Lee *et al. BMC Plant Biology* (2025) 25:682
https://doi.org/10.1186/s12870-025-06657-w

BMC Plant Biology

RESEARCH

Open Access

Weighted gene co-expression network analysis – based selection of hub genes related to phenolic and volatile compounds and seed coat color in sorghum

Ye-Jin Lee[1,2†], Woon Ji Kim[1†], Seung Hyeon Lee[1,3], Jae Hoon Kim[1], Soon-Jae Kwon[1], Joon-Woo Ahn[1], Sang Hoon Kim[1], Jin-Baek Kim[1], Jae Il Lyu[4], Chang-Hyu Bae[2*] and Jaihyunk Ryu[1*]

Abstract

Background Sorghum grains are rich in phenolic compounds, which are noted for their anticancer, antioxidant, and anti-inflammatory properties, as well as volatile compounds (VOCs) that contribute to aroma and fermentation processes. There is a known close relationship between sorghum coat color and phenolic compound content (PCC), particularly flavonoids which are pigments that confer red and purple colors in flowers and seeds.

Results Our results showed that black seeds had the highest total tannin content (TTC) and ketone content, which were measured at 457.7 mg CE g-1 and 96 g 100 g-1, respectively, which were 4.87 and 1.35 – fold higher than those of white seeds. L* showed a negative correlation between TTC (r = -0.770, P < 0.01) and ketone (r = -0.814, P < 0.01), while TFC and a* showed a strong positive correlation (r = 0.829, P < 0.001). RNA sequencing analysis identified 1,422 up-regulated and 1,586 down-regulated differentially expressed genes. Weighted gene co-expression analysis highlighted two color-related gene modules: the magenta 2 module associated with TTC, TPC, VOCs and L* value, and the blue module associated with TFC, and a* values. Hub genes identified within these modules included *ABCB28* in the magenta 2 module, and *PTCD1* and *ANK* in the blue module.

Conclusions We confirmed the relationship between PCC, VOCs, and seed coat color, with darker seed coat colors showing higher tannin, ketone contents and redder colors indicating higher flavonoid content. Network analysis helped pinpoint key genes involved in these traits. This study will provide essential data for improving the food and industrial use of sorghum.

Keywords Sorghum, Seed coat color, Phenolics, Volatile compound, R-seq, Network analysis

†Ye-Jin Lee and Woon Ji Kim contributed equally as first authors.

*Correspondence:
Chang-Hyu Bae
chbae@scnu.ac.kr
Jaihyunk Ryu
jhryu@kaeri.re.kr
[1]Advanced Radiation Technology Institute, Korea Atomic Energy Research Institute, Jeongeup 56212, Republic of Korea
[2]Deparment of Plant Production Sciences, Graduate School, Sunchon National University, Suncheon 57922, Republic of Korea
[3]Department of Integrative Food, Bioscience and Biotechnology, Chonnam National University, Gwangju 61186, Republic of Korea
[4]Department of Agricultural Biotechnology, Rural Development Administration (RDA), National Institute of Agricultural Sciences, Jeonju 54874, Republic of Korea

Background

Sorghum (*Sorghum bicolor* L.) is a widely used crop for feed, food, brewing materials, and biomass, such as biofuel and fiber [1–3]. Sorghum grains contain a variety of secondary metabolites. Compared to other grains, sorghum grains contain a higher content of various phenolic components, such as phenolic acids, 3-deoxyanthocyanidins, condensed tannins, and flavonoids [4, 5]. These phenolic compounds have been shown to reduce the risk of many chronic diseases, such as obesity, diabetes, cancer, and cardiovascular disease [6, 7]. In addition, sorghum seeds contain various VOCs such as esters, alkanes, and alcohols [8]. These volatile compounds affect various flavors and ethanol fermentation when using sorghum seeds as tea or alcohol [8, 9]. Therefore, the metabolites present in sorghum play an important role in enhancing its health benefits and industrial value.

Sorghum grains have both nonpigmented and pigmented seed coats, appearing in various colors such as white, red, brown, and black [10]. The color of sorghum seeds is closely associated with their phenolic content, as darker seeds typically contain higher levels of phenolics [10]. Among these compounds, tannins are strongly correlated with seed coat color, with brown and red grains typically containing more tannins [11]. Sorghum phenolic compounds and seed coat color have been extensively researched, with a focus on the correlation between phenolic compounds and antioxidant activity as well as differences in seed coat color among various sorghum varieties [12]. Studies on sorghum landraces from South Africa have highlighted the physical and nutritional differences based on seed coat color and phenolic content [13]. In this study, unlike previous studies that focused on the PCC according to the seed coat color of sorghum, the content of phenolic and VOCs related to the seed coat color was explored for the first time.

RNA sequencing (RNA-seq) is a ubiquitous tool in molecular biology that is extensively employed to elucidate genomic functions in plants [14]. RNA-seq data are essential for research that is hypothesis-driven as well as for performing data-driven analyses—which generate novel insights and testable hypotheses, thereby directing subsequent functional studies. Primarily, RNA-seq is used to measure gene expression levels and analyze differential gene expressions. RNA-seq is not only applied to model plant species, such as *Arabidopsis* [15, 16] and rice [17, 18], but has also been actively used in recent sorghum research. This technology plays a critical role in comprehensively understanding various aspects of sorghum, including developmental stages, physiological and morphological characteristics, as well as stress tolerance and responses.

Weighted gene co-expression network analysis (WGCNA) is a systems biology method used to explain the patterns of gene correlation across different samples. This approach identifies gene groups (modules) that share similar expression patterns, analyzes the correlation between sample phenotypes and modules, deduces regulatory networks within these modules, and identifies key regulatory genes [19, 20]. The adjacency calculation related to weighted networks provides more biologically relevant information [21]. WGCNA has been successfully employed in various crop studies, including the development of gene models for drought and salt stress experiments in *Arabidopsis* and rice [22–24], identification of desiccation-resistant genes in the herb *Boea hygrometrica* [25], and discovery of key genes involved in the growth and development of potato (*Solanum tuberosum L.*) [26]. Recently, WGCNA has been applied to major crops, such as rice and sorghum, to analyze gene networks related to salt and drought tolerance, contributing to the improvement in important agronomic traits.

The purpose of this study is to elucidate the correlation between the pericarp color and related metabolites in sorghum, and to identify key genes associated with these traits. Differential gene expression analysis (DEGs) was conducted using RNA-seq on four selected lines with different pericarp colors from a mutant sorghum population, and important hub genes were identified using WGCNA and network analysis. This research will provide fundamental data for enhancing the food and industrial applications of sorghum.

Methods

Plant materials

The four sorghum lines used in this study, distinguished by different seed coat colors (S1; white, S2; red, S3; brown, S4; black), were selected from a sorghum mutant population [27]. The mutant population was established by selecting domestically cultivable lines from genetic resources provided by the International Crops Research Institute for the Semi-Arid Tropics and the Rural Development Administration of Korea in 2014 and irradiating them with various doses of gamma rays and protons. The irradiated sorghum seeds were cultivated at the Radiation Breeding Research Farm of the Korea Atomic Energy Research Institute (Jeongeup, 35.51°N and 126.83°E, Republic of Korea). After M_2 generation, elite lines with superior agricultural traits were continuously selected up to the M_7 generation. From this mutant population, four lines varying in seed coat color were chosen for DEG analysis related to the content of metabolite compounds.

Evaluation of sorghum seed coat color and phenolic compound content

To accurately measure the seed coat color of sorghum, a color chart was positioned on top of the seeds, and LED lights were placed on both sides of the camera to capture the image. The background color for the seed coat photography was blue, and the images were taken using a Sony digital camera (a6000, SONY). The captured RGB image data were analyzed by modifying an automated image extraction method initially developed for soybean seeds [28], which was implemented using ImageJ software [29] to process the images and extract data. The extracted RGB values were converted to Lab values for statistical analysis. The a∗ value denotes the color component between red and green, where a positive a* value indicates red, and a negative a∗ value indicates green. The magnitude of this value reflects the intensity of the color. Dried sorghum seed powder, 1 g, was mixed with 5 mL of 80% methanol and extracted at 25 °C for one hour. Total anthocyanin content (TAC) was measured by adding 1% formic acid, ground and extracted in dark conditions at 4 °C, followed by centrifugation at 13,000 rpm for 20 min. In addition, the absorbance of the supernatant was measured at 535 nm using cyanidin chloride (CCE) as the positive control. Total tannin content (TTC) was determined by reacting the supernatant with vanillin-HCl solution at 25 °C for 20 min and measuring the absorbance at 500 nm, with catechin (CE) used as the positive control. Total phenolic content (TPC) was determined using the Folin-Ciocalteu method by reacting the seed extract with 2% sodium carbonate solution and measuring the absorbance at 750 nm, with tannic acid (TAE) as the positive control. Total flavonoid content (TFC) was measured by sequentially adding 1% sodium nitrite, 60% ethanol, 2% aluminum chloride, and 5% sodium hydroxide, and then measuring the absorbance at 405 nm using quercetin (QE) as the positive control. The extraction and quantification followed previous studies, with some modifications to measure each compound's content [30, 31].

GC-MS analysis of volatile compounds

Four samples of fresh seed powder (100 g) were extracted for 4 h using 500 mL of n-hexane: acetone (1:1 v/v) as the solvent. Anhydrous sodium sulphate was used to remove water from the extract, which was then filtered through a PVDF syringe filter (0.45-μM pore size) for gas chromatography-mass spectrometry (GC-MS) analysis. Three replicates of each sample were analyzed. The VOCs compositions were analyzed using GC-MS (Plus-2010, Shimadzu, Kyoto, Japan) equipped with an Rtx-5MS (30 m × 0.32 mm × 50 μm, Shimadzu, Kyoto, Japan) column. 99.99% high-purity helium was used as the carrier gas, and the column flow rate was 1.37 mL/min. The sample was injected in spitless mode. Initially, the oven was set to 40 °C and was then ramped up to 300 °C, increasing at 5 °C per min, followed by a 5 min hold at the final temperature. The parameters set for the mass spectrometry included electron-impact ionization at 70 eV, ion source temperature at 230 °C, and a scan range between 40 and 500. The detected VOCs in sorghum genotypes were tentatively identified using the GC-MS analysis based on an NIST library similarity index greater than 90%. We followed previous studies with some modifications to measure the content of VOCs [32].

RNA extraction

Total RNA was extracted from three biological replicates of sorghum seeds at the 8.5 developmental stage (stage 8; dough stage, and stage 9; physiological maturity), 90 days after emergence, representing four different seed coat colors: S1 (white), S2 (red), S3 (brown), and S4 (black). All seed samples were stored in a deep freezer immediately upon collection. The frozen seeds, weighing 100 mg each, were ground into fine powder under liquid nitrogen. Subsequently, polyphenols and polysaccharides were removed using Fruit-mate (Takara, Shiga, Japan), and RNA was extracted using TRIzol reagent (Invitrogen, Carlsbad, CA, USA). Then, 100 mg of ground seed powder was treated with Fruit-mate, followed by mixing the supernatant with an equal volume of TRIzol reagent to separate the components. Chloroform was then added and vortexed to lyse the cells. Next, isopropanol was added and the mixture was incubated at a low temperature to precipitate the RNA. The RNA pellet was washed with 75% ethanol and resuspended in DEPC-treated water. The concentration and purity of the extracted total RNA were measured using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA).

cDNA library construction and sequencing

Using the TruSeq® sample preparation kit (Illumina, San Diego, CA, USA), paired-end libraries were prepared. The quality and concentration of the libraries were assessed with a 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Sequencing was performed on then HiSeq X Ten system (Illumina) to produce 150 bp paired-end reads. Following cDNA synthesis, PCR amplification of the adapter-ligated cDNA fragments was carried out with adapter-specific primers. High-throughput sequencing of all libraries was conducted on the Illumina HiSeq X platform. The raw sequencing data from three replicates across four genotypes have been stored in the NCBI Sequence Read Archive, accessible via the provided accession number: (NCBI < raw data).

Data preprocessing and classification of DEGs

Adapter sequences were removed from the transcripts using Trimmomatic (v. 0.39) [33]. Quality control of the sequence data was performed using the SLIDINGWINDOW, LEADING, and TRAILING options as follows: (1) window size = 4, mean quality ≥ 15, (2) LEADING, TRAILING ≥ 3, (3) minimum read length ≥ 36 bp.

After these preprocessing steps, the cleaned reads were mapped using HISAT2 software [34], and the expression values (read counts) were calculated based on the total number of reads mapped to each gene using HTSeq (v.0.11.0) [35]. To ensure consistency in the expression data, normalization relative to the total counts was performed using the DESeq library [36]. Gene functions were identified by annotating against the viridiplantae database of amino acid sequences in NCBI NR using BLASTP, with a filter criterion of e-value ≤ $1e^{-10}$ [available online: https://www.ncbi.nlm.nih.gov/ (accessed on 1 June 2024)]. The DEGs among the samples were selected using both a two – fold change method, which identifies a two – fold or greater difference in expression between comparative samples, and a binomial test method with an adjusted P – value (FDR) ≤ 0.01. In this study, genes were named up-regulated if the log_2 (Fold Change) value was greater than 1 and down-regulated if it was less than – 1. To analyze the expression patterns of significantly expressed genes, hierarchical clustering analysis was conducted using the amap and gplot libraries in R. This analysis employed Pearson's correlation to evaluate the similarity of expression patterns among genes, and the "complete" method was used for grouping the genes.

Gene ontology and Kyoto encyclopedia of gene and genome enrichment analyses

Gene Ontology (GO) analysis of the selected DEGs used GO information provided by reference databases and was conducted using in-house scripts. The significance level was set at 0.05, and functional categories were classified into biological processes (BP), cellular components (CC), and molecular functions (MF). Additionally, annotations were performed using the amino acid sequences from the Kyoto Encyclopedia of Gene and Genomes (KEGG) database and BLASTP, applying a filter criterion of e-value ≤ 1e-100 for the best hits.

WCGNA

Based on the RNA-Seq data, a weighted gene co-expression network was constructed using the WGCNA package [37], identifying highly correlated gene clusters (modules). The data were normalized using the DESeq2 package to convert them to FPKM format, followed by log2 transformation. Subsequently, the scale-free topology index was calculated to determine the optimal soft-thresholding power, and this value was used to identify modules through hierarchical clustering and dynamic tree-cutting methods. The correlation between each sorghum trait and the gene modules was assessed by calculating Pearson's correlation coefficients and statistically evaluated. Edge data filtering was performed based on the weights of the edges. In particular, to select edges from the magenta2 and blue modules, the 50th and 98th percentiles of the weight data were calculated, targeting the top 50% and top 2% of the weights, respectively, and these thresholds were used as the basis for filtering. For network analysis, we used the igraph package in the R program to generate undirected graphs and calculated the closeness centrality. The visualization used the Kamada-Kawai layout, and the node sizes were adjusted proportionally to their centrality values. The node colors were assigned using a color palette that transitions from blue to red depending on the centrality values.

Validation of gene expression

To validate the selected hub genes, sorghum seeds experiencing a deepening of seed coat color between the hard dough and physiological maturity stages were collected and immediately frozen in liquid nitrogen. The extracted total RNA was treated with a DNA-free kit (Invitrogen, Grand Island, NY, USA) to remove DNA contamination. Then, the first-strand cDNA was synthesized using the SuperScript III First-Strand Synthesis SuperMix kit (Invitrogen, Grand Island, NY, USA). The transcription levels of DEGs were determined by qRT-PCR using Universal SYBR Green SuperMix (Bio-Rad, Hercules, CA, USA) on the CFX96 RT-PCR system (Bio-Rad), with SAMDC used as the reference gene. The PCR conditions were programmed with an initial denaturation step at 95 °C for 10 min, followed by 40 cycles consisting of 15 s at 95 °C, 15 s at 50 °C, and 30 s at 72 °C. Each sample was analyzed in quintuplicate. Relative transcription levels were calculated according to the $2^{-\Delta\Delta Ct}$ method. Gene-specific primers based on the coding sequences of *Sorghum bicolor* L. were designed using Primer3Plus software; details of the primers are provided in Supplementary Table 10.

Results

Metabolites and seed coat color profiling in sorghum

Sorghum seeds were categorized by color as white, red, brown, and black, and phenolic compounds were analyzed according to the seed coat color, revealing distinct differences across all phenolic compounds (Table 1; Fig. 1a). In the color space, the L* values, indicating lightness, ranged from 3.0 in black seeds (S4) to 73.6 in white seeds (S1). The S2 seeds, with a red seed coat, showed the highest a∗ value of 24.8. In contrast, the S3 seeds, with a brown seed coat, had an a∗ value of 16.1. The a∗ values for the S1 and S4 samples were – 0.4 and – 1.1, respectively, indicating low intensity of red. Additionally, the b∗ value, representing the color components between yellow (positive) and blue (negative), was measured at 12.9 for the S2 samples.

The PCC analysis showed that TTC accounted for the largest proportion (more than 90% on average) of

Table 1 Analysis of sorghum seed coloration and inclusion of seed images with lab color space values

Lines	Phenotype	Color	L*[a]	a*	b*
S1		White	73.6 ± 0.3a[b]	−0.4 ± 1.1c	0.6 ± 0.7b
S2		Red	22.3 ± 0.3b	24.8 ± 1.5a	12.9 ± 5.1a
S3		Brown	8.6 ± 0.1c	16.1 ± 2.0b	0.2 ± 1.5b
S4		Black	3.0 ± 0.3d	−1.1 ± 1.1c	−7.1 ± 1.6c

[a] L*, lightness; a*, red/green value; b*, blue/yellow value

[b] The letters adjacent to average ± standard deviation indicates the result of Fisher's LSD test at the 5% level (n=3)

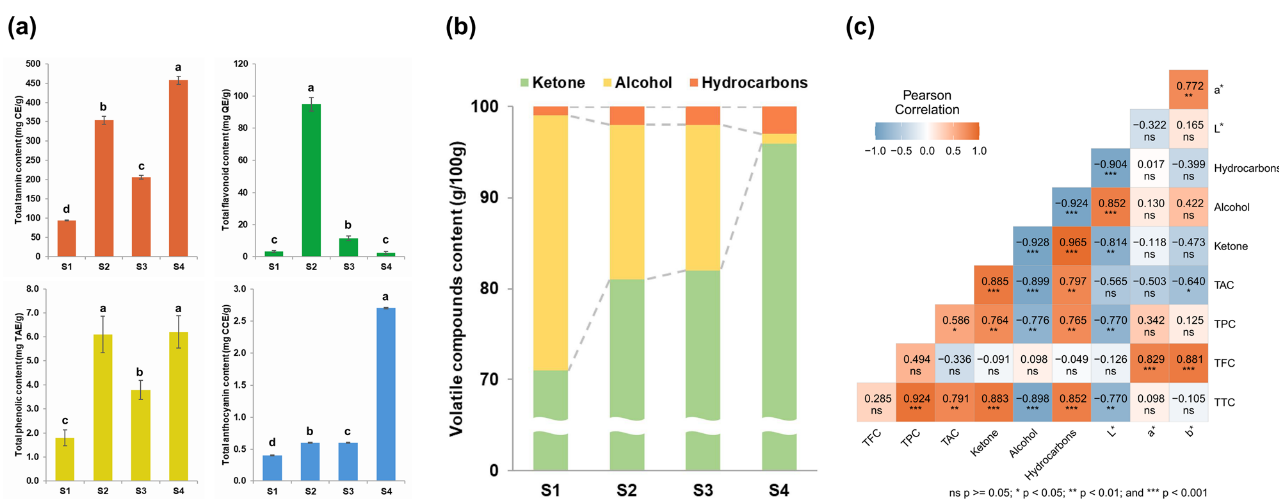

Fig. 1 Phenotypic and metabolite content of sorghum seeds and correlation analysis between each traits. (**a**) PCC in sorghum seed samples. S1, white; S2, red; S3, brown; S4, black; CE, catechin; QE, quercetin E; TAE, tannic acid E; CCE, cyanidin chloride. Lowercase letters above the bars indicate significant differences between genotypes at the 5% level according to Fisher's LSD test ($n = 3$). (**b**) Volatile compound content released from the sorghum seed samples. (**c**) Correlation analysis between metabolite compounds and seed phenotypes. TAC, total anthocyanin content; TFC, total flavonoid content; TPC, total phenolic content; TTC, total tannin content; L*, lightness; a*, red/green value; b*, blue/yellow value; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; and ns, not significant

PCC in all genotypes, while TFC accounted for the second highest proportion (more than 7.4% on average) (Table S1). TTC was highest in S4 (457.7 mg CE g^{-1}) with a 4.87 – fold increase compared to white seeds (S1; 94 mg CE g^{-1}), followed by S2 (353.9 mg CE g^{-1}) with a 3.76 – fold increase and S3 (206.4 mg CE g^{-1}) with a 2.20 – fold increase. TFC was highest in S2 (94.9 mg QE g^{-1}; 28.76 – fold higher), followed by S3 (11.5 mg QE g^{-1}; 3.48 – fold higher) and S4 (2.3 mg QE g^{-1}; 0.70 – fold lower). All genotypes showed significant differences in PCC content based on seed coat color.

Each sample was analysed for VOCs (Table S2). The GC-MS analysis detected six VOCs in the sorghum genotypes, all of which were tentatively identified by mass spectra and retention time based on a NIST library similarity index above 90%. As shown in Fig. 1b, the VOCs belonged to three classes: ketones (4-methoxy-4-methyl-2-pentanone, 3,5-dimethyl-2-cyclohexene-1-one, isophorone, 3,4-dihydro-3,3,6,8-tetramethyl-1(2 H)-naphthalenone), alcohols (2-methyl-2-pentanol), and hydrocarbon (dehydro-cyclolongifolene oxide). Ketones were the dominant VOC category present in all sorghum seeds. The content of ketones in the volatiles for all sorghum genotypes ranged from 71 g 100 g $^{-1}$ (S1) to 96 g 100 g $^{-1}$ (S4). The highest alcohol content (28 g 100 g $^{-1}$) was observed in S1 and the lowest (1 g 100 g $^{-1}$) was observed in S4. As the TTC and TPC of sorghum seeds increased, the alcohol content decreased, and the ketone content increased.

Pearson correlation analysis was conducted to investigate the correlation between PPC, VOCs and Lab values (Fig. 1c). TTC showed a strong positive correlation with TPC ($r = 0.924$, $P < 0.001$), ketone ($r = 0.883$, $P < 0.001$), and HC ($r = 0.852$, $P < 0.001$), and a strong negative correlation with alcohol (r = -0.898, $P < 0.001$). The L* value showed a strong positive correlation with alcohol ($r = 0.852$, $P < 0.001$) and a negative correlation with HC (r = -0.904, $P < 0.001$), ketone (r = -0.814, $P < 0.01$), TTC (r = -0.770, $P < 0.01$), and TPC (r = -0.770, $P < 0.01$). TFC exhibited strong positive correlations with both the a* value ($r = 0.829$, $P < 0.001$) and b* value ($r = 0.881$, $P < 0.001$). These results show that there is a strong correlation between lightness and TTC, TPC, and VOCs, and TFC is strongly positively correlated with the red (a*) and yellow (b*) values. In addition, the interactions between ketones and alcohols with other compounds reveal significant biochemical relationships.

Analysis of transcriptome sequencing data

The RNA-seq analysis included three replicates of sorghum seeds at stage 8.5, when seed coat color is expressed (Fig. 2). In total, 155,341,865 clean reads were produced, having an average length of 151 bp. The percentage of bases with a Phred quality score of 30 (Q30) varied between 89.05% and 91.77%, averaging 90.34% (Table 2), indicating the high quality of the sequencing data. Following the mapping of the collected transcripts, we produced 134,338,393 trimmed reads, averaging 11,194,866 reads for each sample. The reads were aligned to the reference open-reading sequence, achieving an average mapping rate of 86.45%. We then calculated normalized read counts using mapped reads to ascertain gene expression levels. Of the 41,048 reference transcripts (15,601,381 bp long), 32,478 (79.12%) were expressed (Table S3) and annotated according to *SbicolorRio*_v2 in the viridiplantae database of NCBI NR (Table S4). To evaluate the reproducibility of the results among the three biological replicates per sample, we executed a hierarchical clustering analysis to examine gene expression patterns, based on Pearson's correlation coefficient (Fig. 2a). The correlation coefficients among the three replicates in S1, S2, S3, and S4 ranged from 0.94 to 0.98, indicating high reproducibility of the RNA-seq data.

Analysis of differential gene expressions in sorghum seeds with different seed coat color

The DEGs were analyzed by comparing S1 (white) (with light seed color) and S2 (red), S3 (brown), and S4 (black) (with dark seed color) at the 8.5 stage (Fig. 2b). A total of 5,536 DEGs were identified in the comparison between S2 vs. S1, which included 2,779 up-regulated and 2,757 down-regulated genes. In the comparison of S3 vs. S1, 3,560 DEGs were found, of which 1,590 were up-regulated and 1,964 were down-regulated. For S4 vs. S1, the analysis revealed 5,619 DEGs, with 2,991 up-regulated and 2,628 down-regulated. Across these comparisons, 166 up-regulated DEGs were consistent, while individual counts for S2 vs. S1, S3 vs. S1, and S4 vs. S1 were 2,282, 483, and 1,915 respectively (Fig. 2c). Regarding down-regulated DEGs, 282 were consistently noted across all comparisons, and specific counts were 2,096, 467, and 1,126 for S2 vs. S1, S3 vs. S1, and S4 vs. S1 respectively (Fig. 2d). When comparing to S1, we identified 3,008 DEGs that shared similar expression patterns among genes within colored seed coats in categories cluster 2 (1,422 up-regulated DEGs) and cluster 6 (1,586 down-regulated DEGs), (Figure S1a, b). Moreover, 2,142, 1,032, and 1,586 DEGs in categories cluster 3, cluster 5, and cluster 6 respectively mirrored expression patterns correlated with TPC. Particularly, 37 DEGs in cluster 3 were involved in the phenylpropanoid biosynthesis pathway, and 48 in the flavonoid biosynthesis pathway. In category cluster 6, 26 DEGs pertained to the phenylpropanoid pathway and 3 to the flavone/flavonol biosynthesis pathway, as detailed in Table S5.

GO and KEGG enrichment analyses of DEGs

Analysis of DEGs revealed that 1,422 genes were up-regulated (cluster 2) and 1,586 were down-regulated (cluster 6) (Figure S1). GO analysis identified 127 GO terms associated with cluster 2 (BP, 73; CC, 27; MF, 27) and 421 GO terms associated with cluster 6 (BP, 333; CC, 9; MF, 79), with statistical significance noted ($p < 0.01$). For the up-regulated genes in cluster 2, the most critical BP was the SCF-dependent proteasomal ubiquitin-dependent protein degradation process (GO:0031146). The most notable CC was the nucleosome (GO:0000786), and the most significant MF involved was the activity of ABC-type cadmium transporters (GO:0015434) ($p < 0.001$) (Fig. 3a and Table S6). Regarding the down-regulated genes in cluster 6, the most primary BP was the regulation of receptor-mediated endocytosis (GO:0048259). The most significant CC was the replisome (GO:0030894), and the most crucial MF involved was the phosphatidylinositol trisphosphate phosphatase activity (GO:0034594) ($p < 0.001$) (Fig. 3c and Table S6). In addition, cluster 1, containing 3,415 DEGs, showed enrichment in 378 GO terms (BP, 239; CC, 62; MF, 77), whereas cluster 3 (2,142

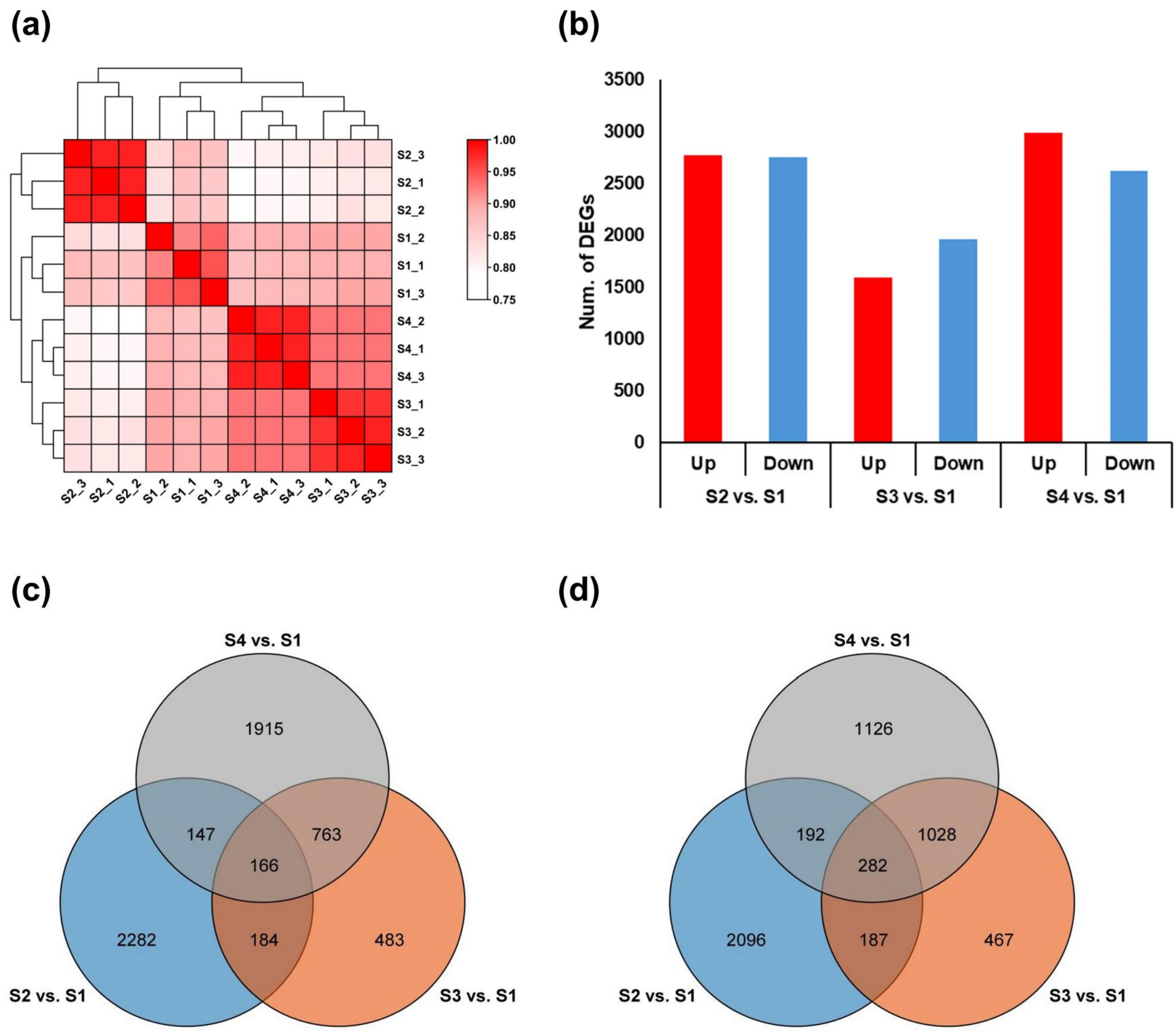

Fig. 2 Comparative analyses of gene expression profiles and differential expressions. (**a**) Hierarchical clustering illustrating the relationships between samples based on Pearson correlation coefficients. (**b**) Comparison of the number of DEGs in S1 with those in S2, S3, and S4. (**c**) Venn diagram representing the up-regulated DEGs, revealed by the comparison of S1 with S2, S3, and S4. (**d**) Venn diagram representing the down-regulated DEGs, revealed by the comparison of S1 with S2, S3, and S4

DEGs) had 268 GO terms (BP, 157; CC, 54; MF, 57). Similarly, cluster 4 (887 DEGs) encompassed 85 GO terms (BP, 26; CC, 10; MF, 49), and cluster 5 (1,032 DEGs) comprised 477 GO terms (BP, 353; CC, 16; MF, 108). In all of these clusters, the most prominent GO terms were intracellular anatomical structure (GO:0005622) and catalytic activity (GO:0003824) ($p < 0.001$) (Table S6).

Functional characterization of DEGs from the comparisons of S2 vs. S1, S3 vs. S1, and S4 vs. S1 also included KEGG enrichment analyses (Fig. 3b, d and Table S7). Analysis identified 102 KEGG pathways associated with up-regulated DEGs (cluster 2), and 118 pathways linked to down-regulated DEGs (cluster 6). These pathways were organized into five primary categories: The largest category, metabolism, encompassed 11 subcategories including metabolic processes and biosynthesis of secondary metabolites. Following this, the second-largest category was genetic information processing, which comprised six subcategories, notably protein folding. The third-largest category, Environmental Information Processing, was subdivided into two subcategories, including Signal Transduction. Cellular processes formed the fourth category, and organismal systems were ranked as the fifth category. Additionally, KEGG analysis of the remaining clusters revealed that cluster 1 had 129 pathways, cluster 3 had 125 pathways, cluster 4 had 74 pathways, and cluster 5 had 110 pathways. In these four

Table 2 Information on RNA-seq data from sorghum seeds at the 8.5 stage

Sample ID	Raw Reads	Avg. Length (bp)	Total Length (bp)	GC (%)	Q30 (%)	Clean Reads	Mapping Rate	
							No. of Reads	Percent (%)
S1_1	15,382,362	151	2,322,736,662	52.08	90.74	13,925,123	12,947,785	92.98
S1_2	13,274,896	151	2,004,509,296	52.02	91.53	12,139,330	11,343,194	93.44
S1_3	14,726,727	151	2,223,735,777	52.78	89.05	12,785,642	11,840,655	92.61
S2_1	15,344,564	151	2,317,029,164	52.26	90.42	13,632,491	12,848,970	94.25
S2_2	13,856,519	151	2,092,334,369	51.19	89.39	11,999,115	11,288,260	94.08
S2_3	14,591,565	151	2,203,326,315	52.14	90.91	13,256,824	12,449,316	93.91
S3_1	14,127,456	151	2,133,245,856	52.48	89.32	12,342,863	8,829,714	71.54
S3_2	15,537,113	151	2,346,104,063	50.99	89.39	13,481,404	10,274,658	76.21
S3_3	13,934,903	151	2,104,170,353	51.67	91.77	12,830,783	9,753,701	76.02
S4_1	14,437,349	151	2,180,039,699	52.05	90.79	13,108,022	11,046,894	84.28
S4_2	14,955,318	151	2,258,253,018	51.78	90.93	13,573,629	11,433,205	84.23
S4_3	13,904,360	151	2,099,558,360	52.04	89.88	12,266,639	10,282,041	83.82
Mean					90.34		11,194,866.08	86.45
Total	174,073,132		26,285,042,932			155,341,865	134,338,393	

GC (%): GC content. Q30 (%): ratio of bases that have Phred quality score of over 30. No. Reads: Number of mapped reads

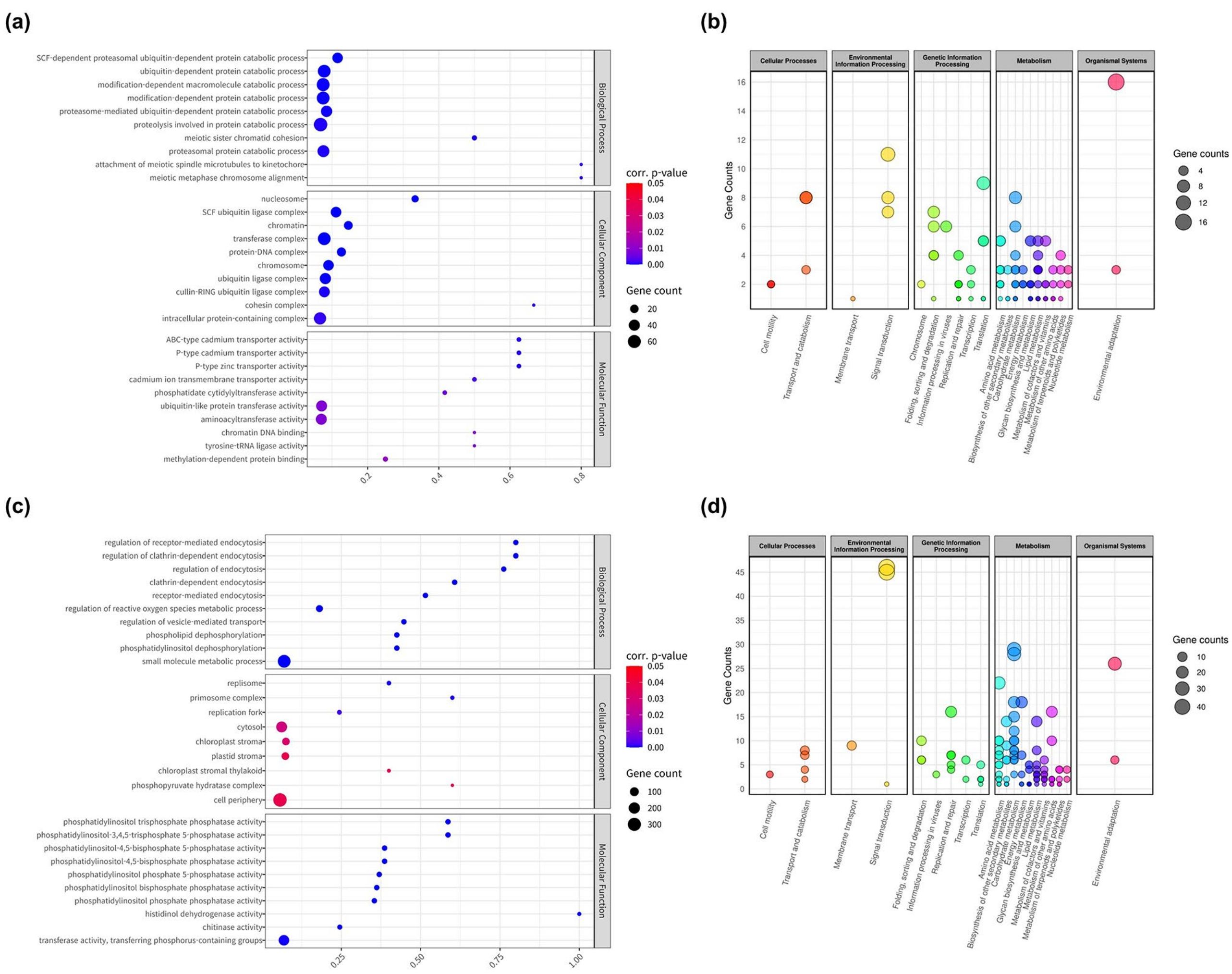

Fig. 3 GO and KEGG enrichment analyses of the DEGs in S1 vs. S2, S3, and S4 comparisons. (**a**, **b**) Enriched GO terms and KEGG categories among the up-regulated DEGs. (**c**, **d**) Enriched GO terms and KEGG categories among the down-regulated DEGs

clusters, the metabolism category was the most prominent (Table S7).

Identification of seed coat color-related gene co-expression modules by WGCNA

To identify the genes associated with seed coat color, we performed WGCNA by integrating data containing 32,498 genes obtained from the S2 vs. S1, S3 vs. S1, and S4 vs. S1 transcriptome comparisons with the physiological parameters PCC, VOCs, and Lab values. To determine the appropriate soft threshold value for WGCNA, the scale-free fit index was targeted to maintain a minimum level of 0.8. According to the graph, a threshold of 6 or higher consistently maintained a high scale-free fit index, which was then used as the basis for configuring the network (Fig. 4a, left). At higher thresholds, the network more stringently selects nodes for connection, incorporating only gene pairs with strong co-expression relationships (Fig. 4a, right). This selective approach enhances the precision of network analysis and enables clearer biological interpretations. The gene clustering dendrogram was constructed based on the correlation of gene expression. Each color represents a specific module, with gray indicating genes that could not be classified into any module. After the preliminary module division, the results were refined based on module eigengene similarity with a threshold > 0.40, leading to the merger of modules with similar expression patterns. Ultimately, this process yielded 14 co-expressed modules (Fig. 4b). Each module contains the following genes: sienna1 (51 genes), darkseagreen2 (82 genes), blue (2,204 genes), darkgoldrod4 (45 genes), dark olivegreen (71 genes), orange4 (50 genes), brown1 (70 genes), mediumpurple (95 genes), blue1 (42 genes), brown3 (41 genes), magenta2 (64 genes), lightcyan (415 genes), pink3 (73 genes), and grey (53 genes). The information on the genes belonging to each module is given in Table S8.

Analysis of sample expression patterns and screening of key modules

Pearson correlation coefficient ($r > 0.5$) and significance of p-value ($p < 0.05$) were used as criteria to assess the strength and significance of the associations between gene modules and traits such as sorghum seed coat color and PCC, and VOCs (Fig. 5a). The results showed that the magenta2 module had strong correlations with TTC ($r = 0.98$, $P < 0.001$), TPC ($r = 0.95$, $P < 0.001$), ketone ($r = 0.78$, $P < 0.01$), alcohol (r = -0.80, $P < 0.01$), and L* value (r = -0.70, $P = 0.01$). The negative correlation between alcohol and L* values indicates a strong adverse effect associated with this module. The blue module exhibited significant associations with TFC ($r = 0.98$, $P < 0.001$), a* value ($r = 0.76$, $P < 0.01$), and b* value ($r = 0.91$, $P < 0.001$). The light cyan module showed a significant correlation with TAC ($r = 0.85$, $P < 0.001$). Through the analysis of module membership versus gene significance, we confirmed how each module's membership correlates with the importance of genes for these traits (Fig. 5b, c). This demonstrates the significant role that each module plays in representing traits.

Network analysis and validation of gene expression

DEGs were selected from the magenta2 and blue modules, and their interactions were analyzed using closeness centrality. As a result, 12 and 16 genes were included in the association analysis for the magenta2 and blue modules, respectively (Table S9). Notably, in the magenta2 module, which showed a strong correlation with TTC, TPC, VOCs, and L* values, the gene *SbRio.02G135800.1*, with a closeness value of 0.774, was identified as a hub gene. This gene encodes the ABC TRANSPORTER B FAMILY MEMBER 28 (*ABCB28*), part of the AT-binding cassette (ABC) transporter family (Fig. 6a and Table S3). In the blue module, *SbRio.02G242300.1* and *SbRio.02G265800.1* were identified as hub genes, with closeness values of 0.491 and 0.518, respectively (Fig. 6b

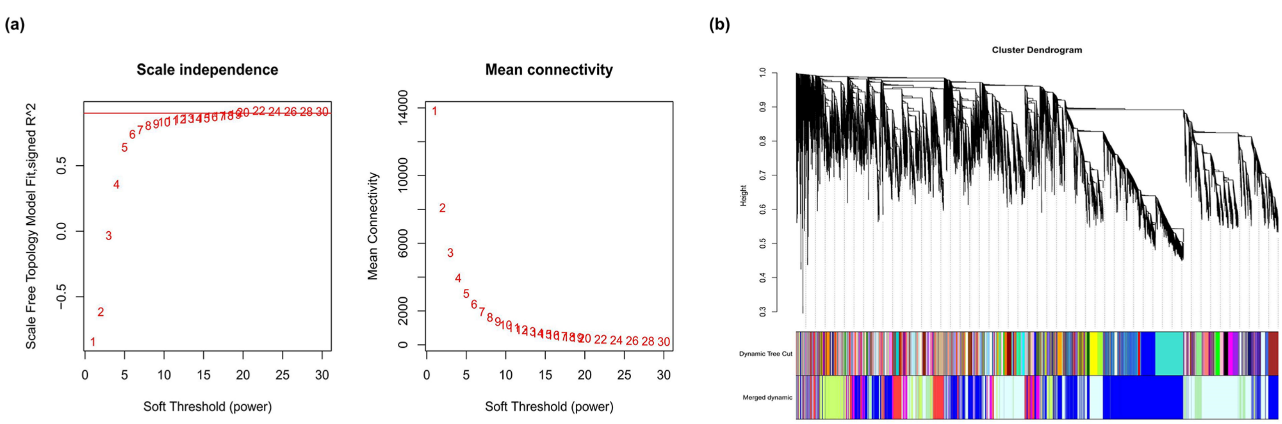

Fig. 4 Co-expression network construction by WGCNA. (**a**) Soft power curves, where the x-axis represents the power value, the y-axis (left) represents the correlation coefficient, and the y-axis (right) represents the average connectivity of genes. (**b**) Gene cluster dendrogram and module colors

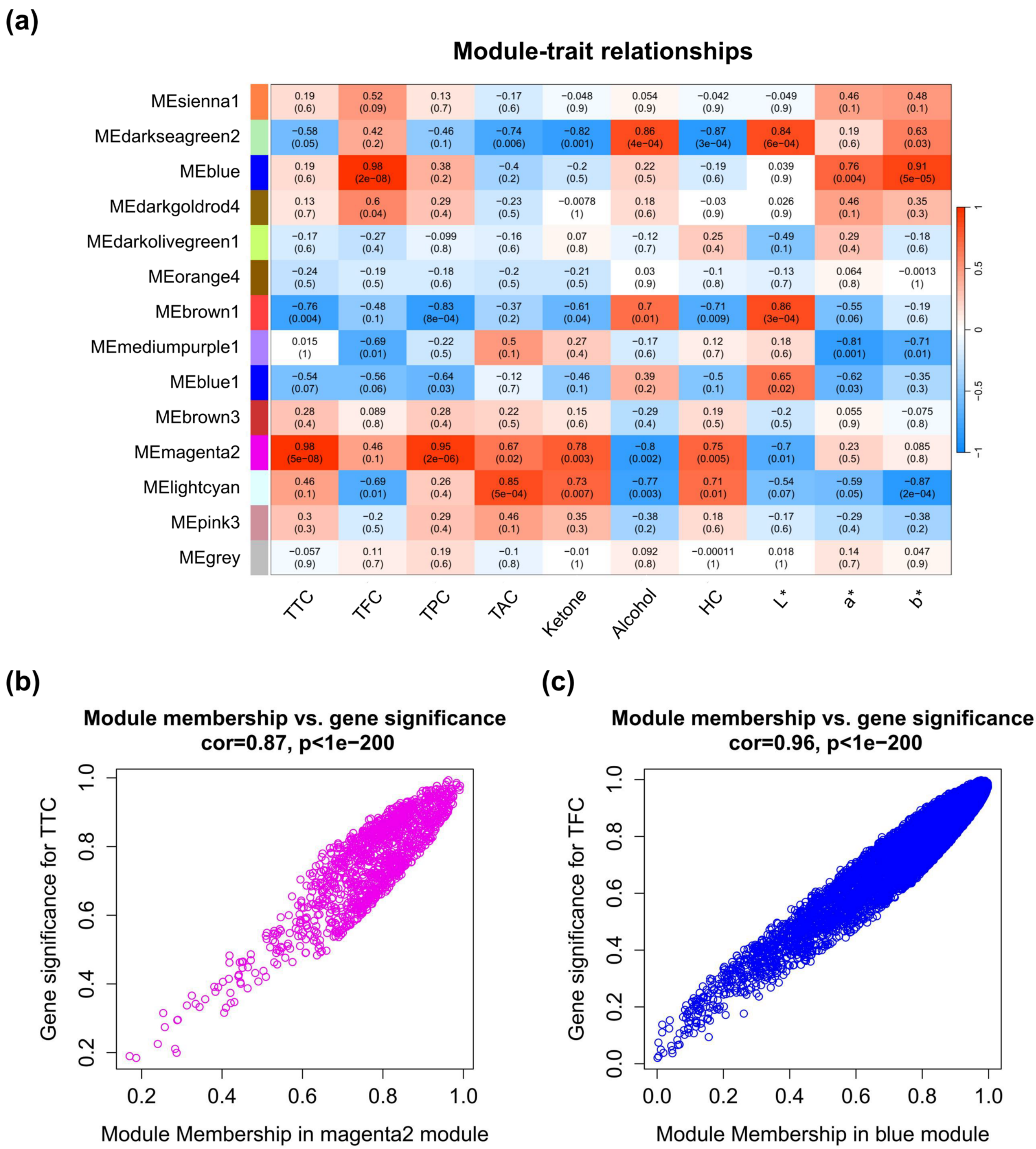

Fig. 5 Correlations between 14 modules and traits were analysed and gene significance of selected key modules were assessed (**a**) module-trait relationships. (**b**, **c**) Correlation analysis of key module membership and gene significance

and Table S3). *SbRio.02G242300.1* encodes the Pentatricopeptide Repeat Containing Domain 1 (*PTCD1*), which plays a role in the processing of mitochondrial RNA and maintaining its stability. *SbRio.02G265800.1* encodes the Ankyrin repeat-containing domain (*ANK*), a protein with ankyrin repeats. qRT-PCR analysis was conducted to verify the expression of hub genes from each module (Table S10). The expression levels of *ABCB28* were 2.62-, 1.66-, and 3.15 – fold higher in S2, S3, and S4, respectively, compared to S1, showing a similar expression pattern to the RNA-seq results (Fig. 6c). *PTCD1* expression increased by 2.03 – fold in S2 compared to S1, while it

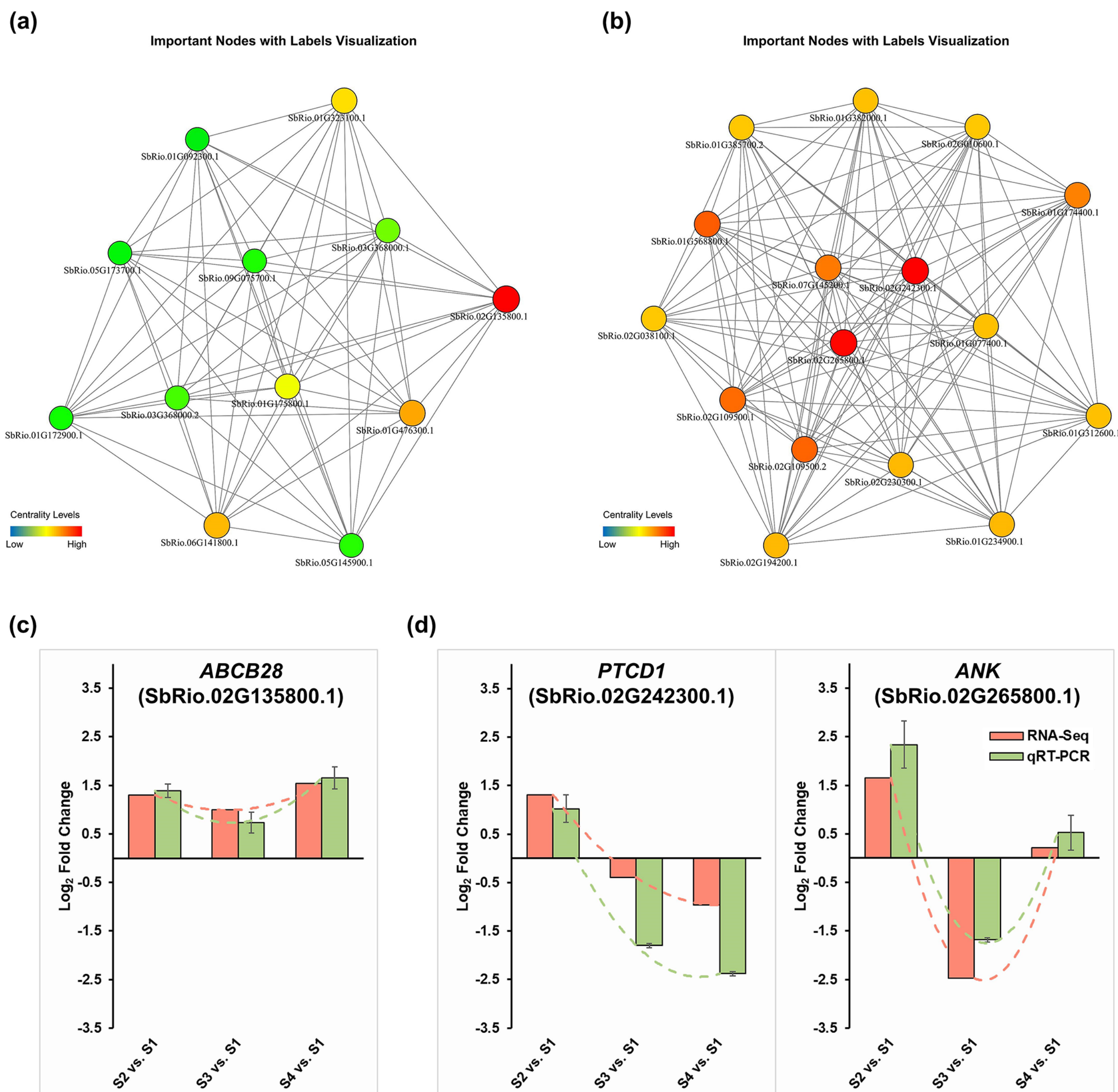

Fig. 6 Co-expression network of magenta2 (**a**) and blue (**b**) modules. (**c**) Expression validation of selected hub genes in Magenta2 module. (**d**) Expression validation of selected hub genes in blue module. Nodes colored closer to red represent higher centrality levels, while those closer to green represent lower centrality levels

was lower in S3 and S4, at 0.29-and 0.19 – fold, respectively. *ANK* expression was 5.06 – fold higher in S2 and 1.44 – fold higher in S4 but only 0.31 – fold higher in S3. Despite the differences, *PTCD1* and *ANK* showed expression patterns similar to the RNA-seq results (Fig. 6d). Additionally, other genes in the magenta2 module, such as *SbRio.01G476300.1* (*HAT22*), *SbRio.06G141800.1*, *SbRio.01G323100.1*, and *SbRio.01G175800.1*, displayed expression patterns consistent with the RNA-seq results (Figure S2). Likewise, in the blue module, genes represented by orange nodes, such as *SbRio.07G145200.1* (*UGT85A24*), *SbRio.01G174400.1* (*F3H*), *SbRio.01G568800.1* (*NADP*), and *SbRio.02G109500.1* (*CHAF1A*), also showed expression patterns similar to the RNA-seq results (Figure S3). The qRT-PCR data validated the RNA-seq results, thereby confirming the reliability of the transcriptomic findings.

Discussion

Sorghum phenolics are predominantly concentrated in the testa and vary in content depending on the sorghum variety and genotype [38]. Tannins and flavonoids are indirectly or directly associated with seed coat color. Sorghum contains high-molecular-weight tannins, and the concentration of tannins varies with the color of the variety [11]. Generally, as the seed coat of sorghum grains darkens, the content of tannins increases [39]. For example, red and brown sorghum have higher contents of biologically active compounds such as tannins [40]. In a preliminary screening of 80 sorghum mutant lines with diverse seed coat colors, we also observed a negative correlation between total tannin content and L* value (r = -0.648, $P<0.001$) (Table S11 and Figure S4). In the current research, the correlation between the numerically quantified Lab values of sorghum seed coat color and phenolic compounds was analyzed, revealing a significant negative correlation between TPC and L* value (lightness) (r = -0.770, $P<0.01$). This indicates that as lightness decreases, tannin content increases. However, seed coat color should not be considered an absolute criterion for determining the tannin content of sorghum [41], and not all sorghum varieties with colored seed coats contain tannins. In particular, the presence of a colored testa (tannin-containing layer) can influence the presence or absence of tannins [42]. Therefore, the high tannin content in the sorghum varieties used in this research is presumed to be due to the presence of colored testa. Although seed coat color is not a definitive indicator of tannin content, it can serve as an indirect marker in the breeding of high-tannin sorghum varieties. Flavonoids, as major secondary metabolites, provide the primary pigmentation in flowers and fruits, such as red and purple colors, through compounds such as anthocyanins and anthoxanthins [43]. Additionally, sorghum varieties with red seed coats showed a high correlation with the a* value (blue/red ratio) ($r=0.829$, $P<0.001$), where a positive a* value indicates redness, closely associated with flavonoid content. This suggests that sorghum varieties with red seed coats can exhibit high flavonoid levels, providing essential foundational data for breeding health-functional and industrially significant sorghum varieties.

In contrast to traditional mutant breeding approaches that prioritize DEGs based on expression levels alone [44, 45], we employed WGCNA to indicate gene modules that correlate more directly with specific phenotypic traits. This methodological advancement facilitates a deeper understanding of genetic networks influencing trait expression, enhancing our ability to identify critical genes for targeted breeding. Our analysis differentiated the phenolic and volatile profiles of four sorghum genotypes with diverse seed coat colors, integrating RNA-seq data with phenotypic traits to discover modules and hub genes critical for metabolite biosynthesis. These findings pave the way for future research to explore the functional roles of candidate genes in phenolic and volatile compound accumulation.

As a result of the WGCNA, we identified the magenta2 module, which is related to TTC, TPC, VOCs and L* values, and the blue module, which has a strong association with TFC and a* and b* values. Based on the DEGs of each module, we performed a closeness centrality network analysis. We applied closeness centrality to identify hub genes that are, on average, located closer to other genes within the network, thereby playing a critical role in metabolic pathways and functional interactions [46]. The closeness centrality analysis revealed 12 genes in the magenta2 module, with *SbRio.02G135800.1* having the highest centrality levels and being selected as the hub gene (Fig. 6a). The hub gene *SbRio.02G135800.1* encodes the *ABCB28* gene. The expression levels of *SbRio.02G135800.1* were 2.62-, 1.66-, and 3.15 – fold higher in S2, S3, and S4, respectively, compared to S1 (Fig. 6c). This *ABCB28* gene belongs to the ABC transporter family, which is known to transport various molecules across cellular membranes, primarily in an ATP-dependent process. Furthermore, ABC transporters play a crucial role in transporting various metabolites across plant cell membranes, particularly flavonoids, which are secondary metabolites in plants [47]. These metabolites include antioxidants that contribute to the plant's physiological response and human health benefits. In particular, multidrug resistance-associated protein-type ABC transporters are involved in flavonoid transport, especially in vacuolar flavonoid accumulation [48]. For example, in maize, the ABC transporter *ZmMRP3* is located on the tonoplast and helps accumulate anthocyanins [49], while in soybean roots, the plasma membrane ABC transporter mediates the secretion of isoflavone genistein [50]. Interestingly, the overexpression of *SbRio.02G135800.1* was associated with increased TTC and TPC levels. *SbRio.02G135800.1* plays a key role in the transport of secondary metabolites which may contribute to seed coat color formation in crops. The increased expression of this gene could have significant effects on seed coat color, which may enhance consumer preference and market value.

Additionally, the strong association of the magenta2 module with VOCs can be explained by underlying biochemical interactions. Our investigation into TTC, TPC, and VOCs showed a significant positive correlation between these two types of compounds (Fig. 1c). These correlations are likely due to shared biosynthetic pathways from which phenolics, and certain ketones are derived: for example, ketones that can act as natural antioxidants similar to phenolics are known to be derived from similar pathways. In adipocytes, compounds such as raspberry ketones inhibit the expression of genes related to lipogenesis and metabolism and exhibit antioxidant functions similar to those of phenolics [51]. Furthermore, in studies involving lemon and orange honey, the positive correlation observed between phenolics and ketones [52], supports our results and suggests a similar mechanism might be at play in sorghum. Furthermore, alcohol production via the phenylpropanoid pathway, which starts from amino acids such as phenylalanine, which also serve as precursors to phenolics, demonstrates biochemical interdependence. This interaction suggests that the synthesis of alcohols might affect or be affected by the production of phenolics, which in turn, responds to various environmental stresses, thus influencing the VOC profile in sorghum.

In the blue module, *SbRio.02G242300.1* and *SbRio.02G265800.1* were selected as hub genes (Fig. 6b). *SbRio.02G242300.1* encodes the *PTCD1* gene, and its expression level was increased by 2.03 – fold at the S2 stage compared to S1, while it decreased to 0.29-and 0.19 – fold at the S3 and S4 stages, respectively (Fig. 6d). The *PTCD1* gene plays a role in mitochondrial RNA processing and stability and has primarily been studied in animals. As such, there is a lack of research regarding *PTCD1* in plants. However, since *PTCD1* contains several pentatricopeptide repeat (PPR) domains, it is likely to perform similar functions to other PPR domain proteins, which are known to be crucial in RNA processing and regulation in both animals and plants. PPR domain proteins are found in plants and are primarily involved in RNA processing, editing, and translation in chloroplasts and mitochondria. They also play diverse roles in plant growth and development, including seed development, photosynthesis, and responses to biotic and abiotic stresses [53]. Many secondary metabolites in plant cells are synthesized from primary metabolites, with precursor compounds being formed through the Krebs cycle and shikimate pathway [54, 55]. In particular, secondary metabolism largely depends on primary metabolism for the necessary enzymes, ATP, and cellular machinery [55]. Among these, ATP is essential in the shikimate pathway because it provides the phosphate group for the phosphorylation of shikimate, a key intermediate leading to the production of chorismite and, ultimately, aromatic amino acids [56]. Meanwhile, chloroplasts serve as a key metabolic center in plants, carrying out various critical biosynthetic pathways, including amino acid and fatty acid biosynthesis, hormone production, and immune responses. Most notably, they perform photosynthesis, converting light energy into carbohydrates [57, 58]. PPR proteins influence chloroplast gene function and expression [59, 60], and thus, may affect photosynthetic efficiency. This, in turn, could directly impact ATP production, indirectly contributing to biochemical processes such as the shikimate pathway that rely on ATP as an energy source. Since phenolic compounds are synthesized through the shikimate pathway, the regulation of PPR genes could indirectly influence the synthesis of phenolic compounds, such as flavonoids.

Finally, *SbRio.02G265800.1* encodes the *ANK* gene. Its expression level increased by 5.06 – fold in S2 and 1.44 – fold in S4 compared to S1, while it was lower in S3 at 0.31 – fold (Fig. 6d). Proteins containing ANK domains are involved in various physiological pathways in plants, including responses to abiotic stress [61], light regulation [62], cell differentiation and development [63, 64], and organogenesis [65, 66]. In particular, *ANK* genes contain domains that play crucial roles in protein–protein interactions, which are essential for cellular signal transduction and developmental processes [67]. Moreover, the expression of *ANK* genes can be regulated by various plant hormones, such as auxin, salicylic acid, and abscisic acid (ABA) [68]. These hormones are key regulators of plant stress responses and development, and the changes in *ANK* gene expression are influenced by these hormonal signaling pathways. Plant hormones also play a critical role in regulating the biosynthesis of secondary metabolites, such as flavonoids [69]. Previous studies have shown that, in *Nitraria tangutorum Bobr*, ABA is crucial for the accumulation and regulation of flavonoids and anthocyanins [70]. Similarly, in *Fragaria* × *ananassa*, ABA promotes *FaMYB10* expression, which accelerates the expression of flavonoid pathway genes, leading to increased anthocyanin accumulation [71]. Auxin, in contrast, regulates flavonoid biosynthesis during plant growth and development, enhancing the plant's adaptive capacity [72]. Thus, hormonal regulation of *ANK* gene expression may impact flavonoid biosynthesis pathways, leading to significant changes in the physiological characteristics and stress responses of plants.

Conclusion

In this study, we analyzed the relationship between sorghum seed coat color, PCC, and VOCs identifying key genetic regions through RNA-seq and WGCNA. PCC and VOCs showed a significant correlation with seed coat color, with darker seeds showing higher phenolic and ketone contents. We identified important hub genes such as *ABCB28*, *PTCD1*, and *ANK* in the key module, providing insights into their roles in regulating phenolic and volatile compoun accumulation. These findings will not only increase our understanding of the molecular mechanisms of sorghum seed coat color and metabolite biosynthesis but also provide essential data for improving the food and industrial uses of sorghum.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-025-06657-w.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Supplementary Material 8

Supplementary Material 9

Supplementary Material 10

Supplementary Material 11

Supplementary Material 12

Supplementary Material 13

Acknowledgements
This work was supported by the research program of Korea Atomic Energy Research Institute (Project No. 523410-24) and National Research Foundation of the Republic of Korea (NRF) grant funded by the Republic of Korea government (RS-2022-00156231).

Author contributions
YJL and JR designed the experiment and drafted the manuscript. JR and SJK developed a mutant sorghum population. WJK, SHL, JIL, and JHK assisted fieldwork and phenotyping. JHK, JWA, and SHK analyzed the data. CHB and JR conceived and revised the manuscript. All authors reviewed the manuscript.

Funding
This work was supported by the research program of Korea Atomic Energy Research Institute (Project No. 523410-24) and National Research Foundation of the Republic of Korea (NRF) grant funded by the Republic of Korea government (RS-2022-00156231).

Data availability
Sequence data that support the findings of this study have been deposited in the NCBI Sequence Read Archive (SRA) repository accession code PRJNA1255307; NCBI accession No. SRR33308694~SRR33308705 (Table S12).

Declarations

Ethics approval and consent to participate
We ensure that all plant materials used in the current study were developed through our previous research [27] and collected from the Radiaition Breeding Farm of the Korea Atomic Energy Research Institute (Jeongeup, 35.51°N and 126.83°E, Republic of Korea), following relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 October 2024 / Accepted: 30 April 2025
Published online: 23 May 2025

References

1. Zhang L, Xu J, Ding Y, Cao N, Gao X, Feng Z, et al. GWAS of grain color and tannin content in Chinese sorghum based on whole-genome sequencing. Theor Appl Genet. 2023;136(4):77.
2. Antonopoulou G, Gavala HN, Skiadas IV, Angelopoulos K, Lyberatos G. Biofuels generation from sweet sorghum: fermentative hydrogen production and anaerobic digestion of the remaining biomass. Bioresour Technol. 2008;99(1):110–9.
3. Dicko MH, Gruppen H, Traoré AS, Voragen AG, Van Berkel WJ. Sorghum grain as human food in Africa: relevance of content of starch and amylase activities. Afr J Biotechnol. 2006;5(5):384–95.
4. Awika JM, Rooney LW. Sorghum phytochemicals and their potential impact on human health. Phytochemistry. 2004;65(9):1199–221.
5. Barros F, Awika J, Rooney LW. Effect of molecular weight profile of sorghum proanthocyanidins on resistant starch formation. J Sci Food Agric. 2014;94(6):1212–7.
6. Awika JM, Yang L, Browning JD, Faraj A. Comparative antioxidant, antiproliferative and phase II enzyme inducing potential of sorghum (Sorghum bicolor) varieties. LWT-Food Sci Technol. 2009;42(6):1041–6.
7. Arbex PM, de Castro Moreira ME, Toledo RCL, de Morais Cardoso L, Pinheiro-Sant'ana HM, dos Benjamin A. Extruded sorghum flour (Sorghum bicolor L.) modulate adiposity and inflammation in high fat diet-induced obese rats. J Funct Foods. 2018;42:346–55.
8. Xiong Y, Zhang P, Luo J, Johnson S, Fang Z. Effect of processing on the phenolic contents, antioxidant activity and volatile compounds of sorghum grain tea. J Cereal Sci. 2019;85:6–14.
9. Szambelan K, Nowak J, Szwengiel A, Jeleń H. Comparison of sorghum and maize Raw distillates: factors affecting ethanol efficiency and volatile by-product profile. J Cereal Sci. 2020;91:102863.
10. Dykes L, Rooney LW, Waniska RD, Rooney WL. Phenolic compounds and antioxidant activity of sorghum grains of varying genotypes. J Agric Food Chem. 2005;53(17):6813–8.
11. Kumari P, Kumar V, Kumar R, Pahuja SK. RETRACTED ARTICLE: Sorghum polyphenols: plant stress, human health benefits, and industrial applications. Planta. 2021;254(3):47.
12. Desta KT, Choi Y-M, Shin M-J, Yoon H, Wang X, Lee Y, et al. Comprehensive evaluation of nutritional components, bioactive metabolites, and antioxidant activities in diverse sorghum (Sorghum bicolor (L.) Moench) landraces. Food Res Int. 2023;173:113390.
13. Bodede O, Mabelebele M. Physical characteristics, nutritional composition and phenolic compounds of some of the sorghum landraces obtained in South Africa. Food Res. 2022;6(4):312–28.
14. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20(11):631–56.
15. Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J. 2016;88(6):1058–70.
16. Loraine AE, McCormick S, Estrada A, Patel K, Qin P. RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. Plant Physiol. 2013;162(2):1092–109.
17. Pradhan SK, Pandit E, Nayak DK, Behera L, Mohapatra T. Genes, pathways and transcription factors involved in seedling stage chilling stress tolerance in indica rice through RNA-Seq analysis. BMC Plant Biol. 2019;19:1–17.
18. Kumar S, Seem K, Kumar S, Mohapatra T. RNA-seq analysis reveals the genes/pathways responsible for genetic plasticity of rice to varying environmental conditions on direct-sowing and transplanting. Sci Rep. 2022;12(1):2241.
19. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinformatics. 2009;10:1–20.
20. Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Res. 2012;22(2):386–97.
21. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics. 2007;8:1–14.
22. Sircar S, Parekh N. Functional characterization of drought-responsive modules and genes in Oryza sativa: a network-based approach. Front Genet. 2015;6:256.
23. Kobayashi Y, Sadhukhan A, Tazib T, Nakano Y, Kusunoki K, Kamara M, et al. Joint genetic and network analyses identify loci associated with root growth under NaCl stress in Arabidopsis thaliana. Plant Cell Environ. 2016;39(4):918–34.
24. Zhu M, Xie H, Wei X, Dossa K, Yu Y, Hui S, et al. WGCNA analysis of salt-responsive core transcriptome identifies novel hub genes in rice. Genes. 2019;10(9):719.
25. Lin C-T, Xu T, Xing S-L, Zhao L, Sun R-Z, Liu Y, et al. Weighted gene co-expression network analysis (WGCNA) reveals the hub role of protein ubiquitination in the acquisition of desiccation tolerance in Boea hygrometrica. Plant Cell Physiol. 2019;60(12):2707–19.

26. Li Z, Wang J, Wang J. Identification of a comprehensive gene co-expression network associated with autotetraploid potato (Solanum tuberosum L.) development using WGCNA analysis. Genes. 2023;14(6):1162.
27. Lee Y-J, Yang B, Kim WJ, Kim J, Kwon S-J, Kim JH, et al. Genome-wide association study (GWAS) of the agronomic traits and phenolic content in sorghum (Sorghum bicolor L.) genotypes. Agronomy. 2023;13(6):1449.
28. Baek J, Lee E, Kim N, Kim SL, Choi I, Ji H, et al. High throughput phenotyping for various traits on soybean seeds using image analysis. Sensors. 2020;20(1):248.
29. Abràmoff MD, Magalhães PJ, Ram SJ. Image processing with ImageJ. Biophotonics Int. 2004;11(7):36–42.
30. Lee S, Choi Y-M, Shin M-J, Yoon H, Wang X, Lee Y, et al. Exploring the potentials of sorghum genotypes: a comprehensive study on nutritional qualities, functional metabolites, and antioxidant capacities. Front Nutr. 2023;10:1238729.
31. Jin-Rui X, Ming-wei Z, Xing-Hua L, Zhang-Xiong L, Rui-fen Z, Ling S, et al. Correlation between antioxidation and the content of total phenolics and anthocyanin in black soybean accessions. Agricultural Sci China. 2007;6(2):150–8.
32. Ryu J, Lyu JI, Kim D-G, Kim J-M, Jo YD, Kang S-Y, et al. Comparative analysis of volatile compounds of gamma-irradiated mutants of Rose (Rosa hybrida). Plants. 2020;9(9):1221.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
34. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
35. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.
36. Anders S, Huber W. Differential expression analysis for sequence count data. Nat Precedings. 2010;11:R106.
37. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:1–13.
38. Ofosu FK, Elahi F, Daliri EB-M, Tyagi A, Chen XQ, Chelliah R, et al. UHPLC-ESI-QTOF-MS/MS characterization, antioxidant and antidiabetic properties of sorghum grains. Food Chem. 2021;337:127788.
39. Qingshan L, Dahlberg JA. Chinese sorghum genetic resources. Econ Bot. 2001;55:401–25.
40. Eastin JD, Lee K-W. Sorghum bicolor. Handbook of flowering. CRC. 2019. 367–75.
41. Boren B, Waniska RD. Sorghum seed color as an indicator of tannin content. J Appl Poult Res. 1992;1(1):117–21.
42. Dykes L, Rooney WL, Rooney LW. Evaluation of phenolics and antioxidant activity of black sorghum hybrids. J Cereal Sci. 2013;58(2):278–83.
43. Cappellini F, Marinelli A, Toccaceli M, Tonelli C, Petroni K. Anthocyanins: from mechanisms of regulation in plants to health benefits in foods. Front Plant Sci. 2021;12:748049.
44. Kim JM, Lee JW, Seo JS, Ha B-K, Kwon S-J. Differentially expressed genes related to isoflavone biosynthesis in a soybean mutant revealed by a comparative transcriptomic analysis. Plants. 2024;13(5):584.
45. Kim D-G, Lyu J-I, Lim Y-J, Kim J-M, Hung N-N, Eom S-H, et al. Differential gene expression associated with altered isoflavone and fatty acid contents in soybean mutant diversity pool. Plants. 2021;10(6):1037.
46. Evans TS, Chen B. Linking the network centrality measures closeness and degree. Commun Phys. 2022;5(1):172.
47. Zhao J, Dixon RA. MATE transporters facilitate vacuolar uptake of epicatechin 3′-O-glucoside for Proanthocyanidin biosynthesis in Medicago truncatula and Arabidopsis. Plant Cell. 2009;21(8):2323–40.
48. Zhao J, Dixon RA. The 'ins' and 'outs' of flavonoid transport. Trends Plant Sci. 2010;15(2):72–80.
49. Goodman CD, Casati P, Walbot V. A multidrug resistance–associated protein involved in anthocyanin transport in Zea mays. Plant Cell. 2004;16(7):1812–26.
50. Sugiyama A, Shitan N, Yazaki K. Involvement of a soybean ATP-binding cassette-type transporter in the secretion of Genistein, a signal flavonoid in legume-Rhizobium symbiosis. Plant Physiol. 2007;144(4):2000–8.
51. Park KS. Raspberry ketone, a naturally occurring phenolic compound, inhibits adipogenic and lipogenic gene expression in 3T3-L1 adipocytes. Pharm Biol. 2015;53(6):870–5.
52. Escriche I, Kadar M, Juan-Borrás M, Domenech E. Using flavonoids, phenolic compounds and headspace volatile profile for botanical authentication of lemon and orange honeys. Food Res Int. 2011;44(5):1504–13.
53. Schmitz-Linneweber C, Small I. Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci. 2008;13(12):663–70.
54. Narayani M, Srivastava S. Elicitation: a stimulation of stress in in vitro plant cell/tissue cultures for enhancement of secondary metabolite production. Phytochem Rev. 2017;16:1227–52.
55. Rahman A, Albadrani GM, Waraich EA, Awan TH, Yavaş İ, Hussain S. Plant secondary metabolites and abiotic stress tolerance: overview and implications. Plant Abiotic Stress Responses and Tolerance Mechanisms; 2023.
56. Maeda H, Dudareva N. The Shikimate pathway and aromatic amino acid biosynthesis in plants. Annu Rev Plant Biol. 2012;63(1):73–105.
57. Hou X, Rivers J, León P, McQuinn RP, Pogson BJ. Synthesis and function of apocarotenoid signals in plants. Trends Plant Sci. 2016;21(9):792–803.
58. Sakamoto W, Miyagishima S-y, Jarvis P. Chloroplast biogenesis: control of plastid development, protein import, division and inheritance. Volume 6. The Arabidopsis book/American Society of Plant Biologists; 2008.
59. Huang W, Zhang Y, Shen L, Fang Q, Liu Q, Gong C, et al. Accumulation of the RNA polymerase subunit RpoB depends on RNA editing by OsPPR16 and affects Chloroplast development during early leaf development in rice. New Phytol. 2020;228(4):1401–16.
60. Zhang J, Xiao J, Li Y, Su B, Xu H, Shan X, et al. PDM3, a pentatricopeptide repeat-containing protein, affects Chloroplast development. J Exp Bot. 2017;68(20):5615–27.
61. Nodzon LA, Xu WH, Wang Y, Pi LY, Chakrabarty PK, Song WY. The ubiquitin ligase XBAT32 regulates lateral root development in Arabidopsis. Plant J. 2004;40(6):996–1006.
62. Zhang H, Scheirer DC, Fowle WH, Goodman HM. Expression of antisense or sense RNA of an Ankyrin repeat-containing gene blocks Chloroplast differentiation in arabidopsis. Plant Cell. 1992;4(12):1575–88.
63. Albert S, Després B, Guilleminot J, Bechtold N, Pelletier G, Delseny M, et al. The EMB 506 gene encodes a novel Ankyrin repeat containing protein that is essential for the normal development of Arabidopsis embryos. Plant J. 1999;17(2):169–79.
64. Hemsley PA, Kemp AC, Grierson CS. The TIP GROWTH DEFECTIVE1 S-acyl transferase regulates plant cell growth in Arabidopsis. Plant Cell. 2005;17(9):2554–63.
65. Huang J, Chen F, Del Casino C, Autino A, Shen M, Yuan S, et al. An Ankyrin repeat-containing protein, characterized as a ubiquitin ligase, is closely associated with membrane-enclosed organelles and required for pollen germination and pollen tube growth in Lily. Plant Physiol. 2006;140(4):1374–83.
66. Garcion C, Guilleminot J, Kroj T, Parcy F, Giraudat J, Devic M. AKRP and EMB506 are two Ankyrin repeat proteins essential for plastid differentiation and plant development in Arabidopsis. Plant J. 2006;48(6):895–906.
67. Gupta T, Chahota R. Unique Ankyrin repeat proteins in the genome of poxviruses-Boon or wane, a critical review. Gene. 2024;927(4):148759.
68. Li L, Yang J, Zhang Q, Xue Q, Li M, Xue Q, et al. Genome-wide identification of Ankyrin (ANK) repeat gene families in three Dendrobium species and the expression of ANK genes in D. officinale under Gibberellin and abscisic acid treatments. BMC Plant Biol. 2024;24(1):762.
69. Lv Z-Y, Sun W-J, Jiang R, Chen J-F, Ying X, Zhang L, et al. Phytohormones jasmonic acid, Salicylic acid, gibberellins, and abscisic acid are key mediators of plant secondary metabolites. World J Traditional Chin Med. 2021;7(3):307–25.
70. Zhang J, Cheng K, Liu X, Dai Z, Zheng L, Wang Y. Exogenous abscisic acid and sodium Nitroprusside regulate flavonoid biosynthesis and photosynthesis of nitraria tangutorum Bobr in alkali stress. Front Plant Sci. 2023;14:1118984.
71. Kadomura-Ishikawa Y, Miyawaki K, Takahashi A, Masuda T, Noji S. Light and abscisic acid independently regulated FaMYB10 in Fragaria× Ananassa fruit. Planta. 2015;241:953–65.
72. Kurepa J, Shull TE, Smalle JA. Friends in arms: flavonoids and the auxin/cytokinin balance in terrestrialization. Plants. 2023;12(3):517.

Publisher's note