CellPress
Partner Journal

Molecular Plant
Correspondence

Rice Gene Index: A comprehensive pan-genome database for comparative and functional genomics of Asian rice

Dear Editor,

Asian rice (*Oryza sativa*) is the staple food for half the world and is a model crop that has been extensively studied. It contributes ~20% of calories to the human diet (Stein et al., 2018). With the increase in global population and rapid changes in climate, rice breeders need to develop new and sustainable cultivars with higher yields, healthier grains, and reduced environmental footprints (Wing et al., 2018). Since the first gold-standard reference genome of rice variety Nipponbare was published (International Rice Genome Sequencing Project, 2005), an increasing number of rice accessions have been sequenced, assembled, and annotated with global efforts. Nowadays, a single reference genome is obviously insufficient to perform the genetic difference analysis for rice accessions. Therefore, the pan-genome has been proposed as a solution, which allows the discovery of more presence-absence variants compared with single-reference genome-based studies (Zhao et al., 2018). Over the past years, several databases, such as RAP-db (https://rapdb.dna.affrc.go.jp), RGAP (http://rice.uga.edu), and Gramene (https://www.gramene.org), have long-term served rice genomic research by providing information based on one or a series of individual reference genomes. To integrate and utilize the genomic information of multiple accessions, we performed comparative analyses and established the user-friendly Rice Gene Index (RGI; https://riceome.hzau.edu.cn) platform. RGI is the first gene-based pan-genome database for rice.

To set up a solid foundation for this database, we selected 16 platinum standard reference genomes of rice accessions that represent the major Asian rice subpopulations when K = 15 (Zhou et al., 2020; Song et al., 2021; Stein et al., 2018), (Figure 1A). Starting with a set of unified *de novo* annotations performed by Gramene (Zhou et al., 2023) of 14 genomes and 4 published annotations including Minghui 63 (MH63), Zhenshan 97, and Nipponbare (RGAP and RAP-db) (Kawahara et al., 2013; Sakai et al., 2013), we incrementally integrated the genes and transcripts identified by newly sequenced isoform sequencing (Iso-Seq) data into the Gramene annotation results as the basics to build homology relationships between 18 annotations (Supplemental Table 1). In addition, a series of Iso-Seq and RNA-Seq data of multiple tissues from selected accessions (Supplemental Tables 2 and 3) were collected and fully presented as baseline information in RGI, which included gene expression, full-length transcripts, and alternative splicing (AS) events. Details on data processing are described in the supplemental methods.

As the primary datasets in RGI, the genome annotations of 16 rice accessions contained an average of 41 346 genes, of which an average of 1178 genes are supplemented by Iso-Seq data (Supplemental Table 4). The GeneTribe pipeline (Chen et al., 2020) identified an average of 33 350 gene pairs between annotations (Supplemental Figure 2), which classified "reciprocal best hits," "single-side best hits," "one-to-many hits," or "singleton hits." By counting unique homolog gene groups, a total of 119 783 non-redundant gene groups were determined to represent the whole Asian rice gene set. To further unify the gene groups in *Oryza sativa*, we defined a unified and sustainable number—Ortholog Gene Index (OGI), which is a homolog group clustered by connected graph methods based on reciprocal best hit relationships, with an updatable score that indicates its representativeness in all accessions. Of the 112 658 OGIs, we classified them into 21 418 OGI core genes (19.01% of OGI) appearing in all rice accessions, 40 141 OGI dispensable genes, and 51 099 OGI accession-specific genes (Supplemental Figure 1A). And we found that the specific genes are younger and shorter (*t*-test, p = 2e−16) than core genes (supplemental information 1).

The first objective of RGI is to logically organize and scientifically index all genes among rice accessions. RGI provides "GeneCard" pages to show comprehensive information for individual genes with convenient links to other modules and outside databases on one page (Figure 1C). By entering a gene ID of rice, through the search box on the homepage, users may browse the "GeneCard" page on three sections: 1) basic information includes sequence, gene function, gene expression, links for accessing various modules and other databases, etc. (Supplemental Figure 4A). 2) "Transcripts" exhibits graph and table of transcript structures. In addition to the baseline expression analysis of all genes, 116 640 AS events at the transcriptome level were extensively revealed by the analysis of different groups (Supplemental Figure 4B; Supplemental Table 5). For example, two AS events were detected for *OsNiR* (OsNip_01g0357100), a critical gene that encodes nitrite reductase in nitrogen assimilation (Yu et al., 2021) (Figure 1D). Additionally, "Homologues" lists all associated homologs of a gene across annotations through a link graph and a table. This section also shows the phylogenetic tree. Furthermore, RGI provides informative pages to show the association graph of genes in each OGI (Supplemental Figure 4C).

Second, RGI provides three ways to search for relationships and comprehensive information for genes.

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

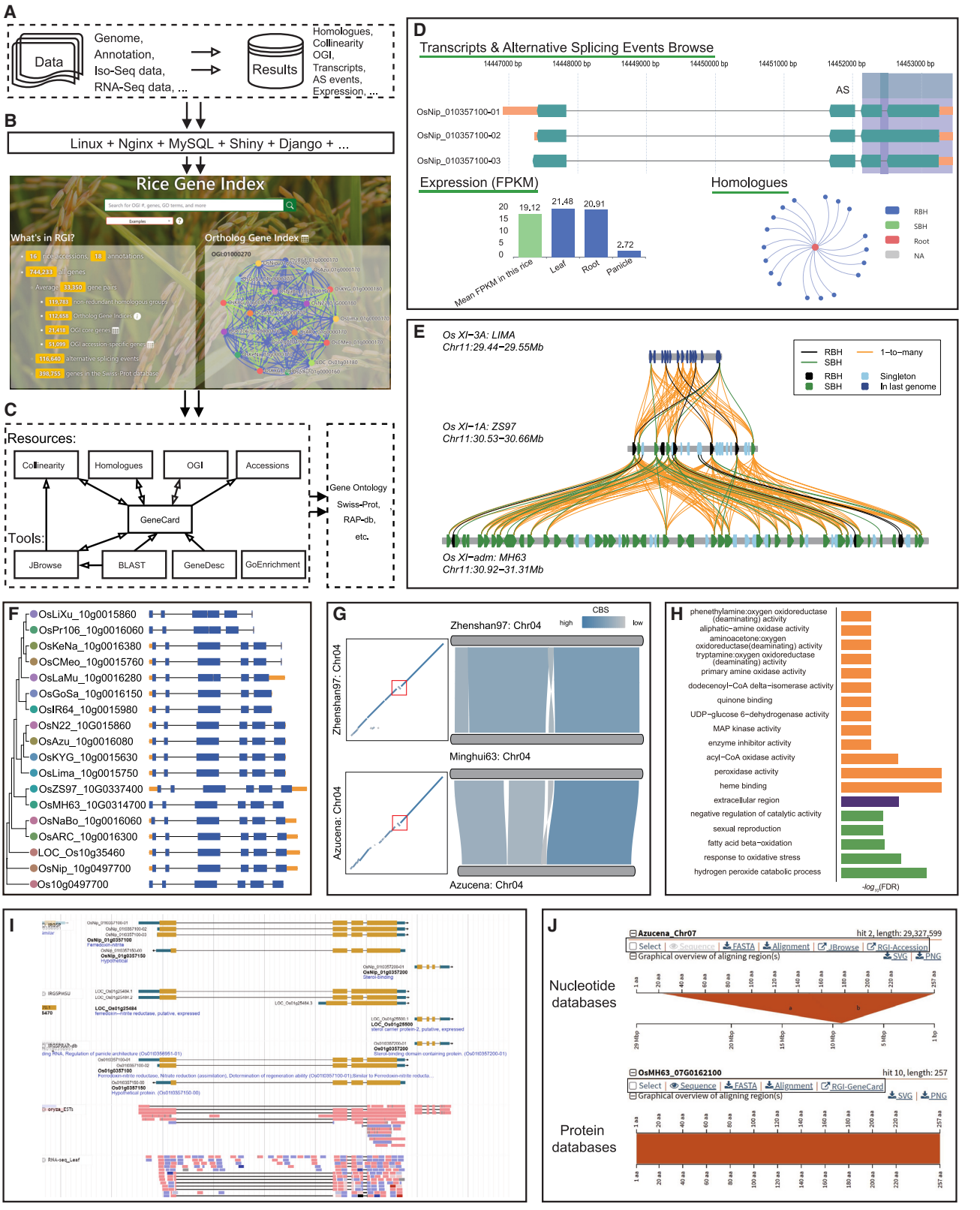

Figure 1. Design and online output examples of RGI.

(A–C) Overview of the construction process of RGI.

(A) The data layer demonstrates the source data and results provided in RGI.

(B) The middleware layer shows the tools used to build the RGI website.

(legend continued on next page)

1) Through keyword-based searches, users can easily search OGI#, gene ID, gene symbol, Gene Ontology, or functional terms in the query box. If users search the famous gene *SD1* in RGI, 306 items will be returned with basic information, which could link to other modules or databases.
2) In the way of sequence-based searches, the classical "BLAST" tool allows users to query amino acid or nucleotide sequences in sequence databases of the whole genome and protein. To easily access other modules, the tool returns gene ID linking to "GeneCard" or chromosome location linking to "JBrowse" when using the protein or nucleotide database, respectively.
3) For association-based searches, the "Homologues" module allows users to query and connect the homologous genes through a given gene ID, which may obtain the homology relationship among annotations. By using TreePlot, users could build the phylogenetic tree with gene structures (Figure 1F) and view multiple sequence alignments of interested genes, as well as the detailed information of each gene. For example, *OsTPP7* (LOC_Os09g20390), an anaerobic germination tolerance gene, was found to be absent in IR64 but present in other accessions by "Homologues" (Supplemental Table 6), and the results were manually verified. This indicates that IR64 has less tolerance to anaerobic germination (Yang et al., 2019).

Third, RGI can visualize the relationship of these annotated genes across accessions at local and global scales corresponding to two modules as follows.

1) At the local scale, the "MicroCollinearity" module enables users to demonstrate genomic collinearities of a gene and its flanking genes in selected accessions (Figure 1E). The homologous relations among genomes help to investigate gene-based variations in the local regions of multiple accessions. Many genes encoding nucleotide-binding site leucine-rich repeat proteins are found in the region close to the end of rice chromosome 11 long arm (Supplemental Figure 5) (Song et al., 2021), and the collinearity comparison results detected by this module show that these nucleotide-binding site leucine-rich repeat genes are significantly more abundant in MH63 than in other accessions, which potentially contribute to MH63's superior resistance to rice diseases.
2) At the global scale, "MacroCollinearity" helps users to explore collinearity between accessions and study rearrangements of rice genome at the whole-chromosome level. With this module, structure variations may be easily detected, and the interactive tool "Dot Plot" was embedded to show the collinearity details and links to associated genome loci on "JBrowse" (Figure 1G). A useful module, "GenePair," is provided to visualize collinearity comparisons of ortholog gene pairs between two accessions on both global and local scales.

All information mentioned above is logically organized and seamlessly integrated by modules and tools in RGI. Four extra modules ("JBrowse" [Figure 1I], "GOEnrichment" [Figure 1H], "GeneDescription," and "Download") were additionally integrated to enhance RGI's serviceability (supplemental information 2). The technical details on RGI construction of RGI are described in supplemental information 3.

Although more than 100 chromosomal-level genomes of Asian rice have been published, most of the relevant databases focus on single genomes for specific domains (e.g., long non-coding RNA, epigenomic, etc.). Two "pan-genome" databases have been published (i.e., RPAN [https://cgm.sjtu.edu.cn/3kricedb/index.php] provides data on individual rice accessions, and Rice RC [http://ricerc.sicau.edu.cn/RiceRC] has a focus on structure variants), while our RGI comprehensively creates and focuses on gene-level relationships across representative Asian rice accessions, establishes a standardized gene index for Asian rice, and provides richer search and visualization capabilities for the whole rice research community.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Molecular Plant Online*.

FUNDING

This research was supported by Fundamental Research Funds for the Central Universities (2662020SKPY010), the Major Project of Hubei Hongshan Laboratory (2022HSZD031), and Huazhong Agricultural University's Start-up Fund to J.Z.

AUTHOR CONTRIBUTIONS

R.A.W., W.G., and J.Z. designed and conceived the research. K.L.M. provided SSD seeds and extracted DNA and RNA for genome and transcript sequencing (both by PacBio and Illumina) for 12 Asian rices. Z.Y., Y. Zhou, Y. Zhang, and M.L. performed the homolog and transcriptome analysis. Z.Y. and Y.C. built the database and managed the computing platforms. Z.Y., Y.C., Y. Zhou, Y.O., D.C., R.M., H.Z., W.X., K.L.M., R.A.W., W.G., and J.Z. wrote and edited the paper. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We are thankful for access to the annotation for the Magic 16 gene structure annotation from the Gramene project, specifically K. Chougule, Z. Lu,

(C) The architecture of RGI shows functions and tools. "GeneCard" is a central module in RGI, which links to other modules via a network.
(D) "GeneCard" page shows *OsNiR*'s transcript structure, AS events, gene expression, and homologs.
(E) "MicroCollinearity" module shows homologous relationships of a disease-resistant gene cluster on the local scale between the MH63 and Zhenshan 97 genomes. The black, green, and yellow lines represent reciprocal best hit, single-side best hit, and one-to-many relationships, respectively. For genes, black, green, blue, and yellow represent genes with reciprocal best hit, single-side best hit, singleton, and one-to-may relationships, respectively.
(F) Gene tree in the "Homologues" module.
(G) "MacroCollinearity" module shows the collinear blocks in chromosome 4 between Zhenshan 97 and MH63 and in chromosome 4 between MH63 and Azucena. Colors indicate the collinear block scores. Inversions were highlighted by red triangle.
(H) Gene Ontology enrichment analysis.
(I) *OsNiR*'s location in JBrowse.
(J) The partial result of searching *GHD7*'s sequence in sequence databases of the whole genome (from 16 assemblies) and protein (from 18 annotations) by the "BLAST" tool. The black triangles show the functions and links to other pages.

and D. Ware, supported by USDA 8062-21000-041-00D. We sincerely thank the computing platform of the National Key Laboratory of Crop Genetic Improvement in HZAU and the ibex & Shaheen Cray XC40 platform at KAUST Supercomputing Laboratory for providing the computational resources. No conflict of interest is declared.

Received: November 15, 2022
Revised: February 7, 2023
Accepted: March 21, 2023
Published: March 24, 2023

Zhichao Yu (于志超)[1,6], Yongming Chen (陈永明)[2,6], Yong Zhou (周勇)[3,6], Yulu Zhang (张雨露)[1], Mengyuan Li (李梦圆)[1], Yidan Ouyang (欧阳亦聃)[1], Dmytro Chebotarov[4], Ramil Mauleon[4], Hu Zhao (赵虎)[1], Weibo Xie (谢为博)[1], Kenneth L. McNally[4], Rod A. Wing[3,5,], Weilong Guo (郭伟龙)[2,*] and Jianwei Zhang (张建伟)[1,*]*

[1]National Key Laboratory of Crop Genetic Improvement, Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China
[2]Frontiers Science Center for Molecular Design Breeding, Key Laboratory of Crop Heterosis and Utilization (MOE), and Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Beijing 100193, China
[3]Center for Desert Agriculture, Biological and Environmental Sciences & Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[4]International Rice Research Institute (IRRI), Los Baños, Laguna 4031, Philippines
[5]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
[6]These authors contributed equally to this article.
***Correspondence: Rod A. Wing (rod.wing@kaust.edu.sa), Weilong Guo (guoweilong@cau.edu.cn), Jianwei Zhang (jzhang@mail.hzau.edu.cn)**
https://doi.org/10.1016/j.molp.2023.03.012

REFERENCES

Chen, Y., Song, W., Xie, X., Wang, Z., Guan, P., Peng, H., Jiao, Y., Ni, Z., Sun, Q., and Guo, W. (2020). A collinearity-incorporating homology inference strategy for connecting emerging assemblies in the triticeae tribe as a pilot practice in the plant pangenomic era. Mol. Plant **13**:1694–1708. https://doi.org/10.1016/j.molp.2020.09.019.

International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. Nature **436**:793–800. https://doi.org/10.1038/nature03895.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., Schwartz, D.C., Tanaka, T., Wu, J., Zhou, S., et al. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice **6**:4. https://doi.org/10.1186/1939-8433-6-4.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.-c., Iwamoto, M., Abe, T., et al. (2013). Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. **54**:e6. https://doi.org/10.1093/pcp/pcs183.

Song, J.-M., Xie, W.-Z., Wang, S., Guo, Y.-X., Koo, D.-H., Kudrna, D., Gong, C., Huang, Y., Feng, J.-W., Zhang, W., et al. (2021). Two gap-free reference genomes and a global view of the centromere architecture in rice. Mol. Plant **14**:1757–1767. https://doi.org/10.1016/j.molp.2021.06.018.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J.L., et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat. Genet. **50**:285–296. https://doi.org/10.1038/s41588-018-0040-0.

Wing, R.A., Purugganan, M.D., and Zhang, Q. (2018). The rice genome revolution: from an ancient grain to Green Super Rice. Nat. Rev. Genet. **19**:505–517. https://doi.org/10.1038/s41576-018-0024-z.

Yang, J., Sun, K., Li, D., Luo, L., Liu, Y., Huang, M., Yang, G., Liu, H., Wang, H., Chen, Z., and Guo, T. (2019). Identification of stable QTLs and candidate genes involved in anaerobic germination tolerance in rice via high-density genetic mapping and RNA-Seq. BMC Genom. **20**:355. https://doi.org/10.1186/s12864-019-5741-y.

Yu, J., Xuan, W., Tian, Y., Fan, L., Sun, J., Tang, W., Chen, G., Wang, B., Liu, Y., Wu, W., et al. (2021). Enhanced OsNLP4-OsNiR cascade confers nitrogen use efficiency by promoting tiller number in rice. Plant Biotechnol. J. **19**:167–176. https://doi.org/10.1111/pbi.13450.

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L., Huang, T., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. **50**:278–284. https://doi.org/10.1038/s41588-018-0041-z.

Zhou, Y., Chebotarov, D., Kudrna, D., Llaca, V., Lee, S., Rajasekar, S., Mohammed, N., Al-Bader, N., Sobel-Sorenson, C., Parakkal, P., et al. (2020). A platinum standard pan-genome resource that represents the population structure of Asian rice. Sci. Data **7**:113. https://doi.org/10.1038/s41597-020-0438-2.

Zhou, Y., Yu, Z., Chebotarov, D., Chougule, K., Lu, Z., Rivera, L.F., Kathiresan, N., Al-Bader, N., Mohammed, N., Alsantely, A., et al. (2023). Pan-genome inversion index reveals evolutionary insights into the subpopulation structure of Asian rice. Nat. Commun. **14**:1567. https://doi.org/10.1038/s41467-023-37004-y.