DOI: 10.1002/tpg2.20448

The Plant Genome

ORIGINAL ARTICLE

Transcriptome-wide expression landscape and starch synthesis pathway co-expression network in sorghum

Zhenbin Hu[1] | **Junhao Chen**[1] | **Marcus O. Olatoye**[2] | **Hengyou Zhang**[3] | **Zhenguo Lin**[1]

[1]Department of Biology, Saint Louis University, Saint Louis, Missouri, USA

[2]USDA-ARS, Forage Seed and Cereal Research Unit, Prosser, Washington, USA

[3]State Key Laboratory of Black Soils Conservation and Utilization, Key Laboratory of Soybean Molecular Design and Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin, China

Correspondence
Zhenguo Lin, Department of Biology, Saint Louis University, Saint Louis, MO 63103, USA.
Email: zhenguo.lin@slu.edu

Hengyou Zhang, State Key Laboratory of Black Soils Conservation and Utilization, Key Laboratory of Soybean Molecular Design and Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin, 150081, China.
Email: zhanghengyou@iga.ac.cn

Assigned to Associate Editor Nils Stein.

Funding information
Natural Science Foundation of Heilongjiang Province of China, Grant/Award Number: JQ2022C005; National Natural Science Foundation of China, Grant/Award Number: 32272176; Innovation Team Project of Northeast Institute of Geography and Agroecology, Grant/Award Number: 2022CXTD03; The U.S. National Science Foundation, Grant/Award Number: 1951332

Abstract

The gene expression landscape across different tissues and developmental stages reflects their biological functions and evolutionary patterns. Integrative and comprehensive analyses of all transcriptomic data in an organism are instrumental to obtaining a comprehensive picture of gene expression landscape. Such studies are still very limited in sorghum, which limits the discovery of the genetic basis underlying complex agricultural traits in sorghum. We characterized the genome-wide expression landscape for sorghum using 873 RNA-sequencing (RNA-seq) datasets representing 19 tissues. Our integrative analysis of these RNA-seq data provides the most comprehensive transcriptomic atlas for sorghum, which will be valuable for the sorghum research community for functional characterizations of sorghum genes. Based on the transcriptome atlas, we identified 595 housekeeping genes (HKGs) and 2080 tissue-specific expression genes (TEGs) for the 19 tissues. We identified different gene features between HKGs and TEGs, and we found that HKGs have experienced stronger selective constraints than TEGs. Furthermore, we built a transcriptome-wide co-expression network (TW-CEN) comprising 35 modules with each module enriched in specific Gene Ontology terms. High-connectivity genes in TW-CEN tend to express at high levels while undergoing intensive selective pressure. We also built global and seed-preferential co-expression networks of starch synthesis pathways, which indicated that photosynthesis and microtubule-based movement play important roles in starch synthesis. The global transcriptome atlas of sorghum

Abbreviations: A3SS, alternative 3′ splice sites; A5SS, alternative 5′ splice sites; AS, alternative splicing; CDS, coding sequence; GO, gene ontology; HKG, housekeeping gene; lncRNA, long noncoding RNA; MXE, mutually exclusive exons; NCBI, National Center for Biotechnology Information; PCA, principal component analysis; PSI, percent spliced in; QTL, quantitative trait locus; RI, retained intron; SE, skip exon; TEG, tissue-specific expression gene; TPM, transcript per million; TW-CEN, transcriptome-wide co-expression network.

generated by this study provides an important functional genomics resource for trait discovery and insight into starch synthesis regulation in sorghum.

Plain Language Summary

To characterize the genome-wide gene expression landscape and provide functional genomics resources for future gene mining for complex traits, we comprehensively analyzed large-scale RNA-seq data from different tissues. We built an atlas of sorghum gene/lncRNA expression and identified housekeeping and tissue-specific expression genes. We observed that gene expression patterns were closely related to gene features and selection pressure. The co-expression network showed that photosynthesis and microtubule-based movement play important roles in starch synthesis. The global transcriptome atlas provides an important functional genomics resource for trait discovery and insight into starch synthesis regulation in sorghum.

1 | INTRODUCTION

Sorghum (*Sorghum bicolor*) is an economically important and dual-purpose crop for both food and bioenergy. It gains growing attention mainly due to its high tolerance to drought and high temperatures. Its genome was assembled using sorghum line BTx623 in 2009, and the assembly was improved later with 34,211 annotated genes (McCormick et al., 2018; Paterson et al., 2009). To improve sorghum yield and local adaptation, genomics-enabled crop breeding approaches have been used to facilitate sorghum selection efficiency (Boyles et al., 2019). For example, a genomics resource has demonstrated its capacity for trait discovery to improve the understanding of the genetic architecture of complex traits (Varshney et al., 2021). Additional functional genomics resources are needed to expand the understanding of the molecular basis of important traits in sorghum to maintain its sustainable role to meet possible food shortages in the next decades.

The RNA-seq technique has been widely used to explore gene expression and their functions under biological phenomena in various plants (Conesa et al., 2016; Stark et al., 2019), including sorghum (Boyles et al., 2019). Large-scale publicly available RNA-seq data provide an opportunity to comprehensively characterize the expressional landscape at different dimensions. Multiple studies have shown housekeeping genes (HKGs), ubiquitously expressed in all situations, have evolved slowly, while narrowly expressed genes or tissue-specific expression genes (TEGs) have experienced fast evolution (J. Yang et al., 2005; L. Zhang & Li, 2004). In addition, gene expression is also associated with the fitness landscape, the relationship between genotypes and the reproductive success, of protein-coding genes (Wu et al., 2022). TEGs would benefit the understanding of the molecular basis for a specific tissue formation and provide a gene resource for modifying a certain tissue using genetic engineering approaches (S.-J. Xiao et al., 2010). It has been demonstrated that it can improve the genomics prediction accuracy when integrating TEGs into the genome prediction model for genomic selection breeding (Fang et al., 2020). In mammals, genotype-tissue expression projects were characterized by integrating large-scale gene expression profiling to gain insight into the transcriptional landscape and identify regulatory elements (Ardlie et al., 2015; Liu et al., 2022; Teng et al., 2024). Thus far, a large amount of RNA-Seq data for sorghum has been generated; however, no efforts have been made to characterize the expression landscape in sorghum. The resources also provided an opportunity to comprehensively characterize the long noncoding RNA (lncRNA) and alternative splicing (AS) (Sun et al., 2020). This knowledge would benefit the discovery of the genomic basis underlying complex and important agricultural traits in sorghum.

Co-expression analysis, which is based on the hypothesis that functionally related genes have similar expression patterns, has been widely used to identify genes in a specific pathway or provide candidate genes for complex traits (Montenegro, 2022; Sarkar et al., 2014; Y. Zhang et al., 2022). Integrating with specific conditions or phenotypes, the identified trait-associated co-expression modules can provide insights into the related pathways or molecular basis of the trait (Hartl et al., 2021; van Dam et al., 2018). In sorghum, co-expression networks were previously used to dissect the biological mechanism of sorghum stem composition (Hennet et al., 2020; X. Xiao et al., 2023). A co-expression database, PlantNexus, was built to facilitate the co-expression analysis for sorghum (Zhou et al., 2022). However, the starch synthesis pathway, the most important nutrition pathway in sorghum, was not studied using the approach of the co-expression

network based on large-scale expression data. Quantitative trait loci (QTLs) were identified in the germplasm population to characterize the natural variation of sorghum seeds' starch content (Ayalew et al., 2022; Rhodes et al., 2017; Zhou et al., 2022), but the underlying regulatory mechanisms and pathways were not fully understood (Q. Xiao et al., 2022). Photosynthesis-related genes were proposed to be promising candidates underlying the starch associations (Ayalew et al., 2022; B. Chen et al., 2019; Rhodes et al., 2017; Q. Xiao et al., 2022), while it remained to be determined in sorghum. Thus far, several starch synthesis pathway genes have been identified (Campbell et al., 2016; Hill et al., 2012; Ke et al., 2022). However, the related regulatory pathways remain underexplored due to the nature of complexity in the metabolism. Lack of such knowledge limited the identification of other starch biosynthesis genes for sorghum nutrient improvement.

In this study, we retrieved 873 transcriptomic datasets representing 19 different tissues (roots, shoots, leaves, seeds, etc.) to characterize the global gene expression landscape and identify possible regulatory pathways of starch synthesis in sorghum. With the systematic analysis, we identified lncRNA, 595 HKGs, and 2080 TEGs. Further analysis showed that broadly expressed genes and HKGs tend to express at relatively high levels and experience intense selection. Reversely, narrow-expressed genes and TEGs are relatively low in expression while evolving fast. We further constructed a transcriptome-wide co-expression network (TW-CEN) that allowed the identification of gene-guided starch synthesis pathways and prioritized candidate genes for further exploration. Our study provides a valuable functional genomics resource for the sorghum community and greatly enhances our understanding of the regulatory basis of starch synthesis in sorghum.

Core Ideas

- Large-scale transcriptome data were analyzed to characterize the sorghum gene and long noncoding RNA expression landscape.
- Housekeeping and tissue-specific genes were identified.
- The co-expression network was constructed to identify the regulatory network of the starch synthesis pathway.

2 | DATA AND METHODS

2.1 | RNA-seq data and processing

In total, 873 RNA-seq datasets were downloaded from the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA) (Table S1). Raw data were converted to FASTQ format using the Fastq-dump function in the SRA Toolkit 3.0.1 (https://github.com/ncbi/sra-tools). Data quality was checked using the FastQC 0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality reads and adapters were filtered using Trimmomatic version 0.38 (Bolger et al., 2014). After filtration, the reads were aligned on the sorghum reference genome (version 3.1) with STAR (version 2.6.0) (Dobin et al., 2013; McCormick et al., 2018; Paterson et al., 2009). The reference genome and annotation were downloaded from Phytozome 13 (https://phytozome.jgi.doe.gov) (Goodstein et al., 2012). The expression level (transcript per million [TPM]) of genes was quantified using StringTie (version 1.3.4) (M. Pertea et al., 2015). Samples with a high map rate (>60%) and gene expression of at least five samples were kept, and gene expression was transferred by $\log_2$(TPM+2) for further analysis.

2.2 | lncRNA identification and expression

To identify the novel transcripts and lncRNA, each transcriptome was assembled, and GTF files were merged using StringTie (version 1.3.4) (M. Pertea et al., 2015). Comparison with reference annotation was performed using the Gffcompare package (G. Pertea & Pertea, 2020). Transcripts with class codes i, u, x, y, and p were identified as potential lncRNA. Further, the coding potential was referred to using Coding Potential Calculator 2 (Kang et al., 2017). The interproscan package was used to identify the potential protein domains with the Pfam argument in *-appl* option (Quevillon et al., 2005). The annotated lncRNA gff was used to refer to the lncRNA expression using StringTie (version 1.3.4) (M. Pertea et al., 2015).

2.3 | Alternative splicing

The rMATs package was used to identify AS with the statoff options (Shen et al., 2014) and five basic AS events: alternative 5′ splice sites (A5SS), alternative 3′ splice sites (A3SS), mutually exclusive exons (MXE), retained intron (RI), and skipped exon (SE).

2.4 | Clustering analysis and tissue relationship

Principal component analysis (PCA) was performed using the *prcomp* function in R (4.1.2). The relationship was visualized using a hierarchical clustering cut tree, which was made using the *dendextend* package.

2.5 | Housekeeping gene and tissue-specificity

The HKGs were identified using the method used in a previous study (Machado et al., 2020), only genes that met all the following three thresholds were identified as HKGs. The method details are as follows:

1. Gene expresses in all tissues or samples and TPM > 1.
2. Genes with MFC − CoV scores falling in the first quartile were classified as HKGs:

$$\mathrm{MFC} = \max\left(\mathrm{TPM}_{\mathrm{all-tissues}}\right) / \min\left(\mathrm{TPM}_{\mathrm{all-tissues}}\right)$$

$$\mathrm{CoV} = \mathrm{sd}\left(\mathrm{TPM}_{\mathrm{all-tissues}}\right) / \mathrm{mean}\left(\mathrm{TPM}_{\mathrm{all-tissues}}\right)$$

$$\mathrm{MFC-CoV\ score} = \mathrm{MFC-CoV}$$

   where MFC is the ratio of the maximum to the minimum of the gene expression and is calculated by dividing the largest by the smallest TPM value.

   CoV was computed by taking the standard deviation divided by the mean expression of a gene.
3. Gene's tissue-specificity index τ was calculated based on previous studies (Kryuchkova-Mostacci & Robinson-Rechavi, 2017; Yanai et al., 2005). The τ score ranged from 0 (broad expression) to 1 (tissue-specific expression). When τ scales <0.35, the corresponding genes were identified as HKGs. The formula was used to calculate the τ scale:

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x}_i)}{n-1};\ \hat{x}_i = \frac{x_i}{max(x_i)}$$

where x_i is the expression of the gene in tissue i, and n is the number of tissues.

2.6 | Tissue-specific expression gene identification

TEGs were identified using tissue-specific expression scores (TS_score). TS_score for each gene was calculated based on a method proposed in a previous study (R. Y. Yang et al., 2018), but without considering the alternative event. The method is as follows:

$$TS\_Score_{g,t} = \frac{\sum_i^N n_i * w_i * (S_{g,i} * log_2 \frac{Exp_{g,t}}{Exp_{g,i}})}{\sum_i^N n_{t,i} * w_i}$$

where $\mathrm{TS\_Score}_{g,t}$ is the tissue-specific score for gene g in tissue t; $\mathrm{Exp}_{g,t}$ is the median TPM gene expression value for gene g in tissue t; $\mathrm{Exp}_{g,i}$ is the median TPM gene expression value for gene g in tissue i; $\log_2(\mathrm{Exp}_{g,t}/\mathrm{Exp}_{g,i})$ is the $\log_2$ transformed gene g expression ratio in between tissues t and i; and w_i is a weight for tissue i to adjust for the global gene expression similarity between tissue i and other tissues. It was calculated as $1/\sum_j^N \mathrm{Cor}_{i,j}$, where $\mathrm{Cor}_{i,j}$ is the Spearman rank correlation coefficient between tissues i and j. When a correlation coefficient ($\mathrm{Cor}_{i,j}$) is less than 0.4, it is set to 0, so that only highly correlated tissues contribute to the weight.

$S_{g,i}$ is a binary flag and set to 1 only when the gene expression difference is statistically significant ($p < 0.01$). The significance is assessed using the linear model function with the rlm function in the MASS package (Venables & Ripley, 2002).

Here, n_i is also weighted to indicate the similarity of tissues i and t using the correlation between two tissues (Spearman). $n_i = 1 - r$, where r is the correlation coefficient. With the n_i flag, highly similar tissues would reduce their contribution to the tissue-specificity calculation for the target.

When a gene with a TS_score >3 for a target tissue, the gene was identified as a tissue-biased expressed gene for the target tissue. If a gene is only expressed in tissue t, the TS_score was assigned as the maximum one in the dataset.

2.7 | Evolutionary pressures: *dN/dS*

The orthofinder package (v2.5.4) (Emms & Kelly, 2019) was used to identify the orthologous group in 10 grass species as follows: *Zea mays* (Schnable et al., 2009), *Brachypodium distachyon* (International Brachypodium Initiative, 2010), *Leersia perrieri* (Stein et al., 2018), *Hordeum vulgare* (International Barley Genome Sequencing Consortium et al., 2012), *Oryza barthii* (Stein et al., 2018), *Setaria italica* (G. Zhang et al., 2012), *Oryza sativa* (Sasaki, 2005), *Aegilops tauschii* (Luo et al., 2017), *Pearl millet* (Varshney et al., 2017), and sorghum (McCormick et al., 2018; Paterson et al., 2009). The protein sequence and CDS (coding sequence) file for all the species except for sorghum were downloaded from Ensembl (Hubbard et al., 2002), and sorghum genome data were downloaded from Phytozome 13 (Goodstein et al., 2012; Paterson et al., 2009). The alignment was performed using the clustalw2 package (version 2.1) (Larkin et al., 2007). The alignment and CDS were merged into the PAML codon format (Suyama et al., 2006). The *dN/dS* (ω) was calculated using the codeML function from the PAML package (4.9) (Z. Yang, 2007).

2.8 | Co-expression network analysis

The row expression matrix was filtered using the goodSamplesGenes function with verbose = 3 from *WGCNA* (1.71) to build expression network (Langfelder & Horvath, 2008). Soft threshold was identified using the *pickSoftThreshold* function in the *WGCNA*. Module was identified using the *cutreeDynamic* function with deepSplit = 4 from the *WGCNA*. Modules were further merged based on similarity <0.25. The Hub gene for each module was identified using the *chooseTopHubInEachModule* function in the WGCNA. The co-expression network was visualized for topological overlap matrix (TOM) > 0.06 using the GGNET package.

2.9 | Gene enrichment analysis

Gene Ontology (GO) annotation for the sorghum gene was downloaded from the Phytozome 13 (Goodstein et al., 2012). Genes without GO term annotation were annotated using the closest orthologous gene from rice (http://rice.plantbiology.msu.edu). GO enrichment analysis was conducted using the *topGO* (4.2) package (version 2.46.0) in R with the "weight01" algorithm and "fisher" statistic; the overrepresented GO terms were identified using $p < 0.01$.

3 | RESULTS AND DISCUSSION

3.1 | Building a comprehensive gene expression dataset in sorghum

To explore the transcriptome-wide expression landscape in sorghum, we comprehensively analyzed 873 RNA-seq datasets retrieved from the NCBI. The RNA-seq datasets covered 19 different sorghum tissues and seed compartments (Figure 1a,b; Table S1). After filtering, a total of ~16.84 billion reads (an average mapping rate of 87.72%) were uniquely mapped to the sorghum reference genome (RTx623, v3.1) using STAR (Dobin et al., 2013; McCormick et al., 2018; Paterson et al., 2009). The uniquely mapped read rates of different RNA-seq datasets range from 14.91% to 97.53% (Figure S1), and 29 datasets with a low uniquely map rate (<60%) were excluded from downstream analysis (Figure S1). The gene expression level per gene was quantified as TPM using Stringtie (M. Pertea et al., 2015). After filtering, we generated a gene expression profile with 31,541 genes across 844 RNA-seq datasets.

3.2 | Transcripts, lncRNA, and alternative splicing

To quantify the transcript using our diverse transcriptome dataset, 844 GTF were merged and compared with reference annotation, resulting in 71,816 transcripts, including 47,196 reference annotated transcripts. 41,434 loci of the 71,816 loci were novel identification/transcript. From those transcripts, 42,973 lncRNA transcripts were identified for 36,848 lncRNA, with an average exon number of 1.56. In total, 77.88% (33,467) of the lncRNA were identified in the intergenic region (Table S2).

In total, 66,750 AS events were identified for 15,163 genes, which accounts for 53.56% of multiexonic genes. Separately, we identified 3037 A3SS from 2211 genes, 1927 A5SS from 1433 genes, 6173 MXE from 3620 genes, 2664 RI from 2025 genes, and 52,949 SE from 14,194. Sobic.001G326900 with 72 AS events represented the gene with highest AS events (Table S3). Although with different AS landscape, SE is dominant AS event in all tissues (Figure S2).

3.3 | Tissue differentiation causes different gene expression profiles

PCA was conducted to identify the drivers of expression variations across samples. Despite the heterogeneity among the RNA-seq datasets from different experiments, we still observed that RNA-seq datasets were clustered by tissue, and some were completely separate from other tissues, which indicates the high quality of the expression data and robustness of our analysis pipeline (Figure 1c,d). For example, in the PCA of coding gene expression, the first PC mainly explains the difference between vegetative and reproductive tissues, and the second PC captures the difference between above-ground tissues and roots (Figure 1c). However, we also observed a mixture among tissues or loose clusters for some tissues. For example, RNA-seq datasets from seedlings were divided into two clusters and clustered with leaf and shoot datasets, separately (Figure 1c), which may result from differences in sampling or tissue definition in different studies (Figure S3). In the present study, we keep seedlings as independent tissues. The PCA of percent spliced in for AS events also showed extended tissue-specific features (Figure 1e; Figure S4).

To further assess if gene expression profiles could represent the tissue relationship, we built a hierarchical clustering cut tree among tissues using tissue expression profiles, which were an average of all $\log_2$(TPM+1) from the same tissue. We found that tissues were mainly clustered based on the similarity of tissue biological function, and five tissue clusters were identified in the present study (Figure 1f). For example, the expression profiles from developing seeds and endosperm were clustered together (Figure 1f). The results indicated that functionally related tissues tend to have similar gene expression patterns. Interestingly, we observed that the gene expression profiles from roots were closely related to those from shoots (Figure 1f). This result was supported by a close relationship between roots and shoots in cotton in coordinating nitrogen usage (J. Chen et al., 2020; Hatrick &

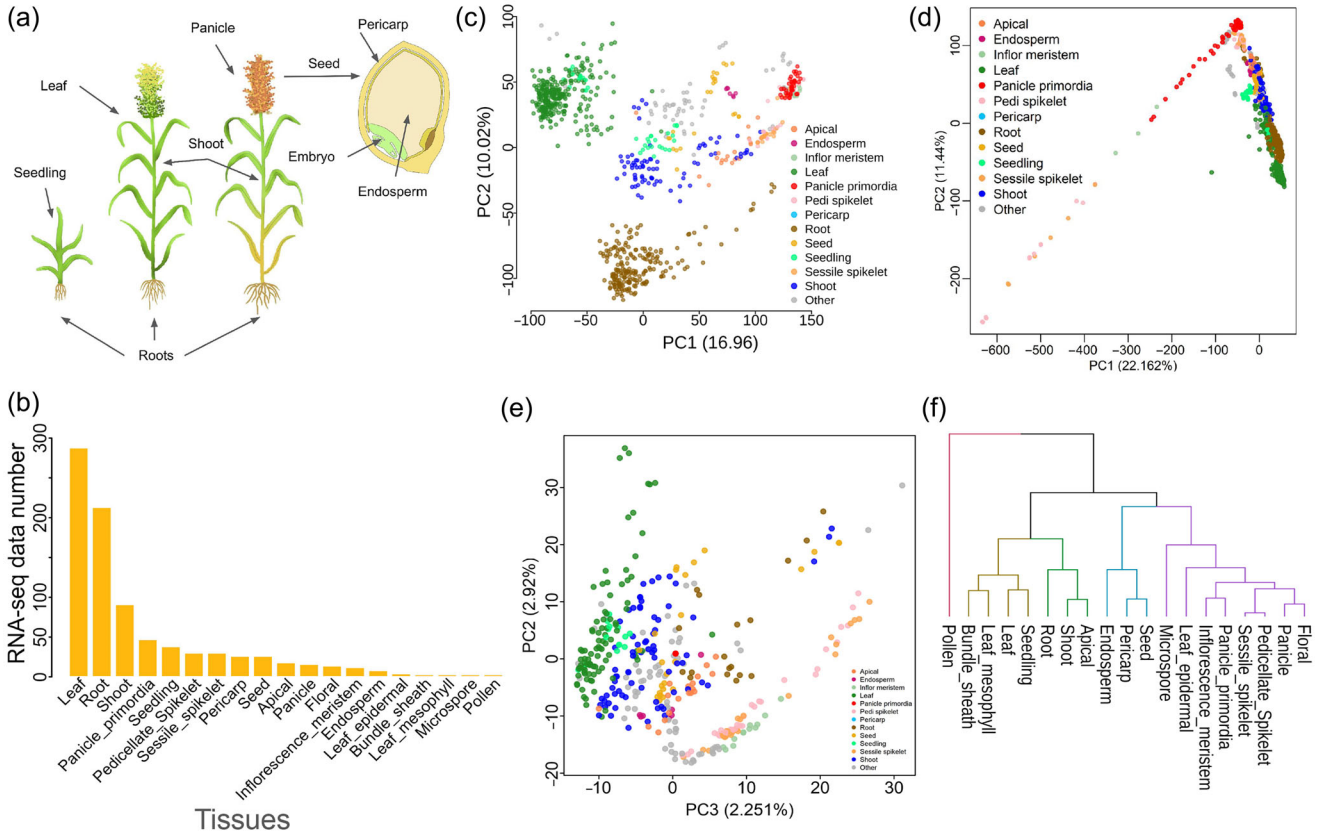

FIGURE 1 Characterization of transcriptome data and expression across tissues in sorghum. (a) Cartoon illustration of developing sorghum plants and tissues. The three plants represent three developmental stages: seedling, flowering, and mature stages. Labeled tissues were represented by the RNA-seq dataset. (b) The counts of RNA-seq dataset across 19 tissues. (c) Principal component analysis of global gene expression patterns. Tissues were indicated with the different colored dots as indexed beside the panel. Unknown tissue RNA-seq datasets were labeled as other. (d) Principal component analysis of global long noncoding RNA (lncRNA) expression patterns. Color same as in (c). (e) Principal component analysis of percent spliced in (PSI) of alternative splicing. Color same as in (c). (d) Transcriptome-wide gene expression based-phylogenetic tree for tissue similarity. The five colors mean five different clusters corresponding to five different biological function units. PC, principal component.

Bowling, 1973). Pollen forms a single cluster that is separated from all other tested tissues (Figure 1e), which may be caused by special transcriptome features of the reproductive cells. The above results showed that tissue differentiation is a major factor of transcriptome differentiation. Overall, the results indicate that the tissues could be well represented by transcriptomic profiles.

3.4 | Gene expression landscape: Expression breadth and tissue specificity

As shown above, the differences in gene expression across tissues were observed. Next, we characterized the genome-wide gene expression landscape across the 19 tissues by quantifying gene expression breadth (described as narrow to broad expression) using the tissue-specific expression index (τ) (Figure S5), expression abundance ($\log_2$(TPM+1)), HKGs, and TEGs. Comparing to coding genes, lncRNA showed lower expression (log2(TPM+1) = 1.44) (Figure 2a) and higher tissue specificity ($\tau = 0.94$) (Figure 2b). Meanwhile, lncRNA (742 basepair) was shorter than coding genes (Figure 2c). The τ has a significant negative correlation with expression level for gene ($r = -0.40$, $p < 0.01$) and lncRNA ($r = -0.75$, $p < 0.01$) (Figure 3a; Figure S6). Genes with $\tau < 0.35$ and expression variability across tissues located in the lower 25 quantiles were identified as HKGs, which were constitutively and stably expressed across 19 tissues (Machado et al., 2020; Yanai et al., 2005). Based on the criteria, we identified 595 HKGs, with an average τ of 0.20 (Figure 3b; Table S4). GO enrichment analysis showed that GO terms associated with basic cell biological activities were highly enriched, such as translation ($p = 1.3 \times 10^{-10}$), mRNA splicing, via spliceosome ($p = 5.3 \times 10^{-4}$) in the biological process GO Term, few binding proteins in molecular function GO Term, and eight cellular component GO Terms (Figure 3c).

With a modified method from a previous study (see Section 2) (R. Y. Yang et al., 2018), we also identified a total of 2648 TEGs, representing 6.08% of sorghum genes (Table

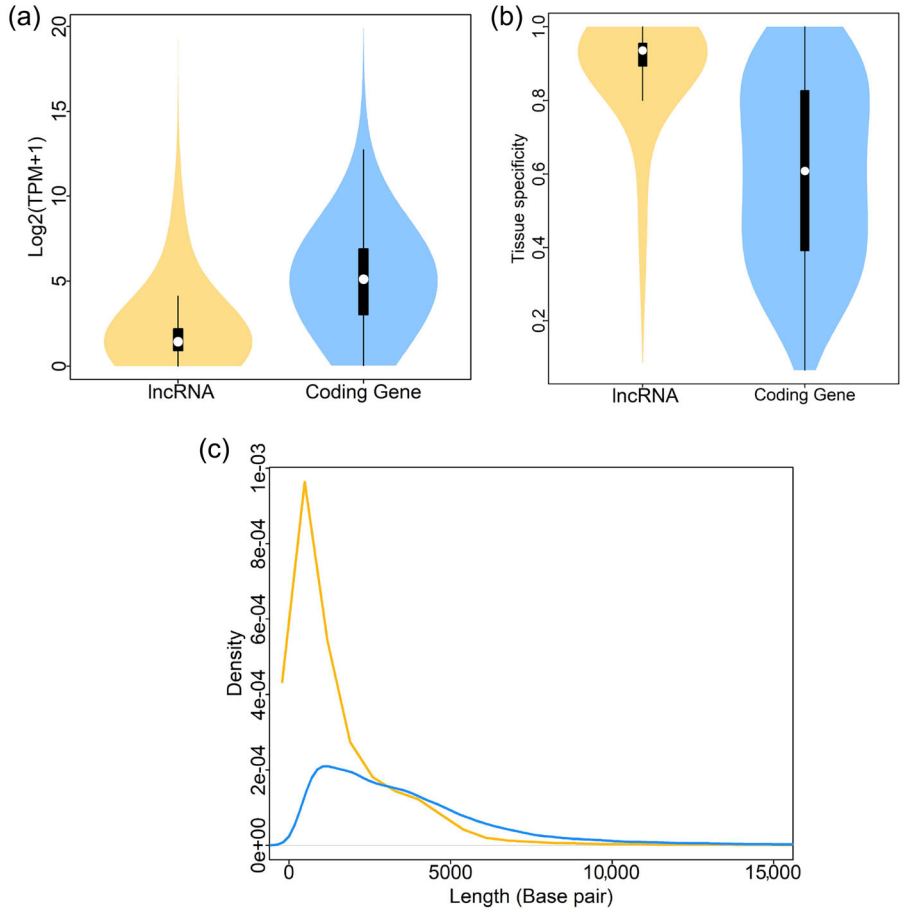

FIGURE 2 Comparison between long noncoding RNA (lncRNA) and coding genes. (a) The comparison of expression between lncRNA and coding gene. (b) The tissue-specific comparison of lncRNA and coding gene. (c) The length of lncRNA and coding gene. Figure shows *x* axis range from 0 to 1500 base pairs. TPM, transcript per million.

S5), for 19 tested tissues. For each tissue, the number of TEG ranges from 8 (floral) to 823 (pollen); the highest number of TEGs for pollen may be related to its specific tissue feature ensuring successful reproduction. The expression levels of TEGs among tissues are significantly different (Figure S7). For example, TEGs for pollen maintain the highest expression levels, while apical TEGs have the lowest expression levels (Figure S7). GO enrichment analysis showed TEGs were mainly associated with tissue development or related tissue biological functions (Table S6). For example, GO terms for TEGs in seeds were highly enriched in reproduction ($p = 2.9 \times 10^{-10}$), embryonic development ($p = 4.0 \times 10^{-8}$), post-embryonic development ($p = 1.4 \times 10^{-7}$), nutrient reservoir activity ($p = 1.7 \times 10^{-4}$) and monolayer-surrounded lipid storage body ($p = 3.3\times 10^{-9}$) (Figure 3d). The identification of TEGs would be beneficial for us to understand tissue development and the molecular basis of tissue-related traits in sorghum, and it also provides gene resources for genetic engineering specific tissues with minimal interference from other tissues.

The comparison of τ also showed a significant difference ($p < 2.2 \times 10^{-16}$, Wilcox test) between HKGs (median, $\tau = 0.20$) and TEGs (median, $\tau = 0.91$) (Figure 3b). Although the HKGs ($\text{Log}_2(\text{TPM+1}) = 7.98$) were expressed significantly higher ($p < 2.2 \times 10^{-16}$) than TEGs ($\text{Log}_2(\text{TPM+1}) = 5.32$) (Figure 3e), the variation in the expression of TEGs (standard deviation = 3.55) was much larger than HKGs (standard deviation = 1.76) (Figure 3e). The results indicate that broadly expressed genes (with high τ score) and HKGs tend to have higher expression levels than narrowly expressed genes on average. This finding was consistent with previous studies in humans (Bentz et al., 2019; Vinogradov, 2004). Interestingly, the highest expressed gene (Sobic.008G043200, $\log_2(\text{TPM+1}) = 19.90$) was specifically expressed in pollen, and it has the highest τ ($\tau = 0.97$) in our dataset (Figure 3a). The varying expression of TEGs may be associated with tissue-specific methylation or chromatin accessibility (Pan et al., 2021). The regulation of gene spatial expression needs to be further explored in the future using Chromatin Immunoprecipitation

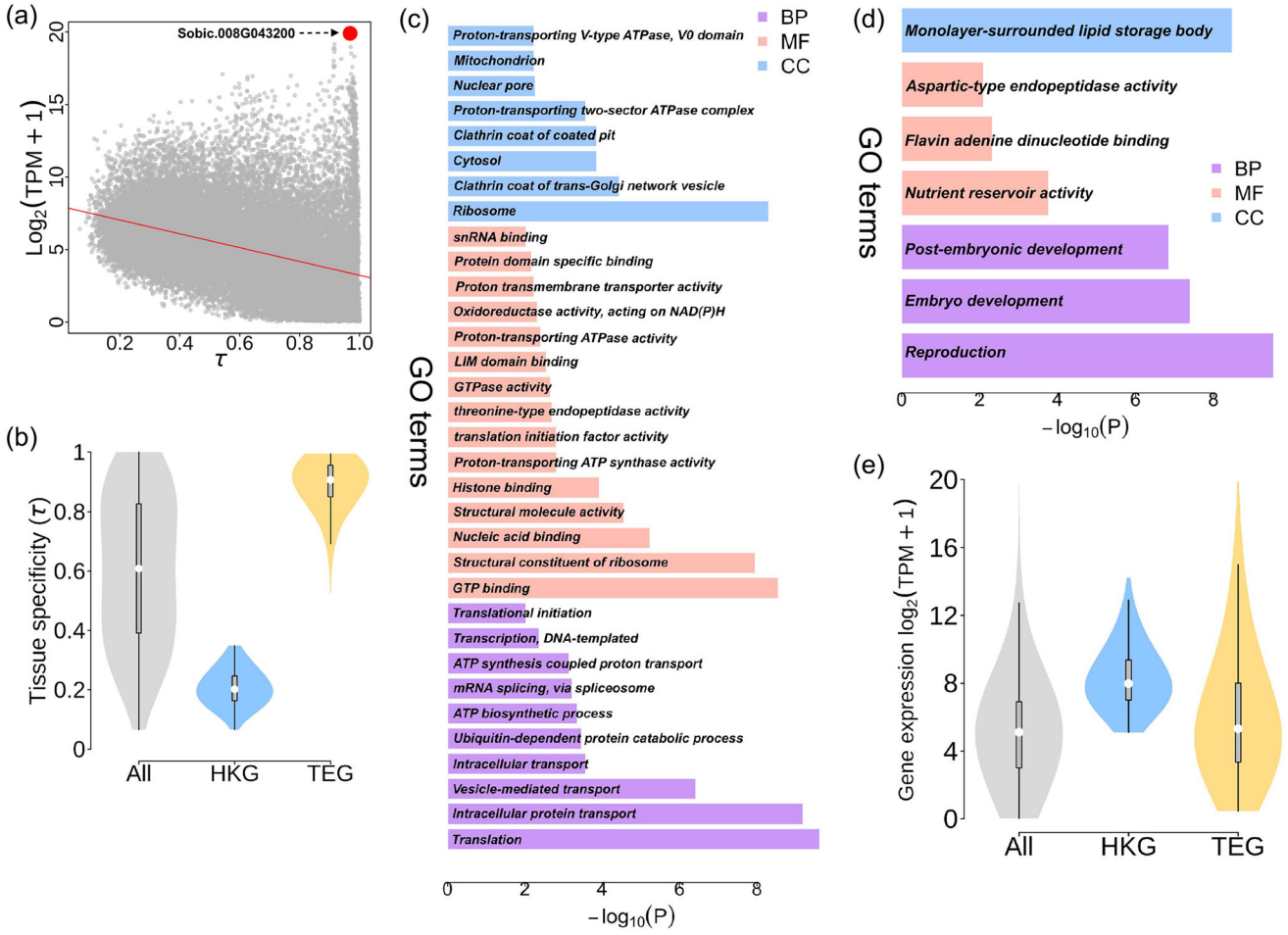

FIGURE 3 Gene expression landscape. (a) Correlation between gene expression tissue specificity and gene expression levels. The red point means the highest expressed gene in our dataset and with a high tissue-specific score. (b) Tissue specificity score and the comparison among all genes (All), tissue-specific expression genes (TEGs), and housekeeping genes (HKGs). (c) Gene Ontology (GO) enrichment analysis for housekeeping genes. BP means biological process; MF means molecular function; CC means cellular component. (d) GO enrichment analysis for seed-specific expressed genes. (e) Gene expression comparison among all genes, TEGs, and HKGs. TPM, transcript per million.

sequencing (ChIP-seq) (Park, 2009), chromatin accessibility (Klemm et al., 2019), and DNA methylation (Moore et al., 2013).

3.5 | Gene expression landscape associated with evolutionary gene features

In mammalian studies, short or compact genes tend to express at high levels (Grishkevich & Yanai, 2014). However, our results showed that gene/lncRNA length was significantly negatively correlated with τ ($\rho = -0.44$, $p < 2.2 \times 10^{-16}$) for gene and ($\rho = -0.38$, $p < 2.2 \times 10^{-16}$) for lncRNA (Figure 4a; Figure S8), but significantly positively correlated with expression levels ($\rho = 0.22$, $p < 2.2 \times 10^{-16}$) for gene and ($\rho = 0.54$, $p < 2.2 \times 10^{-16}$) for lncRNA (Figure 4b; Figure S8). Although the correlation between protein length and expression landscape was diminished, it is still significant statistically (for expression level, $\rho = 0.11$ and $p < 2.2 \times 10^{-16}$; for τ, $\rho = -0.27$; $p < 2.2 \times 10^{-16}$). More specifically, the gene length of HKGs (median, 4233 bp) is significantly ($p < 2.2 \times 10^{-16}$, Wilcox test) longer than that of TEGs (median, 2218 bp). As such, the protein length for HKGs (median, 374 bp) was significantly ($p = 6.85 \times 10^{-3}$) longer than TEGs (median, 366 bp). The difference in the association between gene length and expression between mammals and plants was also discussed in a previous study (H. Yang, 2009). We also observed that intron content (bp) was positively correlated ($\rho = 0.23$, $p < 2.2 \times 10^{-16}$) with gene expression level, but negatively correlated ($\rho = -0.44$, $p < 2.2 \times 10^{-16}$) with τ, implying the positive impact of intron content on gene expression.

Different expression landscape genes undergo different selection pressures during the divergence of sorghum from other species. To characterize the divergence and gene expression landscape, we analyzed the relationship between

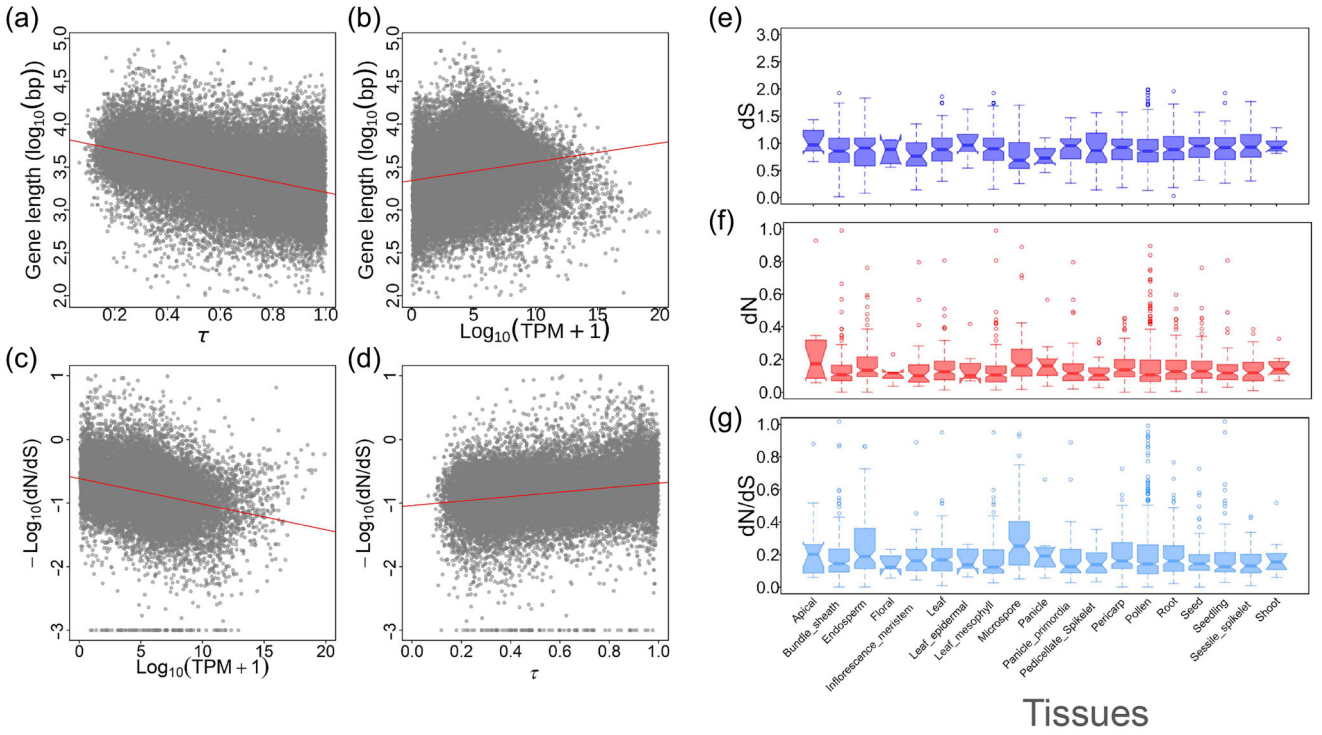

FIGURE 4 The relationship among gene expression, gene features, and selective pressure. The correlations between gene expression and gene length (a), gene expression tissue specificity and gene length (b), gene expression and selective pressure (c), gene expression tissue specificity and selective pressure (d). The *y* axis was log10 transferred *dN/dS*. (e) Selective pressure (*dN/dS*) across 19 tissues using tissue-specific expression genes. To see the variation of *dN/dS*, the *y* axis only show from 0 to 1, the full range figure was showed in Figure S10a. (f) Non-synonymous substitutions (dN) across 19 tissues using tissue-specific expression genes. To see the variation of *dN/dS*, the *y* axis only show from 0 to 1, the full range figure was showed in Figure S10a. (g) Synonymous substitutions (dS) across 19 tissues using tissue-specific expression genes. TPM, transcript per million.

selection pressure as calculated by *dN/dS* (the ratio of non-synonymous to synonymous substitutions, known as ω) and gene expression landscapes. Our result showed a significant negative correlation ($\rho = -0.30$, $p < 2.2 \times 10^{-16}$) between ω and gene expression levels, while a positive correlation ($\rho = 0.20$, $p < 2.2 \times 10^{-16}$) between ω and τ (Figure 4c,d). Not surprisingly, HKGs experienced an intensified purifying selection ($p < 2.2 \times 10^{-16}$) (median, $\omega = 0.09$) than TEGs (median, $\omega = 0.15$) (Figure S9). We identified 33 positively selected genes ($\omega > 1$) in TEGs but did not identify positively selected genes in HKGs, which implies the major cause of speciation from TEGs rather than HKGs. In addition, we identified the significant difference ($p < 2.2 \times 10^{-16}$) between HKGs and TEGs for *dN* and *dS* (Figure 4f; Figure S10a). For example, the ratio (0.46) of *dN* between HKGs (0.05) and TEGs (0.12) was much lower than that (0.72) of *dS* between HKGs (0.63) and TEGs (0.87) (Figure 4g; Figure S10a), indicating that the non-synonymous sites of HKGs experienced more intense selection than TEGs. In support of the finding, similar results were also observed in other species (J. Yang et al., 2005; L. Zhang & Li, 2004). We further analyzed the selective pressure among tissues, the tissue-selection pressure was represented using *dN/dS* of the TEGs. The *dN/dS* ratios showed significant ($p = 2.1 \times 10^{-6}$, Kruskal-Wallis test) differences among tissues (Figure 4e--g; Figure S10a). Specifically, the microspore showed the highest ω (0.25); whereas, leaf mesophyll has the lowest ω (0.12). Our results showed a large variation in ω among tissues, an intensified selection pressure for HKGs rather than TEGs, and a higher selection pressure for reproductive-related tissues than vegetable tissues in sorghum (Figure S10b).

3.6 | Global co-expression network of sorghum genes

Gene co-expression suggests likely co-regulatory relationships or similar molecular functions for the genes. To gain a global view of the co-expression profiles and provide guides for trait gene discovery in sorghum, we built a TWCEN using hierarchical clustering of dissimilarity among the TOM with 22,545 filtered genes from 844 RNA-seq datasets. Those genes passed the filtering of *goodSamplesGenes* from the WGCNA package (Langfelder & Horvath, 2008). A soft threshold of 6 was identified and used to build TWCEN (Figure S11). The 22,545 genes were initially grouped into 88 modules. After merging those modules with high

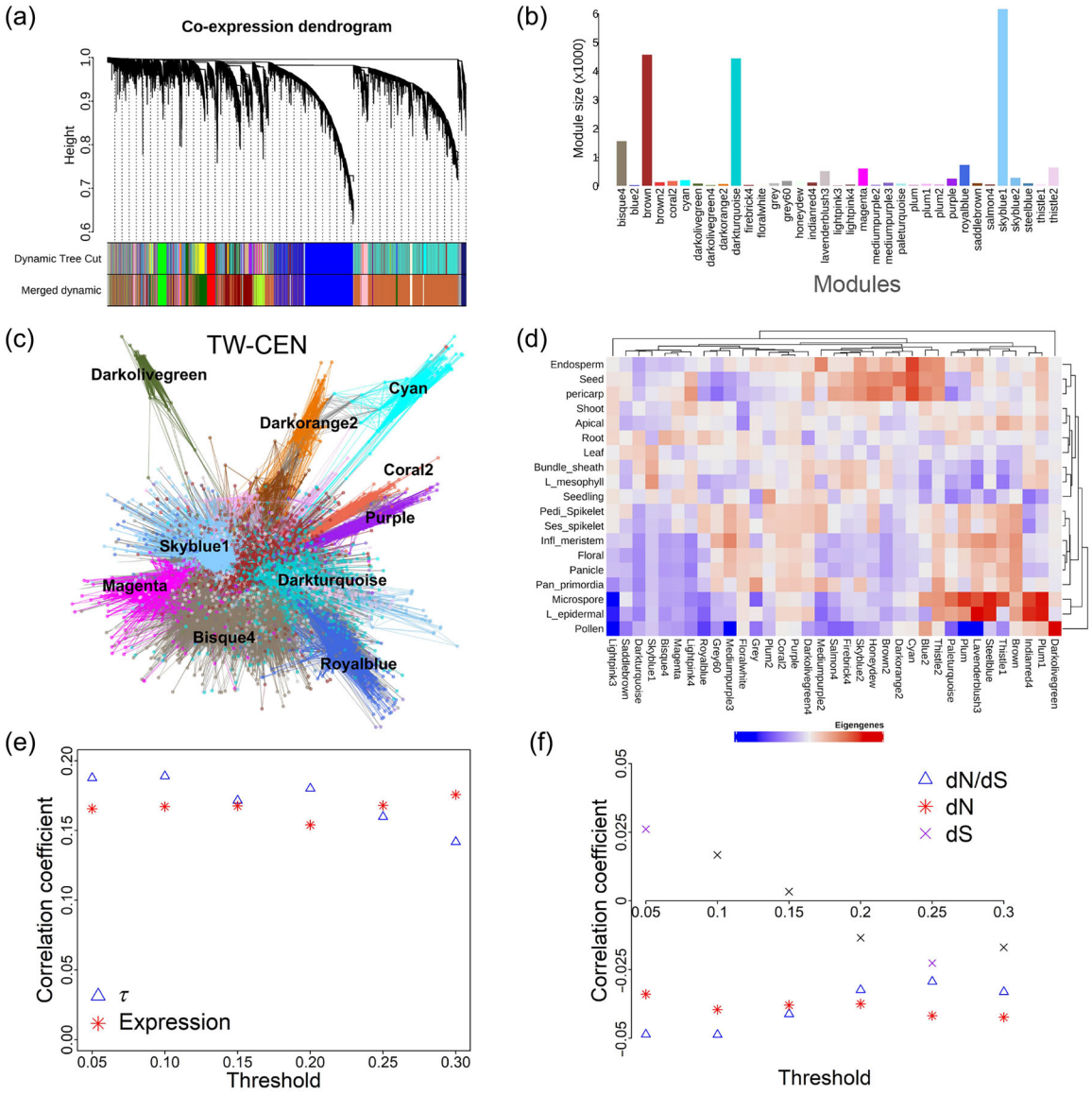

FIGURE 5 Transcriptome-wide co-expression network. (a) The module detection for 22,545 genes and merged dynamic tree. The first and second rows below the dendrogram mean modules were identified using THE *dynamicTreeCut* method and the merged dynamic tree using a cutHeight = 0.25 in *mergeCloseModules*, respectively. (b) The module size of the Transcriptome-wide co-expression network (TW-CEN). The bar color corresponds to the module color in (a). (c) Global co-expression network visualized using ggnet2 based on topological overlap matrix (TOM) > 0.06. The colors represent the corresponding module as shown in (b), and each node means a gene. (d) The heatmap of eigengene across tested tissues. The eigengene scores within the tissue were averaged with a median. (e) The relationship of connectivity and expression landscape using different TOM scores as the thresholds. (f) The correlation of connectivity and selection pressure (*dN/dS*) using different TOM scores as the thresholds.

similarity in expression profiles (<0.25) to condense clusters into more meaningful modules, it was consolidated into 35 modules with a large variation in sizes ranging from 38 (lightpink3) to 6176 genes (skyblue1) (Figure 5a,b). The global weighted co-expression network was visualized in network format based on TOM > 0.06, and it contains 17,895 nodes and 28,081,248 edges (Figure 5c). Eigengenes and tissue relationship analysis showed that the *Darkolivegreen* module genes were highly expressed in pollen (Figure 5d), implying that the genes within the model were likely related to the function specific to pollen development. With the same approach, modules, including cyan, thistle2, blue2, and skyblue2, were identified to be highly expressed within seed-related tissues or compartments (Figure 5d). To explore the potential biological function for each module, we performed GO term enrichment analysis for each of them. For each module, one GO term (*thistle1* module) to 51 GO terms (*skyblue1* module) were enriched (Table S7; Figure S12). For example, the

GO terms enriched for the *lavenderblush3* module are mainly involved in the RNA-related pathways and the GO terms enriched for the *saddlebrown* module are mainly engaged in stress response (Table S7). The results suggested similar functions of those co-expressed genes within the same module.

In total, we identified 34 within-module hub genes for all modules except for the "gray" module using the *chooseTopHubInEachModule* function from the WGCNA package (Table S8). We noticed that seven of the hub genes do not have function annotation in the genome annotation file (Table S9). Considering the function-related genes tend to co-express, we identified the co-express genes for hub genes based on a correlation coefficient greater than 0.50 to annotate those uncharacterized genes. The enriched GO term for the co-express genes of each of the seven hub genes suggests their potential functions (Table S9). For example, the top GO terms from GO domains for Sobic.003G016400 (thistle2 module) were enriched in translation (biological process, $p = 1.20 \times 10^{-11}$), ADP binding (molecular function, $p = 1.40 \times 10^{-17}$), and nucleosomes (cellular component, 2.70×10^{-12}) (Table S9).

To understand the relationship between gene expression landscape and co-expression network, we analyzed the relationship between the connectivity of co-expressed genes, represented using the co-expressed gene number of a gene based on the specific threshold of correlation coefficients of co-expression, and the expression landscape. We observed that connectivity was significantly positively correlated with expression level ($\rho = 0.17$, $p < 0.01$) and τ ($\rho = 0.19$, $p < 0.01$) at 0.1 and the relationship between them was not impacted by the threshold (Figure 5e; Figure S13), indicating that high-expressed gene and narrow-expressed genes tend to have more co-express genes. Interestingly, the connectivity was highly negatively correlated with roots (−0.16) and endosperm (−0.13) (Figure S14), indicating the highly expressed genes in roots and endosperm tend to have low connectivity. Considering the potential impact on other genes' expression of high-connectivity genes, we reasoned that they experienced higher selection pressure than low-connectivity genes. We further analyzed the correlation between selection pressure ($\omega = dN/dS$) and connectivity, and the result showed a weak but significant negative correlation ($\rho = -0.06$, $p < 0.01$: at co-expression coefficient 0.1) between gene connectivity and selection pressure (ω) (Figure 5f), which supports the hypothesis that high-connectivity genes tend to experience more intense purifying selection. This result was supported by similar observations in various plant species such as *Arabidopsis thaliana*, *Glycine max*, *O. sativa*, *Populus* spp., *Solanum lycopersicum*, *Vitis* spp., *Z. mays*, and *Populus tremula* (Mähler et al., 2017; Masalia et al., 2017).

3.7 | Starch pathway genes co-expression networks

Starch is the major content and nutrition in sorghum seed; however, the related regulatory network and pathways were poorly understood. To identify the potential regulation network in sorghum, we built a Starch Synthesis Pathway Co-Expression Network (SSP-CEN) using 102 starch biosynthesis pathway genes in sorghum, as guide genes, that have been characterized in a previous publication (Campbell et al., 2016) (Figure 6a,b). After analysis, the 102 starch biosynthesis pathway genes were distributed in 13 modules from TW-CEN. We noticed that 30.19% (32) of starch pathway genes fell in the skyblue1 module, and 22.64% (24) of starch pathway genes were clustered into the brown module. Both modules represent a total of 52.83% (56) reported starch pathway genes, indicating that the two modules were likely important for starch synthesis. The 102 starch pathway genes exhibited varying expression patterns across the tested tissues, and they can be clustered into four categories based on the clustering analysis (Figure 6a). Overall, genes from Category 1 are lowly expressed in most of the tissues, and the other three were highly expressed in the tissues, particularly seed-related tissues. For example, *Sobic.010G022600* (*SbWx*) encodes glucose-6-phosphate isomerase that is highly expressed in seed compartment endosperms, while *Sobic.004G163700* (*SbSBEII*) encoding starch synthase 2 and *Sobic.010G093400* (*SbSSIIa*) encoding starch branching enzyme were specifically expressed in the endosperm (Figure 6a). Category 2 genes were expressed in most of the tissues, but the expression level was relatively low in reproductive tissues. In contrast, category 3 genes were highly expressed in most of the tissues except for pollen and microspore. Two genes in this category (*SbSH2.1*: Sobic.002G160400 and Sobic.003G230500, encode glucose-1-phosphate adenylyltransferase) were preferentially expressed in seed and endosperm (Figure 6a), suggesting the tissue-specific function in starch biosynthesis. Category 4 genes were broadly expressed in all tested tissues at a relatively higher level than the other three categories (Figure 6a). A large variation in spatial expression patterns in various tissues for the genes indicated a more complex regulatory network for starch synthesis that happened in multiple tissues, including reproductive and vegetable tissues. A similar pattern was also observed in rice and maize (Fu & Xue, 2010; Q. Xiao et al., 2021), suggesting a possible conserved regulatory mechanism across cereal crops.

Here, SSP-CEN was built with a correlation coefficient greater than 0.70. However, only 23 (21.70%) starch pathway genes were present in SSP-CEN, such as *Sobic.010G047700* (starch synthases, *SbSSI*), *Sobic.004G238600* (starch synthases 2, *SSIIb*), *Sobic.006G066800* (starch branching, *SbSbeIIa*), and *Sobic.006G221000* (starch synthases 3, *SbSSIII*).

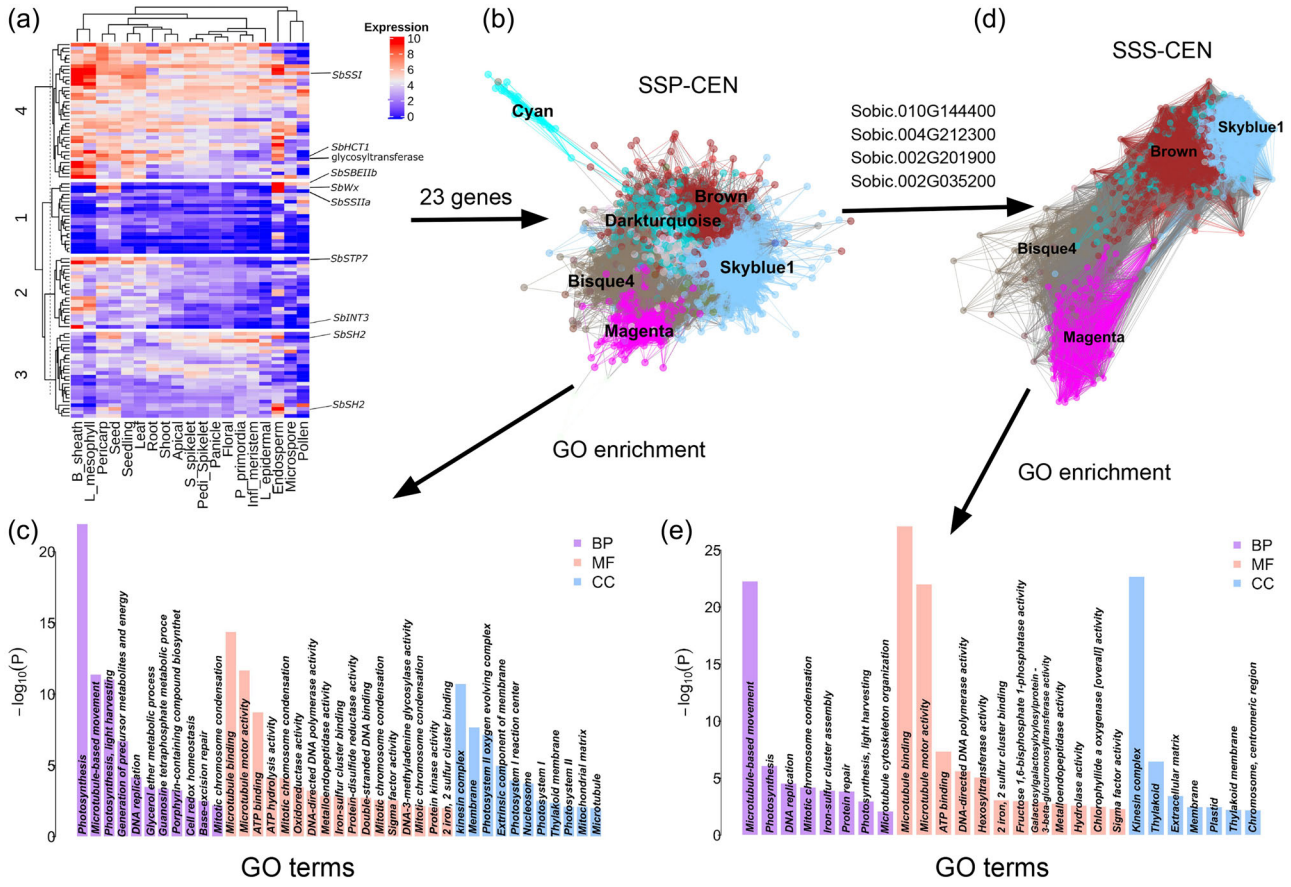

FIGURE 6 Co-expression network of starch synthesis pathway. (a) Expression heatmap of 102 starch pathway genes across 19 tissues. (b) Co-expression network of the starch pathway based on the correlation coefficient > 0.70. The colors represent the co-expression modules. (c) Gene Ontology (GO) enrichment analysis of co-expressed genes in Starch Synthesis Pathway Co-Expression Network (SSP-CEN). BP means biological process; MF means molecular function; CC means cellular component. (d) The co-expression network of seed expression starch pathway genes. (e) GO enrichment analysis of co-expressed genes in seed starch synthesis co-expression network (SSS-CEN).

These genes are clustered in six modules, the majority (14) of which are in category 4 (Figure 6a; Figure S15). In total, 4158 genes were identified to be co-expressed with the aforementioned 23 guide genes, and these genes were used to build starch biosynthesis-associated SSP-CEN (Figure 6b). To gain insights into the potentially related pathways, GO enrichment analysis was performed for the 4158 genes, and it shows that 39 GO terms, including 11 biological process terms, 17 molecular function terms, and 11 cellular component terms, are significantly enriched ($p < 0.01$) (Figure 6c). The multiple photosynthesis-related GO terms were enriched, those terms include photosynthesis, light harvesting, photosystem II oxygen-evolving complex, photosystem I reaction center, and photosystems I and II (Figure 6c), implying the importance of photosynthesis for starch biosynthesis in sorghum (Fünfgeld et al., 2021; Pfister & Zeeman, 2016).

To analyze the seed starch synthesis co-expression, we further narrowed SSP-CEN down to seed starch synthesis co-expression network (SSS-CEN) (Figure 6d) using four starch biosynthesis genes (glycosyltransferase: Sobic.010G144400; *SbHCT1*: Sobic.004G212300; *SbSTP7*: Sobic.002G201900; *SbINT3*: Sobic.002G035200) that were preferentially expressed in seed-related tissues (seed, endosperm, and pericarp) (Figure 6a). In total, the SSS-CEN contains 1261 co-expressed genes (Figure 6d). All of the co-expressed genes were clustered into 11 modules, and the majority are from modules brown (575), skyblue1 (440), magenta (93), bisque4 (63), and darkturquoise (62) (Figure 6d). GO enrichment analysis for SSS-CEN showed that the photosynthesis-related terms were significantly enriched, implying the ability of the seed to photosynthesis (Figure 6e). Interestingly, cellular components movement-related terms were highly enriched for SSS-CEN, such as microtubule-based movement, microtubule cytoskeleton organization, microtubule binding, microtubule motor activity, and kinesin complex. Our results suggested that part of the seed starch was synthesized in the seed.

Twelve GO terms, including the photosynthesis and matter-movement-related terms, were significantly enriched for both starch synthesis networks (Figure 6c,e). The role of

microtubule-based movements, significantly enriched in both starch synthesis co-expression networks (Figure 6c,e), in starch synthesis, was rarely studied in previous studies. In the future, developing metabolic datasets and methods to connect photosynthesis (source) with seed starch content (sink) is needed to facilitate molecular design breeding in sorghum.

3.8 | lncRNA associated with seed starch synthesis

To identify the potential lncRNA associated with starch synthesis, we examined the correlation of expression between seed starch synthesis genes and lncRNA. Eleven and one lncRNA were co-expressed with Sobic.001G083900: *SbPHO1 and Sobic.004G212300: SbHCT1* based on correlation coefficient $r > 0.75$ (Table S10). The highest correlated lncRNA (MSTRG.29733) is localized on chromosome 1 and enrichedly expressed in seed component tissues, such as endosperm (Figure S16). The high correlation between lncRNA and mRNA implies the function of lncRNA in regulating starch content in sorghum seed. However, the fine regulating landscape for starch synthesis and accumulation in sorghum seed needs the single-cell transcriptome and functional genomics (Rich-Griffin et al., 2020).

3.9 | Co-expression gene of seed starch synthesis not colocalized with seed starch QTL for in sorghum

Based on the Guilty-By-Association (GBA) principle, we expected that many co-expression genes in SSP-CEN regulate seed starch contents and also overlap with seed starch content QTLs. To do this, we retrieved a total of 46 seed starch content QTLs that were compiled from previous publications (Table S11). We identified 721 out of the 4158 co-expression genes, including 21 starch synthesis pathway genes, overlapping with QTL or 5669 genes around QTL (±250 kb) windows (Table S11). We tested whether the overlap randomly happened using the χ^2 test, and the results showed no significant ($p = 0.30$, χ^2 test) enrichment was observed. However, the result may be caused by the global co-expression network of starch synthesis and seed-specific starch content QTLs. Nonetheless, we still checked the overlap number (228) of genes between SSS-CEN and seed starch content QTLs. We observed the co-expression genes within SSS-CEN were still not significantly ($p = 0.25$, χ^2 test) enriched in QTL regions. The functional variants are needed for a QTL gene to cause the natural variation, while co-expression networks can only identify similar expressing genes, however, the functional variants were not necessary for co-expression. A previous study in humans showed that the co-expression was non-related to the genetic architecture of neuropsychiatric disease risk (Hartl et al., 2021).

4 | CONCLUSIONS

In this study, we integrated large-scale publicly available RNA-seq data to depict the landscape of transcription-wide transcript expression in sorghum. The dataset and the results provide a valuable genomic resource for sorghum biological and trait discovery. Identification and characterization of lncRNA, HKGs, and TEGs here provide insight into sorghum tissue developmental biology and chances for genetic modification of the specific tissues in sorghum. The transcriptome landscape can be integrated with genome-wide association studies to jointly dissect the genetic architecture of complex traits. The co-expression network helps identify regulatory pathways controlling complex traits. As exemplified by starch synthesis, we reveal that photosynthesis and lncRNA were tightly associated with starch synthesis. Our transcriptome dataset provides a valuable genomics resource to facilitate sorghum genomics-enabled breeding and trait discovery in sorghum to meet the increasing demands of food for humans.

AUTHOR CONTRIBUTIONS

Zhenbin Hu: Formal analysis; writing—original draft. **Junhao Chen**: Formal analysis. **Marcus O Olatoye**: Writing—original draft. **Hengyou Zhang**: Writing—original draft; writing—review and editing. **Zhenguo Lin**: Conceptualization; methodology.

ACKNOWLEDGMENTS

Hengyou Zhang was supported by the Natural Science Foundation of Heilongjiang Province of China (JQ2022C005), the National Natural Science Foundation of China (32272176), the Innovation Team Project of Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences (2022CXTD03). Zhenbin Hu and Zhenguo Lin were supported by the U.S. National Science Foundation (NSF grant 1951332).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The RNA-seq data were downloaded from NCBI and their information can be found in Table S1. Gene expression across tissue can be found at https://zhenbinhu.shinyapps.io/Transcriptome. The scripts were deposited into github (https://github.com/zhenbinHU/SB_transcriptome_atlas).
The gene expression and lncRNA results were deposited into figshare (https://figshare.com/account/projects/197194/articles/25316437)

ORCID

Zhenbin Hu https://orcid.org/0000-0002-1500-1255
Zhenguo Lin https://orcid.org/0000-0002-8400-9138

REFERENCES

Ardlie, K., Deluca, D., Segrè, A., Sullivan, T., Young, T., Gelfand, E., Trowbridge, C., & GTEx Consortium. Emmanouil Dermitzakis. (2015). Human genomics. The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, *348*, 648–660.

Ayalew, H., Peiris, S., Chiluwal, A., Kumar, R., Tiwari, M., Ostmeyer, T., Bean, S., & Jagadish, S. V. K. (2022). Stable sorghum grain quality QTL were identified using SC35 × RTx430 mapping population. *The Plant Genome*, *15*, e20227. https://doi.org/10.1002/tpg2.20227

Bentz, A. B., Thomas, G. W. C., Rusch, D. B., & Rosvall, K. A. (2019). Tissue-specific expression profiles and positive selection analysis in the tree swallow (*Tachycineta bicolor*) using a de novo transcriptome assembly. *Scientific Reports*, *9*, 15849. https://doi.org/10.1038/s41598-019-52312-4

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boyles, R. E., Brenton, Z. W., & Kresovich, S. (2019). Genetic and genomic resources of sorghum to connect genotype with phenotype in contrasting environments. *The Plant Journal*, *97*, 19–39. https://doi.org/10.1111/tpj.14113

Campbell, B. C., Gilding, E. K., Mace, E. S., Tai, S., Tao, Y., Prentis, P. J., Thomelin, P., Jordan, D. R., & Godwin, I. D. (2016). Domestication and the storage starch biosynthesis pathway: Signatures of selection from a whole sorghum genome sequencing strategy. *Plant Biotechnology Journal*, *14*, 2240–2253. https://doi.org/10.1111/pbi.12578

Chen, B., Wang, C., Wang, P., Zhu, Z., Xu, N., Shi, G., Yu, M., Wang, N., Li, J., Hou, J., Li, S., Zhou, Y., Gao, S., Lu, X., & Huang, R. (2019). Genome-wide association study for starch content and constitution in sorghum (*Sorghum bicolor* (L.) Moench). *Journal of Integrative Agriculture*, *18*, 2446–2456. https://doi.org/10.1016/S2095-3119(19)62631-6

Chen, J., Liu, L., Wang, Z., Zhang, Y., Sun, H., Song, S., Bai, Z., Lu, Z., & Li, C. (2020). Nitrogen fertilization increases root growth and coordinates the root–shoot relationship in cotton. *Frontiers in Plant Science*, *11*, 880. https://doi.org/10.3389/fpls.2020.00880

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*, 13. https://doi.org/10.1186/s13059-016-0881-8

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*, 238.

Fang, L., Cai, W., Liu, S., Canela-Xandri, O., Gao, Y., Jiang, J., Rawlik, K., Li, B., Schroeder, S. G., Rosen, B. D., Li, C.-J., Sonstegard, T. S., Alexander, L. J., Van Tassell, C. P., VanRaden, P. M., Cole, J. B., Yu, Y., Zhang, S., Tenesa, A., … Liu, G. E. (2020). Comprehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Research*, *30*, 790–801. https://doi.org/10.1101/gr.250704.119

Fu, F.-F., & Xue, H.-W. (2010). Coexpression analysis identifies Rice Starch Regulator1, a rice AP2/EREBP family transcription factor, as a novel rice starch biosynthesis regulator. *Plant Physiology*, *154*, 927–938. https://doi.org/10.1104/pp.110.159517

Fünfgeld, M. M. F. F., Wang, W., Ishihara, H., Arrivault, S., Feil, R., Smith, A. M., Stitt, M., Lunn, J. E., & Niittylä, T. (2021). The pathway of starch synthesis in *Arabidopsis thaliana* leaves. bioRxiv. https://doi.org/10.1101/2021.01.11.426159

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*, D1178–D1186. https://doi.org/10.1093/nar/gkr944

Grishkevich, V., & Yanai, I. (2014). Gene length and expression level shape genomic novelties. *Genome Research*, *24*, 1497–1503. https://doi.org/10.1101/gr.169722.113

Hartl, C. L., Ramaswami, G., Pembroke, W. G., Muller, S., Pintacuda, G., Saha, A., Parsana, P., Battle, A., Lage, K., & Geschwind, D. H. (2021). Coexpression network architecture reveals the brain-wide and multiregional basis of disease susceptibility. *Nature Neuroscience*, *24*, 1313–1323. https://doi.org/10.1038/s41593-021-00887-5

Hatrick, A. A., & Bowling, D. J. F. (1973). A study of the relationship between root and shoot metabolism. *Journal of Experimental Botany*, *24*, 607–613. https://doi.org/10.1093/jxb/24.3.607

Hennet, L., Berger, A., Trabanco, N., Ricciuti, E., Dufayard, J.-F., Bocs, S., Bastianelli, D., Bonnal, L., Roques, S., Rossini, L., Luquet, D., Terrier, N., & Pot, D. (2020). Transcriptional regulation of sorghum stem composition: Key players identified through co-expression gene network and comparative genomics analyses. *Frontiers in Plant Science*, *11*, 224. https://doi.org/10.3389/fpls.2020.00224

Hill, H., Slade Lee, L., & Henry, R. J. (2012). Variation in sorghum starch synthesis genes associated with differences in starch phenotype. *Food Chemistry*, *131*, 175–183. https://doi.org/10.1016/j.foodchem.2011.08.057

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., … Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research*, *30*, 38–41. https://doi.org/10.1093/nar/30.1.38

Mayer, K. F. X., Waugh, R., Brown, J. W. S., Schulman, A., Langridge, P., Platzer, M., Fincher, G. B., Muehlbauer, G. J., Sato, K., Close, T. J., Wise, R. P., Stein, N., & International Barley Genome Sequencing Consortium. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature*, *491*, 711–716.

International Brachypodium Initiative. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, *463*, 763–768. https://doi.org/10.1038/nature08747

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., & Gao, G. (2017). CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, *45*, W12–W16. https://doi.org/10.1093/nar/gkx428

Ke, F., Zhang, K., Li, Z., Wang, J., Zhang, F., Wu, H., Zhang, Z., Lu, F., Wang, Y., Duan, Y., Liu, Z., Zou, J., & Zhu, K. (2022). Transcriptomic analysis of starch accumulation patterns in different glutinous sorghum seeds. *Scientific Reports*, *12*, 11133. https://doi.org/10.1038/s41598-022-15394-1

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*, 207–220. https://doi.org/10.1038/s41576-018-0089-8

Kryuchkova-Mostacci, N., & Robinson-Rechavi, M. (2017). A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, *18*, 205–214.

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*, 559. https://doi.org/10.1186/1471-2105-9-559

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, *23*, 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., Li, B., Xiang, R., Chamberlain, A. J., Pairo-Castineira, E., D'Mellow, K., Rawlik, K., Xia, C., Yao, Y., Navarro, P., Rocha, D., Li, X., Yan, Z., Li, C., … Fang, L. (2022). A multi-tissue atlas of regulatory variants in cattle. *Nature Genetics*, *54*, 1438–1447. https://doi.org/10.1038/s41588-022-01153-5

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L., Wang, Y., McGuire, P. E., Liu, S., Long, H., Ramasamy, R. K., Rodriguez, J. C., Van, S. L., Yuan, L., Wang, Z., Xia, Z., … Dvořák, J. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature*, *551*, 498–502. https://doi.org/10.1038/nature24486

Machado, F. B., Moharana, K. C., Almeida-Silva, F., Gazara, R. K., Pedrosa-Silva, F., Coelho, F. S., Grativol, C., & Venancio, T. M. (2020). Systematic analysis of 1298 RNA-Seq samples and construction of a comprehensive soybean (*Glycine max*) expression atlas. *The Plant Journal*, *103*, 1894–1909. https://doi.org/10.1111/tpj.14850

Mähler, N., Wang, J., Terebieniec, B. K., Ingvarsson, P. K., Street, N. R., & Hvidsten, T. R. (2017). Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genetics*, *13*, e1006402. https://doi.org/10.1371/journal.pgen.1006402

Masalia, R. R., Bewick, A. J., & Burke, J. M. (2017). Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. *PLoS One*, *12*, e0182289. https://doi.org/10.1371/journal.pone.0182289

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B. D., McKinley, B., Mattison, A., Morishige, D. T., Grimwood, J., Schmutz, J., & Mullet, J. E. (2018). The *Sorghum bicolor* reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, *93*, 338–354. https://doi.org/10.1111/tpj.13781

Montenegro, J. D. (2022). Gene co-expression network analysis. *Methods in Molecular Biology*, *2443*, 387–404. https://doi.org/10.1007/978-1-0716-2067-0_19

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, *38*, 23–38. https://doi.org/10.1038/npp.2012.112

Pan, Z., Yao, Y., Yin, H., Cai, Z., Wang, Y., Bai, L., Kern, C., Halstead, M., Chanthavixay, G., Trakooljul, N., Wimmers, K., Sahana, G., Su, G., Lund, M. S., Fredholm, M., Karlskov-Mortensen, P., Ernst, C. W., Ross, P., Tuggle, C. K., … Zhou, H. (2021). Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nature Communications*, *12*, 5848. https://doi.org/10.1038/s41467-021-26153-7

Park, P. J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*, 669–680. https://doi.org/10.1038/nrg2641

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

Pertea, G., & Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Research*, *9*:304. https://doi.org/10.12688/f1000research.23297.1

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*, 290–295. https://doi.org/10.1038/nbt.3122

Pfister, B., & Zeeman, S. C. (2016). Formation of starch in plant cells. *Cellular and Molecular Life Sciences*, *73*, 2781–2807. https://doi.org/10.1007/s00018-016-2250-x

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: Protein domains identifier. *Nucleic Acids Research*, *33*, W116–W120. https://doi.org/10.1093/nar/gki442

Rhodes, D. H., Hoffmann, L., Rooney, W. L., Herald, T. J., Bean, S., Boyles, R., Brenton, Z. W., & Kresovich, S. (2017). Genetic architecture of kernel composition in global sorghum germplasm. *BMC Genomics*, *18*, 15. https://doi.org/10.1186/s12864-016-3403-x

Rich-Griffin, C., Stechemesser, A., Finch, J., Lucas, E., Ott, S., & Schäfer, P. (2020). Single-cell transcriptomics: A high-resolution avenue for plant functional genomics. *Trends in Plant Science*, *25*, 186–197. https://doi.org/10.1016/j.tplants.2019.10.008

Sarkar, N. K., Kim, Y.-K., & Grover, A. (2014). Coexpression network analysis associated with call of rice seedlings for encountering heat stress. *Plant Molecular Biology*, *84*, 125–143. https://doi.org/10.1007/s11103-013-0123-3

Sasaki, T. (2005). The map-based sequence of the rice genome. *Nature*, *436*, 793–800. https://doi.org/10.1038/nature03895

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., … Wilson, R. K. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, *326*, 1112–1115. https://doi.org/10.1126/science.1178534

Shen, S., Park, J. W., Lu, Z., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., & Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences*, *111*, E5593–E5601. https://doi.org/10.1073/pnas.1419161111

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: The teenage years. *Nature Reviews. Genetics*, *20*, 631–656. https://doi.org/10.1038/s41576-019-0150-2

Stein, J. C., Yu, Y., Copetti, D., Zwickl, D. J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J. L., Wei, S., Wang, J., Liao, Y., Wang, M., Jacquemin, J., Becker, C., Kudrna, D., Zhang, J., Londono, C. E. M., … Wing, R. A. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nature Genetics*, *50*, 285–296. https://doi.org/10.1038/s41588-018-0040-0

Sun, X., Zheng, H., Li, J., Liu, L., Zhang, X., & Sui, N. (2020). Comparative transcriptome analysis reveals new lncRNAs responding to salt stress in sweet sorghum. *Frontiers in Bioengineering and Biotechnology*, *8*, 331. https://doi.org/10.3389/fbioe.2020.00331

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*, W609–W612. https://doi.org/10.1093/nar/gkl315

Teng, J., Gao, Y., Yin, H., Bai, Z., Liu, S., Zeng, H., Bai, L., Cai, Z., Zhao, B., Li, X., Xu, Z., Lin, Q., Pan, Z., Yang, W., Yu, X., Guan, D., Hou, Y., Keel, B. N., Rohrer, G. A., … Fang, L. (2024). A compendium of genetic regulatory effects across pig tissues. *Nature Genetics*, *56*, 112–123. https://doi.org/10.1038/s41588-023-01585-7

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in Bioinformatics*, *19*, 575–592.

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., & Sorrells, M. E. (2021). Designing future crops: Genomics-assisted breeding comes of age. *Trends in Plant Science*, *26*, 631–649. https://doi.org/10.1016/j.tplants.2021.03.010

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., Zhang, H., 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. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nature Biotechnology*, *35*, 969–976. https://doi.org/10.1038/nbt.3943

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.

Vinogradov, A. E. (2004). Compactness of human housekeeping genes: Selection for economy or genomic design? *Trends in Genetics*, *20*, 248–253. https://doi.org/10.1016/j.tig.2004.03.006

Wu, Z., Cai, X., Zhang, X., Liu, Y., Tian, G.-B., Yang, J.-R., & Chen, X. (2022). Expression level is a major modifier of the fitness landscape of a protein coding gene. *Nature Ecology & Evolution*, *6*, 103–115.

Xiao, Q., Huang, T., Cao, W., Ma, K., Liu, T., Xing, F., Ma, Q., Duan, H., Ling, M., Ni, X., & Liu, Z. (2022). Profiling of transcriptional regulators associated with starch biosynthesis in sorghum (*Sorghum bicolor* L.). *Frontiers in Plant Science*, *13*, 999747. https://doi.org/10.3389/fpls.2022.999747

Xiao, Q., Wang, Y., Li, H., Zhang, C., Wei, B., Wang, Y., Huang, H., Li, Y., Yu, G., Liu, H., Zhang, J., Liu, Y., Hu, Y., & Huang, Y. (2021). Transcription factor ZmNAC126 plays an important role in transcriptional regulation of maize starch synthesis-related genes. *The Crop Journal*, *9*, 192–203. https://doi.org/10.1016/j.cj.2020.04.014

Xiao, S.-J., Zhang, C., Zou, Q., & Ji, Z.-L. (2010). TiSGeD: A database for tissue-specific genes. *Bioinformatics*, *26*, 1273–1275. https://doi.org/10.1093/bioinformatics/btq109

Xiao, X., Zhu, M., Liu, Y., Zheng, J., Cui, Y., Xiong, C., Liu, J., Chen, J., & Cai, H. (2023). Phenotypical and gene co-expression network analyses of seed shattering in divergent sorghum (*Sorghum* spp.). *The Crop Journal*, *11*, 478–489. https://doi.org/10.1016/j.cj.2022.08.009

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., & Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, *21*, 650–659.

Yang, H. (2009). In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. *Biology Direct*, *4*, 45. https://doi.org/10.1186/1745-6150-4-45

Yang, J., Su, A. I., & Li, W.-H. (2005). Gene expression evolves faster in narrowly than in broadly expressed mammalian genes. *Molecular Biology and Evolution*, *22*, 2113–2118. https://doi.org/10.1093/molbev/msi206

Yang, R. Y., Quan, J., Sodaei, R., Aguet, F., Segrè, A. V., Allen, J. A., Lanz, T. A., Reinhart, V., Crawford, M., Hasson, S., Consortium, G., Ardlie, K. G., Guigó, R., & Xi, H. S. (2018). A systematic survey of human tissue-specific gene expression and splicing reveals new opportunities for therapeutic target identification and evaluation. bioRxiv. https://doi.org/10.1101/311563

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586–1591. https://doi.org/10.1093/molbev/msm088

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., Xie, M., Zeng, P., Yue, Z., Wang, W., Tao, Y., Bian, C., Han, C., Xia, Q., Peng, X., Cao, R., Yang, X., Zhan, D., Hu, J., … Wang, J. (2012). Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology*, *30*, 549–554. https://doi.org/10.1038/nbt.2195

Zhang, L., & Li, W.-H. (2004). Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Molecular Biology and Evolution*, *21*, 236–239. https://doi.org/10.1093/molbev/msh010

Zhang, Y., Han, E., Peng, Y., Wang, Y., Wang, Y., Geng, Z., Xu, Y., Geng, H., Qian, Y., & Ma, S. (2022). Rice co-expression network analysis identifies gene modules associated with agronomic traits. *Plant Physiology*, *190*, 1526–1542. https://doi.org/10.1093/plphys/kiac339

Zhou, Y., Sukul, A., Mishler-Elmore, J. W., Faik, A., & Held, M. A. (2022). PlantNexus: A gene co-expression network database and visualization tool for barley and sorghum. *Plant & Cell Physiology*, *63*, 565–572.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hu, Z., Chen, J., Olatoye, M. O., Zhang, H., & Lin, Z. (2024). Transcriptome-wide expression landscape and starch synthesis pathway co-expression network in sorghum. *The Plant Genome*, *17*, e20448. https://doi.org/10.1002/tpg2.20448