NAR Genomics and Bioinformatics, 2026, **8**, lqag030
https://doi.org/10.1093/nargab/lqag030
Data Resources

A sorghum-anchored pan-grass syntenic gene set in grasses

Henrique M. Dias[1,2,3,4], Guilherme Y.A. Sagawa[1], Vladimir J. Torres-Rodriguez[2,3], Ravi V. Mural[4], James C. Schnable[2,3], Marie-Anne Van Sluys[1,*]

[1]Departamento de Botânica, Instituto de Biociências, Universidade de São Paulo, SP, Brazil
[2]Center for Plant Science Innovation, University of Nebraska-Lincoln, Lincoln, NE, United States
[3]Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE, United States
[4]Department of Agronomy, Horticulture and Plant Science, South Dakota State University, Brookings, SD, United States
*To whom correspondence should be addressed. Email: mavsluys@usp.br

Abstract

Understanding orthologous and homeologous relationships among genes informs how evolution and artificial selection have shaped plant genomes of related species to adapt to different niches and produce diverse phenotypes. Grasses are ecologically and economically important plants, and many genes are conserved within syntenic blocks across species. Here, we describe the Pan-Grass Syntenic Gene Set (PGSGS), a curated dataset of orthologous and homeologous relationships among 746 743 protein-coding genes from 17 grass genomes anchored to sorghum, including both major and orphan crops, as well as wild grasses. From this analysis, 344 230 genes (46%) were identified as syntelogs, with 27 567 sorghum genes linked to at least one syntelog in other species. Of these, 11 624 syntelogs form a conserved core present in at least 15 species. PGSGS-core syntelogs are enriched for regulatory and cellular functions, form densely interconnected GO networks, and exhibit reduced nucleotide diversity and elevated Tajima's *D* relative to nonsyntenic genes in ∼400 sorghum accessions. Conversely, PACMAD-specific syntelogs ($n = 2266$) is a diverse set of clade-specific retained genes. This dataset provides a foundation for understanding genome evolution after polyploidy and encourages the integration of functional genetic and genomic information across grasses, fulfilling the original promise of the "grasses as a single genetic system."

Introduction

Approximately 50%–60% of the world's calorie intake comes from grasses, including major cereal crops, such as rice, wheat, and maize, as well as minor and orphan crops such as pearl, foxtail, and proso millet [1]. Beyond grains, other grass species also play key roles in agriculture as sources of sugar (sugarcane), forages for livestock, and feedstocks for bioenergy production [2]. Wild grass species are ecologically dominant in vast areas of most continents, and a broad set of molecular, anatomical, and physiological adaptations allowed the members of this clade to thrive in diverse environments [3].

Although grasses exhibit a broadly conserved and collinear genomic architecture, they harbor extensive genetic diversity, encompassing variations in nucleotides [single-nucleotide polymorphisms (SNPs)], differences in gene structure, and the presence or absence of specific genes responsible for particular traits among different species [2]. Grasses exhibit a wide range of photosynthetic and water-use efficiencies [4–6]. In addition, various grass species utilize C3 photosynthesis or three different types of C4 photosynthesis, and different subclades within the grasses exhibit varying capacities to adapt to and thrive in stressful and resource-poor environments [7]. This physiological and ecological diversity is underpinned by genomic divergence and repeated whole-genome duplication (WGD) events that shaped the structure and gene content of modern grasses [8, 9]

Despite their diverse morphologies and ecological niches, the genomes of most grasses exhibit a high degree of genomic colinearity [10–12]. In a typical grass genome, approximately half of all annotated gene models will be conserved at syntenic locations in a pairwise comparison [13]. Both rice and sorghum are diploid relative to the most recent common ancestor of the grasses, and the reference genomes of these two species are widely used in grass comparative genomic analyses. However, as high-contiguity genome assemblies for mesopolyploids (e.g. maize) and neopolyploids (e.g. wheat, sugarcane, switchgrass, miscanthus) have become available, there is a growing need for updated, lineage-balanced frameworks to integrate these complex genomes into comparative analyses [14–16].

These WGDs can create two or more syntenic intervals that are each homologous to a single common region in the genome of sorghum or rice [17, 18]. From this standpoint, comparative genomic analysis is particularly fruitful in the grasses, as synteny can be highly informative both on a single gene level regarding functional data from orthologs in other species and on a genome-wide level in understanding the dynamics of genomic evolution post-polyploidy [19]. Various publicly accessible tools, including Phytozome [20], CoGe [21,

Received: November 14, 2024. **Revised:** February 21, 2026. **Accepted:** February 25, 2026

22], PLAZA [23], Gramene [24], Ensembl Plants [25], can be employed for automated comparative genomics analyses between or among grass genomes and those of other plant species. While these platforms facilitate large-scale comparative analyses, distinguishing orthologs from paralogs in polyploid genomes requires careful parameterization to account for expected copy number ratios and differential fractionation between subgenomes [26]. Differences in gene expression and dosage biases between homeologous chromosomes have been reported across multiple polyploid grasses, underscoring the importance of correctly identifying subgenome-specific syntelogs [27, 28].

In many cases, comparative genomic tools do not explicitly model subgenome identity when categorizing syntelogs in polyploid species, which can limit the precision of orthology assignments across homeologous chromosome sets [29]. Here, we extend the QuotaAlign-based methodology [28, 29] to revise a widely cited syntenic gene set and systematically benchmark multiple analytical pipelines to identify the configuration that most effectively captures syntelogs present in conserved collinear order across chromosomes. We incorporated newly available assemblies of agriculturally, ecologically, and evolutionarily significant grass genomes, including polyploids, and thereby generated the Pan-Grass Syntenic Gene Set (PGSGS) anchored in the sorghum genome. The updated PGSGS integrates a targeted "polish" refinement step to improve syntelog detection and provides an optimized framework for examining synteny of gene-rich chromosomal regions across 17 grass genomes that have undergone one or more WGDs. From 746 743 genes analyzed across these species, 27 567 sorghum genes were linked to at least one syntelog, of which 11 624 constitute a conserved core present in at least 15 of the 17 genomes. Beyond structural comparison, the PGSGS dataset enables downstream analyses of gene function and evolutionary constraint, including enrichment of biological processes and intraspecific variation patterns. Finally, hierarchical clustering of genomes and genes reveals conserved and divergent synteny patterns among diploid and polyploid genomes.

Materials and methods

Genome datasets and annotation sources

A set of 17 grass genomes representing 15 species, all of which are available as part of Phytozome v13 [20], was used for this analysis (Supplementary Table S1). Assemblies were prioritized to balance three objectives: (i) chromosome-scale contiguity (N50 > 20 Mb) to minimize spurious synteny breaks, (ii) annotation completeness (BUSCO > 93% for Embryophyta) to ensure comprehensive gene content, and (iii) phylogenetic representation spanning the major grass lineages (PACMAD and BOP clades). For polyploid species, we prioritized assemblies with well-characterized subgenome structure (e.g. *Triticum aestivum* with A, B, D genomes; *Zea mays* with maize1 and maize2; *Miscanthus sinensis* with misA and misB; and *Saccharum* hybrid (R570) with homologous chromosome sets A–G) to validate our homeolog resolution approach. When multiple assembly versions were available for a species, we selected the most recent chromosome-level assembly. Genome FASTA and GFF3 files were standardized to consistent naming formats before alignment. For each genome, the "cds_primaryTranscriptOnly" fasta file and GFF annotation files were downloaded from Phytozome for downstream analyses, totalling 746 743 genes screened. The workflow for pairwise identification of syntenic orthologous genes between the sorghum reference genome and the other 16 genomes is summarized in Fig. 1.

Synteny pipeline overview

Pairwise genome alignments and synteny inference were performed following the previously published QuotaAlign-based methodology [28]. The PGSGS workflow comprises three main stages (Fig. 1): (i) sequence alignment, (ii) quota-based synteny inference, and (iii) polish refinement.

Stage 1: LASTZ [30] pairwise comparisons were performed using the primary transcript CDS sequences from the species of interest as query and the primary transcript CDS sequences of *Sorghum bicolor* BTx623 v5.1 as reference. LASTZ was run with a seed of match12, a minimum identity of 70%, and a minimum coverage of 50%. The LASTZ output was used to identify preliminary syntenic anchors between *Sorghum* and each genome of interest.

Stage 2: High-confidence anchors were merged into collinear blocks using QuotaAlign v2.0 [29], applying species-specific quota settings (1:1 for diploids, 1:2 for polyploids). Blocks containing fewer than 5 genes were excluded, and adjacent blocks were merged when separated by fewer than 20 genes.

Stage 3: QuotaAlign [29] initially collapses tandemly duplicated genes into single representatives to expedite genome-wide block detection. However, in tandem arrays or clusters of closely related paralogs, the globally optimal gene may not be the true ortholog. To improve detection of these cases, we implemented a local refinement step (polish module; the code is available at figshare DOI: 10.6084/m9.figshare.26793190). For each sorghum reference gene, the algorithm identifies the nearest upstream and downstream genes (within 50 gene positions) that have QuotaAlign assignments. It then examines a ±50-gene window around those anchor positions in the target genome and re-evaluates all LASTZ alignment scores within that valid region. If a different gene within the valid window exhibits a higher LASTZ score than the original QuotaAlign assignment, it replaces the pairing; otherwise, the original assignment is retained. This step refines individual gene-to-gene mappings without altering syntenic block boundaries, improving orthology precision in locally duplicated regions (Fig. 2). The improvement in syntelog recovery through polishing is quantified for each genome in Supplementary Fig. S1.

Benchmarking with existing tools

To evaluate the performance of our three-stage pipeline (LASTZ > QuotaAlign > Polish), we benchmarked it against two widely used synteny detection tools: DAGchainer v1.2 [32] and MCScanX v1.1 [33], using the well-studied sorghum–maize comparison as a study case. To ensure fair comparison, all three tools used identical LASTZ pairwise alignments (minimum coverage 50%, minimum identity 70%) as input.

Tool-specific parameters were set as follows: DAGchainer was run with maximum inter-match distance of 500 kb (–D 500000), maximum gap between anchors of 5 kb (–g 5000), and minimum 5 collinear anchors per block (–A 5). MCScanX was run with minimum syntenic block size of 5 genes (–s 5) and maximum 25 intervening genes between collinear pairs

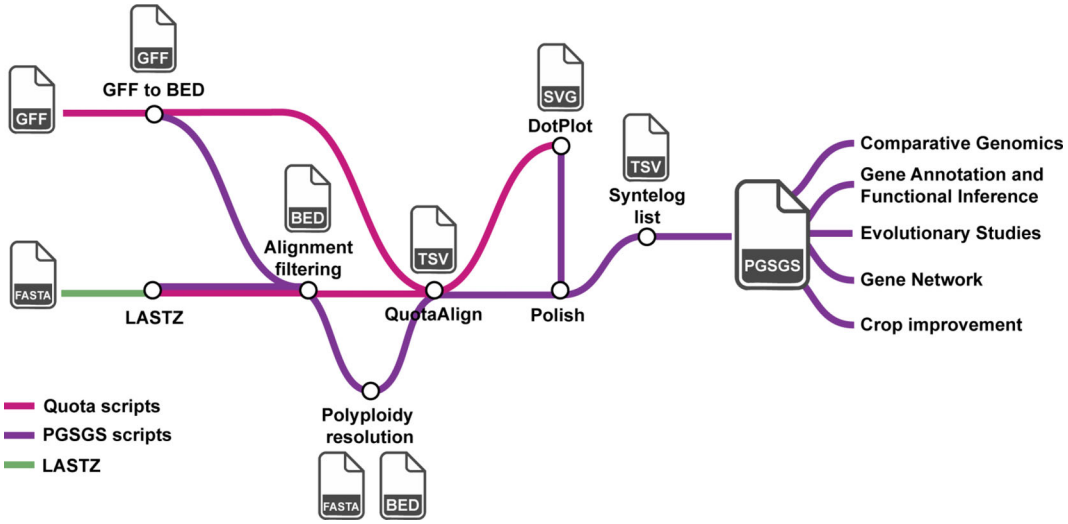


Figure 1. Workflow for comparative genomics and functional analysis of PGSGS. The PGSGS analysis consists of seven major steps, some of which are further divided into sub-components. For each input file, a quality control is performed. This step includes both the application of filters and trimmers. (i) Primary transcript model annotation and coding sequence extraction. Input: .GFF and .fasta files containing the primary transcript model annotations and coding sequences. (ii) Conversion and alignment. Convert GFF to BED: The .GFF annotation file is converted to a .BED format. LASTZ: Align coding sequences using S. bicolor as the reference genome, producing alignment outputs. (iii) Alignment filtering and optimization. Filter alignments to refine results. Code optimization: Improve the code for analyzing polyploid genomes. Generate alignment and annotation outputs. (iv) QuotaAlign and paralog screening. QuotaAlign: Screen and filter for paralogous gene blocks. Polishment data: Further refine paralogous blocks to enhance data quality. (v) Visualization and analysis. Dot plot: Visualize syntenic blocks (collinearity graph). (vi) Orthologs list merging and PGSGS building. Merge ortholog lists to form the PGSGS using S. bicolor as the base. (vii) EggNOG enrichment. Integrate with the EggNOG database for additional functional annotations. (viii) Downstream applications. Comparative genomics: Compare genomic structures across species.

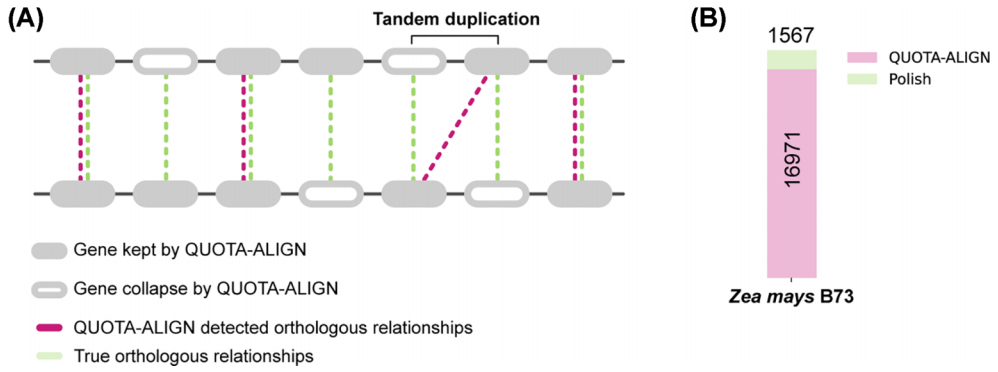


Figure 2. Schematic representation of a sequence homology-based approach to identify syntenic duplicated genes. Pink dashed lines depict the identification of orthologs through pairwise comparison between Genomes A and B using the QUOTA-ALIGN method [31]. Green dashed lines illustrate the refinement of paralog identification arising from duplications within the same genomic environment, employing a sequence homology-based approach.

(default). QuotaAlign was run with quota ratio 1:2 to reflect maize's paleotetraploidy and merge distance –Dm 20, followed by the polish refinement step.

A curated gold-standard set of 382 Sorghum genes was compiled for validation. These genes represent diverse functional categories, including defense (resistance) genes, hormone regulation, transcription factors, sugar transporters, and vitamin metabolism (334 predicted pairs between sorghum and maize; Fig. 3). Gene families were selected to ensure genome-wide coverage across sorghum chromosomes 1–10. Orthology assignments between *Sorghum bicolor* and *Z. mays* were validated using the PLAZA v5 orthology database [23].

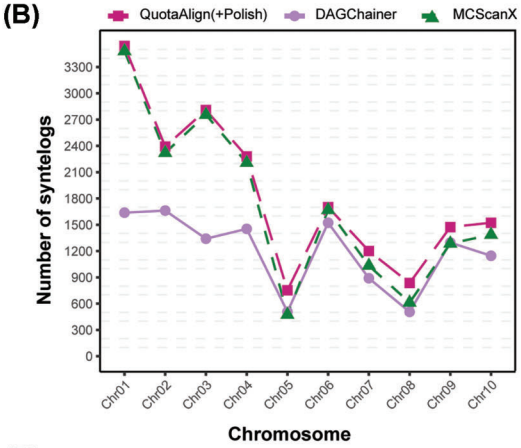
For each tool, we calculated: Recall (sensitivity) = TP / (TP + FN); Precision = TP / (TP + FP); F1 score = 2 × (Recall × Precision) / (Recall + Precision), and Accuracy as (TP + TN) / (TP + TN + FP + FN), where TP = true positives, FP = false positives, FN = false negatives. Complete benchmarking results, including counts of TP, FP, FN, and true negatives (TN) for all three tools, are reported in Supplementary Table S2.

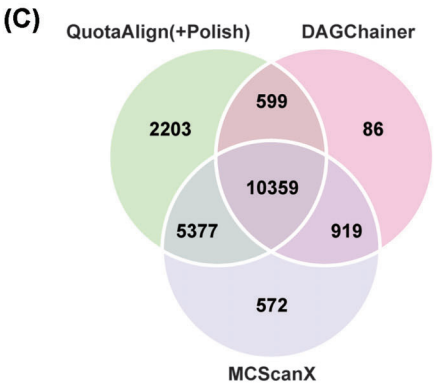
Functional enrichment analyses

Functional annotation of all 27 567 sorghum syntelogs was performed using eggNOG-mapper v2 [34] against the Poaceae taxonomic scope, yielding Gene Ontology (GO) terms, KEGG pathways, and protein family (Pfam) assignments. GO enrichment analysis was conducted using the clusterProfiler R package (v4.0), following the workflow described by Bonnot *et al.* [35], with the complete set of 27 567 syntelogs serving as the background.

(A)

Tool	Max distance	Gene gaps	Min block size	Notes
QuotaAlign (+Polish)	-Dm 20 (max genes)	N/A	-Q 5	Dm controls merging distance; block size is set separately with -Q
DAGChainer	-D 500000 (bp)	-g 5000 (bp)	-A 5	Anchors ≥5; simple bp-based chaining
MCScanX	gene-order based	25	-s 5	Requires ≥5 collinear pairs; less stringent by default





(D)

	Genome coverage (%)	Total syntelogs	Pairs predict./ PLAZA	TP	FP	FN	TN	Precision (%)	Recall (%)	F1	Accuracy (%)
QuotaAlign (+Polish)	57.7	18,559	286/334	213	73	121	137	74	64	0.69	64
DAGChainer	50.1	16,122	62/334	46	16	288	149	74	14	0.23	41
MCScanX	53.6	17,227	136/334	96	40	238	167	71	29	0.41	47

Figure 3. Parameterization and performance of synteny detection tools. (**A**) Detailed parameter settings for QuotaAlign(+Polish) (–Dm 20 genes; –Q 5), DAGChainer (–D 500 kb; –g 5 kb; –A 5 anchors) and MCScanX (–s 5 collinear pairs), with notes on distance thresholds, anchor requirements and chaining logic; (**B**) Venn diagram illustrating overlap of syntelogs called by QuotaAlign(+Polish) ($n = 18\,559$), DAGChainer ($n = 6\,122$), and MCScanX ($n = 17\,227$), with a shared core of 10 359 gene pairs; (**C**) Chromosomal distribution of syntelog counts across chromosomes 1–10 for each algorithm, revealing concordance in high-density regions and divergence at Chr05; (**D**) Performance metrics against the SWEET gene family reference: genome coverage (%), total syntelogs, true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), precision (%), recall (%) and FDR.

To identify functional signatures associated with different levels of evolutionary conservation, we separately tested four gene categories: (i) PGSGS-core: syntelogs present in ≥15 of 17 grass species ($n = 11\,624$), (ii) PACMAD-core: syntelogs present in ≥9 of 11 PACMAD clade species (*Sorghum bicolor*, *Z. mays* B73, *Z. mays* Mo17, *Saccharum* hybrid, *M. sinensis*, *Setaria italica*, *Setaria viridis*, *Panicum virgatum*, *Panicum hallii*, *Urochloa fusca*, *Paspalum vaginatum*; $n = 11\,955$), (iii) PACMAD-specific: syntelogs in PACMAD-core but absent from PGSGS-core ($n = 2266$). This set represents genes under purifying selection within PACMAD grasses but lost or highly divergent in BOP lineages, potentially reflecting clade-specific functional constraints, (iv) Noncore: syntelogs present in <15 species ($n = 15\,943$).

For each category, enrichment of GO Biological Process terms was assessed against the full syntelog background using Fisher's exact test. *P*-values were corrected for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) method. GO terms were considered significantly enriched at FDR < 0.05 and fold-enrichment >1.5. To reduce redundancy in visualization, significantly enriched GO terms were summarized using REVIGO [36] (semantic similarity threshold 0.5), retaining representative terms with the strongest enrichment signals. Graphical representations was generated using clusterProfiler 4.0 [37] and ChIPseeker [38].

Intraespecific diversity and selection analysis

We investigated intraspecific nucleotide diversity and selective constraints across *Sorghum bicolor* genes using publicly available whole-genome resequencing data from the *Sorghum* association panel (PRJEB51985) [39]. Raw variant calls were retrieved as the joint-genotyped dataset aligned to the *Sorghum bicolor* reference genome assembly v3.1 (GCF_000003195.3). Variant quality control retained only biallelic SNPs with per-sample genotype quality (CG) ≥30, missing data ≤20%, and minor allele frequency ≥0.01, yielding a high-confidence SNP set for population-level diversity analysis.

Gene coordinates from the NCBI v3 annotation were used to extract per-gene variant subsets, from which standard population genetic statistics were computed. For each gene, nucleotide diversity (π), Watterson's theta (θW), and Tajima's D were estimated using bcftools (v1.17) and custom Python scripts implementing sliding-window algorithms in accordance with the formulas of Nei and Tajima [40].

To relate genetic diversity to gene evolutionary status, per-gene metrics were intersected with the curated *Sorghum bicolor* v5 gene models from Phytozome (v5.1), which define core and noncore syntelog sets based on PGSGS conservation. Because assemblies v3 and v5 differ structurally, gene correspondence was established through reciprocal best-match mapping between the v3 and v5 predicted proteomes (BLASTP; identity ≥90%, coverage ≥90%). One-to-one matches were retained to ensure consistent gene-level comparisons. Diversity statistics for each v3 gene were then transferred to its corresponding v5 ortholog to allow direct integration with the pan-grass syntenic framework (PGSGS-core).

Finally, distributions of π, θW, and Tajima's *D* were summarized separately for PGSGS-core and noncore gene sets. Comparative analyses tested whether core genes exhibit reduced intraspecific polymorphism and signatures of purifying selection relative to noncore genes, consistent with stronger evolutionary constraint in functionally conserved regions.

Data and code availability

The step-by-step synteny pipeline is fully documented (scripts, helper modules, and parameter files) at FigShare (DOI: 10.6084/m9.figshare.26793190) and Supplementary Files S1 and S2.

Results and discussion

A sequence-homology-based approach to establishing a robust syntenic relationship of duplicated genes

Strategies to identify orthologous genes between species and homeologous regions in a given genome are valuable to study changes in evolutionary rate, functional specialization, pseudogenization, and other factors that can disrupt or cause misleading results in purely tree-based methods. Several tools are available for identifying syntenic regions within or across genomes, given the set of known relationships among genes [21, 29, 41]. However, in many cases, these tools can merge syntenic regions. PGSGS is developed with an initial set of 746 743 protein coding genes from the 17 genomes (Supplementary Table S2). We adopted a three-stage pipeline, employing LASTZ for the identification of similar genes [30] anchoring to the 32 160 protein coding genes from sorghum, QuotaAlign is used for the identification of syntenic regions and the specific identification of syntelogs [29], and a final polishing step to improve orthology assignment, particularly for genes with multiple local duplicate copies within the syntenic regions initially identified by QuotaAlign [28] (Fig. 2). This is particularly crucial for retrieving regions where tandemly duplicated arrays of genes or proximal paralogs have undergone independent evolutionary trajectories over millions of years [42].

The effect of the polishing process is illustrated in Fig. 2. This polishing step is particularly crucial for retrieving regions where tandemly duplicated genes arrays or proximal paralogs have undergone independent evolutionary trajectories over millions of years. To expedite genome-wide comparisons, QuotaAlign(+Polish) consolidates sets of homologous genes within the same genomic intervals and considers only one gene per group when defining genome-wide synteny relationships (Fig. 2A). In some cases, these previously collapsed genes are recent tandem duplications, in other cases, they may represent genes that have subfunctionalized and evolved independently over time. NBS-LRR resistance gene homologs are one such example of tandem duplications and gene expasion followed by neofunctionalization [43].

In the polishing step, the global synteny information from QuotaAlign is used to identify the predicted region in which syntenic orthologs should be present. The raw alignment scores for all possible pairwise gene comparisons are then used to identify pairs of the most similar genes within these syntenic intervals (Fig. 2A). The result is an increase in the number and accuracy of syntenic orthologous gene assignments by ∼10% for the maize B73 genome when using sorghum as a reference (Fig. 2B). This polishing step increased the number and accuracy of syntenic gene assignments in all studied genomes (Supplementary Fig. S1).

Performance evaluation of QuotaAlign for syntelog identification

QuotaAlign(+Polish) was benchmarked against two widely used synteny tools (DAGChainer and MCScanX) on a controlled sorghum–maize comparison. To ensure consistent preprocessing across all three tools, we used identical LASTZ alignments with coverage 50% and identity 70%. DAGChainer v1.2 [32] was configured with a maximum intermatch distance of 500 kb (−D 500000), maximum gap of 5 kb (−g 5000), and minimum of 5 collinear anchors per block (−A 5). MCScanX v1.1 [33] enforced a minimum syntenic block size of 5 genes (−s 5) and allowed up to 25 intervening genes between collinear pairs. QuotaAlign was run with quotas of 1:2 to reflect the two subgenomes of maize, using distance merging parameter −Dm 20 genes and minimum block size −Q 5 (Fig. 3A).

For validation, we compiled a curated gold-standard set of 382 sorghum genes representing diverse functional categories: defense (resistance) genes, hormone regulation, transcription factors, sugar transporters, and vitamin metabolism (an universe of 334 predicted pairs between sorghum and maize; Supplementary Table S3). Orthology assignments between *Sorghum bicolor* and *Z. mays* were validated using the PLAZA v5 orthology database as the reference standard [23]. Each predicted gene pair was classified as TP, FP, FN, or TN. From these counts we derived precision, recall, F1 score (harmonic mean of precision and recall), and overall accuracy.

The chromosomal distribution of syntelogs revealed concordance among methods in high-density regions, with notable divergence at chromossomes 1–4 where QuotaAlign(+Polish) and MCScanX recovered substantially more syntelogs (Fig. 3B and C). The benchmarking revealed substantial differences in tool performance (Fig. 3D). DAGChainer identified only 62 maize orthologs for the 382 sorghum genes, achieving a precision of 74.2% but lowest recall (13.8%), with an F1 score of 0.232 and accuracy of 41.2%. MCScanX recovered 136 maize orthologs with 70.6% precision and improved recall (28.7%), yielding an F1 score of 0.409 and 46.8% accuracy. QuotaAlign(+Polish) identified 286 unique maize orthologs with 213 true positives, achieving 74.5% precision, 63.8% recall, F1 score of 0.687, and 64.3% accuracy, substantially outperforming both alternative approaches (Fig. 3B–D).

Our findings highlight how evolutionary context, algorithmic design, and parameter choices jointly shape the sen-

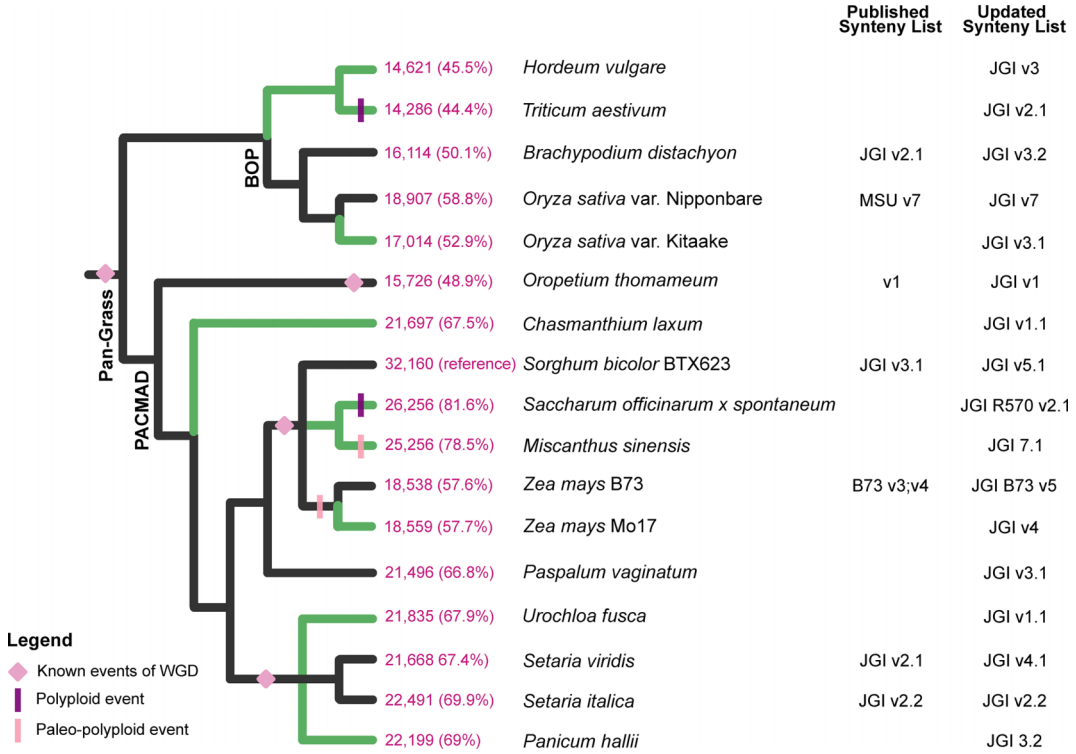


Figure 4. Phylogenetic relationships among genome releases of sequenced grass species and syntelog gene counts within the PGSGS. The PGSGS, with a Sorghum-based foundation, is presented, showcasing syntelog counts in pairwise comparisons. Absolute and relative syntelog numbers are depicted in pink, while green denotes genome versions added in the PGSGS update. The earliest genome versions were updated from the initial PGSGS version (https://doi.org/10.6084/m9.figshare.7926674.v1). Pink diamonds symbolize WGD events referenced by Zhang *et al.* [28], and polyploidy and paleo-polyploidy events are depicted by dashes.

sitivity and specificity of collinearity inference in grasses. The divergent evolutionary trajectories of *Sorghum bicolor* and *Z. mays* underpin many observed differences. Following their ∼12 Ma split, the maize lineage experienced a paleo-tetraploid event ∼5–12 Ma, yielding two subgenomes (maize1 and maize2) that have been fractionated asymmetrically [26, 44, 45]. *Sorghum* chromosomes 1–4 exhibit pronounced asymmetry in syntelog retention between maize1 and maize2, with maize1 generally retaining more syntenic genes, consistent with documented subgenome dominance [26, 44, 45], resulting in uneven gene retention. Methods that do not explicitly model expected syntenic depth (i.e. 1:2 for sorghum:maize) are prone to fragmenting these biased regions or under-detecting true homeologs. DAGChainer's parameters (maximum inter-match distance of 500 kb and minimum 5 anchors per block) constrain detection to closely spaced, contiguous gene arrays. MCScanX allows larger gaps (up to 25 genes) and incorporates a tandem collapse step. QuotaAlign's quota parameter (1:2) explicitly models the expected number of syntenic regions per reference interval, allowing recovery of syntelogs distributed across both maize subgenomes.

The observed recall differences correlate with these algorithmic approaches. DAGChainer recovered 13.8% of reference orthologs, MCScanX recovered 28.7%, and QuotaAlign recovered 63.8%. Precision remained relatively consistent across tools: 74.2% (DAGChainer), 70.6% (MCScanX), and 74.5% (QuotaAlign(+Polish)). The F1 score, which balances precision and recall, was highest for QuotaAlign(+Polish) (0.687) compared to MCScanX (0.409) and DAGChainer (0.232).

Tandem gene duplications represent a methodological challenge, as local gene family expansions, particularly resistance and stress response loci, are common in grasses [43, 46]. DAGChainer does not incorporate specific tandem filtering by default, MCScanX includes a tandem collapse step, and QuotaAlign treats tandem arrays as valid anchor sets within its quota framework. The benchmark dataset included genes from families known to undergo tandem duplication, allowing assessment of tool performance on these complex regions. Structural rearrangements including inversions, translocations, and deletions complicate synteny detection [11]. The sorghum–maize comparison includes documented inversions that can fragment syntenic blocks depending on algorithmic sensitivity to gene orientation. QuotaAlign performs optimization across both DNA strands [29], which may contribute to its broader recovery of syntelogs in rearranged regions.

The pan-grass syntenic gene set spans 17 genomes and 344 230 syntelogs

We prioritized assemblies with high contiguity (N50/L50) and annotation completeness (BUSCO > 93% for Embryophyta), while ensuring broad phylogenetic coverage across the PACMAD and BOP clades and agricultural relevance. The final dataset represents 17 genomes from 15 species (Supplementary Tables S1 and S2), including updates to genome versions from previous pan-grass synteny resources

[28] (Fig. 4). To ensure the pipeline handles extreme polyploid complexity, we included polyploid genomes such as *Saccharum* hybrid R570 (sugarcane), *T. aestivum* (wheat), and *M. sinensis*, testing the quota-based chaining and polishing steps under extensive gene-dosage variation and demonstrating recovery of individual homeologs rather than collapsing tandem or subgenome-derived duplicates.

QuotaAlign(+Polish) was applied to a broad phylogenetic set of grass representatives from the BOP clade, including two rice genomes alongside wheat and barley, while the PACMAD clade comprises both wild and domesticated grasses. To build this dataset, we anchored all pairwise comparisons to the *Sorghum bicolor* BTx623 v5.1 genome. *Sorghum* was selected because it (i) preserves the ancestral 10-chromosome arrangement of the Andropogoneae [26, 47], (ii) has diverged only modestly from maize since their approximately 12 Ma split [44], (iii) lacks the extensive post-tetraploid chromosome fusions that complicate maize's genome [26, 47], and (iv) is closely related to *Saccharum officinarum* × *spontaneum*, a recent hybrid polyploid (*Saccharum* hybrid), and *M. sinensis*, a paleotetraploid.

Out of 32 160 primary-transcript models in *Sorghum bicolor* v5.1, our pipeline linked 27 567 (∼86%) to at least one syntenic ortholog in other grass species (Supplementary Table S3). The PGSGS framework analyzed a total of 746 743 gene models from all 17 genomes and identified 344 230 syntelogs (46%). Each of the 27 567 sorghum genes and their cross-species counterparts was assigned a unique "PGSGS_ID" identifier. Within the PACMAD clade, *Miscanthus* and *Saccharum* retained ∼25 000 *Sorghum* syntelogs, while other PACMAD relatives had 18 000–22 000 syntelogs; outside Panicoideae, 14 000–18 000 syntelogs were conserved (Fig. 4). The decline in syntenic retention with increasing phylogenetic distance indicates that a single anchor genome can capture most conserved orthologs across grasses, though alternative anchors (e.g. *Oryza sativa* for BOP species) may recover additional lineage-specific sets [48].

Following the pangenome concept applied to intraspecific variation, where plant genomes possess both core genes shared among all individuals and accessory genes present in only some individuals [40], we applied a similar framework to classify syntelogs across grass species. We computed the 27 567 sorghum-anchored syntelogs into core and noncore gene sets across the 17 grass genomes. The core set was defined as syntelogs found in at least 15 of 17 genomes, yielding 11 624 PGSGS-core genes (Fig. 5A).

The size of the combined core and noncore gene sets among these grass species continued to expand with the inclusion of additional genomes, indicating its open nature; however, based on the sorghum gene content, a plateau has been reached. The size of the PGSGS-core rapidly declined as more genomes were added, particularly when considering the phylogenetic distance relative to sorghum (Fig. 5B). In addition to the overall core gene set, we computed a clade-specific core set for the PACMAD clade, which is more closely related to sorghum. The PACMAD-core was defined as syntelogs present in at least 9 of 11 PACMAD species, totaling 11 954 syntelogs. The overlap between these two core sets reveals 9689 syntelogs shared between PGSGS-core and PACMAD-core, while 2266 syntelogs are unique to PACMAD-core (present across PACMAD species but not broadly conserved in BOP lineages) (Fig. 5A). This stratification into universal core, clade-specific core, and variable noncore sets provides a framework for investigating functional constraints and adaptive evolution across grass lineages.

Polyploid genome dynamics: subgenome resolution and syntelog distribution

Polyploidy, or WGD, is common in plants [49] and reshapes genome organization and gene content, with significant impacts on evolutionary trajectories and adaptive characteristics [50, 51]. Following polyploidization, duplicated genes often undergo differential retention between subgenomes, a process termed subgenome fractionation, which can lead to subgenome dominance where one subgenome retains more genes than its homeolog [52, 53]. This biased fractionation has been well documented in paleopolyploids such as maize [26, 28] and in more recent allopolyploids including wheat [54], strawberry [55], and *Brassica* sp. [56, 57]. Understanding the mechanisms and patterns governing subgenome evolution and fractionation remains a fundamental challenge in plant genomics, with important implications for crop breeding and trait improvement [39, 58].

Completion of the PGSGS enabled detailed comparison of the sorghum genome with multiple mesopolyploid and neopolyploid grass species, including *Z. mays* (maize paleotetraploidy, ∼12 Mya) [45], *M. sinensis* (recent allotetraploid) [59, 60], *Saccharum officinarum* × *spontaneum* (sugarcane, complex aneuploid, $2n = \sim 100–130$) [31,61,62], and *T. aestivum* (hexaploid wheat, ∼8500 years old) [63, 64]. For recent allopolyploids such as wheat, the subgenomes originated from distinct diploid progenitors, allowing straightforward assignment of each chromosome to its respective subgenome (A, B, or D) based on sequence similarity to known ancestral species [63]. In contrast, for more ancient autopolyploid or paleopolyploid events such as the maize WGD, subgenome identity must be inferred from differential evolutionary rates and fractionation patterns of syntenic genomic segments [26] (Fig. 6A).

The syntelog complement of the paleotetraploid *M. sinensis* genome exhibits remarkable collinearity with sorghum (Fig. 6B), characterized by extensive preservation of gene order with minimal inversions and rearrangements. This high degree of macrosynteny reflects both the relatively recent divergence of *Miscanthus* and *Sorghum* within the Andropogoneae (∼7 Mya) [17, 47, 65] and the preservation of ancestral chromosomal architecture. Critically, the two *Miscanthus* subgenomes (MisA and MisB) demonstrate balanced retention of syntelogs with extensive 2:1 conserved collinear synteny relative to *Sorghum* [53, 54], meaning that for most sorghum genomic regions, corresponding syntenic blocks can be identified on both *Miscanthus* subgenomes in approximately equal proportions. This balanced architecture is consistent with the reported recent allotetraploid origin of *M. sinensis* [53, 54], and suggests limited subgenome fractionation bias, possibly due to insufficient evolutionary time for substantial differential gene loss or functional subfunctionalization to occur [51, 66]. The symmetric retention pattern contrasts with the biased fractionation observed in older polyploid systems and may reflect the genomic consequences of allopolyploidy, where subgenomes from divergent diploid ancestors may experience different selective pressures (Fig. S2; [67]).

The *Saccharum* complex represents one of the most challenging polyploid genome architectures in the grass fam-

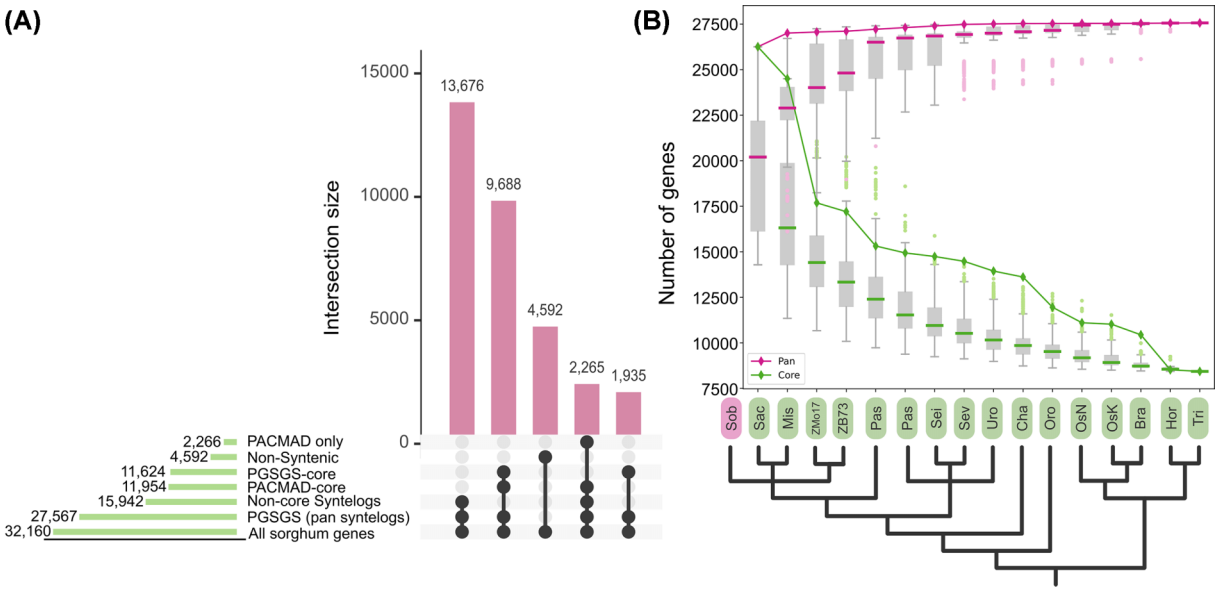


Figure 5. PGSGS overview. (**A**) The Venn diagram delineates the PGSGS, PGSGS-core, PACMAD-core, and noncore syntelogs. (**B**) The PGSGS size increases for each included genome, with gray hues representing the 25th and 75th percentiles, while pink and green lines depict pan- and core-syntelogs median, respectively. Pink and green diamonds correspond to the exact number of syntelogs added by species following the phylogenetic tree.

ily (Fig. 6C). Modern sugarcane cultivars, including the *S. officinarum* × *S. spontaneum* hybrid R570 analyzed here, possess highly polyaneuploid genomes with chromosome numbers typically ranging from 100 to 120 ($2n$ = ~100–120), comprising 70%–80% *S. officinarum*-derived chromosomes, 10%–20% *S. spontaneum*-derived chromosomes, and 10%–20% recombinant chromosomes [31, 61, 68]. This genomic complexity arises from multiple rounds of polyploidization in both parental lineages, followed by interspecific hybridization and ongoing aneuploidy [69, 70]. Despite this extraordinary complexity, syntenic relationships with sorghum remain detectable, though substantially fragmented compared to miscanthus or wheat [61, 71]. The sorghum–sugarcane synteny analysis reveals extensive chromosomal rearrangements, including translocations, inversions, and variable copy number across homeologous chromosome groups, reflecting both the ancient divergence of these lineages within Andropogoneae (~8–9 Mya for *Saccharum* from the *Sorghum* lineage) [62, 72] and the subsequent complex polyploidization history unique to Saccharum [62, 73].

Notably, the identification of syntelog groups in *Saccharum* is complicated by the presence of multiple homeologous copies (often 8–12 copies per gene) derived from different ancestral genomes and hybridization events [61, 74]. However, the maintenance of recognizable syntenic blocks between sorghum and the mosaic *Saccharum* genome demonstrates the deep conservation of grass chromosomal architecture and validates the utility of the sorghum reference for disentangling even highly complex polyploid genomes (Supplementary Fig. S2). The differential retention patterns observed across *Saccharum* homeolog groups suggest potential subgenome-specific selection pressures related to sugar metabolism and biomass production traits that have been targets of crop improvement [62, 74–76], though detailed characterization of subgenome dominance in this system requires further investigation.

In contrast to both the balanced subgenome architecture observed in *Miscanthus* and the extreme complexity of *Saccharum*, comparative analysis with hexaploid wheat reveals an intermediate level of complexity with well-defined subgenome structure [77]. The sorghum–wheat collinearity shows lower conservation compared to the sorghum-miscanthus comparison (Fig. 6C), reflecting the greater evolutionary distance between these lineages, with sorghum and wheat diverging approximately 50–60 Mya [17], and the subsequent complex polyploidization history of wheat involving two sequential allopolyploidization events [63, 64]. Despite this reduced overall synteny, the hexaploid nature of the wheat genome is clearly evident, with three distinct homeologous chromosome sets (A, B, and D subgenomes) each maintaining discernible syntenic correspondence with sorghum chromosomes in roughly equivalent proportions [19, 78].

This tripartite genome structure reflects wheat's allohexaploid origin from three diploid progenitors: *Triticum urartu* (A genome donor), an unknown species related to *Aegilops speltoides* (B genome donor), and *Aegilops tauschii* (D genome donor) [64, 77]. The maintenance of syntenic relationships across all three subgenomes, despite extensive chromosomal rearrangements and the presence of subgenome-specific transposable element expansions [78], underscores sorghum's utility as a reference genome for comparative grass genomics and demonstrates the value of the PGSGS for resolving complex polyploid genome structures across a spectrum of evolutionary timescales and polyploidization modes.

Syntenic conservation predicts selective constraint in *Sorghum* populations

To validate whether syntenic conservation across grass genomes correlates with functional constraint within species, we analyzed nucleotide diversity patterns in 400 sorghum accessions [39], comparing PGSGS-core syntelogs, noncore syntelogs, and nonnsyntenic genes. The relationship between

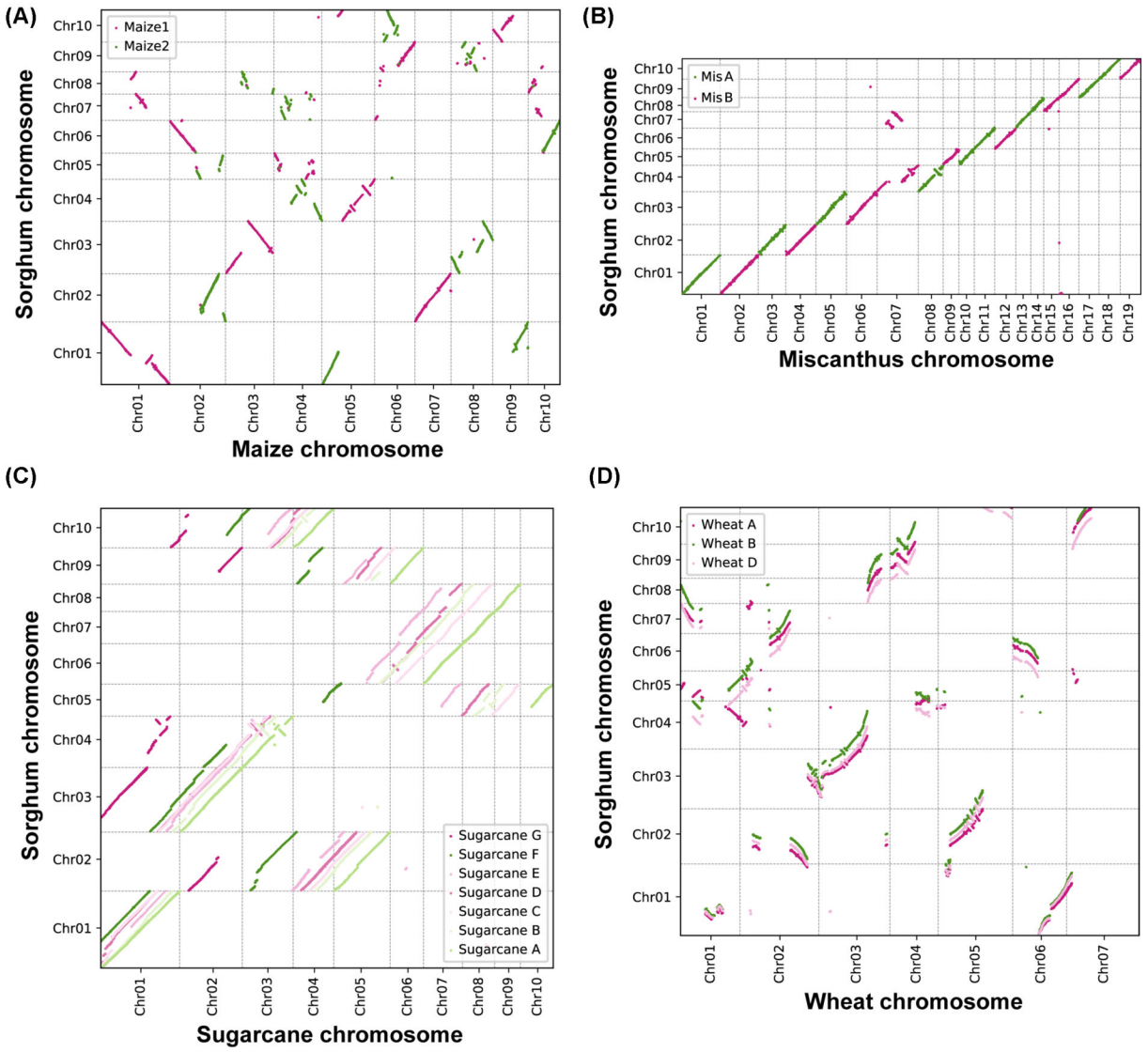


Figure 6. Dot plot illustrating syntelog clusters between Sorghum bicolor and paleo- and polyploid grass genomes. (**A**) Genome collinearity analysis between *S. bicolor* and *Z. mays*, indicating maize 1 and maize 2 subgenomes. (**B**) Genome collinearity analysis between *S. bicolor* and *M. sinensis*, highlighting MisA and MisB subgenomes. (**C**) Genome collinearity analysis between *S. bicolor* and *S. spontaneum* × *officinarum* var. R570, showcasing sugarcane homologous chromosomes A, B, C, D, E, F, and G. Different-colored dots represent subgenomes or homologous chromosomes. (**D**) Genome collinearity analysis between *S. bicolor* and *T. aestivum*, illustrating wheat homologous chromosomes A, B, and D.

cross-species conservation and within-species diversity has been documented in arabidopsis and maize [79, 80], the relationship between syntenic conservation across deep evolutionary time and modern selective constraint remained unexplored at the pan-grass scale. Synteny-based orthology inference provides stronger evidence for functional equivalence than sequence similarity alone, making PGSGS well-suited for linking evolutionary conservation to population genetic signatures.

PGSGS-core syntelogs exhibited significantly reduced nucleotide diversity (median $\pi = 0.0014$) compared to noncore syntelogs ($\pi = 0.0019$), and nonsyntenic genes ($\pi = 0.0022$; Kruskal–Wallis $P < 3.48 \times 10^{-85}$; Supplementary Fig. S3 and Supplementary Table S4), with concordant patterns for Watterson's theta ($\theta W = 0.0012$, 0.0016, and 0.0020, respectively; Supplementary Table S4). This gradient remained significant when controlling for gene length through SNP density normalization (Supplementary Table S4 and Supplementary Fig. S3), demonstrating that syntenic conservation over >70 million years of grass evolution [17] predicts reduced variation. The magnitude of diversity reduction (36.4% for π and 40% for θW comparing core to nonsyntenic genes) indicates purifying selection on PGSGS-core genes. Similar patterns of reduced diversity in evolutionarily conserved genes have been documented in maize [81], consistent with purifying selection acting preferentially on functionally important loci.

Tajima's D distributions revealed that core genes presented elevated positive values (median $D = 0.69$) with only 32.9% exhibiting negative D. In contrast, nonsyntenic genes displayed values closer to neutrality (median $D = 0.39$) with 37.3% showing $D < 0$ (Fig. 7A and Supplementary Table S4). The genome-wide tendency toward positive D across all three gene classes is consistent with demographic processes, particularly the domestication bottleneck documented in sorghum [82–85]. PGSGS-core syntelogs show the most positive D values and the lowest proportion of negative D (32.9%), com-

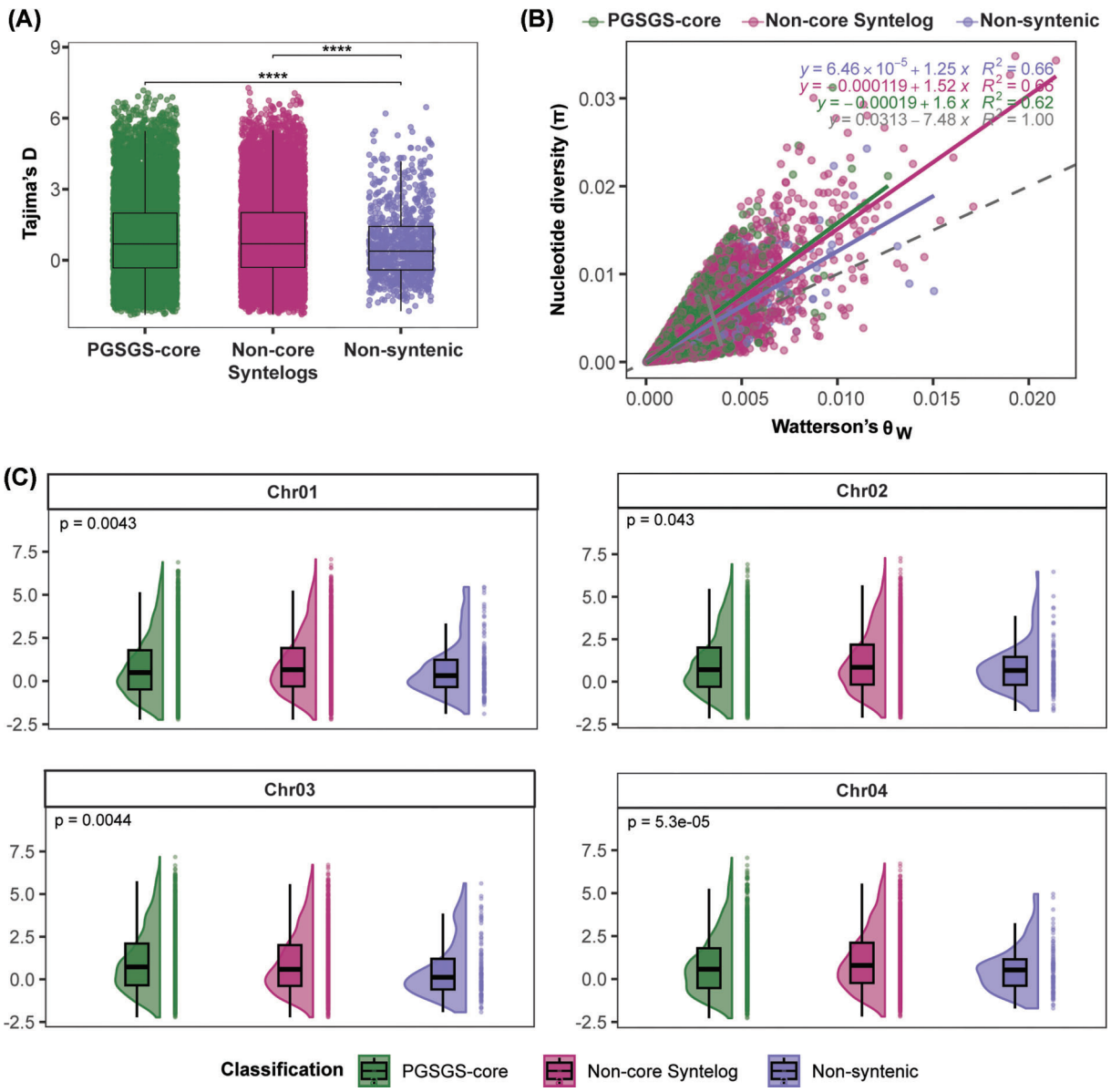


Figure 7. Population genetic signatures across PGSGS categories in sorghum. (**A**) Tajima's *D* distributions across PGSGS categories. Differences among gene categories were tested with a Kruskal–Wallis test followed by Dunn's post-hoc comparisons: PGSGS-core syntelogs versus noncore syntelog ($p = 0.333$); PGSGS-core versus nonsyntenic ($p = 2.16 \times 10^{-5}$); noncore syntelogs versus nonsyntenic ($p = 5.64 \times 10^{-6}$). (**B**) Relationship between nucleotide diversity (π) and Watterson's theta (θW). Each point represents a gene, colored by category. Solid lines: linear regressions for each category; dashed black line: neutral expectation ($\pi = \theta$W). Linear regression equations and R^2 values shown for each category. (**C**) C hromosome-level Tajima's *D* distributions for chromosomes 1–4 across PGSGS-core syntelogs, noncore syntelogs, and nonsyntenic genes. Global differences among categories were tested with Kruskal–Wallis tests followed by Dunn's post-hoc comparisons. Chromosoomes 1–4 are shown representative examples exhibiting the most robust statistical contrasts among the 10 sorghum chromosomes (Kruskal–Wallis $P \leq 0.05$ with significant Dunn's post-hoc pairwise comparisons for all three gene class pairs). Distributions for all 10 chromosomes are provided in Supplementary Figs S4–S7).

pared with noncore syntelogs and nonsyntenic genes (37.3% with $D < 0$). This pattern is consistent with stronger purifying selection on core syntelogs further limiting the accumulation of low-frequency deleterious variants beyond the demographic background. This interpretation is concordant with the π and θW gradients (core < noncore < nonsyntenic), which independently demonstrate reduced diversity in PGSGS-core genes (36.4% for π and 40% for θW comparing core to nonsyntenic genes). The relationship between diversity (π) and θW further supports this interpretation. Genes within synthenic categories (PGSGS-core and noncore syntelogs) have nearly overlappping π–θW regressions (slopes = 1.60 and 1.52; $R^2 = 0.62$ and 0.66, respectively), indicating a consistent departure from neutral expectation shaed by the combination of demographic history and selection constraint. Nonsyntenic genes show a slightly flatter regression (slope = 1.25), lying closer to the neutral line and displaying broader scatter ($R^2 = 0.66$), consistent with heterogeneous evolutionary trajectories among these loci, where relaxed constraint permits greater drift (Fig. 7C).

Patterns of genetic diversity across all 10 sorghum chromosomes (Supplementary Figs S4–S7) support that nonsyntenic genes are less subject to purifying selection than PGSGS core and noncore syntelogs. The four chromosomes (Chr01-

04) were selected because they represent those with the most robust statistical contrasts within the sorghum chromosome set, with consistently significant differences among the three gene classes (Kruskall–Wallis $P \leq 0.05$; Fig. 7C; see Supplementary Figs S4–S7 for all chromosome set). Across these chromosomes, Tajima's D ranges from approximately −2.5 to 7.5, and although all classes tend toward positive values; nonetheless, PGSGS-core and noncore syntelogs closely track each other in central tendency and spread on every chromosome, whereas nonsyntenic genes display a more heterogeneous distribution. Futher integration of breeding markers can contribute to translational valuable marker genes as performed by Boatwright *et al.* [39, 86].

These findings provide insights into grass genome evolution, which the concordance between syntenic conservation and reduced nucleotide diversity indicates that PGSGS-core genes experience sustained purifying selection, consistent with functional constraint. This pattern aligns with observations that highly connected genes in protein interaction networks evolve more slowly [87] and that genes with pleiotropic effects, affecting multiple traits, experience stronger selective constraint [88], suggesting that PGSGS-core genes likely represent functionally central loci within grass metabolic and regulatory networks. The validation of syntenic conservation as a predictor of functional constraint addresses a fundamental challenge in comparative genomics: distinguishing true orthologs with conserved function from paralogs that may have diverged [89]. By demonstrating that positional conservation across >70 million years of grass evolution correlates with selective pressure, PGSGS provides a framework for transferring functional insights across grass species.

Functional stratification distinguishes conserved core from variable syntelogs and nonsyntenic genes

Having established that PGSGS-core syntelogs experience sustained purifying selection in sorghum populations, we characterized the functional basis underlying this constraint through GO enrichment analysis across syntenic categories: PGSGS-core, PACMAD-only (unique syntelogs to PACMAD), and noncore syntelogs; and nonsyntenic genes. The 15-species threshold balances stringency with biological reality, tolerating assembly gaps while capturing genes conserved across both major grass clades in the PGSGS set. While PGSGS-core demonstrated a higher annotation coverage (~80% with UniProt), noncore syntelogs showed ~70% of genes annotated, and nonsyntenic genes only ~15% (Fig. 8A), reflecting experimental tractability and research prioritization toward functionally critical genes [90].

GO enrichment anlaysis, restricted to terms with FDR < 0.05, revealed distinct functional profiles across gene categories (Fig. 8B and Supplementary Table S5). This mirrors previous pan-genome analyses in Arabidopsis, where genes present across all 1135 accessions show fewer mutations in essencial genes [91], while previous population-level analyses in maize, where genes of the dominant subgenome (maize1) [92]. Notably, specific propoerties of enrichment analysis, once PGSGS-core comprises 11 624 genes (~42% of the syntelog background), categories in which core genes are prevalent show diluted fold enrichment values relative to smaller gene sets [93]. The distinguinshing feature of PGSGS-core genes is not any single enriched category, but rather the network density and interconnectedness of their enriched GO terms.

A network topology reinforced this interpretation (Fig. 8C). PGSGS-core syntelogs formed densely (more genes) interconnected modules centered on translation, cell cycle, and chromatin organization, with ~3-fold higher network density than noncore genes and ~20-fold higher than nonsyntenic gens. This clustering aligns with observations that hub genes evolve under stronger constraint due to pleiotropic consequences [94–97]. PGSGS-core genes are part of integrated regulatory architectures whose disruption produces cascading effects across cellular networks, explaining both their evolutionary conservation and selective constraint.

For the 2266 syntelogs unique to PACMAD-core, no GO biological process terms reached statistical significance (FDR < 0.05), precluding robust functional interpretation. This lack of significant enrichment may reflect the limited size of this gene set ($n = 2266$ compared with $n = 11\,624$ for PGSGS-core), which reduces statistical power for enrichment detection. To further explore this gene set, a relaxed threshold (FDR < 0.7) is applied, indicating that PACMAD-specific syntelogs converge on a small number of shared biological functions (Fig. 8B and Supplementary Table S5). Future analyses with expanded species sampling within the PACMAD clade may resolve whether these genes share functional themes that are not detectable at the current sample sizes.

Nonsyntenic genes exhibited significant enrichment (FDR < 0.05) for terms related to central metabolism, flower development, and response to light stimuli, though with relatively low fold enrichment values (Fig. 8B and C, and Supplementary Table S5). The elevated nucleotide diversity and near-neutral Tajima's D (median = 0.39) for nonsyntenic genes are consistent with either relaxed constraint permiting netral drift or diversifying selection at these loci.

This secondary metabolism enrichment contrasts with primary metabolism dominance of PGSGS-core genes, recapitulating bacterial pan-genomes where similarly, accessory genes encode niche-specific adaptations while core genes maintain conserved infrastructure. Rapid evolution of nonsyntenic genes through tandem duplication and neofunctionalization [42, 43, 98, 99] may render these loci incompatible with maintaining syntenic orthology across deep evolutionary time. Together, these results indicate that syntenic. Conservation across grasses correlates with functional network centrality: PGSGS-core genes occupy hub positions in regulatory and cellular networks, which likely explains both their evolutionary conservation over >70 million years and the strong purifying selection observed in sorghum populations. By establishing that positional conservation predicts both functional constraint and selective pressure, PGSGS enables cross-species transfer of functional insights across grass crops.

Limitations and future directions

The development of methods that explicitly model expected copy-number ratios and tolerate structural variation is best suited for comprehensive synteny detection, particularly given the complex genomic architecture of modern grasses, which is likely a result of extensive domestication characterized by ancient polyploidy, biased fractionation, tandem amplification, and rearrangements. The updated PGSGS, based on QuotaAlign(+Polish), is scaled across 17 grass genomes, re-

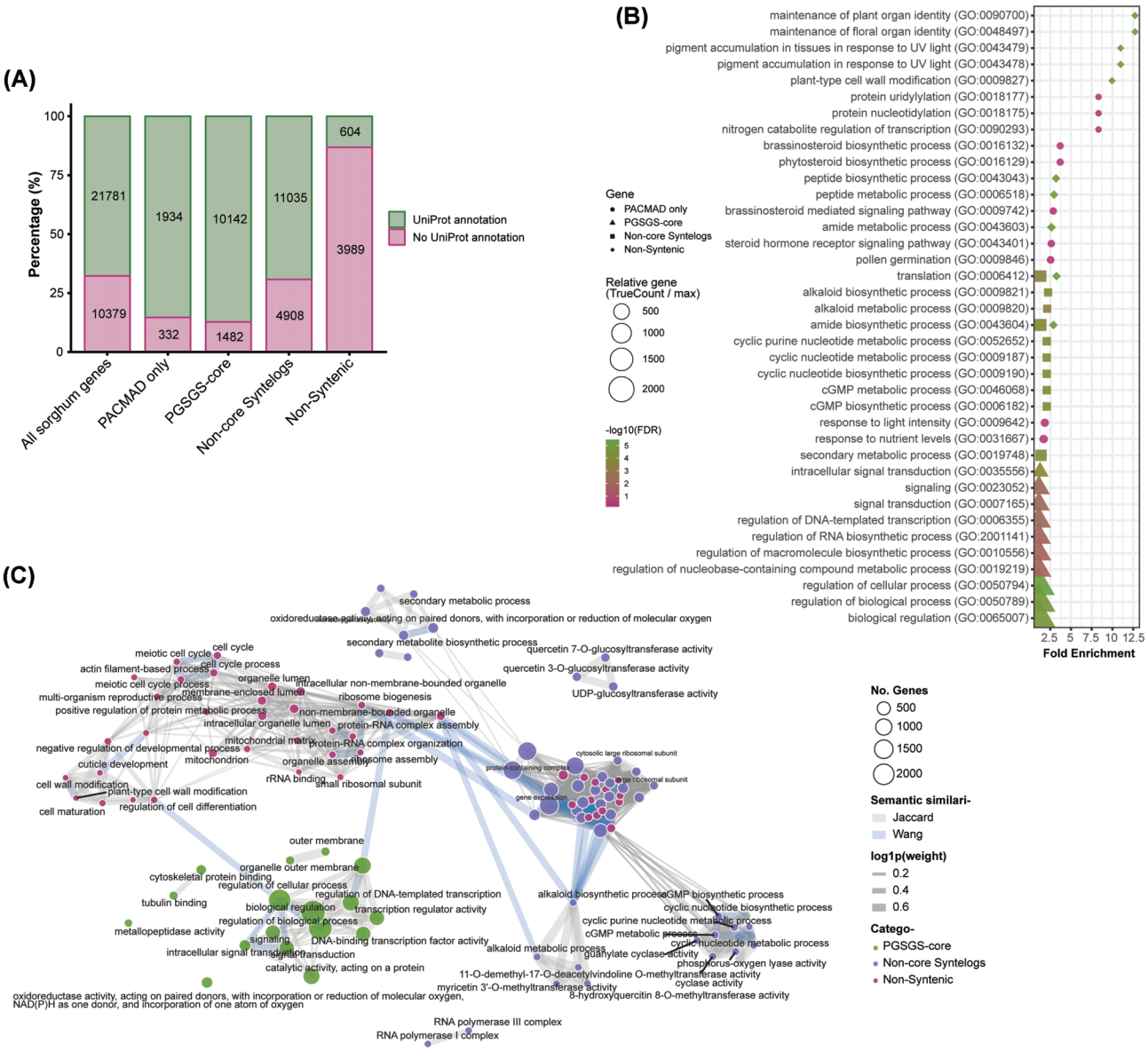


Figure 8. Functional stratification of PGSGS gene categories. (**A**) Distribution of gene counts with UniProt annotation across PGSGS categories. (**B**) GO enrichment for biological processes. Dot plot displays fold enrichment values for significantly overrepresented GO terms (FDR < 0.05) across PGSGS categories. GO terms for the PACMAD-only category are displayed at a relaxed threshold (FDR < 0.70), as no terms reached statistical significance at the more stringent threshold and are included for exploratory purposes only. GO terms are grouped by biological process (BP) category. Complete enrichment statistics are provided in Supplementary Table S5. (**C**) Network representation of functional relationships among enriched GO terms and PGSGS categories. Nodes represent significantly enriched biological process GO terms (FDR < 0.01). Node size corresponds to the number of genes annotated to each term. Edges connect GO terms based on semantic similarity calculated using Jaccard and Wang methods, with edge thickness proportional to similarity scores. Network layout was generated using clusterProfiler 4.0 [37] and ChIPseeker [38].

vealing that 46% of syntelogs (349 890 protein-coding genes) can be a source for functional and evolutionary studies.

Its unique anchoring to sorghum may underdetect synteny in more distantly related BOP clades, and the current species suite omits several newly released assemblies whose integration would strengthen pan-genome completeness. Looking ahead, future updates might include the addition of a BOP-clade anchor (e.g. *Oryza sativa* or *Brachypodium distachyon*) alongside sorghum and the rollout of a dedicated online portal for on-the-fly synteny queries, visualization, and data download. Nevertheless, PGSGS represents a significant advancement in understanding grass genome evolution and enhancing functional genomics research, with clear pathways for extension in species breadth, annotation depth, and user accessibility.

The development of the PGSGS database is a comprehensive resource that enables researchers to conduct comparative genomics, explore evolutionary dynamics, and annotate genes within syntenic regions, leveraging a well-curated *Sorghum bicolor* v.5.1 reference alongside 17 diverse grass genomes to deliver broad phylogenetic coverage. By stratifying syntelogs into core sets enriched for essential processes like growth and primary metabolism and noncore sets tied to stress response and secondary metabolism, PGSGS illuminates both deeply conserved functions and lineage-specific adaptations, thereby accelerating marker development for breeding programs and

enriching pan-genome analyses. The accompanying dataset and GO term summaries provide a valuable first step toward functional genomics in grasses, offering users a ready-to-use annotation resource upon which to build targeted experimental studies [100–102].

Significance statement

Grasses feed the world, but gene functional discoveries in one species rarely translate to others. We identify 11 624 genes maintained in syntenic positions across ≥15 grass species over 70 million years, representing a conserved genomic backbone under sustained purifying selection (36%–40% reduced diversity in sorghum populations). These genes likely perform essential functions conserved across rice, maize, wheat, sorghum, and related crops. The PGSGS enables researchers to transfer functional insights across species, accelerating crop improvement by treating grasses as an integrated genetic system rather than isolated species.

Acknowledgements

The authors extend their heartfelt gratitude to the dedicated teams at GaTE Lab and Schnable Lab for their unwavering motivation and profound curiosity in the realm of plant evolution. We also would like to express our deepest gratitude to the Holland Computing Center at the University of Nebraska-Lincoln (UNL) for their invaluable support and resources. The high-performance computing infrastructure and technical assistance provided by the center were crucial for the completion of this research. Their encouragement has been instrumental in propelling us forward in our pursuit of knowledge, recognizing the transformative impact that science and education have on lives. Special appreciation goes to Zach Shomo for providing unwavering support throughout the internship at the University of Nebraska-Lincoln. Their friendship has been instrumental in shaping a meaningful and enriching experience.

Author contributions: Henrique M. Dias (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Investigation [lead], Resources [equal], Validation [lead], Visualization [lead], Writing – original draft [lead]), Guilherme Y. Sagawa (Data curation [supporting], Formal analysis [supporting], Methodology [supporting]), Vladimir J. Torres-Rodriguez (Formal analysis [supporting], Writing – review & editing [equal]), Ravi Mural (Formal analysis [supporting], Investigation [equal], Writing – review & editing [equal]), James C. Schnable (Formal analysis [equal], Investigation [lead], Methodology [equal], Resources [equal], Supervision [equal], Writing – review & editing [equal]), and Marie-Anne Van Sluys (Conceptualization [equal], Formal analysis [equal], Funding acquisition [lead], Investigation [lead], Project administration [lead], Supervision [lead], Visualization [equal], Writing – review & editing [lead])

Supplementary data

Supplementary data is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

Financial support was obtained from grants FAPESP 2016/17545-8 and CNPq 310779/2017-0 to MAVS, HMD FAPESP 2019/08239-9 and FAPESP 2022/16208-9 scholarship, GYAS FAPESP 2023/03850-7 scholarship. The funders had no role in study design, data collection, analysis, publication decision, or manuscript preparation.

Data availability

The supplementary tables associated with this study are available on Figshare and can be accessed via the following link: 10.6084/m9.figshare.26793190. These data include brief description of the contents and the PGSGS source data. Detailed information of the outputs from PGSGS can be found in Supplementary File S1, and Supplementary File S2 has step-by-step instructions for the PGSGS analysis.

References

1. FAO. *Crop Prospects and Food Situation—Quarterly Global Report*, No. 2. Rome: FAO, 2023.
2. McSteen P, Kellogg EA. Molecular, cellular, and developmental foundations of grass diversity. *Science* 2022;377:599–602.
3. Linder HP, Lehmann CER, Archibald S *et al.* Global grass (P*oaceae*) success underpinned by traits facilitating colonization, persistence and habitat transformation. *Biol Rev* 2018;93:1125–44. https://doi.org/10.1111/brv.12388
4. Wang X, Gowik U, Tang H *et al.* Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol* 2009;10:R68. https://doi.org/10.1186/gb-2009-10-6-r68
5. Chen Z-H, Chen G, Dai F *et al.* Molecular evolution of grass stomata. *Trends Plant Sci* 2017;22:124–39. https://doi.org/10.1016/j.tplants.2016.09.005
6. Griffith DM, Osborne CP, Edwards EJ *et al.* Lineage-based functional types: characterising functional diversity to enhance the representation of ecological behaviour in land surface models. *New Phytol* 2020;228:15–23. https://doi.org/10.1111/nph.16773
7. Pardo J, VanBuren R. Evolutionary innovations driving abiotic stress tolerance in C4 grasses and cereals. *Plant Cell* 2021;33:3391–401. https://doi.org/10.1093/plcell/koab205
8. Jiang X, Assis R. Rapid functional divergence after small-scale gene duplication in grasses. *BMC Evol Biol* 2019;19:97. https://doi.org/10.1186/s12862-019-1415-2
9. Zhang T, Huang W, Zhang L *et al.* Phylogenomic profiles of whole-genome duplications in Poaceae and landscape of differential duplicate retention and losses among major Poaceae lineages. *Nat Commun* 2024;15:3305. https://doi.org/10.1038/s41467-024-47428-9
10. Moore G, Devos KM, Wang Z *et al.* Cereal genome evolution. Grasses, line up and form a circle. *Curr Biol* 1995;5:737–9. https://doi.org/10.1016/S0960-9822(95)00148-5
11. Murat F, Xu J-H, Tannier E *et al.* Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res* 2010;20:1545–57. https://doi.org/10.1101/gr.109744.110
12. Schnable JC, Wang X, Pires JC *et al.* Escape from preferential retention following repeated whole genome duplications in plants. *Front Plant Sci* 2012;3:94. https://doi.org/10.3389/fpls.2012.00094
13. Schnable JC. Genes and gene models, an important distinction. *New Phytol* 2020;228:50–5. https://doi.org/10.1111/nph.16011
14. Jiao C, Xie X, Hao C *et al.* Pan-genome bridges wheat structural variations with habitat and breeding. *Nature* 2025;637:384–93. https://doi.org/10.1038/s41586-024-08277-0
15. Hufford MB, Seetharam AS, Woodhouse MR *et al. De novo* assembly, annotation, and comparative analysis of 26 diverse

maize genomes. *Science* 2021;373:655–62. https://doi.org/10.1126/science.abg5289
16. Healey AL, Garsmeur O, Lovell JT *et al.* The complex polyploid genome architecture of sugarcane. *Nature* 2024;628:804–10. https://doi.org/10.1038/s41586-024-07231-4
17. Paterson AH, Bowers JE, Chapman BA. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* 2004;101:9903–8. https://doi.org/10.1073/pnas.0307901101
18. Wang X, Shi X, Hao B *et al.* Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol* 2005;165:937–46. https://doi.org/10.1111/j.1469-8137.2004.01293.x
19. Salse J, Bolot S, Throude M *et al.* Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 2008;20:11–24. https://doi.org/10.1105/tpc.107.056309
20. Goodstein DM, Shu S, Howson R *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 2012;40:D1178–86. https://doi.org/10.1093/nar/gkr944
21. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* 2008;53:661–73, https://doi.org/10.1111/j.1365-313X.2007.03326.x
22. Lyons E, Pedersen B, Kane J *et al.* Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol* 2008;148:1772–81.
23. Van Bel M, Silvestri F, Weitz EM *et al.* PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Res* 2022;50:D1468–74. https://doi.org/10.1093/nar/gkab1024
24. Tello-Ruiz MK, Jaiswal P, Ware D. Gramene: a resource for comparative analysis of plants genomes and pathways. In Edwards D. (ed.), *Plant Bioinformatics, Methods in Molecular Biology*, Vol. 2443. New York: Springer, 2022, 101–31.
25. Bolser D, Staines DM, Pritchard E *et al.* Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. In Edwards D. (ed), *Plant Bioinformatics, Methods in Molecular Biology*, Vol. 1374. New York: Springer New York, 2016, 115–40.
26. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci USA* 2011;108:4069–74. https://doi.org/10.1073/pnas.1101368108
27. Wang Z, Yang J, Cheng F *et al.* Subgenome dominance and its evolutionary implications in crop domestication and breeding. *Horticulture Research* 2022;9:uhac090. https://doi.org/10.1093/hr/uhac090
28. Zhang Y, Ngu DW, Carvalho D *et al.* Differentially regulated orthologs in sorghum and the subgenomes of maize. *Plant Cell* 2017;29:1938–51. https://doi.org/10.1105/tpc.17.00354
29. Tang H, Lyons E, Pedersen B *et al.* Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* 2011;12:102. https://doi.org/10.1186/1471-2105-12-102
30. Harris RS. *Improved Pairwise Alignment of Genomic DNA*. PA, United States: The Pennsylvania State University, 2007.
31. Zhang J, Zhang X, Tang H *et al.* Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum L. Nat Genet* 2018;50:1565–73. https://doi.org/10.1038/s41588-018-0237-2
32. Haas BJ, Delcher AL, Wortman JR *et al.* DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 2004;20:3643–6. https://doi.org/10.1093/bioinformatics/bth397
33. Wang Y, Tang H, DeBarry JD *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;40:e49. https://doi.org/10.1093/nar/gkr1293
34. Huerta-Cepas J, Szklarczyk D, Heller D *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;47:D309–14. https://doi.org/10.1093/nar/gky1085
35. Bonnot T, Gillard M, Nagel D. A simple protocol for informative visualization of enriched gene ontology terms. *Bioprotocol* 2019;9:e3429. https://doi.org/10.21769/BioProtoc.3429
36. Supek F, Bošnjak M, Škunca N *et al.* REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 2011;6:e21800. https://doi.org/10.1371/journal.pone.0021800
37. Wu T, Hu E, Xu S *et al.* clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* 2021;2:100141. https://doi.org/10.1016/j.xinn.2021.100141
38. Wang Q, Li M, Wu T *et al.* Exploring epigenomic datasets by ChIPseeker. *Curr Protoc* 2022;2:e585. https://doi.org/10.1002/cpz1.585
39. Boatwright JL, Sapkota S, Jin H *et al.* Sorghum association panel whole-genome sequencing establishes cornerstone resource for dissecting genomic diversity. *Plant J* 2022;111:888–904. https://doi.org/10.1111/tpj.15853
40. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989;123:585–95. https://doi.org/10.1093/genetics/123.3.585
41. Della Coletta R, Qiu Y, Ou S *et al.* How the pan-genome is changing crop genomics and improvement. *Genome Biol* 2021;22:3. https://doi.org/10.1186/s13059-020-02224-8
42. Birchler JA, Yang H. The multiple fates of gene duplications: deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* 2022;34:2466–74. https://doi.org/10.1093/plcell/koac076
43. Leister D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet* 2004;20:116–22. https://doi.org/10.1016/j.tig.2004.01.007
44. Swigoňova Z, Lai J, Ma J *et al.* Close split of sorghum and maize genome progenitors. *Genome Res* 2004;14:1916–23.
45. Schnable PS, Ware D, Fulton RS *et al.* The B73 maize genome: complexity, diversity, and dynamics. *Science* 2009;326:1112–5. https://doi.org/10.1126/science.1178534
46. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of gene duplication in plants. *Plant Physiol* 2016;171:2294–316. https://doi.org/10.1104/pp.16.00523
47. Paterson AH, Bowers JE, Bruggmann R *et al.* The sorghum bicolor genome and the diversification of grasses. *Nature* 2009;457:551–6. https://doi.org/10.1038/nature07723
48. Soreng RJ, Peterson PM, Romaschenko K *et al.* A worldwide phylogenetic classification of the Poaceae (Gramineae) II: an update and a comparison of two 2015 classifications. *J Sytematics Evol* 2017;55:259–90. https://doi.org/10.1111/jse.12262
49. Van De Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. *Nat Rev Genet* 2017;18:411–24. https://doi.org/10.1038/nrg.2017.26
50. Freeling M, Thomas BC. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* 2006;16:805–14. https://doi.org/10.1101/gr.3681406
51. Van De Peer Y, Ashman T-L, Soltis PS *et al.* Polyploidy: an evolutionary and ecological force in stressful times. *Plant Cell* 2021;33:11–26. https://doi.org/10.1093/plcell/koaa015
52. Birchler JA, Veitia RA. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci USA* 2012;109:14746–53. https://doi.org/10.1073/pnas.1207726109
53. Wendel JF. Genome evolution in polyploids. In: Doyle JJ, Gaut BS. (eds), *Plant Molecular Evolution*. Dordrecht: Springer, 2000, pp.225–49. https://doi.org/10.1007/978-94-011-4221-2

54. Eckardt NA. Genome dominance and interaction at the gene expression level in allohexaploid wheat. *Plant Cell* 2014;26:1834. https://doi.org/10.1105/tpc.114.127183
55. Edger PP, Poorten TJ, VanBuren R *et al.* Origin and evolution of the octoploid strawberry genome. *Nat Genet* 2019;51:541–7. https://doi.org/10.1038/s41588-019-0356-4
56. Zhang K, Zhang L, Cui Y *et al.* The lack of negative association between TE load and subgenome dominance in synthesized *Brassica* allotetraploids. *Proc Natl Acad Sci USA* 2023;120:e2305208120. https://doi.org/10.1073/pnas.2305208120
57. Cheng F, Sun C, Wu J *et al.* Epigenetic regulation of subgenome dominance following whole genome triplication in *Brassica rapa*. *New Phytol* 2016;211:288–99. https://doi.org/10.1111/nph.13884
58. Glover NM, Redestig H, Dessimoz C. Homoeologs: what are they and how do we infer them? *Trends Plant Sci* 2016;21:609–21. https://doi.org/10.1016/j.tplants.2016.02.005
59. Mitros T, Session AM, James BT *et al.* Genome biology of the paleotetraploid perennial biomass crop Miscanthus. *Nat Commun* 2020;11:5442. https://doi.org/10.1038/s41467-020-18923-6
60. Clark LV, Jin X, Petersen KK *et al.* Population structure of *Miscanthus sacchariflorus* reveals two major polyploidization events, tetraploid-mediated unidirectional introgression from diploid *M. sinensis*, and diversity centred around the Yellow Sea. *Ann Bot* 2019;124:731–48. https://doi.org/10.1093/aob/mcy161
61. Garsmeur O, Droc G, Antonise R *et al.* A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat Commun* 2018;9:2638. https://doi.org/10.1038/s41467-018-05051-5
62. Souza GM, Van Sluys M-A, Lembke CG *et al.* Assembly of the 373k gene space of the polyploid sugarcane genome reveals reservoirs of functional diversity in the world's leading biomass crop. *GigaScience* 2019;8:giz129. https://doi.org/10.1093/gigascience/giz129
63. Brenchley R, Spannagl M, Pfeifer M *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 2012;491:705–10. https://doi.org/10.1038/nature11650
64. Marcussen T, Sandve SR, Heier L *et al.* Ancient hybridizations among the ancestral genomes of bread wheat. *Science* 2014;345:1250092. https://doi.org/10.1126/science.1250092
65. Estep MC, McKain MR, Vela Diaz D *et al.* Allopolyploidy, diversification, and the Miocene grassland expansion. *Proc Natl Acad Sci USA* 2014;111:15149–54. https://doi.org/10.1073/pnas.1404177111
66. Conant GC, Wolfe KH. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 2008;9:938–50. https://doi.org/10.1038/nrg2482
67. Wendel JF, Lisch D, Hu G *et al.* The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. *Curr Opin Genet Dev* 2018;49:1–7. https://doi.org/10.1016/j.gde.2018.01.004
68. D'Hont A, Grivet L, Feldmann P *et al.* Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum spp.*) by molecular cytogenetics. *Molec Gen Genet* 1996;250:405–13.
69. Piperidis N, D'Hont A. Sugarcane genome architecture decrypted with chromosome-specific oligo probes. *Plant J* 2020;103:2039–51. https://doi.org/10.1111/tpj.14881
70. Grivet L, Arruda P. Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr Opin Plant Biol* 2002;5:122–7. https://doi.org/10.1016/S1369-5266(02)00234-0
71. Garsmeur O, Charron C, Bocs S *et al.* High homologous gene conservation despite extreme autopolyploid redundancy in sugarcane. *New Phytol* 2011;189:629–42. https://doi.org/10.1111/j.1469-8137.2010.03497.x
72. Jannoo N, Grivet L, Chantret N *et al.* Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. *Plant J* 2007;50:574–85. https://doi.org/10.1111/j.1365-313X.2007.03082.x
73. Akagi T, Jung K, Masuda K *et al.* Polyploidy before and after domestication of crop species. *Curr Opin Plant Biol* 2022;69:102255. https://doi.org/10.1016/j.pbi.2022.102255
74. Ming R, Moore PH, Wu K *et al.* Sugarcane improvement through breeding and biotechnology. In: Janick J. (ed.), *Plant Breeding Reviews*. Canada : Wiley, 2005, pp.15–118. https://doi.org/10.1002/9780470650349
75. Hoang NV, Furtado A, Mason PJ *et al.* A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and *de novo* assembly from short read sequencing. *BMC Genomics* 2017;18:395. https://doi.org/10.1186/s12864-017-3757-8
76. Li X, Chen X, Fang J *et al.* Whole-genome sequencing of a worldwide collection of sugarcane cultivars (*Saccharum spp.*) reveals the genetic basis of cultivar improvement. *Plant J* 2024;119:2151–67. https://doi.org/10.1111/tpj.16861
77. Levy AA, Feldman M. Evolution and origin of bread wheat. *Plant Cell* 2022;34:2549–67. https://doi.org/10.1093/plcell/koac130
78. Glover NM, Daron J, Pingault L *et al.* Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. *Genome Biol* 2015;16:188. https://doi.org/10.1186/s13059-015-0754-6
79. Gossmann TI, Song B-H, Windsor AJ *et al.* Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol* 2010;27:1822–32. https://doi.org/10.1093/molbev/msq079
80. Wright SI, Gaut BS. Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* 2005;22:506–19. https://doi.org/10.1093/molbev/msi035
81. Hufford MB, Xu X, Van Heerwaarden J *et al.* Comparative population genomics of maize domestication and improvement. *Nat Genet* 2012;44:808–11. https://doi.org/10.1038/ng.2309
82. Smith O, Nicholson WV, Kistler L *et al.* A domestication history of dynamic adaptation and genomic deterioration in Sorghum. *Nat Plants* 2019;5:369–79. https://doi.org/10.1038/s41477-019-0397-9
83. Mace ES, Tai S, Gilding EK *et al.* Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat Commun* 2013;4:2320. https://doi.org/10.1038/ncomms3320
84. Hamblin MT, Casa AM, Sun H *et al.* Challenges of detecting directional selection after a bottleneck: lessons from sorghum bicolor. *Genetics* 2006;173:953–64. https://doi.org/10.1534/genetics.105.054312
85. Wright SI, Gaut BS. Molecular population genetics and the search for adaptive evolution in plants. *Mol Biol Evol* 2005;22:506–19. https://doi.org/10.1093/molbev/msi035
86. Boatwright JL, Brenton ZW, Boyles RE *et al.* Genetic characterization of a *Sorghum bicolor* multiparent mapping population emphasizing carbon-partitioning dynamics. *G3 (Bethesda)* 2021;11:jkab060. https://doi.org/10.1093/g3journal/jkab060
87. Makino T. The evolutionary rate of a protein is influenced by features of the interacting partners. *Mol Biol Evol* 2006;23:784–9. https://doi.org/10.1093/molbev/msj090
88. Wang Z, Liao B-Y, Zhang J. Genomic patterns of pleiotropy and the evolution of complexity. *Proc Natl Acad Sci USA* 2010;107:18034–9. https://doi.org/10.1073/pnas.1004666107
89. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 2005;39:309–38. https://doi.org/10.1146/annurev.genet.39.073003.114725
90. Rhee SY, Mutwil M. Towards revealing the functions of all genes in plants. *Trends Plant Sci* 2014;19:212–21. https://doi.org/10.1016/j.tplants.2013.10.006
91. Monroe JG, Srikant T, Carbonell-Bejerano P *et al.* Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature* 2022;602:101–5. https://doi.org/10.1038/s41586-021-04269-6

92. Pophaly SD, Tellier A. Population level purifying selection and gene expression shape subgenome evolution in maize. *Mol Biol Evol* 2015;msv191, https://doi.org/10.1093/molbev/msv191
93. Gaudet P, Dessimoz C. Gene ontology: pitfalls, biases, and remedies. In Dessimoz C., Škunca N. (eds), *The Gene Ontology Handbook, Methods in Molecular Biology*, Vol. 1446. New York: Springer, 2017, 189–205.
94. Birchler JA, Veitia RA. The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol* 2010;186:54–62. https://doi.org/10.1111/j.1469-8137.2009.03087.x
95. Bertolini E, Rice BR, Braud M *et al.* Regulatory variation controlling architectural pleiotropy in maize. *Nat Commun* 2025;16:2140. https://doi.org/10.1038/s41467-025-56884-w
96. Coate JE, Doyle JJ. Divergent evolutionary fates of major photosynthetic gene networks following gene and whole genome duplications. *Plant Signal Behav* 2011;6:594–7. https://doi.org/10.4161/psb.6.4.15370
97. He X, Zhang J. Toward a molecular understanding of pleiotropy. *Genetics* 2006;173:1885–91. https://doi.org/10.1534/genetics.106.060269
98. Hanada K, Zou C, Lehti-Shiu MD *et al.* Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol* 2008;148:993–1003. https://doi.org/10.1104/pp.108.122457
99. Mizuno H, Katagiri S, Kanamori H *et al.* Evolutionary dynamics and impacts of chromosome regions carrying R-gene clusters in rice. *Sci Rep* 2020;10:872. https://doi.org/10.1038/s41598-020-57729-w
100. Crisp PA, Hammond R, Zhou P *et al.* Variation and inheritance of small RNAs in maize inbreds and F1 hybrids. *Plant Physiol* 2020;182:318–31. https://doi.org/10.1104/pp.19.00817
101. Levy B, Xu Z, Zhao L *et al.* FloraBERT: cross-species transfer learning withattention-based neural networks for geneexpression prediction. *Research Squar, https://doi.org/10.21203/rs.3.rs-1927200/v1*, 30 August 2022, preprint: not peer reviewed.
102. Gomez-Cano F, Rodriguez J, Zhou P *et al.* Prioritizing maize metabolic gene regulators through multi-omic network integration. bioRxiv https://doi.org/10.1101/2024.02.26.582075, 27 February 2024, preprint: not peer reviewed.

Received: November 14, 2024. **Revised:** February 21, 2026. **Accepted:** February 25, 2026