显著多于 A'基因组上丢失的 3242 个基因;且相较于丢失的基因,保留的基因在原基因组上的表达量更高。这些迹象表明,A 亚基因组相对于 D 亚基因组具有明显的基因组优势(图 1F)。

4 结语与展望

4.1 未来泛基因组发展

测序技术在过去的 40 年间飞速发展,积累了海量的数据,包括大规模群体测序和从头组装基因组。在此基础上,泛基因组学应运而生,并且受到学界越来越多的重视^[4,14,65~70],成为作物遗传育种研究的“利器”^[35,40,71]。水稻、玉米、大豆、番茄等作物中不断有泛基因组研究涌现,这些结果或展示了不同研究类群框架下的基因组差异特征,或随着研究技术的提升给出了更高质量的组学参考数据。泛基因组作为一种基于比较基因组的研究方式,研究对象的选择尤为关键。应根据研究目的划定适合的类群范围,挑选代表性个体。泛基因组构建策略的选择应根据样品数量、测序成本以及最终期望呈现的

数据结果综合考虑。图泛基因组作为当下泛基因组研究的前沿和热点,整合构建图泛基因组的算法和软件逐渐多样成熟,但这些算法软件多针对人类泛基因组的研究开发。目前植物研究中主要的泛基因组构建策略多是通过三代测序获得高质量的从头组装染色体水平基因组,再借由比较基因组分析结构变异构建图泛基因组。而图泛基因组本身并不依赖除底盘基因组外其他样品的染色体水平基因组组装,因此,三代测序直接检测结构变异结合底盘基因组构建图泛基因组的方法可能是更低成本及更便利的一种方式。此外,针对植物基因组特征,开发解决重复序列比例大、染色体结构变异复杂、基因组大小差异显著的算法和软件,将能够有效提升植物图泛基因组的精度和构建效率。

未来,对单一物种构建泛基因组或许不是最终的目标,目前已有许多探索正在朝此发展。地球生物基因组计划(Earth BioGenome Project)旨在组装所有已知真核生物的代表性基因组^[72]。类似的还有万种植物基因组计划(The Plant 10000 Genomes Project)等,该项目计划对所有有胚植物、绿藻、原生生物的主要支系的代表性基因组进行测序并展开特征化描述^[73]。此外,泛组学概念并不局限于经典的基因组,泛三维基因组、泛转录组等多层次泛组学是今后值得尝试的方向。

4.2 多维组学数据应用

大数据时代下,新的数据类型不断涌现,其应用和处理场景也日趋复杂。泛基因组研究通常会在一个物种/类群内产生多套参考基因组数据。建立这些基因组间的关联,高效地进行多基因组的联合检索和调用,是后基因组时代迫切的数据需求。图泛基因组是对这类问题很好的回答,但也带来了新的挑战。首先图基因组是与以往不同的数据形式,针对这类数据开发的数据库和前端应用目前仍然有限。如何将这类数据高效地服务于更多研究者,是值得探索的方向。大豆多维组学数据库 SoyOmics 对图泛基因组的单倍型检索和数据可视化提供了实践参考^[74]。此外,全景多维组学的发展,对于当下数据的提炼和整合能力有了更深的要求。通过多维组学数据的联合应用,提升生物信息学分析结果的精度和可信度,从而提高作物遗传解析效率,最终服务于分子

设计育种^[75]。在此过程中,针对多层次组学信号的联合处理与评估,以及多层次组学数据网络的构建,应该成为未来探索的重要方向。

在后基因组时代,泛基因组能够起到对传统基因组的补充和发展作用,其价值和必要性已被证实。在大豆中,泛基因组、变异组、转录组、表观组、表型组等多维度数据已有充分的积累。未来的遗传育种研究应当利用好这些多维组学数据,深度解析重要农艺性状的遗传网络,为分子设计育种提供有力指导,这也是提升大豆产量、改善大豆品质的重要途径。

参考文献(References):

- [1] Clark JW, Donoghue PCJ. Whole-genome duplication and plant macroevolution. *Trends Plant Sci*, 2018, 23(10): 933–945. [\[DOI\]](#)
- [2] Danilevich MF, Tay Fernandez CG, Marsh JJ, Bayer PE, Edwards D. Plant pangenomics: approaches, applications and advancements. *Curr Opin Plant Biol*, 2020, 54: 18–25. [\[DOI\]](#)
- [3] Saxena RK, Edwards D, Varshney RK. Structural variations in plant genomes. *Brief Funct Genomics*, 2014, 13(4): 296–307. [\[DOI\]](#)
- [4] Golicz AA, Batley J, Edwards D. Towards plant pangenomics. *Plant Biotechnol J*, 2016, 14(4): 1099–1105. [\[DOI\]](#)
- [5] Tao YF, Zhao XR, Mace E, Henry R, Jordan D. Exploring and exploiting pan-genomics for crop improvement. *Mol Plant*, 2019, 12(2): 156–169. [\[DOI\]](#)
- [6] Tettelin H, Massignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou LW, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc Natl Acad Sci USA*, 2005, 102(39): 13950–13955. [\[DOI\]](#)
- [7] Baker M. *De novo* genome assembly: what every biologist should know. *Nat Methods*, 2012, 9(4): 333–337. [\[DOI\]](#)
- [8] Edger PP, Poorten TJ, VanBuren R, Hardigan MA, Colle M, McKain MR, Smith RD, Teresi SJ, Nelson ADL, Wai

- CM, Alger EI, Bird KA, Yocca AE, Pumplin N, Ou SJ, Ben-Zvi G, Brodt A, Baruch K, Swale T, Shiue L, Acharya CB, Cole GS, Mower JP, Childs KL, Jiang N, Lyons E, Freeling M, Puzey JR, Knapp SJ. Origin and evolution of the octoploid strawberry genome. *Nat Genet*, 2019, 51(3): 541–547. [DOI]
- [9] Huang SF, Kang MJ, Xu AL. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics*, 2017, 33(16): 2577–2579. [DOI]
- [10] Zhang JS, Zhang XT, Tang HB, Zhang Q, Hua XT, Ma XK, Zhu F, Jones T, Zhu XG, Bowers J, Wai CM, Zheng CF, Shi Y, Chen S, Xu XM, Yue JJ, Nelson DR, Huang LX, Li Z, Xu HM, Zhou D, Wang YJ, Hu WC, Lin JS, Deng YJ, Pandey N, Mancini M, Zerpa D, Nguyen JK, Wang LM, Yu L, Xin YH, Ge LF, Arro J, Han JO, Chakrabarty S, Pushko M, Zhang WP, Ma YH, Ma PP, Lv MJ, Chen FM, Zheng GY, Xu JS, Yang ZH, Deng F, Chen XQ, Liao ZY, Zhang XX, Lin ZC, Lin H, Yan HS, Kuang Z, Zhong WM, Liang PP, Wang GF, Yuan Y, Shi JX, Hou JX, Lin JX, Jin JJ, Cao PJ, Shen QC, Jiang Q, Zhou P, Ma YY, Zhang XD, Xu RR, Liu J, Zhou YM, Jia HF, Ma Q, Qi R, Zhang ZL, Fang JP, Fang HK, Song JJ, Wang MJ, Dong GR, Wang G, Chen Z, Ma T, Liu H, Dhungana SR, Huss SE, Yang XP, Sharma A, Trujillo JH, Martinez MC, Hudson M, Riascos JJ, Schuler M, Chen LQ, Braun DM, Li L, Yu QY, Wang JP, Wang K, Schatz MC, Heckerman D, Van Sluys MA, Souza GM, Moore PH, Sankoff D, VanBuren R, Paterson AH, Nagai C, Ming R. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat Genet*, 2018, 50(11): 1565–1573. [DOI]
- [11] Sherman RM, Salzberg SL. Pan-genomics in the human genome era. *Nat Rev Genet*, 2020, 21(4): 243–254. [DOI]
- [12] Ni LB, Liu YC, Ma X, Liu TF, Yang XY, Wang Z, Liang QJ, Liu SL, Zhang M, Wang Z, Shen YT, Tian ZX. Pan-3D genome analysis reveals structural and functional differentiation of soybean genomes. *Genome Biol*, 2023, 24(1): 12. [DOI]
- [13] Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Peñagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N, Kaeppler SM, Buell CR. Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, 2014, 26(1): 121–135. [DOI]
- [14] Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol*, 2015, 23: 148–154. [DOI]
- [15] De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. *Nat Rev Genet*, 2021, 22(9): 572–587. [DOI]
- [16] Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu SQ, Stritt C, Roulin AC, Schackwitz W, Tyler L, Martin J, Lipzen A, Dochy N, Phillips J, Barry K, Geuten K, Budak H, Juenger TE, Amasino R, Caicedo AL, Goodstein D, Davidson P, Mur LAJ, Figueroa M, Freeling M, Catalan P, Vogel JP. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat Commun*, 2017, 8(1): 2184. [DOI]
- [17] Li YH, Zhou GY, Ma JX, Jiang WK, Jin LG, Zhang ZH, Guo Y, Zhang JB, Sui Y, Zheng LT, Zhang SS, Zuo QY, Shi XH, Li YF, Zhang WK, Hu YY, Kong GY, Hong HL, Tan B, Song J, Liu ZX, Wang YS, Ruan H, Yeung CKL, Liu J, Wang HL, Zhang LJ, Guan RX, Wang KJ, Li WB, Chen SY, Chang RZ, Jiang Z, Jackson SA, Li RQ, Qiu LJ. *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol*, 2014, 32(10): 1045–1052. [DOI]
- [18] Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang CJ, Chougule K, Gao DY, Iwata A, Goicoechea JL, Wei SR, Wang J, Liao Y, Wang MH, Jacquemin J, Becker C, Kudrna D, Zhang JW, Londono CEM, Song X, Lee S, Sanchez P, Zuccolo A, Ammiraju JSS, Talag J, Danowitz A, Rivera LF, Gschwend AR, Noutsos C, Wu CC, Kao SM, Zeng JW, Wei FJ, Zhao Q, Feng Q, El Baidouri M, Carpentier MC, Lasserre E, Cooke R, da Rosa Farias D, da Maia LC, Dos Santos RS, Nyberg KG, McNally KL, Mauleon R, Alexandrov N, Schmutz J, Flowers D, Fan CZ, Weigel D, Jena KK, Wicker T, Chen MS, Han B, Henry R, Hsing YC, Kurata N, de Oliveira AC, Panaud O, Jackson SA, Machado CA, Sanderson MJ, Long MY, Ware D, Wing RA. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat Genet*, 2018, 50(2): 285–296. [DOI]
- [19] Gao L, Gonda I, Sun HH, Ma QY, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu YM, van der Knaap E, Huang SW, Klee HJ, Giovannoni JJ, Fei ZQ. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet*, 2019, 51(6): 1044–1051. [DOI]
- [20] Hübner S, Bercovich N, Todesco M, Mandel JR, Odenheimer J, Ziegler E, Lee JS, Baute GJ, Owens GL, Grassa CJ, Ebert DP, Ostevik KL, Moyers BT, Yakimowski S, Masalia RR, Gao LX, Čalić I, Bowers JE, Kane NC, Swanevelter DZH, Kubach T, Muñoz S, Langlade NB, Burke JM, Rieseberg LH. Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance. *Nat Plants*, 2019, 5(1): 54–62. [DOI]
- [21] Wang WS, Mauleon R, Hu ZQ, Chebotarov D, Tai SS, Wu ZC, Li M, Zheng TQ, Fuentes RR, Zhang F, Mansueto L,

- Copetti D, Sanciango M, Palis KC, Xu JL, Sun C, Fu BY, Zhang HL, Gao YM, Zhao XQ, Shen F, Cui X, Yu H, Li ZC, Chen ML, Detras J, Zhou YL, Zhang XY, Zhao Y, Kudrna D, Wang CC, Li R, Jia B, Lu JY, He XC, Dong ZT, Xu JB, Li YH, Wang M, Shi JX, Li J, Zhang DB, Lee S, Hu WS, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu YC, Yue Z, Naredo MEB, Talag J, Wang XQ, Li JJ, Fang XD, Yin Y, Glaszmann JC, Zhang JW, Li JY, Hamilton RS, Wing RA, Ruan J, Zhang GY, Wei CC, Alexandrov N, McNally KL, Li ZK, Leung H. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, 2018, 557(7703): 43–49. [DOI]
- [22] Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis A, McCombie WR, Parkin IAP, Paterson AH, Pires JC, Sharpe AG, Tang HB, Teakle GR, Town CD, Batley J, Edwards D. The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat Commun*, 2016, 7: 13390. [DOI]
- [23] Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. *De novo* assembly and genotyping of variants using colored *de Bruijn* graphs. *Nat Genet*, 2012, 44(2): 226–232. [DOI]
- [24] Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, Eichler EE. Characterizing the major structural variant alleles of the human genome. *Cell*, 2019, 176(3): 663–675. [DOI]
- [25] Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson BV, Melsted P. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun*, 2019, 10(1): 5402. [DOI]
- [26] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol*, 2018, 36(9): 875–879. [DOI]
- [27] Marcus S, Lee H, Schatz MC. SplitMEM: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics*, 2014, 30(24): 3476–3483. [DOI]
- [28] Zhao YB, Jia XM, Yang JH, Ling YC, Zhang Z, Yu J, Wu JY, Xiao JF. PanGP: a tool for quickly analyzing bacterial pan-genome profile. *Bioinformatics*, 2014, 30(9): 1297–1299. [DOI]
- [29] Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, Sibbesen JA, Hickey G, Chang PC, Carroll A, Gupta N, Gabriel S, Blackwell TW, Ratan A, Taylor KD, Rich SS, Rotter JJ, Haussler D, Garrison E, Paten B. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 2021, 374(6574): abg8871. [DOI]
- [30] Guarracino A, Heumos S, Nahnsen S, Prins P, Garrison E. ODGI: understanding pangenome graphs. *Bioinformatics*, 2022, 38(13): 3319–3326. [DOI]
- [31] Garrison E, Guarracino A, Heumos S, Villani F, Bao ZG, Tattini L, Hagmann J, Vorbrugg S, Marco-Sola S, Kubica C, Ashbrook DG, Thorell K, Rusholme-Pilcher RL, Liti G, Rudbeck E, Nahnsen S, Yang ZY, Moses MN, Nobrega FL, Wu Y, Chen H, de Ligt J, Sudmant PH, Soranzo N, Colonna V, Williams RW, Prins P. Building pangenome graphs. *bioRxiv*, 2023. [DOI]
- [32] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*, 2019, 37(8): 907–915. [DOI]
- [33] Gan XC, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX, Clark RM, Ratsch G, Mott R. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, 2011, 477(7365): 419–423. [DOI]
- [34] Qin P, Lu HW, Du HL, Wang H, Chen WL, Chen Z, He Q, Ou SJ, Zhang HY, Li XZ, Li XX, Li Y, Liao Y, Gao Q, Tu B, Yuan H, Ma BT, Wang YP, Qian YW, Fan SJ, Li WT, Wang J, He M, Yin JJ, Li T, Jiang N, Chen XW, Liang CZ, Li SG. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, 2021, 184(13): 3542–3558. [DOI]
- [35] Zhou Y, Zhang ZY, Bao ZG, Li HB, Lyu YQ, Zan YJ, Wu YY, Cheng L, Fang YH, Wu K, Zhang JZ, Lyu HJ, Lin T, Gao Q, Saha S, Mueller L, Fei ZJ, Städler T, Xu SZ, Zhang ZW, Speed D, Huang SW. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, 2022, 606(7914): 527–534. [DOI]
- [36] Huang Y, He JX, Xu YT, Zheng WK, Wang SH, Chen P, Zeng B, Yang SZ, Jiang XL, Liu ZS, Wang L, Wang X, Liu SJ, Lu ZH, Liu Z, Yu HW, Yue JQ, Gao JY, Zhou XY, Long CR, Zeng XL, Guo YJ, Zhang WF, Xie ZZ, Li CL, Ma ZC, Jiao WB, Zhang F, Larkin RM, Krueger RR, Smith MW, Ming R, Deng XX, Xu Q. 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(11): 1964–1975. [DOI]
- [37] Jin SK, Han ZG, Hu Y, Si ZF, Dai F, He L, Cheng Y, Li YQ, Zhao T, Fang L, Zhang TZ. Structural variation (SV)-based pan-genome and GWAS reveal the impacts of SVs on the speciation and diversification of allotetraploid cottons. *Mol Plant*, 2023, 16(4): 678–693. [DOI]
- [38] Li HB, Wang SH, Chai S, Yang ZQ, Zhang QQ, Xin HJ,

- Xu YC, Lin SG, Chen XX, Yao ZW, Yang QY, Fei ZJ, Huang SW, Zhang ZH. Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat Commun*, 2022, 13(1): 682. [DOI]
- [39] Liu YC, Du HL, Li PC, Shen YT, Peng H, Liu SL, Zhou G-A, Zhang HK, Liu Z, Shi M, Huang XH, Li Y, Zhang M, Wang Z, Zhu BG, Han B, Liang CZ, Tian ZX. Pan-genome of wild and cultivated soybeans. *Cell*, 2020, 182(1): 162–176. [DOI]
- [40] He Q, Tang S, Zhi H, Chen JF, Zhang J, Liang HK, Alam O, Li HB, Zhang H, Xing LH, Li XK, Zhang W, Wang HL, Shi JP, Du HL, Wu HP, Wang LW, Yang P, Xing L, Yan HS, Song ZQ, Liu JR, Wang HG, Tian X, Qiao ZJ, Feng GJ, Guo RF, Zhu WJ, Ren YM, Hao HB, Li MZ, Zhang AY, Guo EH, Yan F, Li QQ, Liu YL, Tian BH, Zhao XQ, Jia RL, Feng BL, Zhang JW, Wei JH, Lai JS, Jia GQ, Purugganan M, Diao XM. A graph-based genome and pan-genome variation of the model plant *Setaria*. *Nat Genet*, 2023, 55(7): 1232–1242. [DOI]
- [41] Chen S, Wang PJ, Kong WL, Chai K, Zhang SC, Yu JX, Wang YB, Jiang MW, Lei WL, Chen X, Wang WL, Gao YY, Qu SY, Wang F, Wang YH, Zhang Q, Gu MY, Fang KX, Ma CL, Sun WJ, Ye NX, Wu HL, Zhang XT. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant *Camellia sinensis*. *Nat Plants*, 2023, 9(12): 1986–1999. [DOI]
- [42] Zhao Q, Feng Q, Lu HY, Li Y, Wang AH, Tian QL, Zhan QL, Lu YQ, Zhang L, Huang T, Wang YC, Fan DL, Zhao Y, Wang ZQ, Zhou CC, Chen JY, Zhu CR, Li WJ, Weng QJ, Xu Q, Wang ZX, Wei XH, Han B, Huang XH. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet*, 2018, 50(2): 278–284. [DOI]
- [43] Song JM, Guan ZL, Hu JL, Guo CC, Yang ZQ, Wang S, Liu DX, Wang B, Lu SP, Zhou R, Xie WZ, Cheng YF, Zhang YT, Liu KD, Yang QY, Chen LL, Guo L. Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat Plants*, 2020, 6(1): 34–45. [DOI]
- [44] Shang LG, Li XX, He HY, Yuan QL, Song YN, Wei ZR, Lin H, Hu M, Zhao FL, Zhang C, Li YH, Gao HS, Wang TY, Liu XP, Zhang H, Zhang Y, Cao SM, Yu XM, Zhang BT, Zhang Y, Tan YQ, Qin M, Ai C, Yang YX, Zhang B, Hu ZQ, Wang HR, Lv Y, Wang YX, Ma J, Wang Q, Lu HW, Wu Z, Liu SL, Sun ZY, Zhang HL, Guo LB, Li ZC, Zhou YF, Li JY, Zhu ZF, Xiong GS, Ruan J, Qian Q. A super pan-genomic landscape of rice. *Cell Res*, 2022, 32(10): 878–896. [DOI]
- [45] Alonge M, Wang XG, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, Levy Y, Harel TH, Shalev-Schlosser G, Amsellem Z, Razifard H, Caicedo AL, Tieman DM, Klee H, Kirsche M, Aganezov S, Ranallo-Benavidez TR, Lemmon ZH, Kim J, Robitaille G, Kramer M, Goodwin S, McCombie WR, Hutton S, Van Eck J, Gillis J, Eshed Y, Sedlazeck FJ, van der Knaap E, Schatz MC, Lippman ZB. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*, 2020, 182(1): 145–161. [DOI]
- [46] Chen JF, Liu Y, Liu MX, Guo WL, Wang YQ, He Q, Chen WY, Liao Y, Zhang W, Gao YZ, Dong KJ, Ren RY, Yang TY, Zhang LY, Qi MY, Li ZG, Zhao M, Wang HG, Wang JJ, Qiao ZJ, Li HQ, Jiang YM, Liu GQ, Song XQ, Deng YR, Li H, Yan F, Dong Y, Li QQ, Li T, Yang WY, Cui JH, Wang HR, Zhou YF, Zhang XM, Jia GQ, Lu P, Zhi H, Tang S, Diao XM. Pangenome analysis reveals genomic variations associated with domestication traits in broomcorn millet. *Nat Genet*, 2023, 55(12): 2243–2254. [DOI]
- [47] Gijzen M, Weng CR, Kuflu K, Woodrow L, Yu KF, Poysa V. Soybean seed lustre phenotype and surface protein cosegregate and map to linkage group E. *Genome*, 2003, 46(4): 659–664. [DOI]
- [48] Jones CD, Begun DJ. Parallel evolution of chimeric fusion genes. *Proc Natl Acad Sci USA*, 2005, 102(32): 11373–11378. [DOI]
- [49] Watanabe S, Hideshima R, Xia ZJ, Tsubokura Y, Sato S, Nakamoto Y, Yamanaka N, Takahashi R, Ishimoto M, Anai T, Tabata S, Harada K. Map-based cloning of the gene associated with the soybean maturity locus *E3*. *Genetics*, 2009, 182(4): 1251–1262. [DOI]
- [50] Tsubokura Y, Watanabe S, Xia ZJ, Kanamori H, Yamagata H, Kaga A, Katayose Y, Abe J, Ishimoto M, Harada K. Natural variation in the genes responsible for maturity loci *E1*, *E2*, *E3* and *E4* in soybean. *Ann Bot*, 2014, 113(3): 429–441. [DOI]
- [51] Lam HM, Xu X, Liu X, Chen WB, Yang GH, Wong FL, Li MW, He WM, Qin N, Wang B, Li J, Jian M, Wang J, Shao GH, Wang J, Sun SSM, Zhang GY. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet*, 2010, 42(12): 1053–1059. [DOI]
- [52] Lu SJ, Zhao XH, Hu YL, Liu SL, Nan HY, Li XM, Fang C, Cao D, Shi XY, Kong LP, Su T, Zhang FG, Li SC, Wang Z, Yuan XH, Cober ER, Weller JL, Liu BH, Hou XL, Tian ZX, Kong FJ. Natural variation at the soybean *J* locus improves adaptation to the tropics and enhances yield. *Nat Genet*, 2017, 49(5): 773–779. [DOI]
- [53] Torkamaneh D, Laroche J, Tardivel A, O'Donoghue L, Cober E, Rajcan I, Belzile F. Comprehensive description of genomewide nucleotide and structural variation in

- short-season soya bean. *Plant Biotechnol J*, 2018, 16(3): 749–759. [DOI]
- [54] Zhou ZK, Jiang Y, Wang Z, Gou ZH, Lyu J, Li WY, Yu YJ, Shu LP, Zhao YJ, Ma YM, Fang C, Shen YT, Liu TF, Li CC, Li Q, Wu M, Wang M, Wu YS, Dong Y, Wan WT, Wang X, Ding ZL, Gao YD, Xiang H, Zhu BG, Lee SH, Wang W, Tian ZX. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol*, 2015, 33(4): 408–414. [DOI]
- [55] Woodworth CM. Inheritance of cotyledon, seed-coat, hilum and pubescence colors in soy-beans. *Genetics*, 1921, 6(6): 487–553. [DOI]
- [56] Tuteja JH, Clough SJ, Chan WC, Vodkin LO. Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in *Glycine max*. *Plant Cell*, 2004, 16(4): 819–835. [DOI]
- [57] Tuteja JH, Zabala G, Varala K, Hudson M, Vodkin LO. Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats. *Plant Cell*, 2009, 21(10): 3063–3077. [DOI]
- [58] Wang CS, Todd JJ, Vodkin LO. Chalcone synthase mRNA and activity are reduced in yellow soybean seed coats with dominant *I* alleles. *Plant Physiol*, 1994, 105(2): 739–748. [DOI]
- [59] Xie M, Chung CYL, Li MW, Wong FL, Wang X, Liu AL, Wang ZL, Leung AKY, Wong TH, Tong SW, Xiao ZX, Fan KJ, Ng MS, Qi XP, Yang LF, Deng TQ, He LJ, Chen L, Fu AS, Ding Q, He JX, Chung G, Isobe S, Tanabata T, Valliyodan B, Nguyen HT, Cannon SB, Foyer CH, Chan TF, Lam HM. A reference-grade wild soybean genome. *Nat Commun*, 2019, 10(1): 1216. [DOI]
- [60] Lin S, Cianzio S, Shoemaker R. Mapping genetic loci for iron deficiency chlorosis in soybean. *Mol Breeding*, 1997, 3(3): 219–229. [DOI]
- [61] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capi P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, 2007, 8(12): 973–982. [DOI]
- [62] Zhuang YB, Wang XT, Li XC, Hu JM, Fan LC, Landis JB, Cannon SB, Grimwood J, Schmutz J, Jackson SA, Doyle JJ, Zhang XS, Zhang DJ, Ma JX. Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. *Nat Plants*, 2022, 8(3): 233–244. [DOI]
- [63] Wendel JF. The wondrous cycles of polyploidy in plants. *Am J Bot*, 2015, 102(11): 1753–1756. [DOI]
- [64] Zhao MX, Zhang B, Lisch D, Ma JX. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell*, 2017, 29(12): 2974–2994. [DOI]
- [65] Ameer A. Goodbye reference, hello genome graphs. *Nat Biotechnol*, 2019, 37(8): 866–868. [DOI]
- [66] Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. Plant pan-genomes are the new reference. *Nature Plants*, 2020, 6: 914–920. [DOI]
- [67] Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic analysis in the age of human genome sequencing. *Cell*, 2019, 177(1): 70–84. [DOI]
- [68] Huang XH, Huang SW, Han B, Li JY. The integrated genomics of crop domestication and breeding. *Cell*, 2022, 185(15): 2828–2839. [DOI]
- [69] Shi JP, Tian ZX, Lai JS, Huang XH. Plant pan-genomics and its applications. *Mol Plant*, 2023, 16(1): 168–186. [DOI]
- [70] Lei L, Goltsman E, Goodstein D, Wu GA, Rokhsar DS, Vogel JP. Plant pan-genomics comes of age. *Annu Rev Plant Biol*, 2021, 72: 411–435. [DOI]
- [71] Yu H, Lin T, Meng XB, Du HL, Zhang JK, Liu GF, Chen MJ, Jing YH, Kou LQ, Li XX, Gao Q, Liang Y, Liu XD, Fan ZL, Liang YT, Cheng ZK, Chen MS, Tian ZX, Wang YH, Chu CC, Zuo JR, Wan JM, Qian Q, Han B, Zuccolo A, Wing RA, Gao CX, Liang CZ, Li JY. A route to *de novo* domestication of wild allotetraploid rice. *Cell*, 2021, 184(5): 1156–1170. e14. [DOI]
- [72] Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington JA, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, Goldstein MM, Grigoriev IV, Hackett KJ, Haussler D, Jarvis ED, Johnson WE, Patrinos A, Richards S, Castilla-Rubio JC, van Sluys MA, Soltis PS, Xu X, Yang HM. Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci USA*, 2018, 115(17): 4325–4333. [DOI]
- [73] Cheng S, Melkonian M, Smith SA, Brockington SF, Archibald JM, Delaux PM, Li F, Melkonian B, Mavrodiev EV, Sun WJ, Fu Y, Yang HM, Soltis DE, Graham SW, Soltis PS, Liu X, Xu X, Wong GKS. 10KP: a phylodiverse genome sequencing plan. *GigaScience*, 2018, 7(3): 1–9. [DOI]
- [74] Liu YC, Zhang Y, Liu XN, Shen YT, Tian DM, Yang XY, Liu SL, Ni LB, Zhang Z, Song SH, Tian ZX. SoyOmics: a deeply integrated database on soybean multi-omics. *Mol Plant*, 2023, 16(5): 794–797. [DOI]
- [75] Han LQ, Zhong WS, Qian J, Jin ML, Tian P, Zhu WC, Zhang HW, Sun YH, Feng JW, Liu XG, Chen G, Farid B, Li RN, Xiong ZM, Tian ZH, Li J, Luo Z, Du DX, Chen SJ, Jin QX, Li JX, Li Z, Liang Y, Jin XM, Peng Y, Zheng C, Ye XN, Yin YJ, Chen H, Li WF, Chen LL, Li Q, Yan JB, Yang F, Li L. A multi-omics integrative network map of maize. *Nat Genet*, 2023, 55(1): 144–153. [DOI]

- [76] Zhou P, Silverstein KAT, Ramaraj T, Guhlin J, Denny R, Liu JQ, Farmer AD, Steele KP, Stupar RM, Miller JR, Tiffin P, Mudge J, Young ND. Exploring structural variation and gene family architecture with *De novo* assemblies of 15 *Medicago* genomes. *BMC Genomics*, 2017, 18(1): 261. [DOI]
- [77] Ou LJ, Li D, Lv JH, Chen WC, Zhang ZQ, Li XF, Yang BZ, Zhou SD, Yang S, Li WG, Gao HZ, Zeng Q, Yu HY, Ouyang B, Li F, Liu F, Zheng JY, Liu YH, Wang J, Wang BB, Dai XZ, Ma YQ, Zou XX. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. *New Phytol*, 2018, 220(2): 360–363. [DOI]
- [78] Yu JY, Golicz AA, Lu K, Dossa K, Zhang YX, Chen JF, Wang LH, You J, Fan DD, Edwards D, Zhang XR. Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnol J*, 2019, 17(5): 881–892. [DOI]
- [79] Jayakodi M, Padmarasu S, Haberer G, Bonthala VS, Gundlach H, Monat C, Lux T, Kamal N, Lang D, Himmelbach A, Ens J, Zhang XQ, Angessa TT, Zhou GF, Tan C, Hill C, Wang PH, Schreiber M, Boston LB, Plott C, Jenkins J, Guo Y, Fiebig A, Budak H, Xu DD, Zhang J, Wang CC, Grimwood J, Schmutz J, Guo GG, Zhang GP, Mochida K, Hirayama T, Sato K, Chalmers KJ, Langridge P, Waugh R, Pozniak CJ, Scholz U, Mayer KFX, Spannagl M, Li C, Mascher M, Stein N. The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, 2020, 588(7837): 284–289. [DOI]
- [80] Varshney RK, Roorkiwal M, Sun S, Bajaj P, Chitkineni A, Thudi M, Singh NP, Du X, Upadhyaya HD, Khan AW, Wang Y, Garg V, Fan Gy, Cowling WA, Crossa J, Gentzbittel L, Voss-Fels KP, Valluri VK, Sinha P, Singh VK, Ben C, Rathore A, Punna R, Singh MK, Tar'an B, Bharadwaj C, Yasin M, Pithia MS, Singh S, Soren KR, Kudapa H, Jarquín D, Cubry P, Hickey LT, Dixit GP, Thuillet AC, Hamwieh A, Kumar S, Deokar AA, Chaturvedi SK, Francis A, Howard R, Chattopadhyay D, Edwards D, Lyons E, Vigouroux Y, Hayes BJ, von Wettberg E, Datta SK, Yang HM, Nguyen HT, Wang J, Siddique KHM, Mohapatra T, Bennetzen JL, Xu X, Liu X. A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature*, 2021, 599(7886): 622–627. [DOI]
- [81] Li JY, Yuan DJ, Wang PC, Wang QQ, Sun ML, Liu ZP, Si H, Xu ZP, Ma YZ, Zhang BY, Pei LL, Tu LL, Zhu LF, Chen LL, Lindsey K, Zhang XL, Jin SX, Wang MJ. Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome Biol*, 2021, 22(1): 119. [DOI]
- [82] Tao YF, Luo H, Xu JB, Cruickshank A, Zhao XR, Teng F, Hathorn A, Wu XY, Liu YM, Shatte T, Jordan D, Jing HC, Mace E. Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat Plants*, 2021, 7(6): 766–773. [DOI]
- [83] Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou SJ, Liu JN, Ricci WA, Guo TT, Olson A, Qiu YJ, Della Coletta R, Tittes S, Hudson AI, Marand AP, Wei SR, Lu ZY, Wang B, Tello-Ruiz MK, Piri RD, Wang N, Kim DW, Zeng YB, O'Connor CH, Li XR, Gilbert AM, Baggs E, Krasileva KV, Portwood JL, 2nd, Cannon EKS, Andorf CM, Manchanda N, Snodgrass SJ, Hufnagel DE, Jiang QH, Pedersen S, Syring ML, Kudrna DA, Llaca V, Fengler K, Schmitz RJ, Ross-Ibarra J, Yu JM, Gent JJ, Hirsch CN, Ware D, Dawe RK. *De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, 2021, 373(6555): 655–662. [DOI]
- [84] Zhang XH, Liu TJ, Wang JL, Wang P, Qiu Y, Zhao W, Pang S, Li XM, Wang HP, Song JP, Zhang WL, Yang WL, Sun YY, Li XX. Pan-genome of *Raphanus* highlights genetic variation and introgression among domesticated, wild, and weedy radishes. *Mol Plant*, 2021, 14(12): 2032–2055. [DOI]
- [85] Li N, He Q, Wang J, Wang BK, Zhao JT, Huang SY, Yang T, Tang YP, Yang SB, Aisimutuola P, Xu RQ, Hu JH, Jia CP, Ma K, Li ZQ, Jiang FL, Gao J, Lan HY, Zhou YF, Zhang XY, Huang SW, Fei ZJ, Wang H, Li HB, Yu QH. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet*, 2023, 55(5): 852–860. [DOI]
- [86] Wang MJ, Li JY, Qi ZY, Long YX, Pei LL, Huang XH, Grover CE, Du XM, Xia CJ, Wang PC, Liu ZP, You JQ, Tian XH, Ma YZ, Wang RP, Chen XY, He X, Fang DD, Sun YQ, Tu LL, Jin SX, Zhu LF, Wendel JF, Zhang XL. Genomic innovation and regulatory rewiring during evolution of the cotton genus *Gossypium*. *Nat Genet*, 2022, 54(12): 1959–1971. [DOI]
- [87] Tang D, Jia YX, Zhang JZ, Li HB, Cheng L, Wang P, Bao ZG, Liu ZH, Feng SS, Zhu XJ, Li DW, Zhu GT, Wang HR, Zhou Y, Zhou YF, Bryan GJ, Buell CR, Zhang CZ, Huang SW. Genome evolution and diversity of wild and cultivated potatoes. *Nature*, 2022, 606(7914): 535–541. [DOI]
- [88] Wang BB, Hou M, Shi JP, Ku LX, Song W, Li CH, Ning Q, Li X, Li CY, Zhao BB, Zhang RY, Xu H, Bai ZJ, Xia ZC, Wang H, Kong DX, Wei HB, Jing YF, Dai ZY, Wang HHL, Zhu XY, Li CH, Sun X, Wang SS, Yao W, Hou GG, Qi Z, Dai H, Li XM, Zheng HK, Zhang ZX, Li Y, Wang TY, Jiang TJ, Wan ZM, Chen YH, Zhao JR, Lai JS, Wang HY. *De novo* genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. *Nat Genet*, 2023, 55(2): 312–323. [DOI]

(责任编辑: 孔凡江)

中国科学院遗传与发育生物学研究所田志喜课题组简介

中国科学院遗传与发育生物学研究所田志喜课题组成立于 2010 年。研究团队致力于大豆功能基因组研究和品种培育,以“中华大豆之崛起”为己任,在多维组学立体解析、农艺性状分子机制挖掘、分子育种等方面开展了全方位系统性的工作,取得了一系列重要理论和实践成果。在 *Cell*、*Nature Biotechnology*、*Nature Genetics*、*Nature Communication*、*PNAS*、*Genome Biology*、*Molecular Plant*、*Plant Cell*、*Plant Biotechnology Journal* 等期刊上发表论文 90 篇,总引用 12,000 余次, h 指数 39。其中 ESI 高被引论文 14, 平均单篇他引 123.39 次。多次应邀在 *Current Opinion in Plant Biology*、*Molecular Plant* 等期刊上撰写综述、评论文章。申请专利 7 项, 培育新品种 9 个。团队承担科技部、农业部、国家自然科学基金委及中国科学院的一系列重大项目。获评 2023 年第四届中国科学院“科苑名匠”。

