植物学报 *Chinese Bulletin of Botany* 2020, **55** (4): 403–406, www.chinbullbotany.com
doi: 10.11983/CBB20096

·热点评述·

360度群体遗传变异扫描——大豆泛基因组研究

祝光涛[1], 黄三文[2*]

[1]云南师范大学马铃薯科学研究院, 昆明 650500; [2]中国农业科学院农业基因组研究所, 岭南现代农业科学与技术广东省实验室深圳分中心, 深圳 518120

摘要 大豆(*Glycine max*)是重要的油料和蛋白作物, 其丰富的遗传变异为生物学性状挖掘和育种改良提供了重要的资源基础。然而, 单个基因组信息无法全面揭示种质资源的遗传变异, 泛基因组研究为解决这一不足提供了新方案。近日, 中国科学院遗传与发育生物学研究所田志喜和梁承志研究团队从2 898份大豆种质中选取26份代表性材料, 并整合已有的3个基因组, 构建了包含野生和栽培大豆的泛基因组和图基因组(graph-based genome), 鉴定了整个群体的绝大多数结构变异数据集, 确定了大豆种质的核心、非必需和个体特异的基因集。利用这些数据系统地揭示了生育期位点*E3*的等位基因变异和基因融合事件、种皮颜色基因*I*的单体型和演化关系以及结构变异对铁离子转运基因表达和地区适应性选择的影响。该研究为作物基因组学研究提供了一个新的模式, 同时将加速推动大豆遗传变异的鉴定、性状解析和种质创新。

关键词 大豆, 泛基因组, 图基因组, 遗传变异, 农艺性状

祝光涛, 黄三文 (2020). 360度群体遗传变异扫描——大豆泛基因组研究. 植物学报 **55**, 403–406.

大豆(*Glycine max*)起源于中国, 大约在5 000年前中国人便完成了大豆的驯化, 开始栽培这种作物, 现今大豆已成为世界上最重要的植物油脂和植物蛋白的提供者(Carter et al., 2004)。据统计, 全世界现有60 000份不同类型的大豆种质资源, 其丰富的遗传变异为重要农艺性状的挖掘和育种提供了宝贵的资源。毋庸置疑, 基因组学的发展为我们开展资源探宝提供了强有力的工具。

在种质资源的群体变异与性状挖掘研究中, 通常需要借助1个参考基因组, 通过将重测序数据比对到参考基因上来鉴定个体间的遗传变异(Huang et al., 2012)。这种变异鉴定的方法受制于参考基因组序列及其与检测个体间的相似性, 参考基因组缺失的基因组信息以及与比对个体差异较大区域的信息将无法有效鉴定, 同时, 大片段的插入、缺失、拷贝数等变异类型也无法有效鉴定, 然而这些基因组信息往往具有重要的生物学功能(Lye and Purugganan, 2019)。因此, 单一参考基因组在揭示种质资源丰富变异的研究中越来越显得“心有余而力不足”。

泛基因组指某个物种内所有个体所包含的全部基因组信息, 泛基因组研究为问题的解决提供了新的方案。随着长片段单分子测序技术(如Pacific Bio-sciences (PacBio)和Oxford Nanopore)和辅助组装技术(如BioNano genome mapping和High-throughput chromosome conformation capture (Hi-C))的升级与发展(Jiao et al., 2017), 低成本、高质量地同时对多个基因组进行拼接成为可能, 使得泛基因组研究备受青睐, 它已逐渐成为群体基因组学研究的新方向(Golicz et al., 2016)。近年来, 研究人员已对大豆、水稻(*Oryza sativa*)和番茄(*Lycopersicon esculentum*)等重要作物陆续开展了泛基因组研究(Li et al., 2014; Zhao et al., 2018; Gao et al., 2019)。

近日, 中国科学院遗传与发育生物学研究所田志喜和梁承志领衔的团队对26份大豆种质进行了基因组组装, 并整合已发表的3个基因组, 构建了迄今为止规模最大的大豆泛基因组图谱(Liu et al., 2020)。为了使选择的26份材料具有更广泛的遗传代表性, 他们对世界范围内2 898份大豆资源的进化关系进行分

收稿日期: 2020-05-26; 接受日期: 2020-06-07
基金项目: 云南省基础研究计划杰出青年基金(2020)
* 通讯作者。E-mail: huangsanwen@caas.net.cn

析。其中，新分析了2 027份材料，将整个群体划分为6个类群，从这6个类群中分别筛选代表性材料，最终选出3份野生种、9份地方种和14份栽培种(图1，不同种群种子形态)。然后，利用PacBio单分子测序数据组装，并结合BioNano genome mapping和Hi-C技术进行辅助组装。此研究中Contig N50长度为18.8–26.8 Mb，基因组组装大小为992.3–1 059.8 Mb，相比之前仅使用二代测序数据对7份野生大豆的组装，基因组的连续性提升了近1 000倍(Li et al., 2014)。基因组注释结果显示，平均每个基因组含有553个microRNAs、171个small nuclear RNAs、439个ribosomal RNA和56 522个编码基因；利用BUSCO单拷贝同源基因评估的基因组完整性为95.6% (Liu et al., 2020)。高质量的基因组拼接为后续遗传变异的挖掘奠定了重要基础。

确定整个大豆类群中基因数目是泛基因研究的一个重点。随着基因组数目的增加，鉴定到的基因家族数目也随之增多，当群体的数目达到25的时候，基因家族的数目到达平台期，说明构建的泛基因组基本囊括了大豆种质的全部基因家族。整个泛基因组包含57 492个基因家族，其中包括20 623个核心基因家族、28 679个非必需基因家族及27个个体特异基因家族。基因结构和选择压分析均表明，核心基因较非必需基因集具有更保守的基因功能。除了泛基因组，此研究还构建了1个图基因组(graph-based genome)，它是由29个材料中的(非相似)序列组成(Liu et al., 2020)。相对于单个参考基因组来说，图基因组可提供较多的(非相似)序列和等位基因变异，为群体中复杂结构变异的鉴定提供了更为完善的(非线性)参考基因组。图基因组的构建是该研究的一大亮点，将有效弥补单一参考基因组的不足。

图1 大豆种子形态的群体变异

左边为野生大豆(拼图为汉字“菽”的象形字)，中间为地方种质(拼图为汉字“菽”)，右边为现代栽培种(拼图为汉字“豆”)。

Figure 1 The seed phenotype variations of soybean germplasm

The left part is wild type (hieroglyphic Chinese writing of “shu”), the middle is landrance type (traditional Chinese character of “shu”), and the right is modern type (traditional Chinese character of “dou”)

通过基因组间的比较分析，鉴定了大量的大片段结构变异，包括723 862个存在/缺失变异(presence/absence variants, PAVs)、27 531个拷贝数目变异(copy number variants, CNVs)、21 886个易位事件(translocation events, TEs)和3 120个倒位事件(inversion events)。这些结构变异的长度集中分布在1–200 kb，存在/缺失变异的总长为4.71 Gb，平均每个基因组长度为167.09 Mb，结构变异的长度约占基因组总长度的16%。这些变异在常规的重测序分析中很难被鉴定出来。泛基因组结合图基因组极大地提高了群体中遗传变异鉴定的精度和范围。

Liu等(2020)研究发现每个基因组PAVs长度的变异与基因组的长度极相关，相关系数高达0.98，这表明PAVs是造成群体内个体间基因组长度差异的重要因素。同时发现78.5%的PAVs来源于重复序列，表明重复序列是个体间基因组分化的重要驱动力。

基因组的结构变异是引起重要性状变异的主要原因，泛基因组为全面解析性状的形成和演化规律提供了全新的视角。对光周期的敏感性直接决定了大豆的地区适应性和产量。*E3*是影响大豆开花和生育期的关键基因，此位点在大豆种质中具有丰富的变异(Watanabe et al., 2009)。该研究还发现*E3*基因的多个等位基因类型，其中1个等位基因*E3-tr-1*为结构变异导致的新融合基因，它的产生是由于*E3*下游的1个13.3 kb的缺失，使*E3*和下游基因(*19G210600*)发生基因融合，形成新的转录本(Liu et al., 2020)。

野生大豆的种皮为黑色，在驯化过程中种皮变为无色，而呈现黄色的种子(Tuteja et al., 2004)。此性状是由位于8号染色体/位点的基因所控制，该区域包含1个编码查耳酮合酶(chalcone synthase, CHS)的基因簇。栽培种此区域的一个反向重复事件引发了对查耳酮合酶基因家族转录后的基因沉默，导致种皮中色素无法积累(Xie et al., 2019)。通过对*I*位点基因组区域序列(约505 kb)的比较分析和分子生物学验证确定了此位点的单体类型，并详细阐明了它们之间的演化关系。群体中的29份材料可以分为5种单体型(H1–H5)，黑色种皮的野生种质均为H1单体型。H2

产生于约4 500年前H1区段23.4 kb片段的1次复制, 并反向插入到CHS基因簇。H3产生于约4 200年前H1几次倒位及复制。H4和H5分别产生于约600年前, 由H3缺失及插入产生(Liu et al., 2020)。

结构变异影响基因的表达, 是产生形态学多样性的重要原因。研究鉴定了17 696个位于基因及附近的结构变异, 利用9个组织的转录组数据进行分析, 发现其中1 021个结构变异与基因的表达存在显著关联。铁离子是植物生长所必需的微量元素, 其吸收效率往往取决于土壤pH值。在高pH值的条件下, 铁元素主要以不溶于水的Fe_2O_3形式存在; 在低pH值的条件下, 铁离子免于被氧化, 更容易被植物根系吸收(Morrissey and Guerinot, 2009)。Liu等(2020)鉴定了1个位于Fe^{2+}/Zn^{2+}调控转运蛋白启动子区域, 长度为1.4 kb的*Mutator*转座子。此变异产生了Hap-1和Hap-2两种单体型。转录组分析发现Hap-2在种子中具有较高的转录水平。2种单体型存在显著的地域差别, Hap-1主要分布在低纬度地区, 而Hap-2主要分布在高纬度地区。我国低纬度地区以铁铝土为主, 具有低pH值, 铁离子更容易被植物根系吸收; 而高纬度地区恰恰相反, 土壤pH值较高(Dai et al., 2009)。具有高转录水平的Hap-2在高纬度区域具有更好的适应性, 2种单体型的地域分化恰好体现了大豆不同变异类型对土壤中离子吸收的适应性选择。

该研究是迄今为止作物学研究领域群体数目最多、基因组组装质量最好的泛基因组解析。全文从大数据的整合到重要生物学性状的解析, 材料选择的广度和数据挖掘的深度均达到了新高度, 将为其它作物的基因组学研究提供有益的借鉴。同时, 近3 000份材料的基因组数据将为大豆的进化解析、性状挖掘、种质创新和育种应用提供重要的数据支撑。该项研究提供的27个代表性参考基因组和泛基因组, 使大豆基因组研究真正进入多参考基因组和泛基因组时代, 使得高效和准确挖掘控制大豆重要性状的结构变异成为可能。该研究是继2010年发布大豆第1个参考基因组(Schmutz et al., 2010)之后, 大豆基因组研究领域的又一个里程碑。

致谢 在写作过程中, 中国农业科学院农业基因组研究所吴瑶瑶和赵建涛博士提供了宝贵的修改建议; 文中图1由中国科学院遗传与发育生物学研究所田志喜研究员提供, 在此一并表示诚挚的谢意。

参考文献

Carter TE Jr, Nelson R, Sneller CH, Cui Z (2004). Soybeans: Improvement, Production and Uses, 3rd edn. Madison: American Society of Agronomy. pp. 97.

Dai W, Huang Y, Wu L, Yu J (2009). Relationships between soil organic matter content (SOM) and pH in topsoil of zonal soils in China. *Acta Pedol Sin* **46**, 851–860.

Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu Y, van der Knaap E, Huang S, Klee HJ, Giovannoni JJ, Fei Z (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet* **51**, 1044–1051.

Golicz AA, Batley J, Edwards D (2016). Towards plant pangenomics. *Plant Biotechnol J* **14**, 1099–1105.

Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501.

Jiao WB, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, Willing EM, Piednoel M, Woetzel S, Madrid-Herrero E, Huettel B, Hümann U, Reinhard R, Koch MA, Swan D, Clavijo B, Coupland G, Schneeberger K (2017). Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* **27**, 778–786.

Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang SS, Zuo Q, Shi XH, Li YF, Zhang WK, Hu Y, Kong G, Hong HL, Tan B, Song J, Liu ZX, Wang Y, Ruan H, Yeung CKL, Liu J, Wang H, Zhang LJ, Guan RX, Wang KJ, Li WB, Chen SY, Chang RZ, Jiang Z, Jackson SA, Li R, Qiu LJ (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat Biotechnol* **32**, 1045–1052.

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, Huang X, Li Y, Zhang M, Wang Z, Zhu B, Han B, Liang C, Tian Z (2020). Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.

Lye ZN, Purugganan MD (2019). Copy number variation in domestication. *Trends Plant Sci* **24**, 352–365.

Morrissey J, Guerinot ML (2009). Iron uptake and transport in plants: the good, the bad, and the ionome. *Chem Rev* **109**, 4553–4567.

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang XC, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010). Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183.

Tuteja JH, Clough SJ, Chan WC, Vodkin LO (2004). Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in *Glycine max*. *Plant Cell* **16**, 819–835.

Watanabe S, Hideshima R, Xia Z, Tsubokura Y, Sato S, Nakamoto Y, Yamanaka N, Takahashi R, Ishimoto M, Anai T, Tabata S, Harada K (2009). Map-based cloning of the gene associated with the soybean maturity locus *E3*. *Genetics* **182**, 1251–1262.

Xie M, Chung CY, Li MW, Wong FL, Wang X, Liu A, Wang Z, Leung AK, Wong TH, Tong SW, Xiao Z, Fan K, Ng MS, Qi X, Yang L, Deng T, He L, Chen L, Fu A, Ding Q, He J, Chung G, Isobe S, Tanabata T, Valliyodan B, Nguyen HT, Cannon SB, Foyer CH, Chan TF, Lam HM (2019). A reference-grade wild soybean genome. *Nat Commun* **10**, 1216.

Zhao Q, Feng Q, Lu H, Li Y, Wang A, Tian Q, Zhan Q, Lu Y, Zhang L, Huang T, Wang Y, Fan D, Zhao Y, Wang Z, Zhou C, Chen J, Zhu C, Li W, Weng Q, Xu Q, Wang ZX, Wei X, Han B, Huang X (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet* **50**, 278–284.

A 360-degree Scanning of Population Genetic Variations—a Pan-genome Study of Soybean

Guangtao Zhu[1], Sanwen Huang[2*]

[1]The CAAS-YNNU Joint Academy of Potato Sciences, Yunnan Normal University, Kunming 650500, China; [2]Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agricultural Science and Technology, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

Abstract Soybean (*Glycine max*) is an important oil and protein crop. The abundancy of genetic diversity within the species provides an essential resource for traits exploration and breeding improvement. However, one reference genome is inadequate for discovering all genetic diversity of a species. Pan-genome provides a new solution to overcome this limitation. Recently, Prof. Zhixi Tian' Group and Prof. Chengzhi Liang' Group from the Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, selected 26 representative soybeans from 2 898 sequenced accessions. Together with three previously published genomes, they constructed a pan-genome and a graph-based genome of wild and cultivated soybean germplasm. The core, dispensable, and private genes as well as all the vast majority of genetic variations within this species were identified and characterized. These data comprehensively revealed allelic variations and gene fusion event of maturity gene *E3*, the haploid types of seed coat color gene *I* and their evolutionary relationship, and structural variations affecting gene expression and regional adaptation selection of ferric ion transporters. This study provide a new mode for crop genomics, and will facilitate genetic variations identification, traits exploration and germplasm innovation of soybean.

Key words soybean, pan-genome, graph-based genome, genetic variation, agronomic traits

Zhu GT, Huang SW (2020). A 360-degree scanning of population genetic variations—a pan-genome study of soybean. *Chin Bull Bot* **55**, 403–406.

* Author for correspondence. E-mail: huangsanwen@caas.net.cn

(责任编辑: 白羽红)