Article:
Bancroft, Ian (2024) A Brassica carinata pan-genome platform for Brassica crop improvement. Plant Communications. 100725. ISSN: 2590-3462

https://doi.org/10.1016/j.xplc.2023.100725

Plant Communications
Correspondence

CellPress
Partner Journal

A *Brassica carinata* pan-genome platform for *Brassica* crop improvement

Dear Editor,

Brassica carinata ($B^cB^cC^cC^cC^c$), an important agricultural crop cultivated for the edible oil, leafy vegetable, high-value protein stockfeed, and biofuel markets, contains favorable genetic variation that can be exploited to develop climate-resilient and water-use-efficient *Brassica* crops (Seepaul et al., 2021). Although several quantitative trait loci (QTLs) for flowering time, seed quality traits, and pod shatter resistance (PSR) have been identified based on high-density genetic linkage maps (Zou et al., 2014; Raman et al., 2017; Zhang et al., 2017), this information alone is insufficient for practical genetic improvement and breeding applications. Unlike efforts devoted to key staple food, fiber, and oilseed crops, few genomic resources have been developed specifically to support the genetic improvement of *B. carinata*. Recently, two genomes of *B. carinata* have been published (Song et al., 2021; Yim et al., 2022). However, a pan-genome is required to represent species-level genetic variation. Here, we assembled two representative high-quality genomes and constructed the first pan-genome of 86 *B. carinata* accessions.

We sequenced two accessions of *B. carinata,* 10167 and C4012, with distinct plant architecture and pod shatter attributes (Figure 1A and Supplemental Figure 1; Supplemental Table 1). The genome size was estimated to be ~1.13–1.31 Gb, with heterozygosity of less than 0.05% (Supplemental Figure 2). The assembled genomes for 10167 and C4012 were 1071 Mb and 1074 Mb, respectively, with a contig N50 of 26.7 and 24.3 Mb. Both genomes had high Benchmarking Universal Single-Copy Orthologs (BUSCO) values (~99.6%), mapping rates (~99.3%), and long terminal repeat (LTR) assembly indices (>15) (Ou et al., 2018) (Supplemental Figure 3; Supplemental Tables 2 and 3), which were equivalent or superior to those of most reported *Brassica* genomes (Yim et al., 2022). To ensure the accuracy of the chromosome-scale assemblies, we successfully anchored sequences by high-through chromosome conformation capture (Hi-C), linkage mapping, and genome-ordered graphical genotypes (He and Bancroft, 2018); 96.2% and 96.8% of the sequences were assembled onto 17 pseudochromosomes for 10167 and C4012, respectively (Supplemental Figures 4 and 5; Supplemental Table 4). Aligned with the internationally agreed nomenclature for *Brassica* chromosomes (Ostergaard and King, 2008), the final chromosome-scale assemblies had a scaffold N50 of 62.4 Mb for 10167 and 64.1 Mb for C4012. Three chromosomes contained only one gap, suggesting that our assemblies nearly achieved chromosome-arm-level contiguity (Supplemental Table 5).

Repetitive sequences accounted for 59% of the *B. carinata* genome, and LTRs were the most abundant, occupying ~26.0%–29.1% (Supplemental Table 6). We identified centromere- and telomere-specific repeats of *B. carinata* and found that the B^c genome had longer centromere regions than C^c (Figure 1B and Supplemental Figure 6; Supplemental Table 7). A total of 102 334 and 103 866 protein-coding genes were annotated for 10167 and C4012, respectively, 97.62% of which were functionally annotated (Supplemental Table 6). With regard to non-coding genes, we identified 12 336 and 13 178 small RNAs for 10167 and C4012, respectively (Supplemental Table 6).

We constructed an initial pan-genome gene set of *B. carinata* based on our two new genome assemblies together with previously published genomes. To select a genome assembly as the reference, we evaluated the four *B. carinata* assemblies using genome-ordered graphical genotypes and BLAST-based collinearity analysis (Supplemental Figures 7 and 8) and prioritized the most accurate genome of C4012 as the reference. Of the 117 225 gene models in the pan-genome, 14 824 (12.65%) were added to the reference genome (Supplemental Tables 8 and 9). In addition to four fully assembled *B. carinata* accessions, 82 representative accessions from different geographic regions were also sequenced and assembled into the final pan-genome (Supplemental Table 10) using the map-to-pan strategy (Glick and Mayrose, 2023). After redundancies were removed, a total length of 1.44 Gb comprising non-reference sequences remained, which contained an additional 8731 protein-coding genes. The final pan-genome size of *B. carinata* was ~2.52 Gb with 127 421 gene models. Modeling of pan-genome size suggested a closed pangenome with a finite number of core genes (88 307) and pan-genes (127 213) (Figure 1C). We categorized genes according to their frequency among the accessions: 88 307 (69.42%) core genes were shared by all accessions, and 21 262 softcore, 16 852 shell, and 792 cloud genes were defined as those present in more than 95%, 5%–95%, and less than 5% of the accessions, respectively (Figure 1D). The core and softcore groups contained highly conserved genes accounting for 86.13% of the genes in the pan-genome, whereas the rest were flexible genes. Enrichment analysis showed that core genes were involved in fatty acid elongation, glycolysis/gluconeogenesis, and fatty acid metabolism, whereas the variable genes were involved in carbon fixation in photosynthesis and linoleic acid metabolism (Supplemental Figure 9).

To demonstrate the usefulness of this pan-genome, we mapped 266 published QTLs (Raman et al., 2017; Zhang et al., 2017) associated with 15 agronomic and seed quality traits of *B. carinata* to the pan-genome (Supplemental Table 11) and prioritized 315 genes in the genomic regions (Supplemental

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

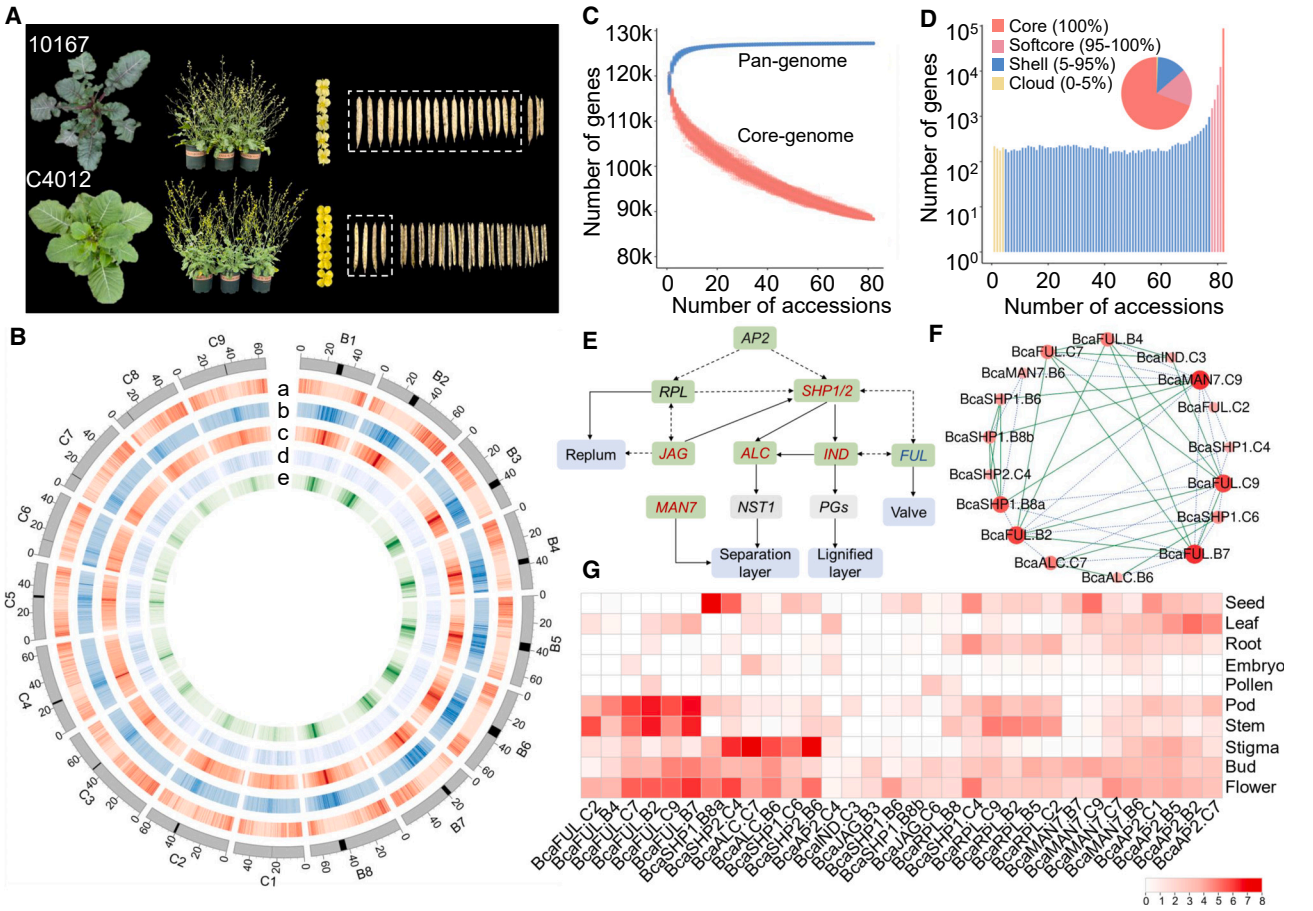

Figure 1. Overview of the pan-genome and genome-wide identification of PSR-related genes in *B. carinata*.

(A) Morphology of *B. carinata* accessions 10167 and C4012: seedling stage, flowering stage, and mature siliques. The mature siliques show the differences in performance of 10167 (pod shatter-resistant accession) and C4012 (shatter-susceptible accession) after five standardized shaking events. Inside and to the right of the white dotted frame are shatter-resistant (intact) and shattered (cracked/broken) siliques.

(B) Genomic features of *B. carinata* C4012. Chromosomes with centromere regions (black band) are shown. a–e: gene density, GC content, repeat density, DNA transposable element density, and LTR density (100-kb windows), respectively.

(C) Simulations of the increase in pan-genome size and the decrease in core genome size.

(D) Composition of the *B. carinata* pan-genome.

(E) Genetic model of PSR genes in *Arabidopsis*. The pod in *Arabidopsis* and *Brassica* comprises three parts (blue background): valve, replum, and dehiscence zone (separation layer and lignified layer). Pod shattering is controlled by several master patterning genes (green background) and their downstream cell wall-modifying genes (gray background). Among these genes, those in red and blue are reported to play negative and positive regulatory roles, respectively, in PSR.

(F) Co-expression of PSR genes from different tissues in *B. carinata*. The green line represents positive regulation, and the blue dashed line represents negative regulation.

(G) Tissue-specific expression patterns of PSR genes from different tissues of *B. carinata*.

Table 12). We also identified 1820 homologs related to glucosinolate metabolism, fatty acid biosynthesis, flowering time, and PSR (Supplemental Table 13). *B. carinata* has been reported to possess higher levels of PSR than other *Brassica* species (Raman et al., 2014, 2017). Using 12 genes of *Arabidopsis* involved in pod dehiscence as reference (Parker et al., 2021), 704 homologous genes were identified in 18 A-, B-, and C-genome *Brassica* species of the triangle of U (Figure 1E; Supplemental Table 14). *FUL* was most conserved with respect to copy number, and *B. carinata* had the most copies of *FUL*. Forty expressed genes involved in PSR were identified in *B. carinata*, and *FUL* had the highest expression in pods (Figure 1G). By contrast, *SHP* and *MAN7* had the highest expression in pods of other *Brassica* species (Supplemental Figures 10 and 11). Our data suggest that variations in the expression of *SHP* and *FUL* homologs may have contributed to PSR differences in *Brassica* (Supplemental Figure 12).

To identify critical genes for PSR, we carried out a bulked segregant analysis and identified a major QTL on B7, along with two minor QTLs on B2 and B8 (Supplemental Figure 13). We identified one *FUL* gene (*BcaFUL.B7*) within the overlapping major QTL region located by bulked segregant analysis and previous QTL mapping (Raman et al., 2017). *BcaFUL.B7* had the highest expression in pods, and co-expression network analysis indicated that it was the hub gene of PSR (Figure 1F). In addition, RNA sequencing showed that *BcaFUL.B7* was downregulated significantly in

B. carinata C4012, which would account for its shatter susceptibility (Supplemental Table 15).

Finally, we analyzed SNPs, insertions or deletions (indels), and structural variations (SVs) to assess unique polymorphisms across the *Brassica* B and C subgenomes (Supplemental Table 16; Supplemental Figure 14A). Among these variations, we identified 5342 SNPs and 9078 indels unique to B^c compared with the B^j and B^n genomes and 4209 SNPs and 3981 indels unique to C^c compared with C^o and C^n. More than half of the SVs (52.5%) were detected in genic regions or their regulatory regions (±1 kb), and 6.0% of the SVs were located in coding regions (Supplemental Figure 14B). Most SV breakpoints were in the regions of TEs and LTRs, and deletions tended to be distributed in LTR/Gypsy repeat-enriched regions (Supplemental Figure 14C). We identified 393 and 401 SV hotspots separately for the B and C subgenomes (Supplemental Table 16). We further identified 8 large inversions (>1 Mb) unique to *B. carinata*; for example, a 1.4-Mb inversion was detected on C3 (Supplemental Figure 14D). Notably, we found that *BcaFAE1.C3* of 10167 was disrupted by a 4.8-kb insertion of an LTR/Copia element (Supplemental Figure 14E), which would account for a significant downregulation of gene expression in pods and a 5% decrease in erucic acid content.

Overall, we report the first pan-genome of *B. carinata* comprising 127 421 gene models constructed from 86 accessions. Our pan-genome platform enabled the identification of genes for favorable traits and identified one hub gene, *BcaFUL.B7*, for PSR by quantitative trait loci sequencing (QTL-seq) and co-expression analysis. We also provided a genomic variation map for the B and C subgenomes of *Brassica* and identified species-specific genomic variations for *B. carinata*. This comprehensive study provides a valuable resource for understanding the genetic architecture of *B. carinata* traits and will usher in the genomics-assisted breeding of *Brassica* crops.

ACCESSION NUMBERS

The genome assembly and annotations have been deposited in the Genome Warehouse at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number PRJCA019810. The genome sequencing data and transcriptome sequencing data have been deposited in the National Genomics Data Center (GSA: CRA012653) and are publicly accessible at https://ngdc.cncb.ac.cn/gsa.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This work was supported by the National Natural Science Foundation of China (31970564, 32171982, and 32000397), the Grains Research and Development Corporation (DAN00208 to H.R.), the UK Biotechnology and Biological Sciences Research Council (BB/L002124/1 and BB/R019819/1 to I.B.), and the Fundamental Research Funds for the Central Universities (2662023PY004).

AUTHOR CONTRIBUTIONS

Y.N. performed the research, analyzed the data, and wrote the paper. Q.L. contributed to the genome assembly. Z.H. performed genome-ordered graphical genotyping, pan-genome analysis, and bulked segregant analysis. R.R. and H.R. developed populations, performed phenotyping and resequencing for the bulked segregant analysis experiment, and revised the manuscript. H.W., X.L., and H.Q. grew and collected the plant materials and performed pod-shatter phenotyping of the *B. carinata* accessions. I.A.P.P. contributed to the design of pan-genome construction and revised the manuscript. I.B. contributed to the design, research, and writing of the pan-genome construction, genome-ordered graphical genotyping, and bulked segregant analysis and revised the manuscript. J.Z. designed the project, analyzed the data, and wrote the manuscript. All authors revised, read, and approved the final manuscript.

ACKNOWLEDGMENTS

We acknowledge Weibo Xie and Daojun Yuan from the College of Plant Science and Technology of Huazhong Agricultural University for providing fresh leaf tissues of *Oryza sativa* (MH63) and *Gossypium barbadense* (Hai7124) for estimation of *B. carinata* genome size by flow cytometry. No conflict of interest is declared.

Received: June 7, 2023
Revised: August 21, 2023
Accepted: September 26, 2023
Published: October 5, 2023

Yan Niu[1,5], Qingqing Liu[1,5], Zhesi He[2], Rosy Raman[3], Hao Wang[1], Xinxin Long[1], Han Qin[1], Harsh Raman[3], Isobel A.P. Parkin[4], Ian Bancroft[2] and Jun Zou[1,]*

[1]National Key Laboratory of Crop Genetic Improvement, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China
[2]Department of Biology, University of York, Heslington, York YO10 5DD, UK
[3]NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Wagga Wagga, NSW 2650, Australia
[4]Agriculture and Agri-Food Canada, Saskatoon Research and Development Centre, Saskatoon, SK, Canada
[5]These authors contributed equally to this article.
***Correspondence: Jun Zou (zoujun@mail.hzau.edu.cn)**
https://doi.org/10.1016/j.xplc.2023.100725

REFERENCES

Glick, L., and Mayrose, I. (2023). The Effect of Methodological Considerations on the Construction of Gene-Based Plant Pan-genomes. Genome Biol Evol **15**:evad121. https://doi.org/10.1093/gbe/evad121.

He, Z., and Bancroft, I. (2018). Organization of the genome sequence of the polyploid crop species *Brassica juncea*. Nat. Genet. **50**:1496–1497. https://doi.org/10.1038/s41588-018-0239-0.

Ostergaard, L., and King, G.J. (2008). Standardized gene nomenclature for the *Brassica* genus. Plant Methods **4**:10. https://doi.org/10.1186/1746-4811-4-10.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. **46**:e126. https://doi.org/10.1093/nar/gky730.

Parker, T.A., Lo, S., and Gepts, P. (2021). Pod shattering in grain legumes: emerging genetic and environment-related patterns. Plant Cell **33**:179–199. https://doi.org/10.1093/plcell/koaa025.

Raman, H., Raman, R., Kilian, A., Detering, F., Carling, J., Coombes, N., Diffey, S., Kadkol, G., Edwards, D., McCully, M., et al. (2014). Genome-Wide Delineation of Natural Variation for Pod Shatter Resistance in *Brassica napus*. PLoS One **9**, e101673, ARTN e101673. https://doi.org/10.1371/journal.pone.0101673.

Raman, R., Qiu, Y., Coombes, N., Song, J., Kilian, A., and Raman, H. (2017). Molecular Diversity Analysis and Genetic Mapping of Pod

Shatter Resistance Loci in *Brassica carinata* L. Front. Plant Sci. **8**:1765. https://doi.org/10.3389/fpls.2017.01765.

Seepaul, R., Kumar, S., Iboyi, J.E., Bashyal, M., Stansly, T.L., Bennett, R., Boote, K.J., Mulvaney, M.J., Small, I.M., George, S., and Wright, D.L. (2021). *Brassica carinata*: Biology and agronomy as a biofuel crop. Gcb Bioenergy **13**:582–599. https://doi.org/10.1111/gcbb.12804.

Song, X., Wei, Y., Xiao, D., Gong, K., Sun, P., Ren, Y., Yuan, J., Wu, T., Yang, Q., Li, X., et al. (2021). *Brassica carinata* genome characterization clarifies U's triangle model of evolution and polyploidy in *Brassica*. Plant Physiol. **186**:388–406. https://doi.org/10.1093/plphys/kiab048.

Yim, W.C., Swain, M.L., Ma, D., An, H., Bird, K.A., Curdie, D.D., Wang, S., Ham, H.D., Luzuriaga-Neira, A., Kirkwood, J.S., et al. (2022). The final piece of the Triangle of U: Evolution of the tetraploid *Brassica carinata* genome. Plant Cell **34**:4143–4172. https://doi.org/10.1093/plcell/koac249.

Zhang, W., Hu, D., Raman, R., Guo, S., Wei, Z., Shen, X., Meng, J., Raman, H., and Zou, J. (2017). Investigation of the Genetic Diversity and Quantitative Trait Loci Accounting for Important Agronomic and Seed Quality Traits in *Brassica carinata*. Front. Plant Sci. **8**:615. https://doi.org/10.3389/fpls.2017.00615.

Zou, J., Raman, H., Guo, S., Hu, D., Wei, Z., Luo, Z., Long, Y., Shi, W., Fu, Z., Du, D., and Meng, J. (2014). Constructing a dense genetic linkage map and mapping QTL for the traits of flower development in *Brassica carinata*. Theor. Appl. Genet. **127**:1593–1605. https://doi.org/10.1007/s00122-014-2321-z.