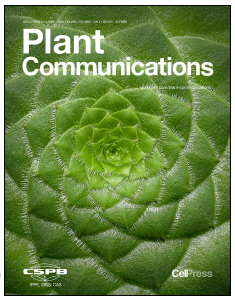


BjuIR: A Multi-omics Database with Various Tools for Accelerating Functional Genomics Research in *Brassica juncea*

Linna Zhang, Jinyuan Xiao, Congyuan Liang, Yifan Chen, Changchun Yu, Xinle Zhao, Jiawei Li, Mingli Yan, Qian Yang, Hao Chen, Zhongsong Liu, Zhengjie Wan, Zhiquan Yang, Qing-Yong Yang

PII: S2590-3462(24)00195-0

DOI: https://doi.org/10.1016/j.xplc.2024.100925

Reference: XPLC 100925

To appear in: *PLANT COMMUNICATIONS*

Please cite this article as: Zhang, L., Xiao, J., Liang, C., Chen, Y., Yu, C., Zhao, X., Li, J., Yan, M., Yang, Q., Chen, H., Liu, Z., Wan, Z., Yang, Z., Yang, Q.-Y., BjuIR: A Multi-omics Database with Various Tools for Accelerating Functional Genomics Research in *Brassica juncea*, *PLANT COMMUNICATIONS* (2024), doi: https://doi.org/10.1016/j.xplc.2024.100925.

BjuIR: A Multi-omics Database with Various Tools for Accelerating Functional Genomics Research in *Brassica juncea*

Linna Zhang[1,2,3,†], Jinyuan Xiao[1,2,†], Congyuan Liang[1,2,3,†], Yifan Chen[1,2,4], Changchun Yu[5], Xinle Zhao[1,2], Jiawei Li[1,2], Mingli Yan[6], Qian Yang[6], Hao Chen[7], Zhongsong Liu[7], Zhengjie Wan[5,*], Zhiquan Yang[1,2,4,*] and Qing-Yong Yang[1,2, 3, 4,*]

[1]National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, College of Informatics, Huazhong Agricultural University, Wuhan, 430070, China

[2]Hubei Key Laboratory of Agricultural Bioinformatics and Hubei Engineering Technology Research Center of Agricultural Big Data, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

[3]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China

[4]Yazhouwan National Laboratory, Sanya 572025, China

[5]National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, College of Horticulture & Forestry Sciences, Huazhong Agricultural University, Wuhan 430070, China

[6]Crop Research Institute, Hunan Academy of Agricultural Sciences, Changsha 410125, China

[7]College of Agronomy, Hunan Agricultural University, Changsha 410128, China

[†]These authors contributed equally to this article.

*Correspondence: Zhengjie Wan (wanzj@mail.hzau.edu.cn), Zhiquan Yang (yang_zq@foxmail.com), Qing-Yong Yang (yqy@mail.hzau.edu.cn)

Dear Editor,

Advancements in high-throughput omics technologies, along with methodologies for integrating multi-omics datasets, have substantially enhanced the efficiency of identifying candidate genes in breeding (Gupta et al., 2019; Gusev et al., 2018). However, this process is often complex and laborious. To address this challenge, databases that integrate extensive data and enable convenient and efficient function genomics studies are being developed (Ma et al., 2021; Yang et al., 2023). *Brassica juncea* (*B. juncea*), commonly known as mustard, is an economically significant agricultural species for its diverse uses, including vegetables, resilient oilseeds, and distinctively flavored condiments (Yang et al., 2018). This diversity of applications has spurred the accumulation of substantial multi-omics data in fundamental research of mustard, yet there lacks a specialized platform to harness these data fully for mustard's genetic improvement. Addressing this gap, we have developed BjuIR (*Brassica juncea* Information Resource, available at https://yanglab.hzau.edu.cn/BjuIR), integrating the most comprehensive mustard omics datasets to date from over 2,000 accessions, including data from genomics, variomics, transcriptomics, phenomics, and metabolomics. BjuIR provides sophisticated analyses for these multi-omics datasets of mustard with user-friendly interfaces, enabling rapid querying of "variant/gene expression-phenotype" associations for the quick identification of candidate genes and greatly benefiting functional genomics research.

DATA AND FUNCTIONAL MODULES IN BjuIR

BjuIR boasts a rich repository of large-scale multi-omics datasets (Figure 1A), encompassing 40 genome assemblies (Supplemental Table 1), 8,869,856 single nucleotide polymorphisms (SNPs) and short insertions/deletions (InDels) across 1,614 accessions (Supplemental Figure 1 and Supplemental Table 2), 941 RNA-seq libraries (Supplemental Table 3), 412 metabolites (Supplemental Table 4), phenotypic data of 628 accessions spanning 16 traits (Supplemental Table 5), and 1,841 mustard-centric literature entries. Various analysis methods were applied to fully explore the value of these datasets and the analysis results are organized and accessible in eight modules within BjuIR (Figure 1B).

The "Genomics" module provides queries for syntenic relationships between genomes and gene annotation; the "Population" module details accession information and provides selective signals of populations; the "Variations" module allows for the exploration of variants, and their associations with phenotypes and gene expression levels; The "Transcriptomics" module features gene expression profiles, co-expression networks and differential expression analysis; the "Phenomics" module presents phenotype data; the "Metabolomics" module provides metabolite information; The "Multi-omics" module facilitates quick queries for "variation-gene expression-phenotype" associations generated from genome-wide association study (GWAS), transcriptome-wide association analysis (TWAS), expression quantitative trait loci (eQTL) mapping analysis, colocalization analysis, and summary-based mendelian randomization (SMR); and the "Literature" module supports literature studies on mustard. Each module is equipped with user-friendly interfaces for result visualization, data downloads, and seamless navigation among modules and other databases.

APPLICATIONS AND ANALYSIS TOOLS IN BjuIR

Comprehensive datasets and thoughtfully designed modules in BjuIR are of great utility for those involved in functional genomics and candidate genes/variations identification efforts.

In the "Genomics" module, users can visualize global genome alignments in a dotplot (Figure 1C) and local genome alignments in Gbrowse in the "Genome synteny" interface (Supplemental Figure 2A) and query homologous gene clusters by entering gene ID or gene name in the "Gene cluster" interface (Supplemental Figure 2B). Additionally, they can explore annotations of gene families and biological pathways for mustard genes in the "Gene family" and "Pathway" interfaces (Supplemental Figure 2C and 2D).

The "Transcriptomics" module offers gene expression profiles across populations and tissues, co-expression analysis, and comparative transcription analysis, as demonstrated in Supplemental Figure 3. These capabilities are exemplified through the gene *BjuA09.WRI1*. By entering the gene BjuVA09G4986 (*WRI1*) in the "Tissue expression profile" and "eFP" interface, users can access the tissue expression profiles of *BjuA09.WRI1* and its homologous genes, presented in a heatmap (Supplemental Figure 3A) and eFP viewer (Figure 1D). Users can also explore gene-gene and lncRNA-mRNA co-expression networks related to *BjuA09.WRI1* in the "Co-expression network" interface (Supplemental Figure 3B and 3C). Moreover, they can access differentially expressed genes/lncRNAs in the "Differential expression" interface (Supplemental Figure 3D and 3E).

The "Population" module provides queries for detailed information of 1,614 accessions, including their subpopulations, usages, and origins (Supplemental Figure 4 and Supplemental Table 2). Furthermore, this module enables the querying of selective signals, such as π, Tajima's D, F_{ST}, and cross-population extended haplotype homozygosity (XP-EHH), which are calculated using variations to identify candidate regions and genes that may be under selection (Supplemental Figure 5).

The "Variations" module provides detailed information on genetic variants and assessment of how specific variants or haplotypes affect phenotypes and gene expression (Supplemental Figure 6). For instance, by inputting the gene ID "BjuVB08G59610" into the "Variations/Single-locus model" interface, the result page displays annotations for all SNPs and InDels within the gene region (Supplemental Figure 6A-6D). Choosing a particular variant, such as "BB_Chr08:62443776" (Supplemental Figure 6D), users can examine its allele frequency across diverse subpopulations and geographical locations (Supplemental Figure 6E and 6F), and explore its correlation with phenotypic traits, such as thousand seed weight, as well as with gene expression levels (Supplemental Figure 6G and 6H).

Integrated analyses of multi-omics data in the "Multi-omics" module vastly improve the efficiency of candidate gene discovery in *B. juncea*. Here, users can submit a gene name, ID, or trait name to unveil associations between variations, gene expression, and phenotypes. This module includes "variation-trait" associations identified by GWAS (Supplemental Figure 7A and Supplemental Table 6), "variation-gene expression" associations from eQTL (Supplemental Figure 7B and Supplemental Table 7), "gene expression-trait" associations identified by TWAS (Supplemental Figure 7C Supplemental Table 8), as well as colocalization analyses (Supplemental Figure 7D-7F and Supplemental Table 9). The reliability of these integrated results is demonstrated by a reproducibility rate of 60.42% in comparison to prior findings (Harper et al., 2020), as listed in Supplemental Table 10. Additionally, the "variation-gene expression-trait" associations are also viewable in a network format, exemplified by entering the "*BIM1*" gene within the "Multi-omics/Association networks" interface, which showcases all related associations in a visual network (Figure 1E).

The "Literature" module offers advanced search capabilities based on keywords, journal names, and publication years, allowing users to efficiently access research advancements related to mustard in a specific field. For instance, by entering the keyword "flowering" in this module, users can retrieve relevant literature on "flowering", with detailed information displayed in a table. In addition, the statistics of

the literature, organized by publication year or journal, are visually presented in line graphs and bar charts. An additional feature is the provision of a keyword co-occurrence network for visualizing trends in studies related to the flowering of mustard (Supplemental Figure 8).

The "Tools" module incorporates 15 essential bioinformatics analysis tools applications, supporting user-initiated analyses, including Gene Ontology (GO) enrichment, linkage disequilibrium (LD) calculations, SNP matching for germplasm identification, sequence extraction, primer design, among other functions (Supplemental Figure 9).

CASE STUDY: MINING NOVEL CANDIDATE VARIANTS AND GENES ASSOCIATED WITH TOCOPHEROL CONTENT USING BjuIR

We illustrate the utility of BjuIR in mining candidate genes and variations using the example of tocopherol, a crucial vitamin E component vital for seed quality and human nutrition. Tocopherol exists in various forms, such as α-tocopherol, which is recognized for having the highest vitamin E activity in mammals and can be derived from γ-tocopherol (Tucker and Townsend, 2005). Initiating with the query "γ-/α-tocopherol content in seed" in the "Multi-omics/GWAS" interface, the analysis rendered a Manhattan plot revealing two genomic loci associated with γ-/α-tocopherol content located on chromosomes AA_Chr02 and AA_Chr06. While the AA_Chr02 locus had been previously reported (Harper et al., 2020), the locus on AA_Chr06 emerged as a novel finding, comprising an LD block between 5.88 to 5.95 Mb (Figure 1F and 1G), within which seven genes reside (Figure 1H and Supplemental Table 11). Specifically, BjuVA06G10820, notable for containing the largest number of GWAS-SNPs in its coding and 3 kb upstream regions, was identified. Its homolog in *Arabidopsis thaliana* (AT1G15125) is known to code for S-adenosyl-L-methionine-dependent methyltransferases that participate in converting γ-tocopherol to α-tocopherol (Tavva et al., 2007), suggesting that BjuVA06G10820 could be a prime candidate gene within the AA_Chr06 locus. Further, haplotype analysis of BjuVA06G10820 through the "Variations" module's "Single-locus model" interface uncovered two prevalent

haplotypes (Figure 1I). Notably, accessions carrying Haplotype1 were significantly associated with reduced γ-/α-tocopherol content compared to those with Haplotype2, delineated in Figure 1J. Such findings will provide a valuable reference for future breeding strategies aimed at boosting α-tocopherol levels in mustard seeds and underscore BjuIR's aptitude for identifying candidate genes and variants associated with specific traits.

In conclusion, BjuIR stands as the most extensive and comprehensive multi-omics database to date for functional genomics research in mustard. Its key features include (1) expedited access to each omics dataset and complete analysis results; (2) quick mining of candidate genes and variants via robust "variant-gene expression-phenotype" associations; (3) multiple user-friendly, online bioinformatic tools; and (4) navigation-friendly interfaces for efficient data mining. With its rich database and thoughtful design, BjuIR proves to be a highly efficient and convenient platform for functional genomics research and candidate gene identification. Looking ahead, BjuIR will persist in incorporating novel omics data, reinforcing its status as an indispensable platform for furthering functional genomics and genetic improvement in mustard.

DATA AVAILABILITY

Sources of all datasets are described in supplemental information. All datasets are available at https://yanglab.hzau.edu.cn/BjuIR/download.

FUNDING

This research was supported by the National Natural Science Foundation of China (32322061 and 32070559); the National Key Research and Development Plan of China (2021YFF1000100); the Fundamental Research Funds for the Central University HZAU (2662023XXPY001); the Hubei Hongshan Laboratory (2021HSZD004); and the Developing Bioinformatics Platform in Hainan Yazhou Bay Seed Lab (no. JBGS-B21HJ0001).

AUTHOR CONTRIBUTIONS

Q.-Y.Y., Z.Y., and Z.W. designed the project. L.Z. and Y.C. collected the datasets. L.Z., C.L., Y.C., J.X., and J.L. performed the bioinformatics analysis. J.X., Y.C., and X.Z. developed the BjuIR database. C.Y. provided the pictures of *B. juncea* germplasms. L.Z., C.L., Z.Y., and Q.-Y.Y. wrote the manuscript. Q.-Y.Y., Z.Y., Z.W., Z.L., H.C., Q.Y., and M.Y. directed the project. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

We thank the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, managed by Hao Liu. No conflict of interest is declared.

REFERENCE

Gupta, P.K., Kulwal, P.L., and Jaiswal, V. (2019). Association mapping in plants in the post-GWAS genomics era. Adv. Genet. **104**: 75-154.

Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Schizophrenia Working Group of the Psychiatric Genomics Consortium, and McCarroll, S., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. **50**: 538-548.

Harper, A.L., He, Z., Langer, S., Havlickova, L., Wang, L., Fellgett, A., Gupta, V., Kumar Pradhan, A., and Bancroft, I. (2020). Validation of an associative transcriptomics platform in the polyploid crop species *Brassica juncea* by dissection of the genetic architecture of agronomic and quality traits. Plant J. **103**: 1885-1893.

Ma, S., Wang, M., Wu, J., Guo, W., Chen, Y., Li, G., Wang, Y., Shi, W., Xia, G., Fu, D., et al. (2021). WheatOmics: A platform combining multiple omics data to accelerate functional genomics studies in wheat. Mol. Plant **14**: 1965-1968.

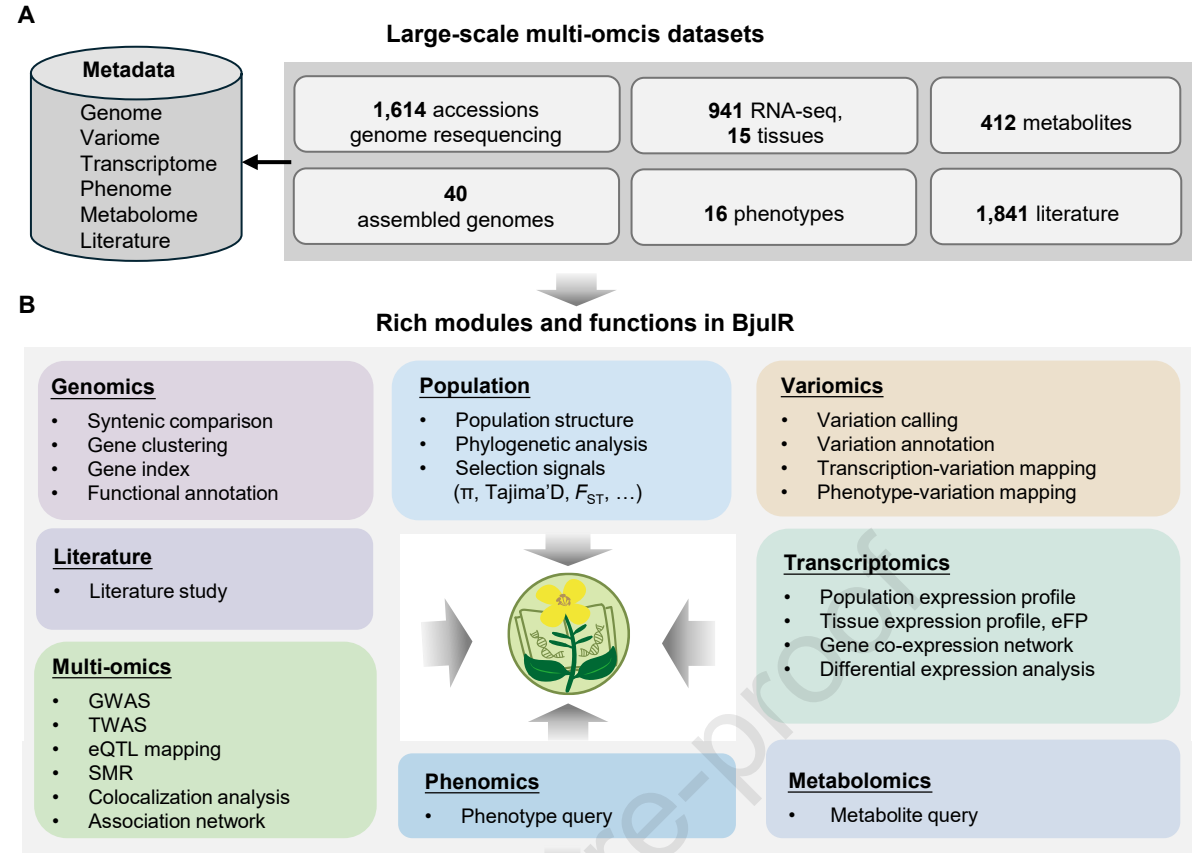
Tavva, V.S., Kim, Y.-H., Kagan, I.A., Dinkins, R.D., Kim, K.-H., and Collins, G.B. (2007). Increased α-tocopherol content in soybean seed overexpressing the Perilla frutescens γ-tocopherol methyltransferase gene. Plant Cell Rep. **26**: 61-70.

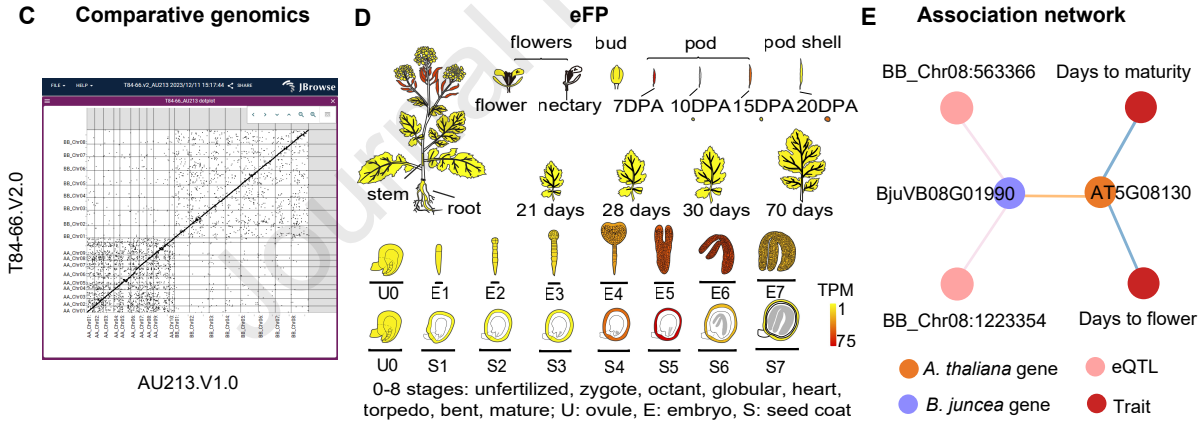
Tucker, J., and Townsend, D. (2005). Alpha-tocopherol: roles in prevention and therapy of human disease. Biomed. Pharmacother. **59**: 380-387.

Yang, J., Zhang, C., Zhao, N., Zhang, L., Hu, Z., Chen, S., and Zhang, M. (2018). Chinese root-type mustard provides phylogenomic insights into the evolution of the multi-use diversified allopolyploid *Brassica juncea*. Mol. Plant **11**: 512-514.

Yang, Z.Q., Wang, S.B., Wei, L.L., Huang, Y.M., Liu, D.X., Jia, Y.P., Luo, C.F., Lin, Y.C., Liang, C.Y., Hu, Y., et al. (2023). BnIR: A multi-omics database with various tools for research and breeding. Mol. Plant **16**: 775-789.

Figure 1. Overview of BjuIR. (A) Large-scale datasets collected in BjuIR. **(B)** Eight modules and their functions in BjuIR. **(C)** Comparative genomics analysis between T84-66.V2.0 and AU213.V1.0 genome in the "Genomics" module. **(D)** eFP viewer displaying tissue-specific expression profiles of the gene BjuVA09G4986 in the "Transcriptomics" module. **(E)** "Variation-gene expression-phenotype" associations related to gene *BIM1* in the "Multi-omics" module. **(F-J)** Identification of novel candidate genes/variants associated with γ-/α-tocopherol content in BjuIR. **(F)** GWAS for γ-/α-tocopherol content in seed. The *P*-value threshold was set at 6.54e-6 based on 1/n, where n represents the number of independent SNPs (n =152,884). **(G)** Local Manhattan plot of GWAS for γ-/α-tocopherol content and heatmap of linkage disequilibrium (LD) blocks. The color of the dots represents the degree of LD with the lead SNP. **(H)** Genes in the LD block are significantly associated with γ-/α-tocopherol content. **(I)** Haplotypes formed by combinations of GWAS-SNPs on the coding region of BjuVA06G10820 and its 3 kb upstream flanking region. **(J)** Comparison of γ-/α-tocopherol content between accessions with different haplotypes. * indicates $P < 0.05$ (Wilcoxon rank sum test).





Case study: mining new candidate variants and genes associated with tocopherol content using BjuIR database