Received: 16 April 2020 | Revised: 9 September 2020 | Accepted: 23 September 2020

DOI: 10.1111/1755-0998.13271

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

The elephant grass (*Cenchrus purpureus*) genome provides insights into anthocyanidin accumulation and fast growth

Qi Yan[1] | Fan Wu[1] | Pan Xu[1] | Zongyi Sun[3] | Jie Li[1] | Lijuan Gao[1] | Liyan Lu[1] | Dongdong Chen[2] | Meki Muktar[4] | Chris Jones[4] | Xianfeng Yi[2] | Jiyu Zhang[1]

[1]State Key Laboratory of Grassland Agro-Ecosystems, Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs, Engineering Research Center of Grassland Industry, Ministry of Education, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou, China

[2]Guangxi Institute of Animal Sciences, Nanning, China

[3]Nextomics Biosciences Institute, Wuhan, China

[4]Feed and Forage Development, International Livestock Research Institute, Nairobi, Kenya

Correspondence
Chris Jones, Feed and Forage Development, International Livestock Research Institute, Nairobi, Kenya.
Email: c.s.jones@cgiar.org

Xianfeng Yi, Guangxi Institute of Animal Sciences, Nanning 530001, China.
Email: 1154128631@qq.com

Jiyu Zhang, State Key Laboratory of Grassland Agro-Ecosystems, Key Laboratory of Grassland Livestock Industry Innovation, Ministry of Agriculture and Rural Affairs, Engineering Research Center of Grassland Industry, Ministry of Education, College of Pastoral Agriculture Science and Technology, Lanzhou University, Lanzhou 730020, China.
Email: zhangjy@lzu.edu.cn)

Funding information
Key Research and Development Plan of Guangxi Science and Technology, Grant/Award Number: Guike AB19245024; Guangxi Science and Technology, Grant/Award Number: Guike AD17129043; Program for Changjiang Scholars and Innovative Research Team in University, Grant/Award Number: IRT_17R50; Guangxi Science and Technology Major Project, Grant/Award Number: Guike AA16380026; The 111 Project, Grant/Award Number: B12002

Abstract

Elephant grass ($2n = 4x = 28$; *Cenchrus purpureus* Schumach.), also known as Napier grass, is an important forage grass and potential energy crop in tropical and subtropical regions of Asia, Africa and America. However, no study has yet reported a genome assembly for elephant grass at the chromosome scale. Here, we report a high-quality chromosome-scale genome of elephant grass with a total size of 1.97 Gb and a 1.5% heterozygosity rate, obtained using short-read sequencing, single-molecule long-read sequencing and Hi-C chromosome conformation capture. Evolutionary analysis showed that subgenome A' of elephant grass and pearl millet may have originated from a common ancestor more than 3.22 million years ago (MYA). Further, allotetraploid formation occurred at approximately 6.61 MYA. Syntenic analyses within elephant grass and with other grass species indicated that elephant grass has experienced chromosomal rearrangements. We found that some key enzyme-encoding gene families related to the biosynthesis of anthocyanidins and flavonoids were expanded and highly expressed in leaves, which probably drives the production of these major anthocyanidin compounds and explains why this elephant grass cultivar has a high anthocyanidin content. In addition, we found a high copy number and transcript levels of genes involved in C_4 photosynthesis and hormone signal transduction pathways that may contribute to the fast growth of elephant grass. The availability of elephant grass genome data advances our knowledge of the genetic evolution of elephant grass and will contribute to further biological research and breeding as well as for other polyploid plants in the genus *Cenchrus*.

KEYWORDS

anthocyanidin biosynthesis, C_4 photosynthesis, comparative genomic, elephant grass, plant hormone, reference genome

Qi Yan, Fan Wu, Pan Xu, and Zongyi Sun contributed equally to this work.

1 | INTRODUCTION

Elephant grass (*Cenchrus purpureus* Schumach. syn. *Pennisetum purpureum* (Schumach.) Morrone; $2n = 4x = 28$) is a perennial C_4 plant native to sub-Saharan Africa (Farrelletal.,2002). Elephant grass belongs to the subfamily Panicoideae of the family Poaceae and is one of the most important forage species and potential energy grasses in tropical and subtropical regions of Asia, Africa and America (Fang, 2015; Mapato & Wanapat, 2018; Strezov, Evans, & Hayman, 2008). Elephant grass is an excellent fodder crop with a yield of up to 150 tons green matter per hectare each year and is capable of withstanding repeated cuttings (four to six cuts per year), resisting high temperatures, drought stress, low soil fertility and biotic stress (Kebede et al., 2017; Liu et al., 2008). Its excellent reproductive and adaptive characteristics have led to its widespread use as a cut-carry feed. Furthermore, recent studies have demonstrated that elephant grass, as a lignocellulosic plant, has high potential for bioenergy and paper production (Daud et al., 2014). For example, the alcohol production and calorific value of elephant grass are three and 0.7 times those of switchgrass and coal, respectively (Cardona et al., 2014). In addition, elephant grass can act as an ecological grass to improve soil fertility and protect against soil erosion based on its root and tiller development (Zhran & Lotfy, 2014).

Elephant grass (A'A'BB) and pearl millet (*Cenchrus americanus* Morrone syn. *Pennisetum glaucum*; $2n = 2x = 14$; AA) are economically important species in the genus *Cenchrus* serving mainly as forage grasses or cereals. A close relationship and common origin between elephant grass and pearl millet was suggested by mitochondrial DNA, chloroplast DNA and repetitive DNA sequences (Reis et al., 2014). The chromosomes in the A' genome of elephant grass are believed to be homologous to those of the A genome of pearl millet (Gupta & Mhere, 1997). With the aid of pearl millet genome sequences, modern molecular genetic techniques such as whole-genome resequencing and genotyping-by-sequencing have been used to identify key single nucleotide polymorphisms (SNPs) and loci associated with economically important traits, which will undoubtedly enhance the efficiency of molecular breeding (Pucher, 2018; Varshney et al., 2018). However, to our knowledge, no currently polyploid genomes have been reported in the genus *Cenchrus*, which limits comparative genomic studies to some extent. Genome surveys and high-density genetic maps (Paudel et al., 2018; Wang et al., 2018), RNA sequencing (RNA-seq) and metabonomic studies (Zhou et al., 2019; Zhou, et al., 2018), and the application of molecular markers from simple sequence repeats (SSRs; Zhou, et al., 2018) to genotyping-by-sequencing (Muktar et al., 2019) across elephant grass genotypes have enabled rapid progress in elephant grass genomics and breeding studies (Rocha et al., 2019). However, the lack of reference genomes and availability of only short-read resequencing data has led to some limitations in identifying key genes and characterizing genomic variations that may substantially contribute to genome evolution and the genetics of economically important traits in elephant grass such as purple leaves, fast growth and bioenergy production.

FIGURE 1 *Cenchrus purpureus* cultivar Purple

Due to the self-incompatibility and obligate outcrossing nature of elephant grass, its genotypes exhibit high levels of heterozygosity. Here, *C. purpureus* cv. Purple was chosen for the draft genome assembly of elephant grass; this cultivar has several desirable traits, including purple leaves, high biomass production, regeneration ability and easy establishment (Figure 1). It is a large herbaceous plant (2.5–3.6 m) and can be used for bioethanol production, paper production and phytoremediation and as an ornamental plant. Anthocyanins provide multiple benefits to plants, conferring protection against biotic and abiotic stressors (Mazumder, 2015). Additionally, anthocyanin-rich plant material has powerful antioxidant properties, which is good for both humans and animals that consume it (Kruger et al., 2014). The leaf anthocyanin content of *C. purpureus* cv. Purple (~2.25 mg per 100 g) is higher than that of some types of grape skins and teas (Yi et al., 2016). In addition, animals that were fed on *C. purpureus* cv. Purple were shown to have better growth and productivity compared with animals fed on a green cultivar (Yao et al., 2016). Owing to the high heterozygosity and large genome size of *C. purpureus*, we used short-read sequencing, single-molecule long-read sequencing of Oxford Nanopore Technologies (ONT) and high-throughput chromosome conformation capture (Hi-C) approaches to assemble a high-quality genome. Investigating this reference genome may provide a foundation for functional genomics to improve the molecular basis of its economically valuable traits and to elucidate the evolution of the genus *Cenchrus*.

2 | MATERIALS AND METHODS

2.1 | DNA and RNA sampling and sequencing

Young leaves from individual plants of the elephant grass cultivar *Cenchrus purpureus* cv. Purple, at 12 weeks of age, were collected from the glasshouse of Lanzhou University and were plucked and frozen in liquid nitrogen. Genomic DNA was extracted from the leaf tissue using a DNeasy Plant Maxi kit (Qiagen). The quality of the DNA was checked using a 2100 Bioanalyzer (Agilent Technologies), and high-integrity

DNA molecules were measured using 1% agarose gel electrophoresis. For genome sequencing, next-generation sequence data were obtained on the Illumina Navoseq 6000 platform with a 400-bp insert size, and third-generation data were obtained on the Nanopore PromethION platform (Leamon, 2018). For the Hi-C data, the Hi-C library was prepared following a standard procedure and sequenced using the Illumina HiSeq platform (Illumina) with a 400-bp insert size. For RNA-seq, 15 samples of *C. purpureus* cv. Purple (at the flowering stage) were collected from the field in Nanning (108°33′N, 22°84′E, Nanning city, Guangxi province, China), including shoot, root, leaf, flower and stem tip samples. Before their use for sequencing, plants were regenerated for two generations through cuttings. RNA was extracted from the samples using TRNzol Universal Reagent (Cat. no. DP424, Tiangen). The cDNA library was prepared using the TruSeq Sample Preparation Kit (Illumina), and paired-end sequencing with a length of 125 bp was conducted on the HiSeq 2500 platform (Illumina). Clean data were obtained by removing reads containing adapters or poly-N sequences and low-quality reads from the raw data.

2.2 | Genome assembly and pseudochromosome construction

Root tips were excised and treated using routine methods for chromosome counting (Yang et al., 2017). The K-mer method was used to estimate the elephant grass genome size using the quality-filtered reads, sequenced on the Illumina X Ten platform (Liu et al., 2012). Genome size was estimated based on the following formula: genome size = modified K-mer number/average K-mer depth (total K-mers were filtered to remove incorrect K-mers to obtain the modified K-mers) with KmerFreq_AR (Luo et al., 2012). For heterozygosity, *Arabidopsis* genomic data were used for the simulation of paired-end reads, which was carried out by the profile-based Illumina pair-end Reads Simulator (pIRS), and then fitting curves, constructed according to the K-mer distribution curve of elephant grass (Galaxy et al., 2012). When the two K-mer curves were consistent, the heterozygosity of *Arabidopsis* was representative of that of elephant grass.

Nanopore reads which passed the quality criteria (those with a read quality score of less than 7 were discarded) were corrected using NEXTDENOVO (version 1.0; https://github.com/Nextomics/NextDenovo.git) with specific parameters (read_cutoff = 2k, seed_cutoff = 20k) to obtain consensus sequences, and the initial genome (G1) was then assembled with SMARTDENVO (version 1.0.0; https://github.com/ruanjue/smartdenovo; wtpre -J 3000, wtzmo -k 21 -z 10 -Z 19 -U −1 -m 0.1 -A 1000) using the consensus sequences. To acquire more accurate genome sequences, next-generation sequencing (NGS) data were mapped to G1 with BWA MEM (0.7.17-r1188) using default parameters, and G1 was then polished using PILON (version 1.22; --fix bases; Walker et al., 2014). The polishing process involved three iterations, and the accurate genome (G2) was finally assembled. As the elephant grass genome was highly heterozygous (heterozygosity = 1.5%), the G2 size was larger than the expected assembly size. Some redundant sequences were removed from G2 using REDUNDANS (version 0.13c) with specific parameters (identity = 0.824; coverage = 0.8), and we thereby obtained the nonredundant genome (G3; Pryszcz & Gabaldón, 2016). To evaluate the completeness of the G3 genome, Benchmarking Universal Single-Copy Orthologs (BUSCO, version 4.0.2) and Core Eukaryotic Genes Mapping Approach (CEGMA, version 2.5) were applied using default parameters to search the annotated genes in the assembly (Simão et al., 2015). Additionally, the RNA-seq data were aligned against the nonredundant genome using the HISAT2 (version 2.1.0) program with default parameters (Kim et al., 2015).

To obtain the chromosome-level genome assembly, we used LACHESIS software to cluster, order and orient the contigs by the Hi-C data (Dekker et al., 2002). First, 1,850 contigs were sorted and anchored to 14 chromosomes using LACHESIS with specific parameters (CLUSTERMINRESITES = 100; CLUSTERMAXLINKDENSITY = 2; CLUSTERNONINFORMATIVERATIO = 1.5;4. ORDERMINNRESINTRUNK = 60; ORDERMINNRESINSHREDS = 60; Burton et al., 2013). We further corrected the misassembled contigs based on the interaction strength among the contigs and a linkage map of elephant grass using JUICEBOX (Figures S3 and S4). The genetic linkage map (Paudel et al., 2018) of elephant grass was aligned to raw chromosomes using BLASTN (E-value ≤ 1e−5; Figure S4).

2.3 | Repeat annotation

Repeat sequences can be classified into three types based on repeat degree: SSRs, moderately repetitive sequences and highly repetitive sequences. SSRs consist of 1–6 bp of DNA and are widely distributed in genomes. To identify the SSRs in the elephant grass genome, the MIcroSAtellite Identification Tool (MISA; https://webblast.ipk-gatersleben.de/misa/) was used to distinguish and locate both simple and compound (where two or more microsatellites are located directly adjacent to each other) SSRs (Beier et al., 2017). For other repeat types, a combination of de novo-based and homology-based strategies was utilized at both the DNA and the protein levels to identify transposable elements (TEs). First, REPEATMODELER (version 1.0.8), LTR_FINDER (version 1.07), LTR_RETRIVER (version 2.8) and MITE-HUNTER were used to search the genome for repeat sequences, and the identified repeat sequences were then used to construct a de novo repeat library at the DNA level (Han & Wessler, 2010; Saha et al., 2008; Xu & Wang, 2007). Next, a custom TE library was integrated with the de novo repeat library and Rebase to produce the final repeat library (Jurka et al., 2005). Finally, all potential TE sequences were searched for in the final repeat library using REPEATMASKER (version 4.0.6; Tarailo-Graovac & Chen, 2009).

2.4 | Gene prediction and functional annotation

Gene structure prediction depended on the application of three methods: ab initio prediction, homology-based prediction and

RNA-seq-assisted prediction (Yandell & Ence, 2012). For ab initio prediction, AUGUSTUS (version 3.3.1) and GLIMMERHMM (version 3.0.4) were used for de novo-based gene prediction with the default parameters to predict the genes of the elephant grass genome (Hoff & Stanke, 2018; Nachtweide et al., 2016; Stanke et al., 2008). Additionally, the filtered proteins (incomplete and wrong) of three species (*Sorghum bicolor* GCF_000003195.3, *Setaria italica* GCF_000263155.2 and *Zea mays* GCF_000005005.2) were used for homology-based prediction with GEMOMA (version 1.5.3) and GENEWISE (version 2.4.1) using default settings (Haas et al., 2003; Keilwagen et al., 2016,2018). Then, PASA (version 2.0.2) was used for RNA-seq-based gene prediction (Haas et al., 2003). Finally, the results from the three approaches were integrated using EVIDENCEMODELER (EVM; version.1.1) to obtain the elephant grass raw gene set (Haas et al., 2008). To obtain a precise gene set, some genes whose sequences included transposable elements were filtered with TRANSPOSONPSI software (http://transposonpsi.sourceforge.net). To assess the completeness of the gene set, BUSCO (version 4.0.2) was used to evaluate the gene set based on the encoded proteins using embryophyta_odb10.

To obtain the functions of the genes, genes were annotated using two strategies based on protein sequences. First, the predicted protein sequences were aligned to the SwissProt protein databases using BLASTP (https://blast.ncbi.nlm.nih.gov/Blast.cgi) under the best match parameter (Mercier & Bougueleret, 2007). The gene pathways of the predicted sequences were extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) Automatic Annotation Server (version 2.1; Koech, 2019). Then, the annotation of motifs and domains was performed using INTERPROSCAN (version 5.32-71.0) to search against open databases of InterPro, including the member databases of Pfam, ProDom, PRINTS, PANTHER, SMRT and PROSITE (Hunter et al., 2008). Gene Ontology (GO) IDs (Harris et al., 2004) for each gene were determined using the BLAST2GO (version 1.44) pipeline (Conesa et al., 2005).

Additionally, the annotation of the noncoding RNA gene set was performed. The data set was aligned to the Rfam (version 11.0) noncoding database to annotate genes encoding ribosomal RNA (rRNA), small nuclei RNA (snRNA) and micro RNA (miRNA) first (Griffiths-Jones et al., 2005). Then, the transfer RNA (tRNA) sequences were predicted using TRNASCAN-SE (version 1.3.1; Lowe & Eddy, 1997). rRNA and its subunits were predicted by RNAMMER (version 1.2; Lagesen et al., 2007).

2.5 | Phylogenetic analysis and divergence time estimation

For the phylogenetic analysis of elephant grass, nine additional species (*Brachypodium distachyon* [GCF_000005505.3], *Dichanthelium oligosanthes* [GCA_001633215.2], *Oryza sativa* [phytozome], *Cenchrus americanus* [http://cegsb.icrisat.org/ipmgsc/genome.html], *Sorghum bicolor* [GCF_000003195.3], *Setaria italica* [GCF_000263155.2], *Triticum uratu* [GCA_000347455.1], *Zea mays* [GCF_000005005.2], and one outgroup species, *Arabidopsis thaliana* [GCA_000001735.2]) were selected. To identify gene families, the ORTHOFINDER (version 2.3.14) pipeline (Emms & Kelly, 2019) was sequentially applied to the 10 genomes with all-to-all BLASTP (E-value ≤ 1e−5), reciprocity best hit, pairs connected by orthology and in-paraolgy, normalize the E-value and cluster pairs by ORTHOFINDER. Finally, genes were classified into orthologues, paralogues and single-copy orthologues (only one gene in each species). To construct the phylogenetic tree, single-copy orthologous genes were used; each gene family nucleotide sequence was aligned using MAFFT (Dm, 2013), and the alignments were curated with GBLOCKS (version 0.91b; Castresana, 2000). Then, the alignment (four-fold degenerate positions) used to compute the tree and infer the divergence dates with the optimal model and 2,000 bootstrap replicates using IQ-TREE (version 1.6.12; http://www.iqtree.org/). Finally, MCMCTREE in PAML (version 4.9e) was used to estimate the divergence times and 95% confidence intervals (CIs) of elephant grass and other plants (Yang, 1997). Three fossil calibration times were obtained from the TimeTree database (http://www.timetree.org/), including the divergence times of *A. thaliana* (148–173 million years ago [MYA]) and *O. sativa* (40–53 MYA).

CAFE version 4.0.1 (http://sourceforge.net/projects/cafehahnlab/) was used to detect the expanded and contracted gene families according to the ORTHOFINDER gene family results with the default parameters (Emms & Kelly, 2019).

Synonymous (Ks) and nonsynonymous (Ka) substitution rates were then estimated using PAML codon substitution models and likelihood ratio tests (codeml) based on the branch site model (Kimura, 1980). The likelihood ratio test (LRT) of the *p*-values was used to further verify the significant genes under positive selection.

2.6 | Whole-genome duplication and synteny analysis

The four-fold synonymous third-codon transversion (4DTv) estimation was applied to detect whole-genome duplication (WGD) events in elephant grass. First, the protein sequences of elephant grass, *C. americanus* and *S. italica* were aligned against seft with BLASTP (E-value ≤ 1e−10; https://blast.ncbi.nlm.nih.gov/Blast.cgi). Then, the collinear blocks of these plants were identified with MCSCANX (Tang et al., 2008). The WGD events in each plant species were evaluated based on their 4DTv distribution (Xu et al., 2011). The synteny blocks between the elephant grass and *C. americanus* genome were identified and represented by MINIMAP2 (version 2.17) with the –cx asm5 parameter (Li, 2017).

2.7 | Genome-wide expression dominance analysis and phylogenetic analyses of genes

Protein-coding genes from the two subgenomes of elephant grass were clustered by ORTHOFINDER with default parameters. On the best reciprocal BLAST matches between the A and B subgenomes of elephant grass, we identified 7,809 single-copy genes that had a 1:1

correspondence across the two homologous subgenomes. To investigate the expression dominance of these genes from the two subgenomes, we calculated the FPKM (fragments per kilobase of transcript per million) values of the homologous genes in leaf, shoot, stem tip, root and flower. We identified differentially expressed genes (DEGs) using the DESEQ2 (version 3.11) software. We filtered the DEGs with a minimum of two-fold differential expression (|log2 A vs. B FPKM |>1) and a significant p_{adj} value ($p_{adj} < 0.05$) in DESEQ2. Then, we performed GO analysis and KEGG enrichment of the DEGs. The sequence alignment and phylogenetic tree construction was performed and illustrated in IQ-TREE (version 1.6.12) with optimal model (2,000 bootstraps) and FIGTREE (version 1.4.3), respectively.

3 | RESULTS AND DISCUSSION

3.1 | Genome sequencing and assembly

To estimate the elephant grass genome size and heterozygosity, 102.9 Gb of Illumina clean reads were used for K-mer analysis (Table S1). The results showed that the number of modified 17-mers and the peak depth were 90,669,357,231 and 46, respectively (Figure S1; Table S2). The estimated genome size and heterozygosity rate were calculated to be 1,971,072,983 bp and 1.5%, respectively, which is consistent with previous studies (Wang et al., 2018).

A total of 290.9 Gb of data were obtained from three ONT flow cells, among which 203.6 Gb of data were collected after filtering (~100 × coverage; Table S3). Average read length and N50 length were 21.3 kb and 28.2 kb, respectively (Table S3). NEXTDENOVO and SMARTDENOVO dramatically improve the assembly quality and reduce computing resource usage by combining correction and assembly and were used to correct reads and perform the assembly, respectively. The initial genome (G1) obtained from these Nanopore data was 2,281,167,711 bp in length, which was larger than the genome size estimated via K-mer analysis (Table 1). To improve Nanopore sequencing read-level accuracy, an assembly polishing approach was applied, in which PILON used Illumina short reads to correct ONT reads. As shown in Table 1, the polished genome (G2) was 2,318,534,166 bp in size (Table 2), which was also larger than the genome size estimated via K-mer analysis. Regarding the 1.5% heterozygosity within the elephant grass genome, there were some redundant sequences according to the genome assembly. To obtain the nonredundant genome (G3), these redundant sequences were removed from G2 using the REDUNDANS software. Finally, we obtained a genome of 1,966,924,190 bp containing 2,059 contigs, which was close to the genome size estimated via 17-mer analysis (1,971,072,983 bp), suggesting that the nonredundant genome was appropriate. The contig N50 length and the longest contig length of the resulting assembled genome were 1.83 Mb and 15.1 Mb, respectively (Table 1).

Elephant grass has 14 pairs of chromosomes based on cytological observations (Figure S2). The 236.8 Gb of data obtained from the Hi-C library for scaffold extension at the chromosome level were used for the genome assembly (Table S4). A total of 96.65% (1,901,035,793 bp) of the total contig bases (1,966,924,190 bp) were anchored and oriented to the 14 chromosomes, with a contig N50 of 1.97 Mb and a chromosome N50 of 150 Mb (Table 1; Figure S3 and Table S5). To verify the Hi-C assembly, we mapped a composite genetic linkage map of the elephant grass (Paudel et al., 2018) to our assembly and found that the genetic map supports chromosomal assignment and order (Figure S4). Chromosome length ranged from 66.0 Mb to 199.1 Mb (Table S6). The genome exhibited a relatively high GC content (46.95%; Table 1; Figure S5), which is near the upper limit of the range reported in monocots (33.6%–48.9%; Smarda et al., 2014). A high GC content is reported to be associated with plant adaptation to abiotic stress (Costa et al., 2017). Furthermore, the completeness of the genome assembly was assessed with the BUSCO and CEGMA databases and RNA-seq data. The CEGMA assessment revealed that 93.9% of proteins were completely present and 96.7% of proteins were partially present, while the BUSCO data set indicated 97.8% complete and 12.4% fragmented Viridiplantae BUSCOs, with only 0.43% missing, which was made up of the nonredundant genome sequences (Tables S7 and S8). Assembly base accuracy was also assessed based on Illumina short read mapping. In total, 90.15%

TABLE 1 Genome assembly statistics and postprocessing of elephant grass

Type	G1	G2	G3	G3 + Hi-C
Total assembly size of contigs (bp)	2,281,167,711	2,318,534,166	1,966,924,190	
Number of contigs	2,455	2,455	2,059	
Contig N50 (bp)	1,997,163	2,031,655	1,829,308	
Longest contig (bp)	14,778,734	15,071,384	15,071,384	
Total assembly size of scaffolds (bp)				1,901,035,793[a]
Chromosome N50 (bp)				150,585,890
Chromosome numbers				14
GC content				46.95%

[a]Unanchored contig base count is not included.

of the clean reads were mapped to the genome assembly (Table S9). All of these evaluations indicated the high completeness, high continuity and high base accuracy of the present genome assembly.

3.2 | Genome annotation

The elephant grass genome contained 66.32% repetitive sequences, which was similar to the proportion of repeat elements found in the 1.58-Gb *Cenchrus americanus* genome (~77.2%) and greater than the proportions in the ~400-Mb *Setaria italica* (~46%) and 466-Mb *Oryza sativa* (~42%) genomes (Bennetzenetal.,2012; Varshney et al., 2018; Yu et al., 2001; Table 2). Among these repeat sequences, retroelements and DNA transposons accounted for 55.29% and 7.19% of the genome, respectively. In a similar way to the pattern in many other plant genomes, long terminal repeats (LTRs) were the most abundant repeat class and comprised 53.22% of the elephant grass genome (Figure 2 and Table 2). The Gypsy and Copia superfamilies were the dominant types (35.01% of the elephant grass genome).

TABLE 2 Summary statistics of annotated repeats

Type	Number	Length (bp)	Percentage of genome (%)
Retroelements	1,233,887	1,087,594,215	55.29
LINEs	56,423	38,660,338	1.97
SINEs	6,183	2,039,450	0.1
LTR elements	1,171,281	1,046,894,427	53.22
Gypsy	463,658	513,499,133	26.11
Copia	192,858	175,104,557	8.9
DNA transposons	433,671	141,468,364	7.19
Unknown	329,551	74,844,241	3.81
Total	1,997,109	1,303,906,820	66.29

For genome annotation, a total of 65,927 protein-coding genes were identified in the elephant grass genome (Table 3). Compared with other published Poaceae genomes, the number of genes in elephant grass is greater than that in *C. americanus*, *S. italica*, *S. bicolor* and *Z. mays* (Bennetzen et al., 2012; A. Paterson et al., 2009; Varshney et al., 2018; Yu et al., 2001). The average length and average intron length of the elephant grass genes were 4.1 kb and 0.65 kb (5.33 exon per gene), respectively (Table 3). For the completeness of protein-coding genes, 97.1% and 0.9% of the "total complete BUSCOs" and "fragmented BUSCOs" were identified by BUSCO annotation, respectively (Table S10). A total of 64,630 (98.03%) protein-coding genes were assigned functions, and 86.65% and 79.86% of these genes exhibited homology and conserved protein domains in the Swiss-Prot and InterProScan databases respectively (Table S11). Most of the genes were annotated with the non-redundant protein sequence database (NR; 84.35%), and 52.15% of the genes were classified according to GO terms, with

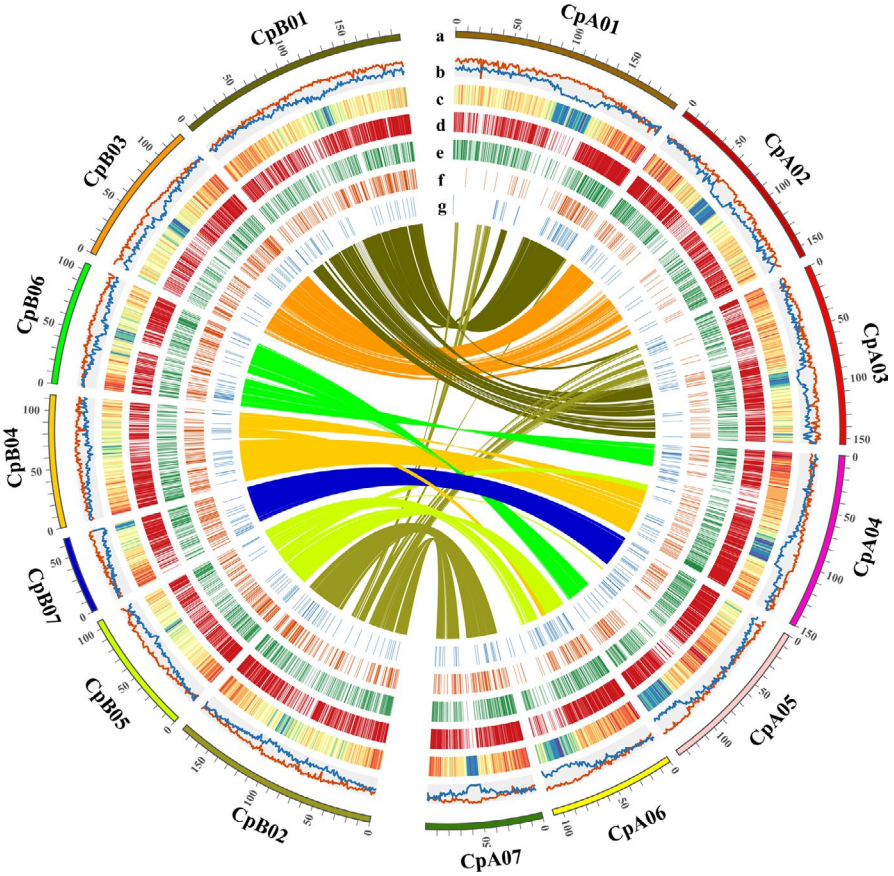

FIGURE 2 Genomic landscape of the 14 assembled elephant grass chromosomes. (a) Fourteen assembled elephant grass chromosomes. (b) Gene (orange) and repeat density (blue; orientation is inward). (c) LTR density (1-Mb nonoverlapping windows; blue represents high-density regions). (d) Expanded gene locations in chromosomes. € Unique gene locations in chromosomes. (f) Contracted gene locations in chromosomes. (g) Positively selected gene locations in chromosomes. Central coloured lines represent syntenic links between A' and B subgenomes

TABLE 3 Summary statistics of protein-coding genes in elephant grass

	Cenchrus purpureus	*Cenchrus americanus*	*Setaria italica*
Number of genes	65,927	38,579	27,422
Average gene length (bp)	4,110.44	2,420.19	3,008.75
Average CDS length (bp)	1,294.61	1,014.71	1,335.69
Average exons per gene	5.33	4.09	5.16
Average exon length (bp)	243.09	248.06	259.03
Average introns per gene	4.33	3.09	4.16
Average intron length (bp)	650.97	454.77	402.52

31.8% being mapped to known plant biological pathways based on the KEGG pathway database (Table S11). In addition, we predicted 2,058 tRNAs, 349 rRNAs, 5,635 snRNAs and 2,543 miRNAs in the elephant grass genome (Table S12).

3.3 | Comparative genomics, genome evolutionary and WGD analysis

We clustered the annotated genes into gene families among elephant grass and eight other Poaceae species with *Arabidopsis* as the outgroup using ORTHOFINDER. A total of 743 single-copy genes were identified among the 10 species, which were used for phylogenetic tree construction (Table S13). The results suggested that elephant grass (Cp) is related to *C. americanus* (Ca), *Dichanthelium oligosanthes* and *Setaria italica* (Si), which belong to the tribe Paniceae (Figure 3a). We also found that elephant grass diverged phylogenetically from Ca ~ 3.22 (1.71–5.66) MYA after the divergence of Si (genus *Setaria*) at 10.44 (6.85–16.35) MYA (Figure 3a). The divergence times were close to the time at which Ca diverged phylogenetically from Si (~15 MYA; Varshney et al., 2018). The allotetraploid elephant grass A'A'BB genome originated ~6.61 (4.11–10.92) MYA.

We identified homologous gene pairs in Cp, Ca and Si and estimated species divergence times using the 4DTv distance. The results indicated that all gene pairs showed a shallow peak at 0.38, probably reflecting the rho (ρ) WGD event (Paterson et al., 2004; Figure 3b). We found that the recent tetraploidization of Cp occurred based on the 4DTv and Ks results (Figure 3b; Figure S6).

To determine the chromosome structure in Cp, we performed an intragenome synteny analysis. A large number of synteny blocks existed between pairs of Cp chromosomes based on homologous genes; for example, chromosomes 08 (CpB03), 03 (CpB02), 11 (CpB05) and 09 (CpB04) corresponded closely to chromosomes 04 (CpA02), 13 (CpA07), 10 (CpA06) and 06 (CpA04), respectively (Figure 2). The results also showed that chromosome 01 (CpA01) and chromosome 05 (CpA03) shared syntenic regions with chromosome 02 (CpB01). Likewise, chromosome 12 (CpB06) and chromosome 14 (CpB07) shared syntenic regions with chromosome 07 (CpA05). These chromosomes exhibited a higher level of structural variation which strongly suggests the existence of subgenomes in Cp. Previous studies have indicated that the chromosomes in the A' genome of Cp ($2n = 4x = 28$, A'A'BB) are homologous to those of the A genome of Ca ($2n = 2x = 14$, AA; Gupta & Mhere, 1997). To investigate the relationship between Cp and Ca, we conducted a phylogenomic study between the genomes. A total of 34,377 pairs of collinear genes were identified between Cp and Ca. We found that some Ca chromosomes corresponded to a pair of Cp chromosomes based on considerable collinearity (Figure 4a; Figure S7a). For example, chromosomes 03 (CpB02) and 13 (CpA07) corresponded to chromosome 6 of Ca (Ca06). However, some Ca genomic regions presented more than two corresponding regions in the Cp genome. For example, chromosome 3 of Ca (Ca03) exhibited regions corresponding to chromosomes 07 (CpA05), 12 (CpB06) and 14 (CpB07) of Cp (Figure 4a; Figure S7a). Chromosome 05 (CpA03) of Cp contained regions corresponding to chromosomes 1 and 4 of Ca (Ca01 and Ca04). These results indicate that possible chromosomal rearrangements had occurred in elephant grass, which is consistent with a previous study (Paudel et al., 2018). Compared to other chromosomes of Cp, chromosomes 01, 04, 05, 06, 07 and 10 had a higher number of collinear genes between Cp and Ca. We further clarified the homologous genome sequences (synteny blocks) using collinear analysis. Most of chromosomes 01, 04, 05, 06, 07, 10 and 13 of Cp were collinear with chromosomes 01, 02, 03, 04, 05, 06 and 07 of Ca respectively (Figure S8). We thus assigned and denoted the assembled seven chromosomes as A01–A07 and other chromosomes as B01–B07. The phylogenetic tree suggested that the A' genome of Cp was related to Ca, which provides further evidence that the A' genome of Cp is homologous to the A genome of Ca. To verify these results, we selected Si ($2n = 2x = 18$), which is closely related to Cp, to perform a collinearity analysis. We found a total of 36,753 pairs of collinear genes between Cp and Si (Figure S7b). Based on the results of the synteny analysis, we also identified the subgenomes of Cp, which supports the results of the analysis of Cp and between Cp and Ca.

3.4 | Comparative genomics of gene families

Based on sequence homology among 10 plant species, we assigned 28,520 genes from subgenome A' and 34,360 genes from subgenome B of elephant grass to 15,681 and 16,716 families, respectively. A total of 365 gene families (222 in subgenome A'; 143 in subgenome B) were unique to the A' and B subgenomes of elephant grass (Table S14), respectively. Furthermore, we selected four other Paniceae species to identify unique and shared gene families. The results showed that a total of 10,281 gene clusters were shared by

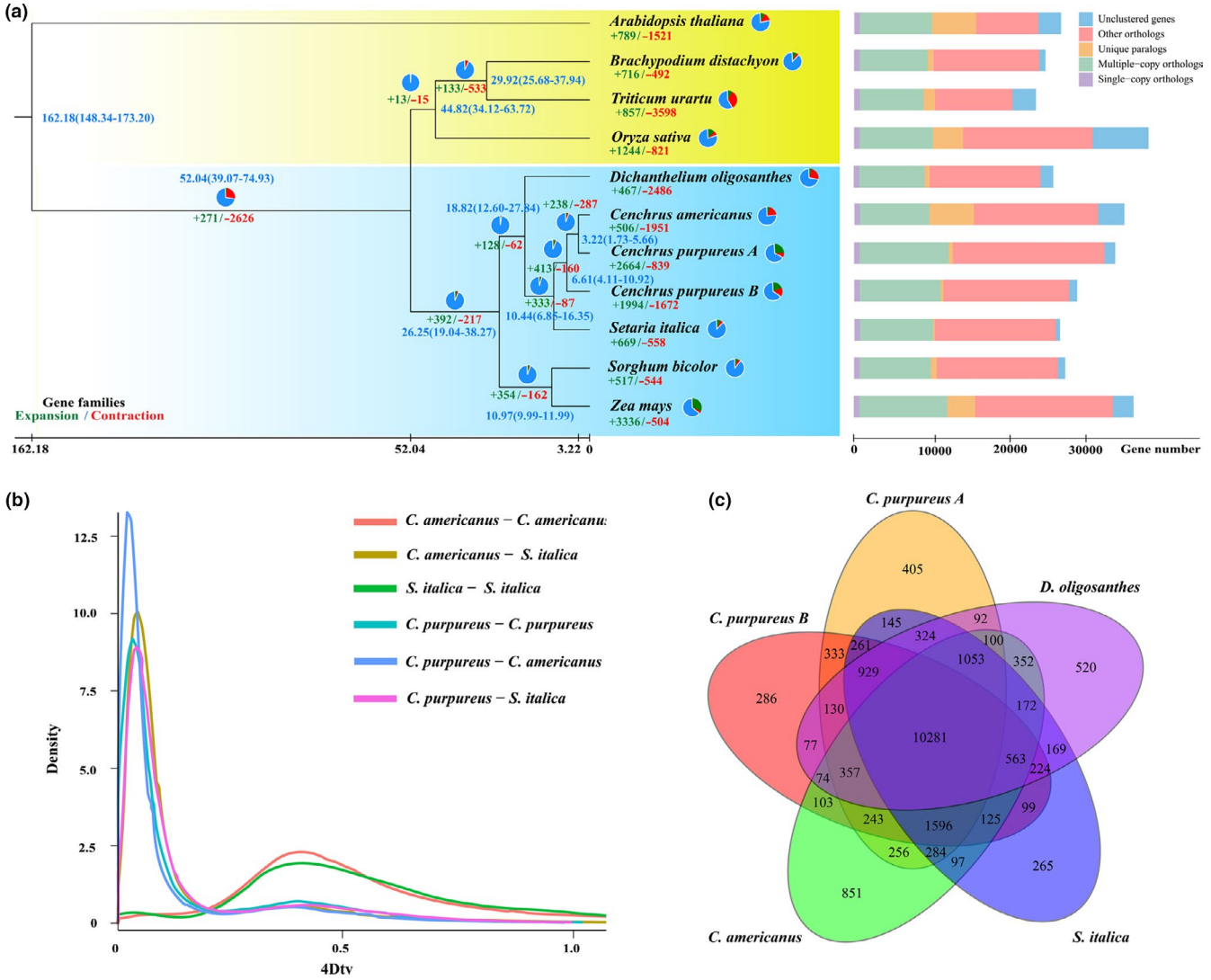

FIGURE 3 Evolution of the elephant grass genome. (a) Phylogenetic relationship of elephant grass with nine other plant species. C_4 species are shown with blue background and C_3 species with yellow background. Divergence times are labelled in blue; gene family expansion and contraction are enumerated below the species names in green and red; gene categories used from all the species are shown on the right. (b) Distribution of 4DTv distance between syntenic orthologous genes. Ca, *Cenchrus americanus*; Cp, *Cenchrus purpureus*; Si, *Setaria italica*. (c) Venn diagram of shared orthologous gene families in four species. The number of gene families is listed for each component

the four species, and 405 and 286 elephant grass-specific clusters were identified in the A' and B subgenomes, respectively (Figure 3c).

The expansion and contraction of gene families is thought to be important in adaptive phenotypic diversification (Hahn et al., 2005). Based on sequence homology, we identified 2,663 (1,994 in subgenome A'; 669 in subgenome B) and 2,512 (839 in subgenome A'; 1,673 in subgenome B) gene families showing expansion or contraction, respectively, after divergence from *C. americanus* (Figure 2; Table S15; Figure 3a). GO enrichment analysis revealed that the expanded genes of subgenome A' were significantly enriched ($p < 0.05$) in functions associated with the regulation of organ growth, cotyledon morphogenesis, wax biosynthesis process and auxin biosynthesis process (Table S16). However, the GO terms of the biosynthesis processes associate with auxin, abscisic acid and other phytohormones were found in expanded genes of subgenome B (Table S17). KEGG analysis showed that the expanded genes of subgenome A' were involved in plant hormone signal transduction, glycolysis/gluconeogenesis, and flavone and flavonol biosynthesis (Figure S9a). Anthocyanin biosynthesis, plant–pathogen interaction and nitrogen metabolism were uniquely enriched in the expanded genes of subgenome A'.

In addition to the expansion and contraction of gene families, genes showing positive selection commonly contributed to adaptive phenotypic evolution and adaptation. A total of 432 (248 in subgenome A'; 184 in subgenome B) positively selected genes were identified in the elephant grass genome compared to those of other species. We performed a KEGG enrichment analysis of these positively selected genes in subgenome A' and identified some KEGG pathways that were significantly enriched, which were related to starch and sucrose metabolism, the phosphatidylinositol signalling

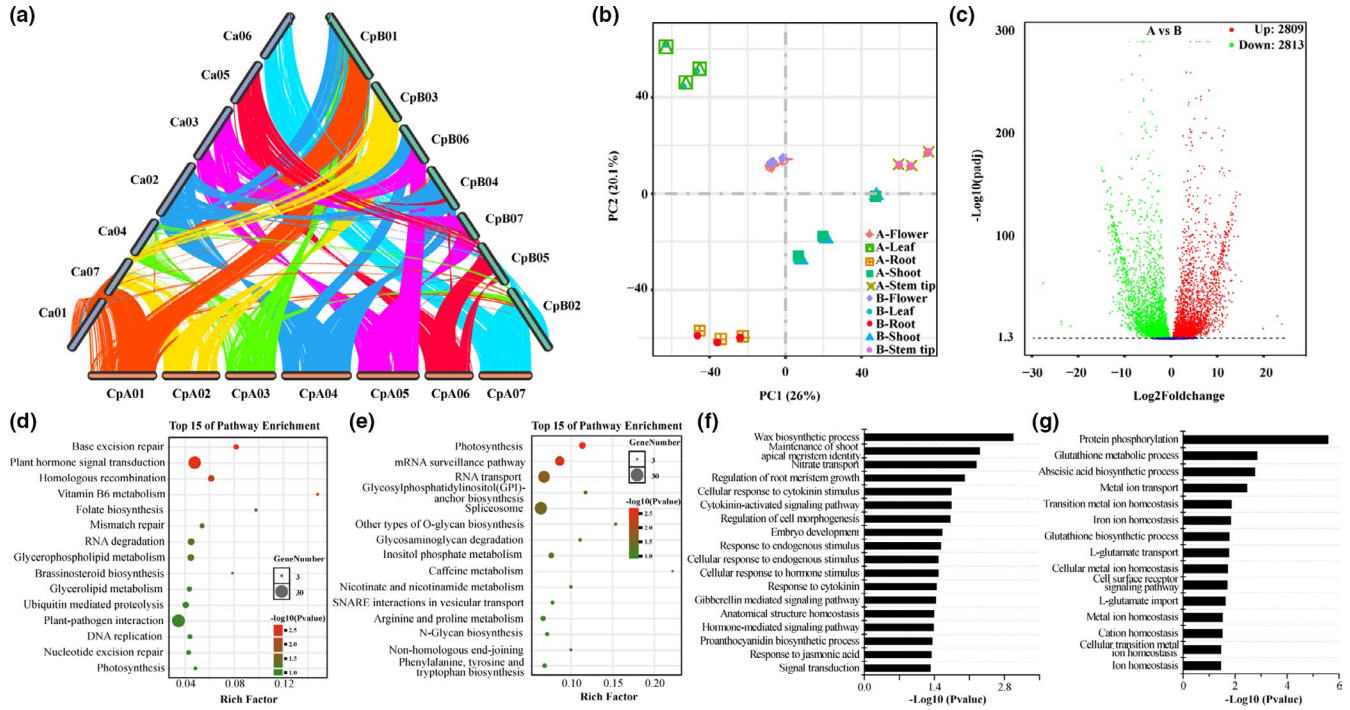

FIGURE 4 Characterization of expression dominance in subgenomes of elephant grass. (a) Syntenic analysis of the elephant grass (Cp) and the pearl millet (Ca) genomes. Ca are pearl millet chromosomes, Cp are elephant grass chromosomes. (b) Principal components analysis of the expression of single-copy genes between the two subgenomes in five tissues. (c) Volcano plot of differentially expressed genes between the two subgenomes (subgenome A' vs. subgenome B) in five tissues. (d,e) Enriched KEGG pathways of up-regulated DEGs and down-regulated DEGs, respectively. (f,g) Enriched GO terms of down-regulated DEGs and up-regulated DEGs, respectively

system, and steroid biosynthesis (Figure S9e). Furthermore, glutathione metabolism, nitrogen metabolism, ubiquitin-mediated proteolysis, and pentose and glucuronate interconversions were particularly represented in the positively selected genes of subgenome B (Figure S9f).

Furthermore, 1,946 and 1,384 elephant grass genes were found to belong to unique families in subgenome A' and subgenome B, respectively. Compared to subgenome B, unique genes in subgenome A' were involved in flavone and flavonol biosynthesis, oxidative phosphorylation, and phagosome (Figure 2 and 3c; Figure S9c). Genes associated with the ribosome and spliceosome pathways were significantly enriched in subgenome B (Figure S9d).These positively selected, unique and expanded genes might contribute to the adaptability of elephant grass. In addition, 247 genes from contracted gene families of subgenome A' were enriched in some KEGG pathways, including ribosome and monoterpenoid biosynthesis (Figure S9g). Interestingly, common pathways were also identified in subgenome A' of elephant grass (Figure S9i).

3.5 | Genome-wide expression dominance

To investigate homologous genome dominance, we conducted transcriptional analyses of homologous genes in different tissues of elephant grass, including the leaf, shoot, stem tip, root and flower (Table S9). The homologous gene expression levels of the two subgenomes were similar in different tissues, which suggested that there was no significant genome dominance (Figure 4b). To further investigate the potential of homologous genome dominance, we extracted 7,809 single-copy genes from the two elephant grass subgenomes. We analysed the genes expressed from the two subgenomes and defined the genes with a greater than two-fold difference in expression within a subgenome as DEGs. Of 7,809 genes, 64.6% were differentially expressed (Table S18). Among them, 2,813 and 2,109 genes were up-regulated and down-regulated in subgenome A' compared to subgenome B, respectively (Figure 4c; Table S18). Furthermore, the numbers of DEGs were similar among the examined tissues (Table S18).

We further analysed the gene function of DEGs between subgenome A' and subgenome B by KEGG analysis. The results indicated that genes involved in plant hormone signal transduction, homologous recombination and base excision repair had a higher level of expression in subgenome B than in subgenome A', whereas the photosynthesis and mRNA surveillance pathways were more represented in subgenome A' (Figure 4d,e). GO annotation showed that those significantly up-regulated DEGs were mainly enriched in protein phosphorylation, glutathione metabolic and biosynthesis processes, and ion transport (Figure 4g). The highly expressed genes from subgenome A' were mainly associated with development, hormone signalling and response to stimulus, including the regulation of root meristem growth, regulation of cell morphogenesis, signal transduction, hormone-mediated

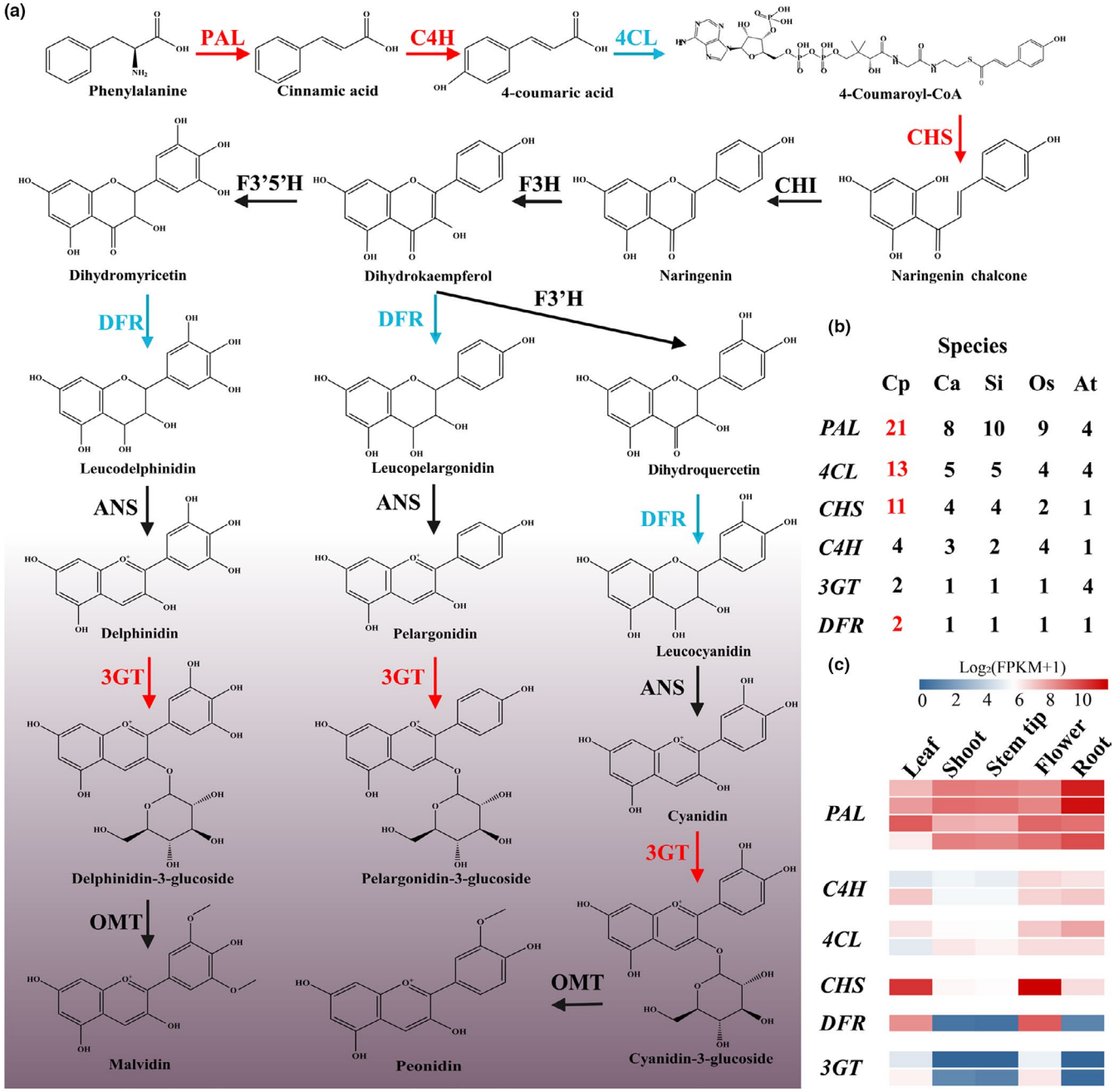

FIGURE 5 A schematic presentation of the phenylpropanoid pathway and flavonoid biosynthetic pathway leading to anthocyanins in elephant grass. (a) Diagram depicting the main genes and metabolic pathway involved in anthocyanin accumulation. Expanded genes are shown in red, and both expanded and positively selected in blue. Enzyme abbreviations: PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase, 4CI, 4-coumarate CoA ligase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-hydroxylase; F3′H, flavonoid 3′-hydroxylase; F3′5′H, flavonoid 3′5′-hydroxylase; DFR, dihydroflavonol 4-reductase; ANR, anthocyanidin reductase; ANS, anthocyanidin synthase; 3GT, anthocyanidin 3-*O*-glucosyltransferase; OMT, *O*-methyl transferase. (b) Gene copy number of key genes involved in anthocyanin accumulation in elephant grass (Cp), *Cenchrus americanus* (Ca), *Setaria italica* (Si), *Oryza sativa* (Os) and *Arabidopsis thaliana* (At). Gene copy numbers that are at least two-fold higher in elephant grass than in other species are labelled in red. (c) Heatmap showing the expression level of candidate genes involved in anthocyanin accumulation in different tissues of elephant grass

signalling pathway and response to endogenous stimulus (Figure 4f). These results indicate that the complementary functions of subgenomes play an important role in improving the development and adaptation of elephant grass, which was also found in other allopolyploid plant species (Grover et al., 2012; Yoo et al., 2013).

3.6 | Characterization of putative genes in the anthocyanidin biosynthesis pathway

To evaluate the potential of elephant grass for the genetic dissection of agriculturally important traits, we focused on the purple phenotype

of the leaves, which is associated with increased forage quality and may improve the growth and productivity of animals (Bariexca et al., 2019). The phenylpropanoid, flavonoid and anthocyanin biosynthesis pathways are the three major biosynthetic pathways related to leaf pigmentation (Jaakola, 2013). The genes encoding the key enzymes involved in the phenylpropanoid pathway were expanded in the "Purple" elephant grass genome, including phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H) and 4-coumarate CoA ligase (4Cl; Figure 5a). The number of *PAL* and *4Cl* genes in elephant grass (Cp) was significantly higher than in *C. americanus* (Ca), *S. italica* (Si), *O. sativa* (Os) and *Arabidopsis* (At; Figure 5b). We identified 21 copies of *PAL*, four copies of *C4H* and 13 copies of *4Cl* in Cp. In addition, the *4Cl* family included positively selected genes. In this pathway, chalcone synthase (CHS) plays a core role in naringenin chalcone biosynthesis by condensing 4-coumaroyl-CoA, produced via the phenylpropanoid pathway (Gao et al., 2019; Figure 5a). Interestingly, we found that the *CHS* gene family was expanded and contained 11 copies in Cp, the number of which was higher than the four copies identified in Ca and the one copy in At (Figure 5b). Flavanone 3-hydroxylase (F3H), flavonoid 3′-hydroxylase (F3′H) and flavonoid 3′,5′-hydroxylase (F3′5′H) can catalyse the production of dihydroflavonols, which are precursors of leucoanthocyanins (Cao et al., 2018). Dihydroflavonol 4-reductase (DFR) was the first enzyme shown to produce leucoanthocyanins (Zhu et al., 2018), and *DFR* genes have been expanded and positively selected in Cp (Figure 5a,b). Anthocyanidins are converted from leucoanthocyanins by anthocyanidin synthase (ANS) and further glycosylated by anthocyanidin 3-*O*-glucosyltransferase (3GT; Zhang et al., 2016). In Cp, the *3GT* family was also identified as an expanded family and had two copies (Figure 5b).

We identified candidate genes involved in the anthocyanidin biosynthesis pathway in elephant grass based on their expression in different tissues (Figure S10). The result indicated that the transcript levels of one candidate *CHS* (*CpA0101784*) were over 14- and 21-fold higher in leaf than in root and shoot (Figure 5c). We also found that key genes involved in anthocyanidin metabolism, such as two *DFR* genes (*CpB0203538* and *CpA0702380*) and two *3GT* genes (*CpA0402271* and *CpB0403234*), were more highly expressed in leaves and flowers of elephant grass. *PAL*, *C4H* and *4Cl* were expressed at similar levels in tissues of elephant grass (Figure 5c; Figure S11). In addition, the leaves of the elephant grass "Purple" cultivar are purple, while the whole plant of the "Mott" cultivar is green. The results of comparative metabolome analysis indicated that "Purple" exhibits higher malvidin, peonidin and pelargonidin levels than "Mott." Expression of the *CHS*, *ANS*, *DFR* and *4Cl* genes was also found to be significantly different between the "Purple" and "Mott" cultivars (Zhou et al., 2019). These observed results supported our results and suggested that the high copy number and high expression levels of key genes involved in the anthocyanidin biosynthesis pathway drive the production of these major anthocyanidin compounds in the "Purple" cultivar.

Location analysis showed that some gene families were unevenly represented on the two subgenomes or rearranged chromosomes of Cp (Figure S11). For example, the number of *4Cl* and *CHS* genes in subgenome A' was over two-fold higher than in subgenome B and 18 of 21 *PAL* genes were distributed in the rearranged chromosomes A05, B06 and B07 (Figure S11). We further performed phylogenetic analyses on these expanded genes using the coding sequences from elephant grass (Figure S10). The results indicated that all of the expanded genes come from different clades. The *PAL* elephant grass genes were divided into two clades, and the most highly expressed genes were from the first clade (Figure S10). Furthermore, five candidate *4Cl* elephant grass genes represented three major lineages. However, all of the highly expressed genes were found to be in clade 2 (Figure S10).

3.7 | Genes involved in C_4 photosynthesis

C_4 plants are typically more efficient in carbon fixation and have higher water-use efficiency, contributing to their ability to survive in drier environments. C_4 plants can be divided into three subtypes based on different decarboxylation enzymes in the bundle sheath (BS) cells, including nicotinamide adenine dinucleotide-dependent malic enzyme (NAD-ME), nicotinamide adenine dinucleotide phosphate-dependent malic enzyme (NADP-ME) and phosphoenolpyruvate carboxykinase (PEPCK; Figure 6a). In the NADP-ME subtype, malate (Mal), converted from oxaloacetate (OAA) by NADP-dependent malate dehydrogenase (NADP-MDH) in the chloroplasts of mesophyll (M) cells, is the main metabolite transported from M cells to BS cells; in the chloroplast of BS cells, Mal is decarboxylated by NADP-ME (Rao & Dixon, 2019). However, the main metabolite which is transported from M cells to BS cells is OAA, which is converted to Mal in the mitochondria of BS cells; Mal is further decarboxylated by NADP-ME (Figure 6a). In addition, OAA could also be directly decarboxylated by PEPCK in the BS cytosol, which is defined as the PEPCK subtype (Rao & Dixon, 2019).

We analysed the nine main gene families involved in C_4 carbon fixation, including enzymes and metabolite transporters, and found that they were expanded in elephant grass and have a higher copy number in Cp than in Ca, Si and Os (Figure 6b). For example, the *NAD-MDH* gene is strongly expanded and has eight copies in Cp, while there is only one copy in Ca and three copies in Si. The *NADP-ME*, *NAD-MDH* and *NAD-ME* genes are also strongly expanded in Cp compared to the other species (Figure 6b). We further identified genes coding for candidate enzyme involved in carbon fixation in elephant grass based on gene expression in five tissues (Figure 6c; Figure S12). Transcript levels of three candidate *NADP-MDH* genes (*CpB0102961*, *CpA0301816* and *CpB0301137*) were over 74-, 284- and 1550-fold higher in the leaves than in the roots, respectively (Figure 6c; Figure S12). Similar results were also reported for the analysis of C_4 genes in broomcorn millet (Zou et al., 2019). For the NADP-ME subtype, four candidate *NADP-ME* genes (*Cp0001672*, *CpA0704015*, *CpB0200469* and *Cp0001733*) were more highly expressed in the leaves. Transcript levels of one candidate *PEPCK* gene was over 2,000- and 80-fold higher in the leaves than in the roots and flowers, respectively. We also found that the three candidate *NAD-MDH* genes of the NAD-ME subtypes (*CpA0405043*, *CpB0400697* and *CpB0400712*) were more highly expressed in the

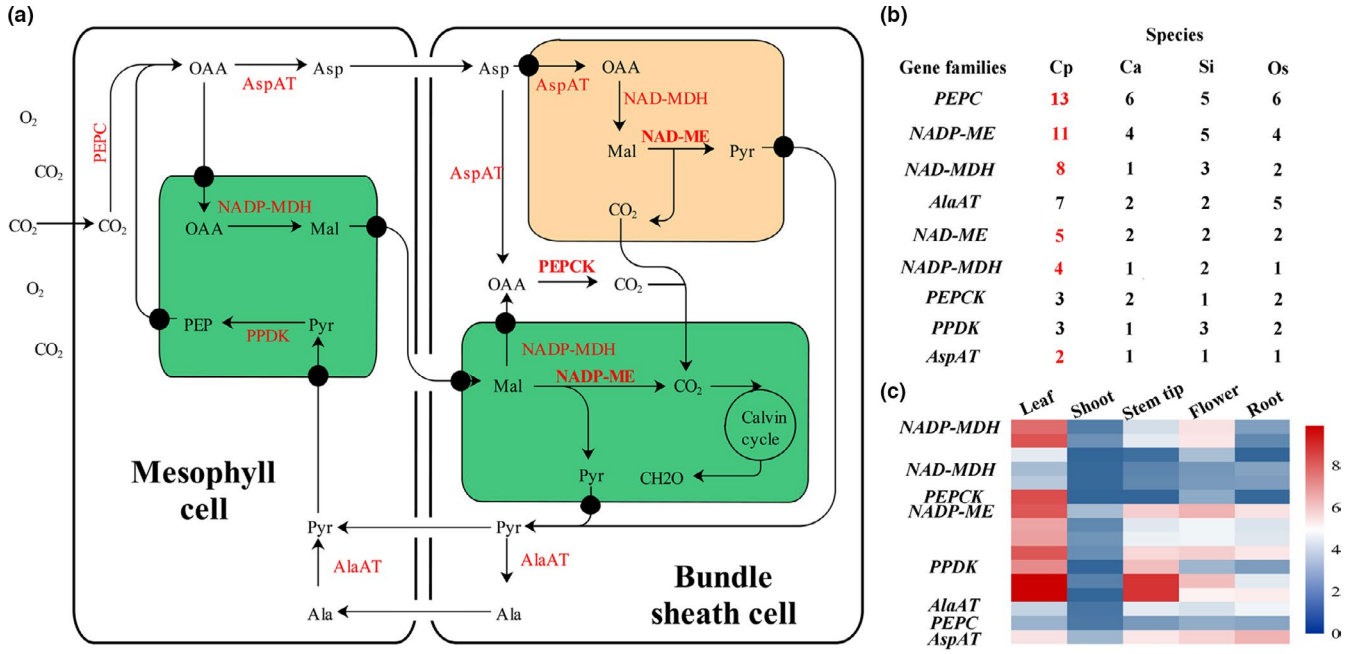

FIGURE 6 A proposed model of C_4 photosynthesis in elephant grass. (a) Diagram depicting the main proteins and metabolic fluxes involved in C_4 photosynthesis. Expanded genes are shown in red. Chloroplast and mitochondria are shown in green and brown, respectively. Abbreviations for metabolites and enzymes: CO_2, carbon dioxide; Ala, alanine; Asp, aspartate; Mal, malate; Pyr, pyruvate; OAA, oxaloacetate; PEP, phosphoenolpyruvate; PEPC, phosphoenolpyruvate carboxylase; PPDK, pyruvate/orthophosphate dikinase; AspAT, aspartate aminotransferase; AlaAT, alanine aminotransferase; NADPMDH, NADP-dependent malate dehydrogenase; NADP-ME, NADP-dependent malic enzyme; NAD-MDH, NAD-dependent malate dehydrogenase; NAD-ME, NAD-dependent malic enzyme; PEPCK, phosphoenolpyruvate carboxykinase. Metabolite transporters are presented by a dark circle. (b) Gene copy number of key genes involved in C_4 photosynthesis in elephant grass (Cp), *Cenchrus americanus* (Ca), *Setaria italica* (Si) and *Oryza sativa* (Os). Gene copy numbers that are at least two-fold higher in elephant grass than in other species are labelled in red. (c) Heatmap showing the expression level of candidate genes involved in C_4 photosynthesis in different tissues of elephant grass

leaves than in other tissues (Figure 6c; Figure S12). These results suggest a mixed C_4 model that contains features from the traditional subtypes in elephant grass. In broomcorn millet, key candidate genes of all three C_4 subtypes were also identified (Zou et al., 2019).

Analysis of gene location and synteny of C_4 genes showed that these genes were evenly distributed in homologous chromosomes of the two subgenomes, which suggests that the expansion was probably generated by the recent tetraploidization of elephant grass (Figure S13). We performed phylogenetic analyses on the C_4-related genes using the coding sequences from elephant grass. The results suggested that all of the C_4 genes represented different clades. For example, the *NADP-MDH* genes in elephant grass were divided into two clades, and all of the highly expressed genes were from the first clade (Figure S12). Furthermore, the 11 candidate *NADP-ME* genes represented two major lineages. Among them, all of the lowly expressed genes were from clade 1 (Figure S12).

3.8 | Expansion of genes involved in the plant hormone signal transduction pathway

Phytohormones are naturally occurring organic substances that govern every aspect of plant biological process at extremely low concentrations, including developmental processes, signalling networks, and responses to biotic and abiotic stresses (Muday et al., 2012). Phytohormones are traditionally classified into five major classes: auxins, cytokinins (CKs), gibberellins (GAs), ethylene and abscisic acid. In addition, other compounds such as brassinosteroids (BRs) have also been recognized as plant hormones.

Among the plant hormones, auxin has been shown to mediate cell enlargement and adventitious root development. The hormone signal transduction pathway, which included some key genes, plays an important role in the process. The auxin influx carrier (*AUX1*) gene encodes a component of the auxin influx carrier, which is involved in auxin transportation (Figure 7). Transport inhibitor response (TIR1) binds the transported auxin and degrades the auxin/indole-3-acetic acid (Aux/IAA) transcriptional repressor that improves the activity of auxin response factor (ARF) transcription factors (Ori, 2019). ARFs bind to the auxin-responsive *cis*-acting element in early auxin response genes including *Aux/IAA*, *Small auxin upregulated* (*SAUR*) and *Gretchenhagen-3* (*GH3*; Hage n& Guilfoyle, 2002; Figure 7). Overexpression *SAUR* genes is sufficient to induce cell elongation and growth (Stortenbeker & Bemer, 2018). The *gh3* mutants exhibit reduced lateral root number, and auxin-deficient traits in *Arabidopsis* (Zhang et al., 2007). We found that these key gene families of the auxin signal transduction pathway were expanded in elephant grass. For example, we discovered that the *AUX/IAA* gene family was strongly expanded in Cp (six copies), whereas there is

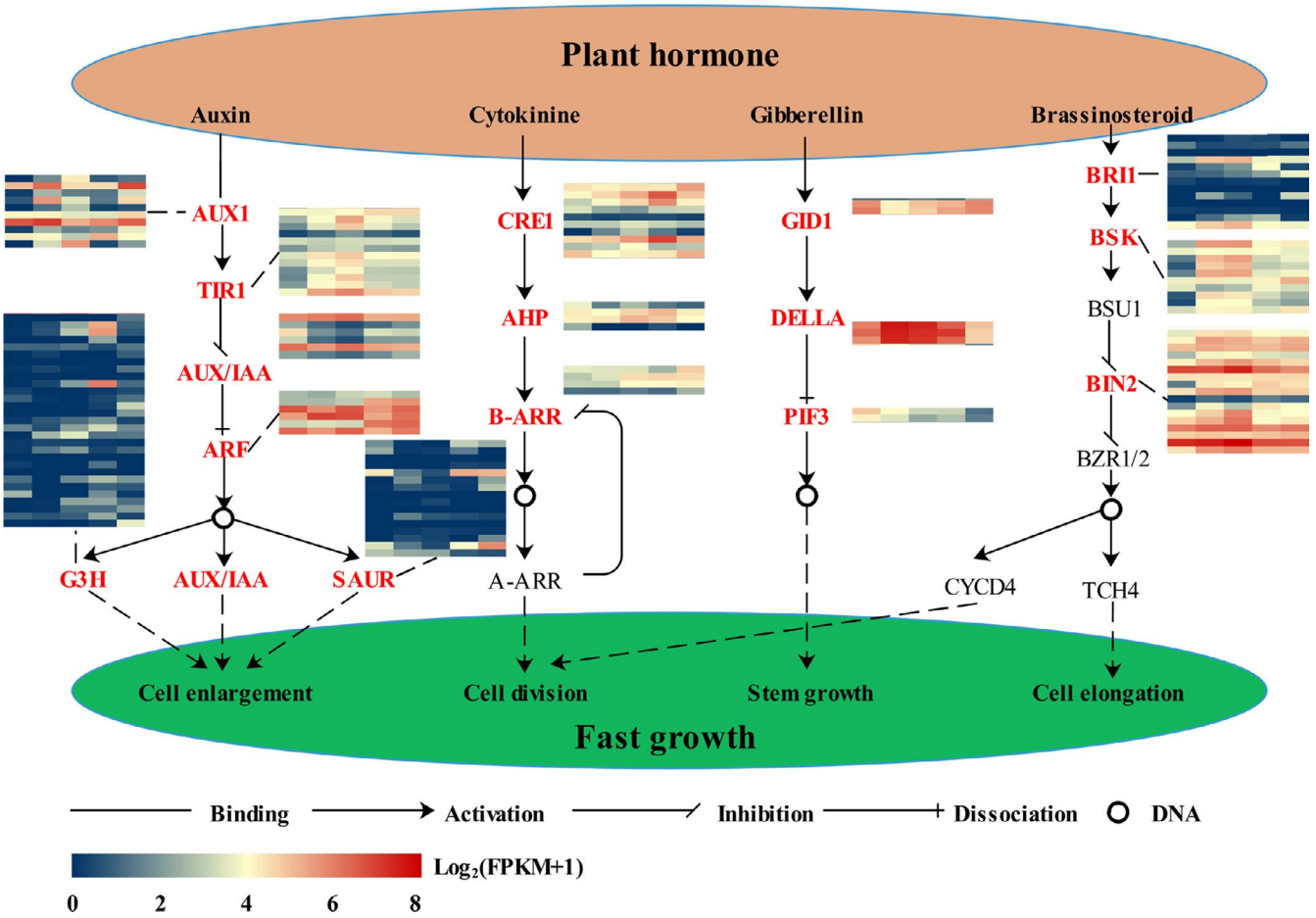

FIGURE 7 Plant hormone signal transduction pathway in elephant grass. A proposed model of the plant hormone signal transduction pathway in elephant grass. Expanded genes are shown in red. The heatmaps show the expression level of genes involved in the plant hormone signal transduction pathway in different elephant grass tissues. Columns and rows correspond to the tissues and gene copies, respectively (left to right represents leaf, shoot, stem tip, flower and root). Abbreviations: AUX1, auxin influx carrier; TIR1, transport inhibitor response; ARF, auxin response factor; Aux/IAAs, auxin/indole-3-acetic acid; SAUR, small auxin upregulated; GH3, Gretchenhagen-3; CRE1, cytokinin receptor 1; AHP, Arabidopsis histidine phosphotransfer proteins; BRR, B-type Arabidopsis response regulators; ARR, Arabidopsis response regulators; GID1, GA-insensitive dwarf 1; PIF3, phytochrome-interacting factor 3; BRI1, Brassinosteroid insensitive 1; BSK, Brassinosteroid -signalling kinases; BIN2, Brassinosteroid insensitive 2; BZR1/2, Brassinazole-resistant 1/2; BSU1 bril suppressor1 phosphatase; TCH4, Touch 4; CYCD4, Cyclin D 4. Heatmap showing the expression level of candidate genes involved in plant hormone signal transduction pathway in tissues of elephant grass (left to right represents leaf, shoot, stem tip, flower and root)

one copy in Ca and two copies in Si (Table S14). We also identified 16 copies of *SAUR*, 29 copies of *G3H* and 12 copies of *TIR1*. The analysis of transcript levels showed that these genes were highly expressed in tissues of Cp (Figure 7; Figure S14). In addition, we also found that some genes, such as *AUX1* (*CpA0602793*, *CpA0601023*) and *TIR1* (*CpA0104594*, *CpB0700127*), were highly expressed in the stem tip (Figure 7; Figure S14). CKs also play essential roles in many aspects of plant development through regulation of cell division (Artner & Benkova, 2019). The CK signal transduction pathway contains four signalling gene families, including *cytokinin receptor 1* (*CRE1s*), *Arabidopsis histidine phosphotransfer proteins* (*AHPs*), B-type *Arabidopsis response regulators* (*BRRs*) and *Arabidopsis response regulators* (*ARRs*; Wybouw & Rybel, 2018). We analysed the copy number of these genes and found that, with the exception of *ARRs*, there is a higher copy number in Cp than in Ca, Si and Os, the ratio of which was usually over two-fold (Table S19). We also found that most of these genes had high expression in Cp and some genes were more highly expressed in the stem tip, such as *CpB0400426* (*CRE1*) and *CpB0602799* (*CRE1*; Figure 7; Figure S14).

GAs are one of the plant hormones that stimulate plant development, including stem elongation and fertility. In GA signalling, *GA-insensitive dwarf1* (*GID1*) is a GA receptor, which further interacts with DELLA repressors. *DELLA* belongs to a subfamily of GRAS transcription factors, which are regulated by the expression of the *phytochrome-interacting factor 3* (*PIF3*; Nelson & Steber, 2016; Figure 7). In the Cp genome, the *DELLA*, *GID1* and *PIF3* families are also expanded and were expressed at similar levels in different tissues (Figure 7; Table S19). BRs are growth-promoting steroid hormones that regulate cell division and cell elongation. In the BR signal pathway, *Brassinosteroid insensitive 1* (*BRI1*), a member of the leucine-rich

repeat receptor-like kinases (LRR RLKs) gene family, perceives BRs at the cell surface and regulates the BR-signalling kinases (BSKs; Clouse, 2011). *Brassinosteroid insensitive 2* (*BIN2*), identified as a downstream negative regulator of the BR signal pathway, inhibited the activation of *Brassinazole-resistant 1/2* (*BZR1/2*; Yin et al., 2005). In this study, copy number determination revealed that *BRI1*, *BSK* and *BIN2* have a higher copy number in Cp than in Ca, Si and Os (Table S19). Expression analysis indicated that some *BSKs* were more highly expressed in the stem tip and shoot, such as *CpB0400240* and *CpB0402659* (Figure 7; Figure S14).

We further performed phylogenetic analyses on growth-related genes using the coding sequences from elephant grass. The results also suggested that these gene families represented different clades, especially BRI1, ARF and TIR1 (Figure S14). The elephant grass *ARF* genes were from two clades, and all of the highly expressed genes in all tissues were from the second clade (Figure S14). Furthermore, 13 candidate NADP-MEs came from two major lineages. Interestingly, two highly expressed genes in different tissues were from clade 2 (Figure S14). We also found that expanded genes were evenly distributed across homologous chromosomes based on location and synteny analysis in elephant grass (Figure S15).

4 | CONCLUSIONS

In this study, we have assembled a high-quality chromosome-level genome of elephant grass showing a high rate of heterozygosity; this is the first published polyploid genome for the genus *Cenchrus*. A primary assembly was performed with short reads, long sequence reads produced by nanopore sequencing technology and Hi-C chromatin contact maps. Our well-annotated genome allowed us to identify genes and pathways related to leaf colour. We demonstrated that the expansion of anthocyanidin biosynthesis pathways has resulted in anthocyanidin accumulation in the elephant grass cultivar "Purple." In addition, our analysis revealed a high copy number and high transcript levels of genes involved in C_4 photosynthesis and hormone signal transduction that may contribute to the fast growth of elephant grass. The assembled elephant grass genome could provide a system for studying the diversity, speciation and evolution of this family, and it offers an important resource for understanding the mechanism of economically important traits and adaption. It also provides new resources for exploring other species in the genus *Cenchrus*, which have great economic, ecological and research value.

ACKNOWLEDGEMENTS

The research was supported by the Key Research and Development Plan of Guangxi Science and Technology (Grant No. Guike AB19245024), Guangxi Science and Technology (Grant No. Guike AD17129043), Program for Changjiang Scholars and Innovative Research Team in University (IRT_17R50), Guangxi Science and Technology Major Project (Grant No. Guike AA16380026), and the 111 Project (B12002).

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

J.Z., X.Y. and Q.Y. conceived the project, J.L., L.G., D.C. and L.L. collected the samples, P.X., Z.S., F.W. and Q.Y. performed the genome assembly and data analysis, Q.Y. wrote the manuscript, and J.Z., C.J. and M.M. revised the manuscript. All authors reviewed the manuscript.

DATA AVAILABILITY STATEMENT

The whole genome sequence data (including Illumina short-gun reads, Nanopore reads and Hi-C interaction reads), and transcriptomes of different tissues used in this study have been deposited in the NCBI, under accession nos. PRJNA649020 and PRJNA649119. The final assembly genome and genome annotation information have been deposited in the National Genomics Data Center (https://bigd.big.ac.cn/), under accession no. GWHAORA00000000 that is publicly accessible at https://bigd.big.ac.cn/gwh.

ORCID

Jiyu Zhang https://orcid.org/0000-0002-3642-373X

REFERENCES

Artner, C., & Benkova, E. (2019). Ethylene and cytokinin: Partners in root growth regulation. *Molecular Plant*, *12*, 1312–1314. https://doi.org/10.1016/j.molp.2019.09.003

Asem, I. D., Imotomba, R. K., Mazumder, P. B., & Laishram, J. M. (2015). Anthocyanin content in the black scented rice (Chakhao): Its impact on human health and plant defense. *Symbiosis*, *66*(1), 47–54. https://doi.org/10.1007/s13199-015-0329-z

Bariexca, T., Ezdebski, J., Redan, B., & Vinson, J. (2019). Pure polyphenols and cranberry juice high in anthocyanins increase antioxidant capacity in animal organs. *Foods*, *8*, 340. https://doi.org/10.3390/foods8080340

Beier, S., Thiel, T., Münch, T., Scholz, U., & Mascher, M. (2017). MISAweb: A web server for microsatellite prediction. *Bioinformatics*, *33*, 2583–2585. https://doi.org/10.1093/bioinformatics/btx198

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., Estep, M., Feng, L., Vaughn, J. N., Grimwood, J., Jenkins, J., Barry, K., Lindquist, E., Hellsten, U., Deshpande, S., Wang, X., Wu, X., Mitros, T., Triplett, J., … Devos, K. M. (2012). Reference genome sequence of the model plant Setaria. *Nature Biotechnology*, *30*, 555–561. https://doi.org/10.1038/nbt.2196

Burton, J., Adey, A., Patwardhan, R., Qiu, R., Kitzman, J., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*, 1119–1125. https://doi.org/10.1038/nbt.2727

Cao, Y., Xing, M., Xu, C., & Li, X. (2018). Biosynthesis of flavonol and its regulation in plants. *Acta Horticulturae Sinica*, *45*, 177–192. https://doi.org/10.16420/j.issn.0513-353x.2017-0306

Cardona, E., Jorge, R., Juan, P., & Luis, R. (2014). Effects of the pretreatment method on enzymatic hydrolysis and ethanol fermentability of the cellulosic fraction from elephant grass. *Fuel*, *118*, 41–47. https://doi.org/10.1016/j.fuel.2013.10.055

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*, 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Clouse, S. D. (2011). Brassinosteroid signal transduction: From receptor kinase activation to transcriptional networks regulating plant development. *The Plant Cell*, *23*(4), 1219–1230. https://doi.org/10.1105/tpc.111.084475

Conesa, A., Götz, S., García-Gómez, J., Terol, J., Talon, M., & Robles, M. (2005). BLAST2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*, 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Costa, M.-C., Artur, M. A. S., Maia, J., Jonkheer, E., Derks, M. F. L., Nijveen, H., Williams, B., Mundree, S. G., Jiménez-Gómez, J. M., Hesselink, T., Schijlen, E. G. W. M., Ligterink, W., Oliver, M. J., Farrant, J. M., & Hilhorst, H. W. M. (2017). A footprint of desiccation tolerance in the genome of *Xerophyta viscosa*. *Nature Plants*, *3*, 38. https://doi.org/10.1038/nplants.2017.38

Daud, Z., Hatta, M., Kassim, A., Mohd Aripin, A., & Awang, H. (2014). Analysis of Napier grass (*Pennisetum purpureum*) as a potential alternative fibre in paper industry. *Material Research Innovations*, *18*, 18–20. https://doi.org/10.1179/1432891714Z.000000000925

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, *295*, 1306–1311. https://doi.org/10.1126/science.1067799

Dm, S. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), https://doi.org/10.1186/s13059-019-1832-y

Fang, Z. (2015). Effects on feeding goats with Pennisetum.purpureum Schum cv.Taiwan. *Animal Husbandry & Feed Science*, *7*(5), 264–266. https://doi.org/10.19578/j.cnki.ahfs.2015.05.003

Farrell, G., Simons, S. A., & Hillocks, R. J. (2002). Pests, diseases and weeds of Napier grass, *Pennisetum purpureum*: A review. *International Journal of Pest Management*, *48*(1), 39–48. https://doi.org/10.1080/09670870110065578

Gao, J., Shen, L. I., Yuan, J., Zheng, H., Su, Q., Yang, W., Zhang, L., Nnaemeka, V. E., Sun, J., Ke, L., & Sun, Y. (2019). Functional analysis of *GhCHS*, *GhANR* and *GhLAR* in colored fiber formation of *Gossypium hirsutum* L. *BMC Plant Biology*, *19*, 455. https://doi.org/10.1186/s12870-019-2065-7

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S., & Bateman, A. (2005). Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, *33*, D121–124. https://doi.org/10.1093/nar/gki081

Grover, C. E., Gallagher, J. P., Szadkowski, E. P., Yoo, M. J., Flagel, L. E., & Wendel, J. F. (2012). Homoeolog expression bias and expression level dominance in allopolyploids. *New Phytologist*, *196*(4), 966–971. https://doi.org/10.1111/j.1469-8137.2012.04365.x

Gupta, S. C., & Mhere, O. (1997). Identification of superior pearl millet by Napier hybrids and Napier in Zimbabwe. *African Crop Science Journal*, *5*, 5. https://doi.org/10.4314/acsj.v5i3.27840

Haas, B., Delcher, A., Mount, S., Wortman, J., Smith, R., Hannick, L., & White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*, 5654–5666. https://doi.org/10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*, R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hagen, G., & Guilfoyle, T. (2002). Auxin-responsive gene expression: Genes, promoters and regulatory factors. *Plant Molecular Biology*, *49*, 373–385. https://doi.org/10.1023/A:1015207114117

Hahn, M., Bie, T., Stajich, J., Nguyen, C., & Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research*, *15*, 1153–1160. https://doi.org/10.1101/gr.3567505

Han, Y., & Wessler, S. (2010). MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Research*, *38*, e199. https://doi.org/10.1093/nar/gkq862

Harris, M., Deegan, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., & White, R. (2004). The Gene Oncology (GO) database and informatics resource. *Nucleic Acids Research*, *32*, 258–261. https://doi.org/10.1093/nar/gkh036

Hoff, K., & Stanke, M. (2018). Predicting genes in single genomes with AUGUSTUS. *Current Protocols in Bioinformatics*, *65*, https://doi.org/10.1002/cpbi.57

Hu, X., Yuan, J., Shi, Y., Lu, J., Liu, B., Li, Z., Chen, Y., Mu, D., Zhang, H., Li, N., Yue, Z., Bai, F., Li, H., & Fan, W. (2012). pIRS: Profile based Illumina pair-end Reads Simulator. *Bioinformatics*, *28*, 1533–1535. https://doi.org/10.1093/bioinformatics/bts187

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., … Yeats, C. (2008). InterPro: The integrative protein signature database. *Nucleic Acids Research*, *37*, D211–D215. https://doi.org/10.1093/nar/gkn785

Jaakola, L. (2013). New insights into the regulation of anthocyanin biosynthesis in fruit. *Trends in Plant Science*, *18*, 3. https://doi.org/10.1016/j.tplants.2013.06.003

Jurka, J., Kapitonov, V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*, 462–467. https://doi.org/10.1159/000084979

Kebede, G., Feyissa, F., Assefa, G., Alemayehu, M., Kehaliew, A. … Abera, M. (2017). Agronomic performance, dry matter yield stability and herbage quality of Napier grass (Pennisetum purpureum (L.) Schumach) accessions in different agro-ecological zones of Ethiopia. *The Journal of Agricultural and Crop Research*, *5*, 49–65.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*, *19*, 5. https://doi.org/10.1186/s12859-018-2203-5

Keilwagen, J., Wenk, M., Erickson, J., Schattat, M., Grau, J., & Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research*, *44*, gkw092. https://doi.org/10.1093/nar/gkw092

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120. https://doi.org/10.1007/BF01731581

Koech, R. (2019). Genome-enabled prediction models for black tea (*Camellia sisnesnsis*) quality and drought tolerance traits. *Plant Breeding*, *00*, 1–13. https://doi.org/10.1111/pbr.12813

Kruger, M., Davies, N., Myburgh, K., & Lecour, S. (2014). Proanthocyanidins, anthocyanins and cardiovascular diseases. *Food Research International*, *59*, 46. https://doi.org/10.1016/j.foodres.2014.01.046

Lagesen, K., Hallin, P., Rødland, E., Stærfeldt, H., Rognes, T., & Ussery, D. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*, 3100–3108. https://doi.org/10.1093/nar/gkm160

Le Mercier, P., & Bougueleret, L. (2007). The universal protein resource (UniProt). *Nucleic Acids Research*, *35*, gkl929. https://doi.org/10.1093/nar/gkl929

Li, H. (2017). Minimap2: Fast pairwise alignment for long DNA sequences. *Bioinformatics*, *34*(18):3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Liu, X., Shen, Y., He, Y., & Al, E. (2008). Tolerance to copper stress in Elephantgrass (Pennisetum purpureum) under soil culture. XXI International Grassland Congress & the VIII International Rangeland Congress.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Li, Z., Chen, Y., Mu, D., & Fan, W. (2012). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Preprint at, http://arxiv.org/abs/1308.2012

Lowe, T., & Eddy, S. (1997). tRNAscan-SE: A program for improved detection of transfer RNA Genes in genomic sequence. *Nucleic Acids Research*, *25*, 955–964. https://doi.org/10.1093/nar/25.5.0955

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q. I., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B. O., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*, 18. https://doi.org/10.1186/2047-217X-1-18

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2018). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. https://doi.org/10.1038/nature03959

Muday, G., Rahman, A., & Binder, B. (2012). Auxin and ethylene: Collaborators or competitors? *Trends in Plant Science*, *17*, 181–195. https://doi.org/10.1016/j.tplants.2012.02.001

Muktar, M. S., Teshome, A., Hanson, J., Negawo, A. T., Habte, E., Domelevo Entfellner, J.-B., Lee, K.-W., & Jones, C. S. (2019). Genotyping by sequencing provides new insights into the diversity of Napier grass (*Cenchrus purpureus*) and reveals variation in genome-wide LD patterns between collections. *Scientific Reports*, *9*, 1–15. https://doi.org/10.1038/s41598-019-43406-0

Nachtweide, S., Romoth, L., Gerischer, L., & Stanke, M. (2016). Simultaneous gene finding in multiple genomes. *Bioinformatics*, *3*, btw494. https://doi.org/10.1093/bioinformatics/btw494

Nelson, S., & Steber, C. (2016). Gibberellin hormone signal perception: Down-regulating DELLA repressors of plant growth and development. *Annual Plant Reviews*, *49*, Chapter: 6. https://doi.org/10.1002/9781119210436.ch6

Ori, N. (2019). Dissecting the biological functions of *ARF* and *Aux/IAA* genes. *The Plant Cell*, *31*(6), 1210–1211. https://doi.org/10.1105/tpc.19.00330

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., Poliakov, A., Schmutz, J., Spannagl, M., Tang, H., Wang, X., Wicker, T., Bharti, A. K., Chapman, J., Feltus, F. A., Gowik, U., … Rokhsar, D. S. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, *457*, 551–556. https://doi.org/10.1038/nature07723

Paterson, A. H., Bowers, J., & Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 9903–9908. https://doi.org/10.1073/pnas.0307901101

Paudel, D., Kannan, B., Yang, X., Harris-Shultz, K., Thudi, M., Varshney, R. K., Altpeter, F., & Wang, J. (2018). Surveying the genome and constructing a high-density genetic map of napiergrass (*Cenchrus purpureus* Schumach). *Scientific Reports*, *8*(1), 14419. https://doi.org/10.1038/s41598-018-32674-x

Pryszcz, L., & Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*, gkw294. https://doi.org/10.1093/nar/gkw294

Pucher, A. (2018). Pearl millet breeding in West Africa. Thesis for Doctoral.

Rao, X., & Dixon, R. (2019). Corrigendum: Corrigendum: The differences between NAD-ME and NADP-ME subtypes of C4 photosynthesis: More than decarboxylating enzymes. *Frontiers in Plant Science*, *10*, 247. https://doi.org/10.3389/fpls.2019.00247

Reis, G., Mesquita, A., Torres, G., Andrade-Vieira, L., Pereira, A., & Davide, L. (2014). Genomic homeology between *Pennisetum purpureum* and *Pennisetum glaucum* (Poaceae). *Comparative Cytogenetics*, *8*, 199–209. https://doi.org/10.3897/CompCytogen.v8i3.7732

Rocha, J. R. D. A. S. D. C., Marçal, T. D. S., Salvador, F. V., da Silva, A. C., Carneiro, P. C. S., de Resende, M. D. V., Carneiro, J. D. C., Azevedo, A. L. S., Pereira, J. F., & Machado, J. C. (2019). Unraveling candidate genes underlying biomass digestibility in elephant grass (*Cenchrus purpureus*). *BMC Plant Biology*, *19*(1), 548. https://doi.org/10.1186/s12870-019-2180-5

Saha, S., Bridges, S., Magbanua, Z., & Peterson, D. (2008). Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Tropical Plant Biology*, *1*, 85–96. https://doi.org/10.1007/s12042-007-9007-5

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., Tichý, L., Grulich, V., & Rotreklová, O. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, E4096–E4102. https://doi.org/10.1073/pnas.1321152111

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve. *Bioinformatics*, *24*, 637–644. https://doi.org/10.1093/bioinformatics/btn013

Stortenbeker, N., & Bemer, M. (2018). The *SAUR* gene family: The plant's toolbox for adaptation of growth and development. *Journal of Experimental Botany*, *70*, ery332. https://doi.org/10.1093/jxb/ery332

Tang, H., Wang, X., Bowers, J., Ming, R., Alam, M., & Paterson, A. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gaps. *Genome Research*, *18*, 1944–1954. https://doi.org/10.1101/gr.080978.108

Tarailo-Graovac, M., & Chen, N. (2009). Using repeatmasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, *25*(1), 4–10. https://doi.org/10.1002/0471250953.bi0410s25

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., Zhang, H. E., Zhao, Y., Wang, X., Rathore, A., Srivastava, R. K., Chitikineni, A., Fan, G., Bajaj, P., Punnuri, S., Gupta, S. K., Wang, H., Jiang, Y., Couderc, M., … Xu, X. (2018). Erratum: *Pearl millet* genome sequence provides a resource to improve agronomic traits in arid environments. *Nature Biotechnology*, *36*, 368. https://doi.org/10.1038/nbt0418-368d

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, *9*, e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, C., Yan, H., Li, J., Zhou, S., Liu, T., Zhang, X., & Huang, L. (2018). Genome survey sequencing of purple elephant grass (*Pennisetum purpureum Schum* 'Zise') and identification of its SSR markers. *Molecular Breeding*, *38*, 94. https://doi.org/10.1007/s11032-018-0849-3

Wybouw, B., & Rybel, B. (2018). Cytokinin – A developing story. *Trends in Plant Science*, *24*, https://doi.org/10.1016/j.tplants.2018.10.012

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., & Visser, R. (2011). Genome sequence and analysis of tuber crop potato. *Nature*, *475*, 189–195. https://doi.org/10.1038/nature10158

Xu, Z., & Wang, H. (2007). LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, *35*, W265–W268. https://doi.org/10.1093/nar/gkm286

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*, 329–342. https://doi.org/10.1038/nrg3174

Yang, X., Hu, R., Yin, H., Jenkins, J., Shu, S., Tang, H., & Tuskan, G. A. (2017). The Kalanchoe genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nature Communications*, *8*(1), 1899. https://doi.org/10.1038/s41467-017-01491-7

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*, *13*, 555–556. https://doi.org/10.1093/bioinformatics/13.5.555

Yao, N., Xian-Feng, Y. I., Lai, Z. Q., Liang, Y. L., Deng, S. Y., & Lai, D. W. (2016). Effects of *Pennisetum purpureum* Schumab cv. Purple on growth performance and serum biochemical parameters of meat geese. *Journal of Southern Agriculture*, *47*(12), 2163–2168.

Yi, X., Lai, Z., Yao, N., Cai, X., Wei, J., Lai, D., & Liang, Y. (2016). Planting performance of pennisetum purpureum schumab cv. purple in the southern region of China. *Agricultural Science & Technology*, *17*(3), 667–671. https://doi.org/10.16175/j.cnki.1009-4229.2016.03.041

Yin, Y., Vafeados, D., Tao, Y., Yoshida, S., Asami, T., & Chory, J. (2005). A new class of transcription factors mediates brassinosteroid-regulated gene expression in Arabidopsis. *Cell*, *120*, 249–259. https://doi.org/10.1016/j.cell.2004.11.044

Yoo, M. J., Szadkowski, E., & Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, *110*(2), 171–180. https://doi.org/10.1038/hdy.2012.94

Yu, J., Hu, S., Wang, J., Li, S., Wong, K.-S., Liu, B., Deng, Y., Dai, L. I., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., … Yang, H. (2001). A draft sequence of the rice (*Oryza sativa* ssp.indica) genome. *Chinese Science Bulletin*, *46*(23), 1937–1942. https://doi.org/10.1007/BF02901901

Zhang, Y. Z., Xu, S. Z., Cheng, Y. W., Ya, H. Y., & Han, J. M. (2016). Transcriptome analysis and anthocyanin-related genes in red leaf lettuce. *Genetics and Molecular Research*, *15*, 15017023. https://doi.org/10.4238/gmr.15017023

Zhang, Z., Li, Q., Li, Z., Staswick, P., Wang, M., Zhu, Y., & He, Z. (2007). Dual regulation role of *GH3.5* in salicylic acid and auxin signaling during Arabidopsis-Pseudomonas syringae interaction. *Plant Physiology*, *145*, 450–464. https://doi.org/10.1104/pp.107.106021

Zhou, S., Chen, J., Lai, Y., Yin, G., Chen, P., Pennerman, K. K., Yan, H., Wu, B., Zhang, H., Yi, X., Wang, C., Fu, M., Zhang, X., Huang, L., Ma, X., Peng, Y., Yan, Y., Nie, G., & Liu, L. (2019). Integrative analysis of metabolome and transcriptome reveals anthocyanins biosynthesis regulation in grass species *Pennisetum purpureum*. *Industrial Crops and Products*, *138*, 111470. https://doi.org/10.1016/j.indcrop.2019.111470

Zhou, S., Wang, C., Frazier, T. P., Yan, H., Chen, P., Chen, Z., Huang, L., Zhang, X., Peng, Y., Ma, X., & Yan, Y. (2018). The first Illumina-based de novo transcriptome analysis and molecular marker development in Napier grass (*Pennisetum purpureum*). *Molecular Breeding*, *38*, 8. https://doi.org/10.1007/s11032-018-0852-8

Zhou, S., Wang, C., Yin, G., Zhang, Y., Shen, X., Pennerman, K., & Huang, L. (2018). Phylogenetics and diversity analysis of *Pennisetum* species using Hemarthria EST-SSR markers. *Grassland Science*, *65*(1), 13-22. https://doi.org/10.1111/grs.12208

Zhran, M., & Lotfy, S. M. (2014). Phytoremediation of contaminated soil with cobalt and chromium. *Journal of Geochemical Exploration*, *144*, 367–373. https://doi.org/10.1016/j.gexplo.2013.07.003

Zhu, Y., Peng, Q., Li, K., & Xie, D.-Y. (2018). Molecular cloning and functional characterization of a dihydroflavonol 4-reductase from *Vitis bellula*. *Molecules*, *23*, 861. https://doi.org/10.3390/molecules23040861

Zou, C., Li, L., Miki, D., Li, D., Tang, Q., Xiao, L., Rajput, S., Deng, P., Peng, L. I., Jia, W., Huang, R. U., Zhang, M., Sun, Y., Hu, J., Fu, X., Schnable, P. S., Chang, Y., Li, F., Zhang, H., … Zhang, H. (2019). The genome of broomcorn millet. *Nature Communications*, *10*, 5. https://doi.org/10.1038/s41467-019-08409-5

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Yan Q, Wu F, Xu P, et al. The elephant grass (*Cenchrus purpureus*) genome provides insights into anthocyanidin accumulation and fast growth. *Mol Ecol Resour.* 2021;21:526–542. https://doi.org/10.1111/1755-0998.13271