

Brief Communication

PGCP: A comprehensive database of plant genomes for comparative phylogenomics

Xinkai Zhou^{1,†}, Hai-Yun Fan^{2,†}, Xing-Yu Feng^{1,†}, Zhonghao Ruan^{1,†}, Jinkai Yuan¹, Quan Han¹, Zhaohui He¹, Yuxin You¹, Haoyu Chao³ , Ming Chen³, Zhu-Qing Shao^{1,*} , Jia-Yu Xue^{2,*}  and Dijun Chen^{1,*} 

¹State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Nanjing University, Nanjing, China

²Bioinformatics Center, College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing, China

³Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China

Received 17 January 2025;

revised 28 March 2025;

accepted 15 April 2025.

*Correspondence (Tel/fax +86 25 89682705; email zhuqingshao@nju.edu.cn (Z.-Q.S.); Tel/fax +86 25 84395036; email xuejy@njau.edu.cn (J.-Y.X.); Tel/fax +86 25 89682705; email dijunchen@nju.edu.cn (D.C.))

[†]These authors contributed equally to this work.

Keywords: plant genomes, database, comparative phylogenomics.

With the rapid advancement of sequencing and genome assembly technologies, thousands of plant genomes have been sequenced, assembled and annotated (Marks *et al.*, 2021; Xie *et al.*, 2024), providing a rich and invaluable resource for comparative genomic and evolutionary studies in plants. However, these genomes are often dispersed across various databases, creating barriers to efficient data access, integration and comprehensive analysis. Existing plant genome databases, such as Phytozome (Goodstein *et al.*, 2012) and Ensembl Plants (Bolser *et al.*, 2017), offer significant contributions but are limited in scope, typically covering only a limited number of species and a few representative genomes. The lack of a centralized platform for accessing and analysing diverse plant genomes makes it challenging to perform large-scale comparative studies, impeding the identification of conserved and divergent genomic features and hampering the discovery of broader evolutionary and functional patterns across the plant kingdom. Therefore, there is a pressing need for a comprehensive, user-friendly and integrative database that brings together high-quality genomic data from a wide range of plant species.

Here, we introduce PGCP (Plant Genomes for Comparative Phylogenomics), a comprehensive database that integrates extensive genomic data across the plant phylogeny. PGCP consolidates 719 publicly available genomes from various platforms and published studies, ensuring the inclusion of high-quality assemblies and complete annotation information (Figure 1a). After rigorous quality control, we obtained a total of 26 600 642 protein-coding genes from these genomes. To enhance data consistency and usability, we performed a uniform and standardized re-building of genome annotations across the database. Gene and chromosome identifiers (IDs) were transformed into PGCP-specific identifiers, with records carefully filtered to maintain data integrity and confidentiality. All the protein-coding genes were further annotated with various functional attributes, including the prediction of transcription factors (TFs) and transcriptional regulators (TRs), using iTAK (Zheng *et al.*, 2016). Additionally, gene ontology (GO) terms and protein domain annotations were assigned using InterProScan

(Jones *et al.*, 2014), providing a comprehensive functional landscape of the gene datasets (Figure 1b).

To uncover hidden patterns within extensive genomic datasets, we conducted a large-scale comparative phylogenomic analysis using OrthoFinder (Emms and Kelly, 2019). From the 719 genomes collected in PGCP, we identified 237 400 orthogroups, each comprising genes from at least three different species. This comprehensive resource enables the exploration of complex evolutionary relationships by comparing homologous genes across plant species, tracing gene family origination and divergence that highlight adaptive evolution, identifying lineage-specific genes contributing to unique species characteristics and predicting gene functions based on conserved and divergent features (Figure 1c). Such analyses provide an in-depth understanding of the genetic foundations of plant diversity, offering valuable information on evolutionary trajectories, adaptive strategies and functional innovations in the plant kingdom. In a nutshell, the integration of these orthogroups significantly enhances the utility of the PGCP database, making it an invaluable resource for advancing evolutionary research.

The PGCP database also offers a wide range of visualization features and powerful analytical tools to help users explore the wealth of genomic data effectively.

1. **Genome View and Gene Annotation View:** The 'Genome View' provides comprehensive information on genome assembly and annotation details, and links to internal genome browser and external databases, along with a searchable list of annotated genes (Figure 1d). From this list, users can select specific genes to access the 'Gene Annotation View', which offers detailed sequence information, functional annotations and homologue relationships for genes from any other species. This view allows for in-depth investigation of gene functions, evolutionary relationships and comparative analyses, making it easier for researchers to identify comprehensive features of selected genes across species (Figure 1e).
2. **BLAST Tools for multiple genomes:** PGCP provides robust 'BLAST Tools' that enable cross-species sequence search. Users can easily select reference genomes for alignment from an intuitive species relationship tree, streamlining the process of identifying homologous sequences across different plants. BLAST results remain accessible for 1 week, with an option to receive detailed results via email, facilitating seamless and efficient cross-genome comparisons for various research needs (Figure 1f).
3. **Functional Enrichment Analysis:** The platform provides an online 'Functional Enrichment Analysis' tool, enabling users to conduct functional analysis for any species within the



Figure 1 Overview of the Plant Genomes for Comparative Phylogenomics (PGCP) database. (a) Compilation of plant genome data, including both high-quality genome sequences and gene feature annotations. (b) Statistical summary of the genomic datasets and various functional annotations of the protein-coding genes. (c) Identification of orthogroups by OrthoFinder, showcasing comprehensive phylogenomic analyses that enable the exploration of evolutionary relationships, gene family dynamics and functional divergence across plant species. (d–i) Screenshots of the pages of ‘Genome View’ (d), ‘Gene Annotation View’ (e), ‘BLAST Tools’ (f), ‘Functional Enrichment Analysis’ (g), ‘Synteny View’ (h) and ‘Primer Designer’ (i).

database. Results are visually presented in enrichment plots and maps, providing a visual overview of enriched functions and pathways. Detailed tabular data are also available, allowing for in-depth exploration (Figure 1q).

4. **Synteny View for chromosome-level genomes:** The 'Synteny View' feature enables users to compare up to five genomes simultaneously, displaying synteny relationships between selected genomes using SynVisio (Bandi and

Gutwin, 2020). This tool is particularly useful for studying structural variation or even indicating whole-genome duplications among species (Figure 1h).

5. **Primer Designer tool:** The 'Primer Designer' tool allows users to design primers for any region within the genomes listed in PGCP. Visualization options are provided, aiding in the selection of optimal primers for experimental applications (Figure 1i).

In summary, PGCP aims to provide an extensive and accessible platform that integrates genomic data across a wide array of plant species, encompassing well-studied model plants, important crops and diversified wild plants. By offering a unified repository of high-resolution genome sequences, annotations and comparative genomics data, PGCP facilitates large-scale comparative analyses and enables researchers to explore plant phylogenetics, gene family dynamics and functional genomics across the plant kingdom. The database includes advanced analytical tools that support the identification of homologues, phylogenetic reconstruction and the discovery of key genes involved in adaptive and agronomic traits. As a unified platform, PGCP is designed to be an essential resource for researchers in plant biology, fostering new insights into plant evolution, biodiversity and the molecular mechanisms that drive trait diversity. In the future, we plan to expand the PGCP database by incorporating additional plant genome resources to enhance its comprehensiveness. As part of our ongoing development, we will work towards systematically including genomes from GWH (Ma *et al.*, 2025) and other relevant databases to provide a more extensive and up-to-date resource for the plant research community.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 32070656 to D.C. and 32070243 to Z.-Q.S.) and the Fundamental Research Funds for the Central Universities (to J.-Y.X.). The authors acknowledge the Center for Information Technology and the High-Performance Computing Center of Nanjing University for providing high performance computing (HPC) resources.

Author contributions

D.C., J.-Y.X. and Z.-Q.S. conceived and supervised the study. H.-Y.F., X.-Y.F. and X.Z. collected genome sequence, gene

annotation and meta data with help from J.Y., Q.H. and Y.Y. X.Z. and Z.R. processed the genomic data and constructed the database with support from H.C. and M.C. D.C. wrote the manuscript. All authors reviewed and approved the submitted manuscript.

Conflict of interest

The authors declare no competing interests.

Data availability statement

The PGCP database is available at <https://biobigdata.nju.edu.cn/PGCP/>.

References

- Bandi, V. and Gutwin, C. (2020) Interactive exploration of genomic conservation. In: Levin, D.I.W., Chevalier, F. and Jacobson, A. (Eds.) *Proceedings of Graphics Interface 2020*, pp. 74–83. Kelowna, BC: Canadian Human-Computer Communications Society/Société canadienne du dialogue humain-machine.
- Bolser, D.M., Staines, D.M., Perry, E. and Kersey, P.J. (2017) Ensembl plants: Integrating tools for visualizing, mining, and analyzing plant genomic data. *Methods Mol. Biol.* **1533**, 1–31.
- Emms, D.M. and Kelly, S. (2019) OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14.
- Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–D1186.
- Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
- Ma, Y., Zhao, X., Jia, Y., Han, Z., Yu, C., Fan, Z., Zhang, Z. *et al.* (2025) The updated genome warehouse: enhancing data value, security, and usability to address data expansion. *Genomics Proteomics Bioinformatics*, qzaf010. <https://doi.org/10.1093/GPBJNL/QZAF010>
- Marks, R.A., Hotaling, S., Frandsen, P.B. and VanBuren, R. (2021) Representation and participation across 20 years of plant genome sequencing. *Nat. Plants*, **7**, 1571–1578.
- Xie, L., Gong, X., Yang, K., Huang, Y., Zhang, S., Shen, L., Sun, Y. *et al.* (2024) Technology-enabled great leap in deciphering plant genomes. *Nat. Plants*, **10**, 551–566.
- Zheng, Y., Jiao, C., Sun, H., Rosli, H.G., Pombo, M.A., Zhang, P., Banf, M. *et al.* (2016) iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant*, **9**, 1667–1670.