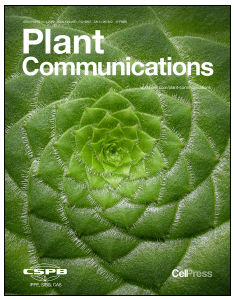
Journal Pre-proof



Comparative pan-genomics reveals pervasive lineage-specific innovations and improves cross-species synteny inference in cereals

Tashi Dorjee, Yanbo Wang, Shichao Sun, Chuanzheng Wei, Adrian Hathorn, Sofie Pearson, David Jordan, Emma Mace, Yongfu Tao

PII: S2590-3462(26)00384-6

DOI: https://doi.org/10.1016/j.xplc.2026.102076

Reference: XPLC 102076

To appear in: *Plant Communications*

Received Date: 1 April 2026

Revised Date: 15 June 2026

Accepted Date: 12 August 2026

Please cite this article as: Dorjee, T., Wang, Y., Sun, S., Wei, C., Hathorn, A., Pearson, S., Jordan, D., Mace, E., Tao, Y., Comparative pan-genomics reveals pervasive lineage-specific innovations and improves cross-species synteny inference in cereals, *Plant Communications* (2026), doi: https://doi.org/10.1016/j.xplc.2026.102076.

Comparative pan-genomics reveals pervasive lineage-specific innovations and improves cross-species synteny inference in cereals

Tashi Dorjee[1,†], Yanbo Wang[1,†], Shichao Sun[1], Chuanzheng Wei[1], Adrian Hathorn[2,3], Sofie Pearson[2,3], David Jordan[2], Emma Mace[2,3*], Yongfu Tao[1,*]

[1]State Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, Guangdong 518120, China

[2]Queensland Alliance for Agriculture and Food Innovation (QAAFI), The University of Queensland, Hermitage Research Facility, Warwick, QLD 4370, Australia

[3]Queensland Department of Primary Industries (DPI), Hermitage Research Facility, Warwick, QLD 4370, Australia

*Correspondence (taoyongfu@caas.cn and emma.mace@uq.edu.au)

†These authors contributed equally to this work.

Dear Editor,

Comparative genomics has long advanced our understanding of genome composition, function and evolution by revealing conserved genomic components and lineage-specific innovations (Koonin et al., 2000). Its analytical framework is built on two complementary pillars: sequence homology, which defines sequence-level similarity, and synteny conservation, which reflects the preserved ordering of genes across divergent genomes. Traditional reference-based comparative analyses have highlighted widespread conservation of gene order and function among cereal crops (Tang et al., 2008). However, recent pan-genome studies across major crops have uncovered extensive structural variation and widespread gene presence-absence variation, demonstrating that a single reference is insufficient to capture the full gene repertoire of a species (Guo et al., 2025; Jayakodi et al., 2024; Jiao et al., 2025; Sun et al., 2026). By integrating complete gene complements and their population frequency spectra, pan-genomes provide a holistic representation of intraspecific genomic diversity, emerging as the new foundational resource for crop genomic analyses

(Matthews et al., 2024). This paradigm shift raises fundamental questions about how pan-genomic diversity reshapes comparative genomics, an area that remains poorly explored. The availability of high-quality pan-genomes for three major cereal crops of global agricultural importance, namely sorghum (Tao et al., 2021), maize (Hufford et al., 2021), and rice (Qin et al., 2021), offers an unprecedented opportunity to address these questions with direct relevance to crop biology and improvement.

To examine how the pan-genome framework impacts sequence homology-based comparative genomics, we first conducted protein-based gene family analysis using both the reference genomes and pan-genomes for rice, sorghum, and maize. Comparative analysis of the three reference genomes showed that the majority of gene families (>83%) in the three crops have homologs in at least one other species (Figure 1A; Dataset 1), consistent with previous reports that cereals share a large proportion of their gene content (Bennetzen, 2007). In contrast, pan-genome-based comparisons identified substantial proportions (>38%) of species-specific gene families in all three species (Figure 1A), exposing major limitations of functional inference based on single reference genomes. These species-specific gene families have fewer member genes than shared gene families and collectively contribute ~10% of the total gene complement across the three pan-genomes (Figure 1B). GO enrichment analysis showed that species-specific genes in rice, sorghum, and maize were enriched in processes related to adaptation and regulation, including the regulation of protein modification and homeostasis, developmental differentiation, and responses to environmental stresses (Supplemental Tables 1, 2, and 3). Consistently, expression analysis revealed that species-specific genes show stronger tissue specificity than shared genes (Supplemental Figure 1), suggesting a greater involvement in specialized functions.

These species-specific gene families were further classified into core, shell, and cloud categories according to their occurrence frequency within each crop (Datasets 2, 3, and 4). A total of 1,144, 489, and 327 core gene families were identified in rice, sorghum, and maize, respectively. Core gene families were markedly underrepresented among species-specific gene families, with only 4.1%, 3.7%, and 1.8% classified as core in rice, sorghum, and maize, respectively (Figure 1C). These species-specific core genes exhibit strong sequence conservation within species and are likely associated with the development of hallmark, species-defining traits. Rice-specific core genes are

enriched for functions related to oxidative stress response and protease inhibition (Supplemental Figure 2A), and include *OsSub1B* (LOC_Os09g11460), which is involved in submergence adaptation (Supplemental Table 4, Dataset 5). Maize-specific core genes are enriched for seed storage protein accumulation (Supplemental Figure 2B), and include α-zein genes, such as *az19D2* (Zm00001eb030160) (Supplemental Table 4, Dataset 5). Sorghum-specific core genes are enriched for functions related to abiotic stress responses and secondary metabolism (Supplemental Figure 2C), and include genes encoding key enzymes in the dhurrin and sorgoleone biosynthetic pathways, such as *GSTL1* (SORBI_3002G421200) and *SbOMT3* (SORBI_3006G007900) (Supplemental Table 4, Dataset 5). Overall, the majority of species-specific gene families (>92% in all three species) were classified into the shell category (Figure 1C).

SbOMT3 encodes an O-methyltransferase that catalyzes the penultimate step in the sorgoleone biosynthetic pathway (Supplemental Figure 3). Sorgoleone is a hydrophobic p-benzoquinone exuded from sorghum root hairs that potently suppresses the growth of competing weeds (de Oliveira et al., 2024). Microsynteny analysis further supported the absence of a syntenic *SbOMT3* ortholog in maize and rice, as no homologous sequence was detected in the corresponding genomic regions (Figure 1D). Homology searches further revealed that *SbOMT3* homologs are restricted to the Panicoideae subfamily and are broadly distributed across the Andropogoneae, Paniceae, and Paspaleae tribes (Figure 1E, Supplemental Figure 4), suggesting that *SbOMT3* was likely present in the common ancestor of Panicoideae and was subsequently lost in the maize lineage. Phylogenetic analysis showed that the sorghum SbOMT3 protein clustered closely with homologous proteins from *Erianthus*, *Miscanthus*, and *Saccharum,* with over 90% protein sequence similarity despite loss of synteny (Supplemental Figures 5 and 6). Sequence comparison further revealed a higher abundance of ATAT-box and TA-ATAT-box motifs in the sorghum *SbOMT3* promoter relative to other Saccharinae species, consistent with its uniquely elevated and root-specific expression pattern (Figures 1F and 1G). Together, these regulatory and sequence differences likely contribute to divergence in sorgoleone biosynthesis between sorghum and other Saccharinae lineages. Overall, our analysis demonstrates that comparative pan-genomic analysis of closely related species could provide a powerful framework to understand the evolution of lineage-specific traits in crops.

Figure 1. Comparative analysis of species-specific and shared gene repertoires and synteny of key loci under single-reference and pan-genome frameworks.

(A) Venn diagrams showing the distribution of shared and species-specific gene families among sorghum, maize, and rice based on reference genomes (left) and pan-genomes (right).

(B) Comparison of the number of genes per orthogroup (OG) between species-specific gene families (Unique) and shared gene families (Shared). ** indicates $P < 0.01$; **** indicates $P < 0.0001$.

(C) Frequency-based classification of species-specific gene families in sorghum, maize, and rice pan-genomes. Gene families present in all accessions are defined as core gene families, those found in only one accession are defined as cloud gene families, and the remainder are defined as shell gene families.

(D) Microsynteny of the *SbOMT3* locus among two sorghum accessions and the rice (NIP) and maize (B73) genomes. Pink lines indicate syntenic regions surrounding the *SbOMT3* locus.

(E) Phylogenetic tree constructed using selected Poaceae genomes; taxa containing *SbOMT3* homologs are highlighted in red.

(F) Abundance of the top five cis-regulatory elements within the 2-kb promoter regions upstream of *OMT3* in *S. bicolor*, *S. spontaneum*, *E. rufipilus*, and *M. lutarioriparius*.

(G) Expression patterns of *OMT3* in root, stem, leaf, and inflorescence of *S. bicolor*, *S. spontaneum*, *E. rufipilus*, and *M. lutarioriparius* (values shown as $\log_2(\text{TPM} + 1)$).

(H) Comparison of syntenic relationships between rice chromosome 5 and sorghum chromosome 9 using the reference genome (top) and the common-path genome (bottom). Gray lines connect syntenic homologous genes between the two chromosomes. Blue triangles indicate genomic intervals where syntenic relationships were improved or newly recovered in the common-path genome relative to the reference-genome analysis. The corresponding genomic coordinates are shown adjacent to each marker.

(I) Microsynteny of the *OsNCED5* locus from the rice NIP genome with the maize B73, sorghum common-path genome, and sorghum BTx623 genomes. Red lines indicate syntenic regions corresponding to *OsNCED5*, while pink bubbles represent structural variations. The syntenic relationship is conserved in maize B73 and the sorghum

common-path genome, whereas the corresponding syntenic *OsNCED5* ortholog is absent in BTx623, indicating a loss of the syntenic ortholog at this locus.

(J) Comparison of interspecific expression correlations between syntenic homologous gene pairs (Syntenic homologs) and non-syntenic homologous gene pairs (Non-syntenic homologs) across tissues in rice, sorghum, and maize. ** indicates $P < 0.01$; **** indicates $P < 0.0001$.

The pan-genome-based comparison identified over 13% more shared gene families than analyses relying on a single reference genome (Figure 1A; Dataset 6), resulting in a higher proportion of genes from each of the three reference genomes having homologs in the other species (Supplemental Figure 7). The shared genes are enriched with functions related to essential biological processes in plants, including basic metabolism, genetic information processing, and cell structure maintenance (Supplemental Table 5). The shared gene families identified by pan-genome comparison are predominantly core genes within each species (Supplemental Figure 8), with 98.1% present in over half of individuals (Supplemental Figure 9). Among the 7,920 core gene families shared across the three species, 1,465 single-copy gene families were identified across the three species, suggesting their highly conserved roles in the genome (Supplemental Figure 10A). We further identified core gene families that are significantly expanded in one species relative to the other two, with 387, 510, and 2,446 gene families detected in rice, sorghum, and maize, respectively (Supplemental Figure 10B). These expanded core gene families likely underpin functional divergence and adaptive evolution among the crops.

On average, around 10% of genes in each genome of sorghum, maize, and rice are present in less than half of genotypes in their respective pan-genomes (Supplemental Figure 9). These low-frequency genes are, on average, less likely to have counterparts in the other species and are often lineage-specific. However, a subset of them does possess detectable homologs across species, and their inclusion in the pan-genome analysis contributes to the observed increase in shared gene families. These low-frequency genes could otherwise complicate the inference of cross-species gene order conservation when using single-reference genomes.

To mitigate the impact of low-frequency genes on synteny inference, we proposed a synthetic “common-path” genome that incorporates the major haplotype (frequency

≥ 50%) at each structural variation locus within the sequence-based pan-genome to represent the ancestral genome for cross-species comparison. We constructed sequence-based pan-genomes for rice, maize, and sorghum and derived their corresponding synthetic common-path genomes for synteny analysis. Compared with single reference-genome-based synteny analysis, the common-path genome approach consistently identified more syntenic blocks and syntenic genes across all species pairs, and exhibited higher syntenic gene density and greater coverage of syntenic blocks in pairwise comparisons among rice, maize, and sorghum (Supplemental Figures 11A–11D; Dataset 7). These findings suggest that the pan-genome-derived common-path genomes could restore syntenic relationships lost owing to structural variation by capturing ancestral haplotypes. This improvement was particularly evident in the syntenic regions between rice chromosome 5 and sorghum chromosome 9 (Figure 1H). Overall, these results indicate that integrating pan-genome information effectively enhances the detection sensitivity and genomic coverage of cross-species synteny inference.

Using the common-path genomes, we identified a syntenic relationship between the rice drought-tolerance gene *OsNCED5* (LOC_Os12g42280) (Huang et al., 2019) and its sorghum counterpart. The two proteins share 91.7% sequence identity and exhibit strong conservation of their flanking syntenic regions (Figure 1I; Supplemental Table 4). Although *OsNCED5* is present in all 33 rice genomes analyzed, no ortholog of this gene was identified in two sorghum accessions, including the reference genome. The absence of a homologous sequence in these genomes prevents the identification of the corresponding syntenic relationship in reference-based analyses. Notably, *OsNCED5* co-localizes with a drought-tolerance quantitative trait locus (QTL) in sorghum (Sabadin et al., 2012), supporting the potential functional relevance of this gene in sorghum drought tolerance.

To test whether the syntenic relationship could better predict functional similarity than homology alone, we compared interspecific expression correlations between syntenic homologous gene pairs and non-syntenic homologous gene pairs within each gene family. Our results showed that syntenic gene pairs displayed significantly higher expression similarity across species (Figure 1J), suggesting that syntenic genes are more likely to retain functional similarity than non-syntenic homologous genes.

In summary, our comparative pan-genomics analyses reveal extensive lineage-

specific innovation in cereal crops and establish a framework for comparative genomics and crop improvement that is readily transferable as pan-genomes expand. However, several limitations should be noted. These cross-species comparisons in this study are heavily influenced by the quantity, representativeness, and quality of available genomes. In addition, excluding low-frequency genes may result in the loss of rare yet beneficial alleles. Moreover, the genetic divergence between species or conspecific plant samples is best viewed as a continuum shaped by evolutionary processes, making inferred patterns sensitive to sampling and dataset composition. Future incorporation of more high-quality and representative genomes across broader phylogenetic and population scales will help validate and refine these findings, advancing our understanding of genome evolution.

Funding

This work was supported by the Special Funds of the National Natural Science Foundation of China (32441040).

Author contributions

T.D. and Y.W. contributed equally to this work and share first authorship. T.D., Y.W., and C.W. performed the formal analysis and visualization. S.S. curated and preprocessed the genomic data. Y.T. conceptualized the study and supervised the project. Y.T., E.M., A.H., S.P., and D.J. provided intellectual and analytical input and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Prof. Xingtan Zhang (Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences) and Prof. Haibao Tang (Fujian Agriculture and Forestry University, China) for their thoughtful discussions. The authors have no conflicts of interest to declare.

Data and code availability

The genome assemblies and gene annotations used in this study are publicly available from the following sources: Sorghum (*Sorghum bicolor*): SorghumBase (https://www.sorghumbase.org/). Maize (*Zea mays*): MaizeGDB (https://www.maizegdb.org/). Rice (*Oryza sativa*): Rice Resource Center (https://ricerc.sicau.edu.cn/). All code is available via our repository at

https://github.com/Dorjeetashi/grass-pan-genome. The associated dataset has been deposited in Figshare and is available at https://doi.org/10.6084/m9.figshare.31914012.

References

Bennetzen, J.L. (2007). Patterns in grass genome evolution. Curr. Opin. Plant Biol. **10**:176–181. https://doi.org/10.1016/j.pbi.2007.01.010.

de Oliveira, I.F., Gomes, T.C., Simeone, M.L.F., Karam, D., and de Sousa Tinoco, S.M. (2024). Sorgoleone unveiled: exploring its biosynthesis, functional perspectives and applications. Braz. J. Bot. **47**:723–733. https://doi.org/10.1007/s40415-024-01026-7.

Guo, D., Li, Y., Lu, H., Zhao, Y., Kurata, N., Wei, X., Wang, A., Wang, Y., Zhan, Q., Fan, D., et al. (2025). A pangenome reference of wild and cultivated rice. Nature **642**:662–671. https://doi.org/10.1038/s41586-025-08883-6.

Huang, Y., Jiao, Y., Xie, N., Guo, Y., Zhang, F., Xiang, Z., Wang, R., Wang, F., Gao, Q., Tian, L., et al. (2019). *OsNCED5*, a 9-cis-epoxycarotenoid dioxygenase gene, regulates salt and water stress tolerance and leaf senescence in rice. Plant Sci. **287**:110188. https://doi.org/10.1016/j.plantsci.2019.110188.

Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A., Qiu, Y., et al. (2021). De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science **373**:655–662. https://doi.org/10.1126/science.abg5289.

Jayakodi, M., Lu, Q., Pidon, H., Rabanus-Wallace, M.T., Bayer, M., Lux, T., Guo, Y., Jaegle, B., Badea, A., Bekele, W., et al. (2024). Structural variation in the pangenome of wild and domesticated barley. Nature **636**:654–662. https://doi.org/10.1038/s41586-024-08187-1.

Jiao, C., Xie, X., Hao, C., Chen, L., Xie, Y., Garg, V., Zhao, L., Wang, Z., Zhang, Y., Li, T., et al. (2025). Pan-genome bridges wheat structural variations with habitat and breeding. Nature **637**:384–393. https://doi.org/10.1038/s41586-024-08277-0.

Koonin, E.V., Aravind, L., and Kondrashov, A.S. (2000). The impact of comparative genomics on our understanding of evolution. Cell **101**:573–576. https://doi.org/10.1016/s0092-8674(00)80867-3.

Matthews, C.A., Watson-Haigh, N.S., Burton, R.A., and Sheppard, A.E. (2024). A gentle introduction to pangenomics. Brief. Bioinform. **25**:bbae588.

https://doi.org/10.1093/bib/bbae588.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., et al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. Cell **184**:3542–3558.e16. https://doi.org/10.1016/j.cell.2021.04.046.

Sabadin, P., Malosetti, M., Boer, M., Tardin, F., Santos, F., Guimaraes, C., Gomide, R., Andrade, C., Albuquerque, P., Caniato, F., et al. (2012). Studying the genetic basis of drought tolerance in sorghum by managed stress trials and adjustments for phenological and plant height differences. Theor. Appl. Genet. **124**:1389–1402. https://doi.org/10.1007/s00122-012-1795-9.

Sun, S., Wei, C., Pearson, S., Cruickshank, A., Wang, Y., Wang, Y., Dorjee, T., Zhou, Y., Jordan, D., Yang, Y., et al. (2026). Pan-genome analysis uncovers extensive structural variations in Africa's indigenous legume crop cowpea. Nat. Commun. **17**:6656. https://doi.org/10.1038/s41467-026-73494-2.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science **320**:486–488. https://doi.org/10.1126/science.1153917.

Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., Hathorn, A., Wu, X., Liu, Y., Shatte, T., et al. (2021). Extensive variation within the pan-genome of cultivated and wild sorghum. Nat. Plants **7**:766–773. https://doi.org/10.1038/s41477-021-00925-x.