Journal of Genetics and Genomics 50 (2023) 921–923

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Research communications

Triticeae-BGC: a web-based platform for detecting, annotating and evolutionary analysis of biosynthetic gene clusters in Triticeae

Triticeae species encompass many important crops including wheat, barley, and rye, which are essential for ensuring human survival and world food security. Expansions of genes involved in stress responses are common in Triticeae genomes, which contributed to the high adaptability of Triticeae species. Secondary metabolites are key weapons for plants to deal with changing environments. However, metabolic enzymes generally underwent diversifying selection depending on the environments, a large fraction of which are species-specific whose functions are difficult to deduce merely based on sequence homology. During the evolution process, the catalytic enzymes of some functionally important metabolites are clustered on chromosomes, namely, biosynthetic gene clusters (BGCs), which tend to have coordinated expression to ensure catalytic efficiency. Elucidating BGCs can promote the synthesis of disease-resistant chemical weapons (Field and Osbourn, 2008; Boycheva et al., 2014; Cimermancic et al., 2014; Doroghazi et al., 2014; Medema et al., 2014; Ziemert et al., 2014; van der Lee and Medema, 2016). Detecting and analyzing the similarities and differences of these BGCs across wheat and related species provide important information for pinpointing candidate defense genes and metabolites, and promoting the understanding of versatile defense mechanisms. However, detection, annotation, and evolutionary analysis of BGCs are challenged by the large and complex genomes of Triticeae species.

Here, via integrating genomic and transcriptomic data, along with enzymatic and metabolite information, we established a process for identification, functional annotation, and evolutionary analysis of whole genome BGCs (Fig. 1A). Firstly, PlantiSMASH algorithm (Kautsar et al., 2017) is applied to predict gene clusters in nine Triticeae and related species including *Triticum aestivum* (common wheat), *Triticum durum*, *Triticum dicoccoides*, *Triticum urartu*, *Aegilops tauschii*, *Secale cereale*, *Brachypodium distachyon*, *Zea mays*, and *Hordeum vulgare*. The types of product were predicted based on the types of enzymes in BGCs, including alkaloid, polyketone, terpenoid, glycoside, and lignin. Given that there is a high probability that some enzymes are located in nearby loci without functional relevance, which may result in high false positives for the prediction, we retrieved functional information including co-expression and co-pathway to provide additional evidence for identifying functional BGCs. Co-expression screens for clusters with two or more genes located in the same co-expression module, and co-pathway screens for two or more genes predicted to be involved in the same metabolic pathway. These two screens help guarantee identification of functional BGCs. All six gene clusters that have been experimentally validated in wheat (Polturak et al., 2022) are covered in the wheat gene clusters following the co-expression screen.

Due to the wide variety of habitats of different Triticeae species, for example, between common wheat and different diploid progenitors, the encoded defense metabolite BGCs underwent positive selection and may vary greatly among species. Comparison of the types of BGCs enzymes within and between Triticeae species can shed light on the evolutionary dynamics of plant responses to environments. Here, we compared BGCs from two perspectives, namely, the functional protein domain composition and the sequence collinearity across BGCs.

i) Comparison of protein domain composition and sequences across BGCs. Similar functional domain composition reflects similar types of enzymes within the BGCs under comparison. Briefly, protein structural domains were identified for each gene cluster. Jaccard index was calculated to measure the structural domain similarity (enzyme family composition) between gene clusters; domain sequence similarity (DSS) index was applied to quantify the sequence similarity of protein domains between gene clusters. This type of comparison provides useful clues for predicting the functional and evolutionary relationships of gene clusters from a phylogenetic perspective (Fig. S1).
ii) Collinearity analysis across gene clusters. Sequence collinearity reflects the genome structure and evolutionary history of clusters across different species. We calculated collinearity intervals of predicted gene clusters across nine species (Fig. S1). The MCscan function of the python-based jcvi module is applied for visualization of the collinearity across species (Wang et al., 2012). These evolutionary analyses of gene clusters provide important information on the function, origin, and evolutionary history of natural product biosynthesis pathways.

All the information and tools are implemented in a web-based platform with customized filtering and interactive visualization options (Fig. 1B), which is freely available at http://119.78.67.240:3838/Triticeae-BGC/. Relevant information is readily accessible by entering gene or chromosome locations. The current version of the Triticeae-BGC provides four main functions (Fig. 1C–1E): (i) predicting whether the input gene from Triticeae species localized within a BGC, other members in the cluster, the function, pathway, and expression responses to various pathogens, and the type of metabolite product; (ii) customized filtering of common wheat BGCs via co-expression or/and co-pathway information; (iii) identifying collinearity regions of input gene or gene clusters among specified species; (iv) comparing protein domain composition of input gene cluster across species. A detailed manual illustrating the input and output is available on the platform website.

https://doi.org/10.1016/j.jgg.2023.09.014

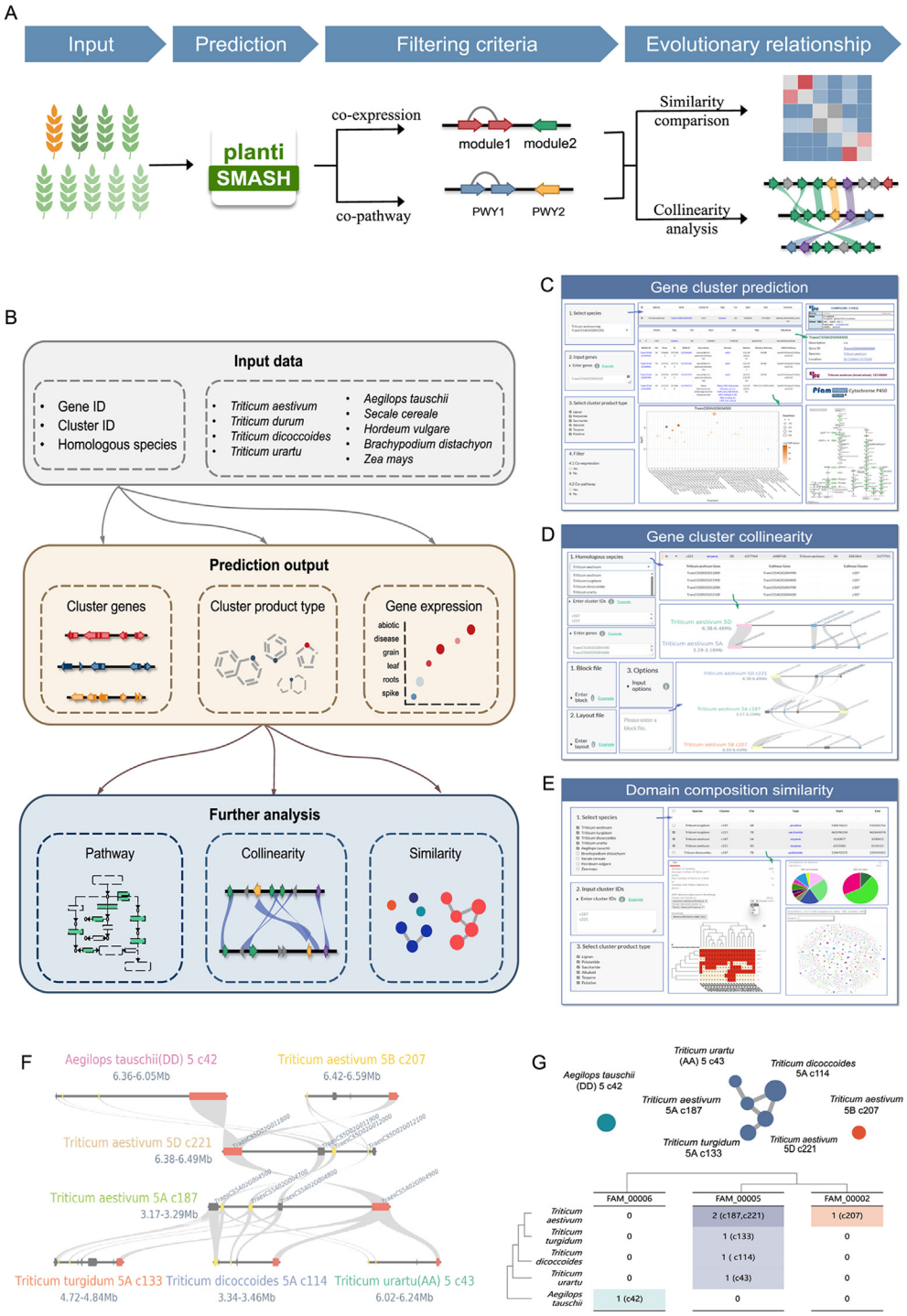

Fig. 1. Web-based interface for prediction, functional annotation, and evolutionary analysis of Triticeae BGCs. **A**: The workflow of BGC prediction and evolutionary analysis. Gene cluster prediction was performed genome-wide via plantiSMASH algorithm (Kautsar et al., 2017) based on the type of enzymes and the chromosome linkage. Information including co-expression (gene clusters with at least two metabolic genes in the same co-expression module) and co-pathway (gene clusters containing at least two metabolic genes in the same metabolic pathway) in common wheat were included for increasing accuracy of prediction. All results were implemented in a web-based platform with customized filtering options. The type of metabolite, gene function, and domain information, as well as expression responses to various pathogens, were provided, facilitating prediction of BGCs functions. Furthermore, domain similarity and sequence collinearity of BGCs across species could be compared via the platform, which helps elucidate the evolutionary dynamics of BGCs. **B**: Input and output. Gene clusters and types of product could be retrieved by querying species and gene IDs. Members in the BGCs, pathways involved in, and expression responses to different pathogens are provided. **C**: BGC co-expression and co-pathway information for common wheat BGCs. **D**: The collinear regions of input gene or gene cluster on related species could be retrieved for visualization. **E**: Domain composition similarity between colinear BGCs. **F**: Schematic diagram of the syntenic regions across wheat species for homologous gene clusters AABBDD_5A_cluster187 and AABBDD_5D_cluster221. **G**: Top, network depicting the protein composition similarity of the BGC shown in panel A across wheat species. Each circle represents the BGC in one species. Circles connected by gray edges represent significant homology cutoff > 0.7. Bottom, evolutionary tree of wheat species constructed based on the homology of BGC shown in panel (**A**). BGC, biosynthetic gene cluster.

We illustrate the usage of Triticeae-BGC for evolutionary comparison of BCGs across Triticeae species. A total of 332 gene clusters in common wheat were predicted based on plantiSMASH (all BGCs are available for download from the website), among which 247 passed the co-expression filtering step on the website, including all six functional BGCs producing defense-related metabolites as recently reported (Polturak et al., 2022). Two of these BGCs are homologous between subgenomes A and D (Fig. 1F). The enzyme composition is largely similar among subgenomes. The two gene clusters were found to be highly induced by biotic stress (Fig.S2), suggesting that this pair of homologous gene clusters potentially involved in defenses.

Regions in other wheat species syntenic to these two BGCs were retrieved via the collinearity page on the website, including putative diploid progenitors of A and D subgenomes (*Triticum Urartu* and *Aegilops tauschii*), and tetraploid wheat species (*Triticum dicoccoides* and *Triticum turgidum*). Comparison of the protein domain composition across these syntenic regions suggested that the protein composition in diploid D and subgenome B in common wheat are distantly related to other genomes (Fig. 1G). Thus, it is likely that this cluster originated in diploid progenitor of subgenome A.

Detection of agronomically important genes from mutants or natural variants is an important task in molecular research on crops. However, because of the high linkage equilibrium of wheat due to self-pollination and low population recombination frequency (Zhang et al., 2010), the candidate interval for gene mapping is generally large. Integrating functional annotation is needed to pinpoint candidate genes. However, inferring functional information depends heavily on sequence homology with reported genes, whereas defense-related genes, especially metabolic genes, are highly diverse. BGCs represent an important group of defense genes. De novo detection of BGCs provides an important resource for functional annotation and narrowing down key candidates. Additionally, for BGCs of interest, the platform provides transcriptional and evolutionary analyses that help infer the origin and function of BGC members.

BGCs are generally subject to diversifying selection based on their environments, enabling the production of metabolites responsive to specific pathogens. Elucidating their synthetic pathways could enrich chemical tools for dealing with specific environments. The BGCs provided in the platform represent a subset of Triticeae BGCs, and more species- and population-specific BGCs are to be detected when more genomes and expression data become available. In addition, given that functional BGCs may not necessarily be co-expressed or exist in the same pre-defined pathway, it is recommended that the users adjust filtering options and combine transcriptional and evolutionary analysis provided by the platform to obtain more candidates for follow-up analysis.

Conflict of interest

The authors declare no conflict of interests.

Acknowledgments

We thank Prof. Zhenhua Liu from Shanghai Jiao Tong University for helpful comments. This study was supported by the National Natural Science Foundation of China (32270628, 32270629), State Key Laboratory of Crop Gene Exploration and Utilization in Southwest (SKL-KF202305), and State Key Laboratory of Genetic Engineering (SKLGE-2312).

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jgg.2023.09.014.

References

Boycheva, S., Daviet, L., Wolfender, J.-L., Fitzpatrick, T.B., 2014. The rise of operon-like gene clusters in plants. Trends Plant Sci. 19, 447–459.
Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., et al., 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158, 412–421.
Doroghazi, J.R., Albright, J.C., Goering, A.W., Ju, K.-S., Haines, R.R., Tchalukov, K.A., Labeda, D.P., Kelleher, N.L., Metcalf, W.W., 2014. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat. Chem. Biol. 10, 963–968.
Field, B., Osbourn, A.E., 2008. Metabolic diversification–independent assembly of operon-like gene clusters in different plants. Science 320, 543–547.
Kautsar, S.A., Suarez Duran, H.G., Blin, K., Osbourn, A., Medema, M.H., 2017. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 45, W55–W63.
Medema, M.H., Cimermancic, P., Sali, A., Takano, E., Fischbach, M.A., 2014. A systematic computational analysis of biosynthetic gene cluster evolution: lessons for engineering biosynthesis. PLoS Comput. Biol. 10, e1004016.
Polturak, G., Dippe, M., Stephenson, M.J., Chandra Misra, R., Owen, C., Ramirez-Gonzalez, R.H., Haidoulis, J.F., Schoonbeek, H.-J., Chartrain, L., Borrill, P., et al., 2022. Pathogen-induced biosynthetic pathways encode defense-related molecules in bread wheat. Proc. Natl. Acad. Sci. U. S. A. 119, e2123299119.
van der Lee, T.A.J., Medema, M.H., 2016. Computational strategies for genome-based natural product discovery and engineering in fungi. Fungal Genet. Biol. 89, 29–36.
Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., et al., 2012. *MCScanX*: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40, e49.
Zhang, D., Bai, G., Zhu, C., Yu, J., Carver, B.F., 2010. Genetic diversity, population structure, and linkage disequilibrium in U.S. Elite winter wheat. Plant Genome 3, 117–127.
Ziemert, N., Lechner, A., Wietz, M., Millán-Aguiñaga, N., Chavarria, K.L., Jensen, P.R., 2014. Diversity and evolution of secondary metabolism in the marine actinomycete genus Salinispora. Proc. Natl. Acad. Sci. U. S. A. 111, E1130–E1139.

Mingxu Li[1]
State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Department of Biochemistry, Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai 200438, China

Haoyu Wang[1]
National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China

State Key Laboratory of Crop Stress Adaptation and Improvement, School of Life Sciences, College of Agriculture, Henan University, Kaifeng, Henan 457004, China

Shilong Tian
National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China

Yan Zhu*, Yijing Zhang*
State Key Laboratory of Genetic Engineering, Collaborative Innovation Center of Genetics and Development, Department of Biochemistry, Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai 200438, China

* Corresponding authors.
E-mail addresses: zhu_yan@fudan.edu.cn (Y. Zhu), zhangyijing@fudan.edu.cn (Y. Zhang).

23 August 2023
Available online 6 October 2023

[1] These authors contributed equally to this work.