Mol Breeding (2023) 43:71
https://doi.org/10.1007/s11032-023-01414-z

# An omics strategy increasingly improves the discovery of genetic loci and genes for seed-coat color formation in soybean

**Jian Song · Ruixin Xu · Qingyuan Guo · Caiyu Wu · Yinghui Li · Xuewen Wang · Jun Wang · Li-Juan Qiu**

Received: 1 March 2023 / Accepted: 13 August 2023 / Published online: 31 August 2023
© The Author(s), under exclusive licence to Springer Nature B.V. 2023

**Abstract** The phenotypic color of seeds is a complex agronomic trait and has economic and biological significance. The genetic control and molecular regulation mechanisms have been extensively studied. Here, we used a multi-omics strategy to explore the color formation in soybean seeds at a big data scale. We identified 13 large quantitative trait loci (QTL) for color with bulk segregating analysis in recombinant inbreeding lines. GWAS analysis of colors and decomposed attributes in 763 germplasms revealed associated SNP sites perfectly falling in five major QTL, suggesting inherited regulation on color during natural selection. Further transcriptomics analysis before and after color accumulation revealed 182 differentially expression genes (DEGs) in the five QTL, including known genes *CHS*, *MYB*, and *F3′H* involved in pigment accumulation. More DEGs with consistently upregulation or downregulation were identified as shared regulatory genes for two or more color formations while some DEGs were only for a specific color formation. For example, five upregulated DEGs in QTL *qSC-3* were in flavonoid biosynthesis responsible for black and brown seed. The DEG (*Glyma.08G085400*) was identified in the purple seed only, which encodes gibberellin 2-beta-dioxygenase in the metabolism of colorful terpenoids. The candidate genes are involved in flavonoid biosynthesis, transcription factor regulation, gibberellin and terpenoid metabolism, photosynthesis, ascorbate and aldarate metabolism, and lipid metabolism. Seven differentially expressed transcription factors were also

**Key message** The genetic basis of soybean seed coat color by BSA mapping of segregation population and GWAS of 763 germplasms

Jian Song, Ruixin Xu and Qingyuan Guo contributed equally to this work.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11032-023-01414-z.

J. Song · R. Xu · Q. Guo · C. Wu · J. Wang (✉)
Yangtze University, Jingzhou 434025, Hubei, P.R. China
e-mail: wangjagri@yangtzeu.edu.cn

Y. Li · L.-J. Qiu (✉)
The National Key Facility for Crop Gene Resources and Genetic Improvement (NFCRI)/Key Laboratory of Crop Gene Resource and Germplasm Enhancement (MOA)/Key Laboratory of Soybean Biology (Beijing) (MOA), Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing 100081, China
e-mail: qiulijuan@caas.cn

X. Wang (✉)
Department of Genetics, University of Georgia, Athens, GA 30602, USA
e-mail: Xuewen.Wang@unthsc.edu

speculated that may regulate color formation, including a known MYB. The finds expand QTL and gene candidates for color formation, which could guide to breed better cultivars with designed colors.

**Keywords** Seed color · GWAS · Bulk sequencing · Differential expression · Regulation · Soybean

## Introduction

Seed coat color is an important quantitative agronomic trait and attribute determining the outward appearance of soybean and considered as a useful phenotypic marker in breeding. Soybean seed coat presents a range of colors including yellow, green, brown, purple red, black, and bicolor, and seed coat color is also an evolutionary trait within the soja subgenus and it was changed from black in wild soybean to various colors in cultivated soybeans during domestication (Song et al. 2016). The formation of seed coat color is closely related to flavonoids and anthocyanins, which have become important substances used for disease prevention and treatment in human beings (Santos-Buelga et al. 2014; Wallace and Giusti 2015; Kim et al. 2020; Yuan et al. 2022), and for plant tolerance to stress (Ballaré 2003; Yang et al. 2010). In soybean, seed coat color genes also involve in a variety of physiological and developmental functions including the cleavage of the seed coat, the expansion and infiltration of the seed coat, and promotion of soybean seed dormancy (Zabala and Vodkin 2003; Sun et al. 2015; Wang et al. 2018).

Studies on soybean seed coat color have been started as early as the discovery of Mendel's law of inheritance. So far, at least seven loci (*I*, *R*, *T*, *O*, *W1*, *K1*, *G*) have been found to be involved in controlling the formation of color traits (Yang et al. 2010; Cho et al. 2017; Wang et al. 2018; Song et al. 2016; Palmer et al. 2004), and two loci (*D1*, *D2*) affect yellow or green cotyledon and then seed coat (Song et al. 2017; Nakano et al. 2014; Fang et al. 2014). Most of these known loci were involved in the flavonoid pathway. Three loci (*I*, *R*, and *T*) mainly regulate the color of seed coat by controlling the synthesis of pigment (Yang et al. 2010; Cho et al. 2017; Song et al. 2016). The *I* locus encodes the chalone synthase (CHS), which controls the formation of anthocyanins through the post-transcriptional gene silencing mechanism (Cho et al. 2013; Senda et al. 2004; Senda et al. 2013; Todd 1996; Wang et al. 2018; Zhou et al. 2015; Tuteja et al. 2009). *R* locus may be an MYB transcription factor regulating the synthesis pathway (Zabala et al. 2014; Gillman et al. 2011; Gao et al. 2021). Locus *T* encodes flavonoid 3′-hydroxylase (F3′H) (Zabala and Vodkin 2003; Toda et al. 2002). *O* and *W1* affect seed coat color only in homozygous recessive *ir* and *it* genotypes (Zabala and Vodkin 2007; Song et al. 2016). *O* locus has been found to correspond to an anthocyanidin reductase gene (ANR) (Xie et al. 2003; Yang et al. 2010). *W1* encodes flavonoid 3′5′-hydroxylase (F3′5′H) (Zabala and Vodkin 2007). A complex genetic basis was found from an interaction between loci. *K1* encodes AGO5 protein which interacts with locus *I* to form pigmentation in different regions (seed coat, hilum, and saddle type region) by RNA-mediated post-transcriptional gene silencing on CHS (Cho et al. 2017). *G* encodes the terminal protease protein of CAAX amino acid, which has been proved to control the separation between yellow and green colors (Wang et al. 2018).

With advances of next-generation sequencing technology and the soybean genome assembly (Xie et al. 2019), high-density single nucleotide polymorphism (SNP) variants have speeded up the marker-assisted selection breeding process. Genome-wide association studies (GWAS) resolve SNPs associated with complex genetic traits. The combined linkage analysis and GWAS (Dobbels et al. 2017) have improved the power for fast mining important yield traits in soybean, quality traits such as glycinin and β-conglycinin (Zhang et al. 2021) and agronomic traits (Sonah et al. 2014).

In this study, to extensively explore molecular mechanisms of genes controlling the seed coat color in soybean, we used bulked sequencing analysis to map the quantitative trait loci (QTL) in the fourth generation of recombinant inbreed lines (RILs) derived from a cross between ZD41 (yellow seed coat) and ZCXHD (brown seed coat). The color traits in 763 accessions of germplasms were characterized and decomposed into multiple color attributes. SNPs in these germplasms were used to conduct GWAS for color-associated loci, which overlapped with QTL and narrowed down the QTL into small regions. Genes in QTL were screened and their expressional regulation was examined with transcriptome analysis via RNA-Seq before and after color accumulation.

Differentially expressed genes (DEGs) have been identified as the candidates for color regulation, which included shared gene regulation for all colors and also for a single specific color and its attributes. The identified genes include those in previously known loci and more novel ones for color formation. This study resolved systematic analysis and more genes for color formation and improved our understanding of the molecular basis of color regulation in soybean.

## Materials and methods

### Plant materials and field trails

The plant materials include a linkage panel and an association panel. The linkage panel consisted of 598 F4 RILs derived from a cross between ZD41 (yellow seed coat) and ZCXHD (brown seed coat) were developed at Yangtze University, China (30.37°N, 112.06°E), by phenotypic selection and self-fertility of specific lines for several generations. The RILs and their parents were planted in 2018. The association panel was composed of 763 accessions, including 591 accessions from China and 172 accessions from abroad (Stable 1). All accessions were planted in 2019, Jingzhou, Hubei (30.37°N, 112.06°E). The arrangement of plantation was set to 2 m-long row with 0.5-m row spacing and 0.1 m of distance between individuals.

Fresh seeds were collected after the pods were bulged. The developmental stage of a seed was determined by the weight of fresh seeds. Two hundred to 300 mg/seed was defined as before pigment accumulation (nc), and 400–500 mg/seed as after the pigment accumulation (c). The F5 RILs and their parents were planted and harvested from Jingzhou in 2019.

### Phenotypic data and analysis

The segregating RHL population was selected as mixing pool materials using traditional seed coat color identification methods by comparing with the RAL K7 color card. The seed coat colors were divided into four bulked lines of black, purple, brown, and yellow.

For the association population, the Tomato Analyzer-Color Test (TACT) 4.0 software (Darrigues et al. 2008) was used to convert the color of an image taken from seed coat into nine color attribute values including red (R), green (G), blue (B), Luminosity (Lum), light-and-shade drawing (L), red-green (a), yellow-blue (b), Hue (Hue) and color saturation (Chroma). A dimension reduction was performed by principal component analysis. The 1st or 2nd principal components obtained were also used as phenotypic values of seed coat color (Sadohara et al. 2021; Rodríguez et al. 2010).

### Bulk segregation analysis

The F4 RILs were divided into four groups by the seed coat color. The groups are termed yellow (Y), brown (BR), purple (P), and black (BL). DNA was extracted from 20 plants of each group using the CTAB method (Saghai-Maroof 1985), normalized to 50 ng/μL, and mixed as a bulk of pooled DNA. Each bulked DNA was sequenced to generate average 20 × depth coverage with an Illumina HiSeq4000 according to manufacturer's standard DNA sequencing recommendation. Reads were cleaned and mapped against Glycine max Wm82.a2.v1 reference genome from Phytozome using BWA with default parameters (Langmead and Salzberg 2012). The GATK (Genome Analysis Toolkit, version 4.2) was used to call SNPs and small indels less than 50 bp (Mckenna et al. 2010). The Euclidean distance (ED) algorithm was applied to evaluate the regions associated with traits. The SNP genotype and depth between different mixing pools were used to calculate the ED value at each site. The 2nd power of the original ED was taken as the correlation value to eliminate background noise (Hill et al. 2013). Then, the distance method was used to fit the ED value, and the median+3SD of the fitted value of all loci was taken as the correlation threshold. Finally, the candidate regions related to traits were obtained.

### Genotype analysis and genome-wide association mapping

SNP variant data from all 763 accessions were examined in 2214 soybeans by CAAS (Li and Lam 2022). Plink2 was used to remove locus with minor allele frequency less than 0.05, missing rate above 20% as well as with multiple allele loci for genotype data (Purcell et al. 2007). The remaining SNPs were used for subsequent analysis. Genotypic prediction of missing loci was carried out with Beagle.21.

Principal components were calculated with a PCA module in GCTA software for the correction of population structure (Price et al. 2006). TASSEL (version 5) was used to analyze kinship among individuals. GWAS was conducted with the mixed linear model integrated into the top three principal components and kinship in TASSEL. The critical threshold of significantly correlated SNPs was set to $P = 0.05/N$, where $N$ is the total number of SNPs used in GWAS. The final $P$ value = $2.49e^{-8}$, where $N$ equals 2,006,127.

### RNA extraction and RNA-Seq transcriptomics analysis

Seed samples were collected before and after color accumulation, separately. Total RNAs from each sample were extracted with TRIzol Reagent (Life Technologies, Carlsbad, CA, USA), and enriched mRNA was sequenced on an Illumina HiSeq 2500 and analyzed with methods as described previously (Wang et al. 2019). Briefly, the transcriptome data were preprocessed and mapped to reference genome Wm82.a2.v1, and transcript abundance was calculated in fragments per kilobase of transcript per million mapped reads (FPKM). Differentially expressed genes (DEGs) were identified as at least twofolds of change and adjusted $P$ value < 0.01. Gene Ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment for DEGs were analyzed. RNA-Seq data were generated from eight samples with three biological replicates for each, totaling 24 samples (Stable 5). Samples BLc2, BLnc1, BRc3, and Pc2 had problematic correlations compared with corresponding biological replicates and were excluded from subsequence analyses.

### Identification of candidate genes in an LD block

The linkage disequilibrium (LD) analysis was conducted with the mentioned Tassel package with the SNPs from the germplasm population. The LD critical threshold was set to LD coefficient $R$-squared less than 0.2. Based on the LD critical decay distance, upstream and downstream flanking a region on a chromosome were explored to identify the DEG list. Since the GWAS region was much smaller than the corresponding BSA region, further detailed linkage disequilibrium (LD) blocks at each GWAS site were estimated with LDBlockShow (Dong et al. 2020) with default settings.

The fixation index (FST) was calculated from the published genome sequence data of 2214 soybean accessions (Li et al. 2022) using VCFtools (0.1.13) (Danecek et al. 2011) with a window size of 100 bp, and coding regions with FST ≥ 0.6 were assigned as potential domestication genes (Chen et al. 2022b).

DEGs in overlapped regions of the QTL intervals and within the reasonable LD blocks were annotated to identify the most likely molecular role in seed coat color regulation. The annotation was analyzed against GO (http://geneontology.org/) and KEGG (https://www.genome.jp/). The function of the key differentially regulated gene was further examined through a literature search if evidences from existing publications are available.

## Results

### Phenotypic variation of seed coat colors in two panels

The seed coat color of parental lines ZD41 and ZCXHD is brown and yellow, respectively, whereas no difference between embryos was observed. Seed coat color of the F1 plant was yellow and four different colors were observed in the F2 generation. A total of 404, 83, 76, and 29 individuals showed yellow, brown, black, and bicolor in seed coat, respectively. The browns include light brown, brown, dark brown, and purplish red.

Regarding the seed coat color, the 763 resequencing accessions are composed of 277 yellow, 157 green, 39 purple, 76 brown, 14 dark brown, and 200 black (Stable 1). High-resolution photos were taken from the seeds of each accession, and the color of seed- oat color was recorded as high-resolution photos and then further decomposed into nine color attributes of R, G, B, L, a, b, Luminosity, Hue, and Chroma with the TACT software. Significant differences between colors were observed (Fig. 1a). Principal component analysis on color attributes showed that PC1 (87.9%) and PC2 (11.4%) could reflect 99.2% of the variation. PC1 represents more yellow, green, and brown, while PC2 represents dark brown, purple, and black (Fig. 1b).

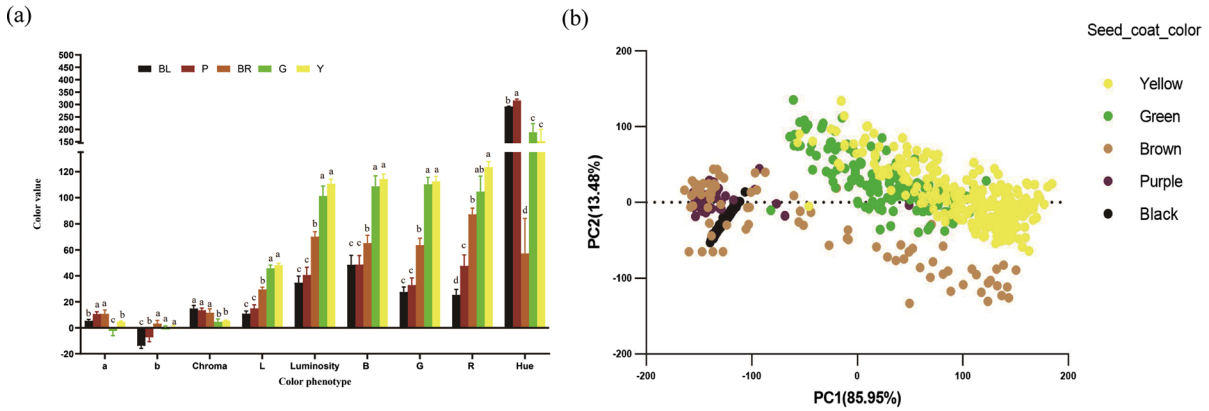


**Fig. 1** Variation of seed coat color and principal component analysis. ***a*** *Bar plot showing the differences in color attributes of seed coat color; X-axis for color attributes or phenotypes.* ***b*** *Scatter plot of first two components of principal component analysis.* Seed coat color: BL (black seed coat), P (purple red seed coat), BR (brown seed coat), G (green seed coat), and Y (yellow seed coat). Color attributes: R (red), G (green), B (blue), Luminosity (Lum), light-and-shade drawing (L), A (red-green), B (yellow-blue), Hue, Chroma, PC1, PC2. a, b, c, d: significant difference at 0.05 level

## Identification of loci associated with seed coat color by BSA

Four DNA bulks from F4 RHL individuals with yellow, brown, black, and purple seed colors were sequenced independently with Illumina HiSeq4000. In total, ~341 million clean reads, at least 79 million reads per DAN bulk, were obtained with an average read depth ~20 × of the Williams 82 reference genome. After mapping to the reference genome, more than 98% of genome regions were covered by reads at least 1 time. SNPs and small InDels (<50 bp) were identified between different DNA bulks with GATK packages, including 274K-309K SNPs and 91K-97K InDels (Stable 2).

Based on the identified InDels and SNPs, 13 loci were identified from the color pool analyses. The ED analysis between four color pools revealed four overlapped significantly associated loci on chromosomes 6, 8, and 9 (Fig. 2). In the brown pool, two significant loci were mapped on chromosome 6 (Fig. 2a). One locus at 14,620,000–20,740,000 bp with a total length of 6.12 Mb contains 365 predicted genes. In this region, the cloned *T* locus *Glyma.06G202300* encoding F3′H has been reported to be related to the formation of brown seed coat. The other associated locus is at 37,630,000–47,320,000 bp, spanning 9.69 Mb, and has 505 genes. Comparative analysis of the yellow pool with the other three color pools revealed overlapped significant association loci in the region between 5,940,000 bp and 10,780,000 bp on chromosome 8, harboring 631 genes in this region. The chalcone synthase (CHS)–related genes reported to control yellow seed coat formation were located in this region (Fig. 2b). Comparative analysis between the yellow pool and brown pool, brown pool and black pool, black pool and purple pool, and purple pool and yellow pool showed that there were common significant association loci in the range of 45,600,600–47,900,000 bp (2.3 Mb) on chromosome 9 containing 277 genes (Fig. 2c). The *R* locus candidate gene *Glyma.09G235100* controlling anthocyanin synthesis is located in this region. Besides, the association analysis also found nine other loci with ED values higher than the association threshold on chromosomes 3, 4, 7, 8, 12, 15, and 17, respectively. These loci and four common significant association loci were used as candidate intervals for QTL control of seed coat color (Stable 3).

## GWAS analysis of seed color with genome-wide SNPs

The SNP dataset of 763 natural germplasms was retrieved from publicly available resequencing reads. The GWAS analysis was conducted for the filtered SNP dataset and seed color phenotype plus derived nine color attributes with the Tassel package (version

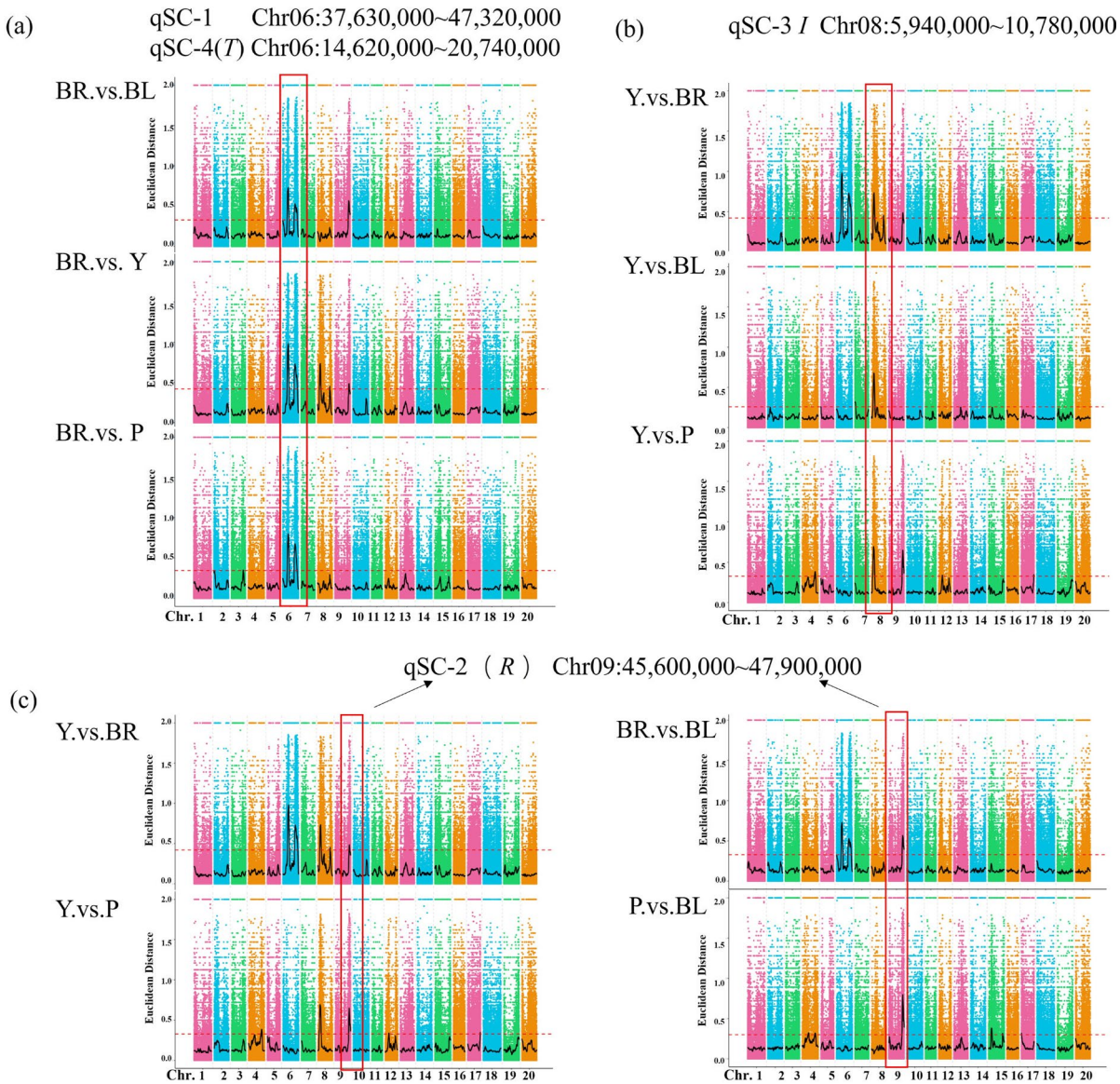


**Fig. 2** Overlapped associated loci between different color bulk pools. **a** Manhattan plot of the associated loci on chromosome six between the yellow pool and other color pools; **b** Manhattan plot of the association on chromosome eight between the brown pool and other color pools; **c** Manhattan plot of the association on chromosome nine between remaining color pools

4). To obtain SNPs controlling seed coat color traits, we first filtered the SNPs and removed the loci with missing rate > 10% and minor allele frequency < 0.05 and then imputed to fill in possible remaining missing data based on population. Population structure and relatedness were then predicted in the preprocessed SNP dataset. Principal component analysis revealed that the first three PCs explained most of the population structure and no significant outlier samples. The kinship analysis showed that most of the materials had no obvious relatedness, and only a small portion of the 763 samples were closely related (Sfigure 1). Genome-wide association analysis was performed for 2,006,128 SNPs and 10 seed coat color–derived phenotypes with the top three PCs and kinship Fig. 3. GWAS results revealed a clear shift

from a random distribution for all traits at a *P* value less than ~0.001 in the statistical QQ plot, indicating a significant correlation between phenotypes and genotypes by natural selection (Fig. 4). A total of 22 significantly associated SNP sites (threshold $P <$ 2.49e−08) were discovered for the color trait while only some of these SNP loci were associated with its derived color attributes (Fig. 3). All associated SNPs were distributed in all chromosomes except for chromosomes 2 and 11. No SNP association was found for the red attribute. The detailed associated loci are available in the supplementary file (Stable 4). An identical significant SNP association with all 10 phenotypes was detected on chromosome 8 (Chr8: 7,820,505–9,957,944). In addition, SNPS significantly associated with more than three phenotypic values were detected on chromosomes 1, 6, 10, and 20. Color and Hue were the top two phenotypes with more abundant associated SNPs than others (Fig. 3). Genome-wide association analysis of 10 seed coat color phenotypes identified a total of 783 SNP significant loci. Of those, 611 SNPs were located on chromosome 8 (Stable 4).

## Overlapped loci for seed color between GWAS and BSA analysis

Regions identified by the BSA-seq analysis spanned very large physical fragments on chromosomes. To narrow down the associated regions for seed color traits, we integrated the BSA and GWAS results and discovered the overlapped loci. In total, five regions were shared by both NSA and GWAS results, including highly overlapped *qSC-2* on chromosome 9, *qSC-3* on chromosome 8, *qSC-4* on chromosome 5, *qSC-5* on chromosome 4, and *qSC-6* on chromosome 7. The GWAS regions were much smaller than BSA-mapped regions, except for the *qSC-2* region, which largely narrowed down the bin size of candidate genes on the chromosome (Table 1). This suggests that these

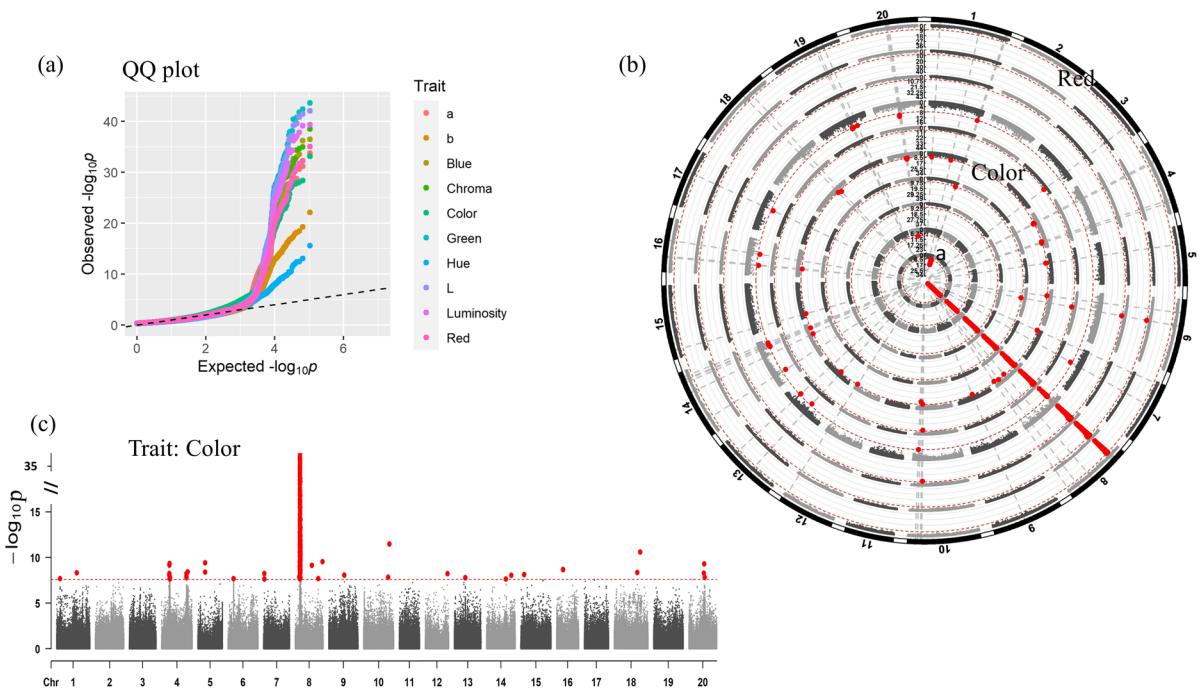


**Fig. 3** GWAS DNA variants' association with seed color attributes from 763 natural germplasm. **a** The QQ plot shows the difference from the random association of the black line. **b** The comparison of associated SNPs with the color trait and its attributes. Red dots indicate the significant association loci at the threshold $P <$ 2.49e−8. The trait names from the central circle to the outward circle are a, b, Blue, Chroma, Color, Green, Hue, L, Luminosity, and Red, respectively. The red dashed line above the Manhattan plot is the threshold line ($P <$ 2.49e−8). The radial gray dash line represents an identical SNP position across traits. **c** A Manhattan plot showing the association of SNP sites with the color trait. The dots above the horizontal threshold line ($P <$ 2.49e−8) were statistically significant SNP sites

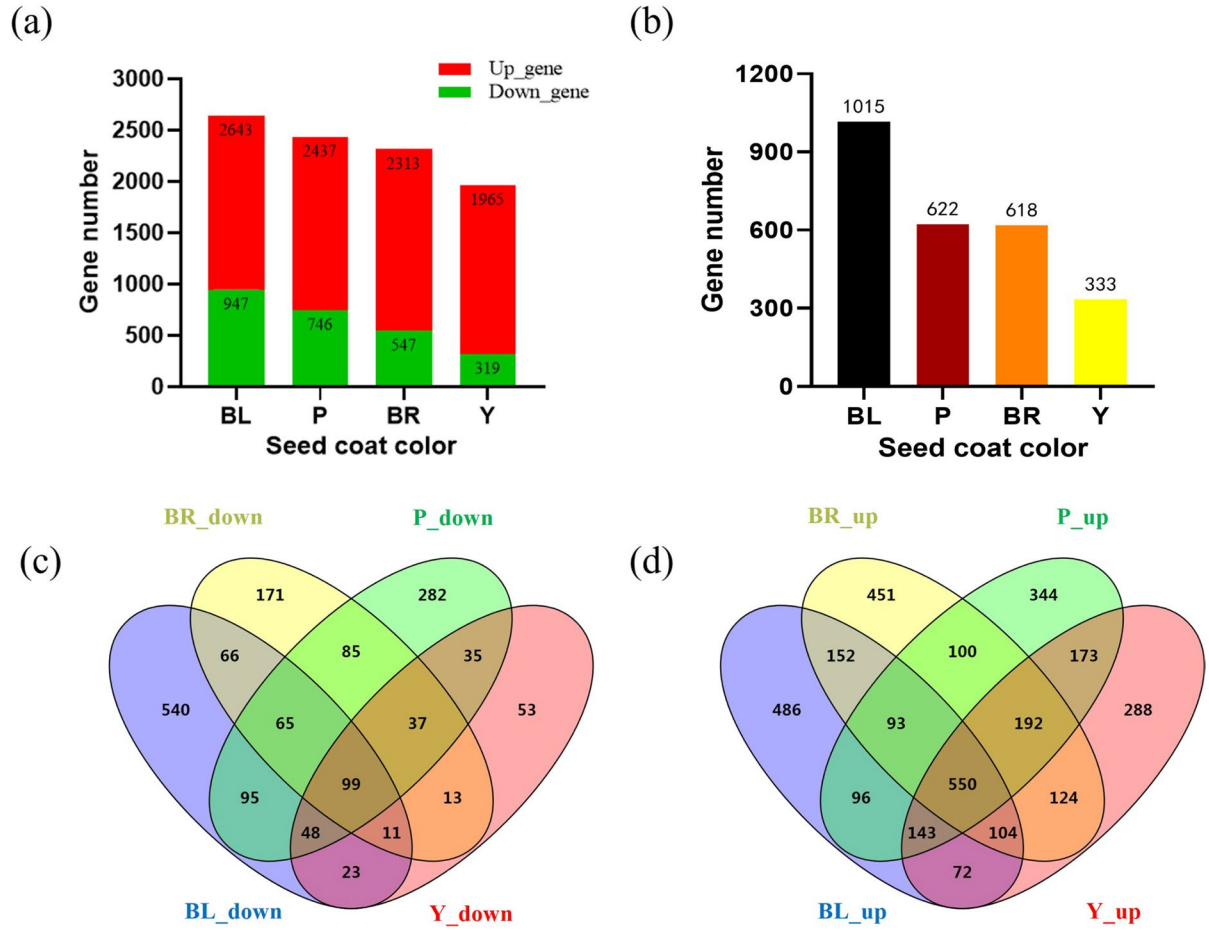


**Fig. 4** Differentially expressed genes during the seed coat color formation. **a** Bar diagrams of differentially upregulated and downregulated genes during seed coat color formation; **b** bar diagrams of specific differentially expressed genes in different color seed coats; **c** Venn diagram of differential genes downregulated across the formation of different seed coat colors; **d** Venn diagram of differential genes upregulated across formation of different seed coat colors

**Table 1** Candidate QTL for joint analysis of GWAS and BSA

The positions of GWAS are the start and end coordinates on a chromosome of the significant associated SNPs in Glycine max genome assembly Wm82.a2.v1. If there is only one associated SNP, the start and end positions are the same.

| Chromosome | QTL | Source | Start position (bp) | End position (bp) |
|---|---|---|---|---|
| chr4 | qSC-5 | BSA | 41,620,000 | 44,050,000 |
| | | GWAS | 42,698,572 | 44,620,540 |
| chr6 | qSC-4 | BSA | 14,620,000 | 20,740,000 |
| | | GWAS | 18,887,736 | 18,887,736 |
| chr7 | qSC-6 | BSA | 1 | 2,140,000 |
| | | GWAS | 1,293,309 | 1,293,309 |
| chr8 | qSC-3 | BSA | 5,940,000 | 10,780,000 |
| | | GWAS | 7,820,505 | 9,957,944 |
| chr9 | qSC-2 | BSA | 45,600,000 | 47,900,000 |
| | | GWAS | 46,691,031 | 46,691,031 |

regions in the BSA bulk dataset are also inherited in natural germplasms. *qSC-2*, *qSC-3,* and *qSC-4* were located in the previously known *R*, *I,* and *T* loci. No overlapping of the remaining regions was found.

## Expressional regulation of QTL candidate genes via transcriptome analysis

To gain the expressional regulation on seed coat color, we generated 54 million clean reads per sample via RNA-Seq technology to examine the gene expression in four types of colored seeds, including yellow, brown, purple, and black before and after pigment accumulation (Stable 5). The gene expression level was calculated in FPKM after mapping to the reference genome Wm82 and in total 9358 DEGs (at least two-folds of change, $P < 0.01$) were identified in colored seeds compared with those before visible pigment accumulation. More DEGs were downregulated than upregulated, which may be associated with the mature stage (Fig. 4a). Analysis of DEGs unique to each seed coat color revealed 1015, 622, 618, and 333 DEGs during the formation of black, purple, brown, and yellow seed coat colors (Fig. 4b). Ninety-nine downregulated (Fig. 4c) and 550 upregulated (Fig. 4d) DEGs were common across all four types of seed color formation.

DEGs located in the overlapped loci (Table 1) can be the critically regulated genes responsible for the formation of seed color. We examined these DEGs in each locus, especially the shared DEGs by several colors. The DEGs identified specifically for each color are also highlighted for our future analysis (Stable 6). Overall, highly consistent upregulation and downregulation patterns of these DEGs across samples were revealed (Fig. 5a).

In the *qSC-2* region on chromosome 9, 25 DEGs were identified during the accumulation of four colors of yellow, brown, purple, and black (Stable 6). Among those, DEGs *Glyma.09G238300* and *Glyma.09G249500* were the commonly upregulated, suggesting a shared regulation for the color pigment accumulation. Three downregulated DEGs (*Glyma.09G233800*, *Glyma.09G241300*, and *Glyma.09G250800*) and three upregulated DEGs (*Glyma.09G235100*, *Glyma.09G241800*, and *Glyma.09G250400*) shared a similar expressional pattern during the formation of yellow, brown, and purple (Fig. 5a). We found that *Glyma.09G235100* was the previously reported *R* locus with an MYB transcription factor gene (Zabala et al. 2014; Gillman et al. 2011), suggesting consistent results in our study. No strong ($R^2 \geq 0.8$) LD blocks were identified around the GWAS associated SNP within the *qSC-2* region. The physical distance between the known MYB and the top significant SNP was around 1 Mbp (Fig. 5b).

In the *qSC-3* region on chromosome 8, 78 DEGs were identified as candidate genes during the accumulation of four colors in the seed coat (Stable 6). Pathway analysis against the Kyoto Encyclopedia of Genes and Genomes (KEGG) revealed six DEGs (*Glyma.08G109100*, *Glyma.08G109200*, *Glyma.08G109500*, *Glyma.08G110400*, *Glyma.08G110500*, and *Glyma.08G110700*) in the flavonoid biosynthesis pathway. The *Glyma.08G109100* was upregulated but was significant only in purple and brown accumulation, while the remaining five DEGs were significantly upregulated during the pigmentation accumulation of black and brown seed coat only (Fig. 5a). These six DEGs are located between 8.3 Mbp and 8.5 Mbp on chromosome 8, which is in the middle of GWAS associating SNPs between 7.8 Mbp and 9.9 Mbp. The LD blocks around the GWAS-associated SNPs on this chromosome were found to be around 20 Kbp, which only linked these six candidate DEGs, indicating the most likely genes for color accumulation on chromosome 8 (Fig. 5c).

In the *qSC-4* region on chromosome 6 (Fig. 5d), 33 DEGs were identified as candidates for the color traits. Of those, DEG *Glyma.06G202300*, a known *T* locus encoding F3′H, was downregulated in all colors but was significant only in yellow, brown, and purple seed coats. Two DEGs *Glyma.06G182200* and *Glyma.06G187100* were upregulated in all four types of colored seed coats (Fig. 5a). The large LD block size for this GWAS-associated SNP was ~35 Kbp, much bigger than other sites, which may indicate more distantly DEGs should regulate the pigment accumulation (Fig. 5d).

In the *qSC-5* region on chromosome 4 (Fig. 5e), 17 DEGs were identified as candidates for the color traits. However, no DEGs were shared by all four colors. Three downregulated DEGs (*Glyma.04G167900*, *Glyma.04G173100*, and *Glyma.04G177600*) were shared by yellow, brown, and purple traits (Fig. 5a). These DEGs were located between 42.1M and 44.2M, which was perfectly located at GWAS SNP sites between 42.6M and 44.6M on chromosome 4. The estimated LD distances were ~15 Kbp around the associated SNPs (Fig. 5e).

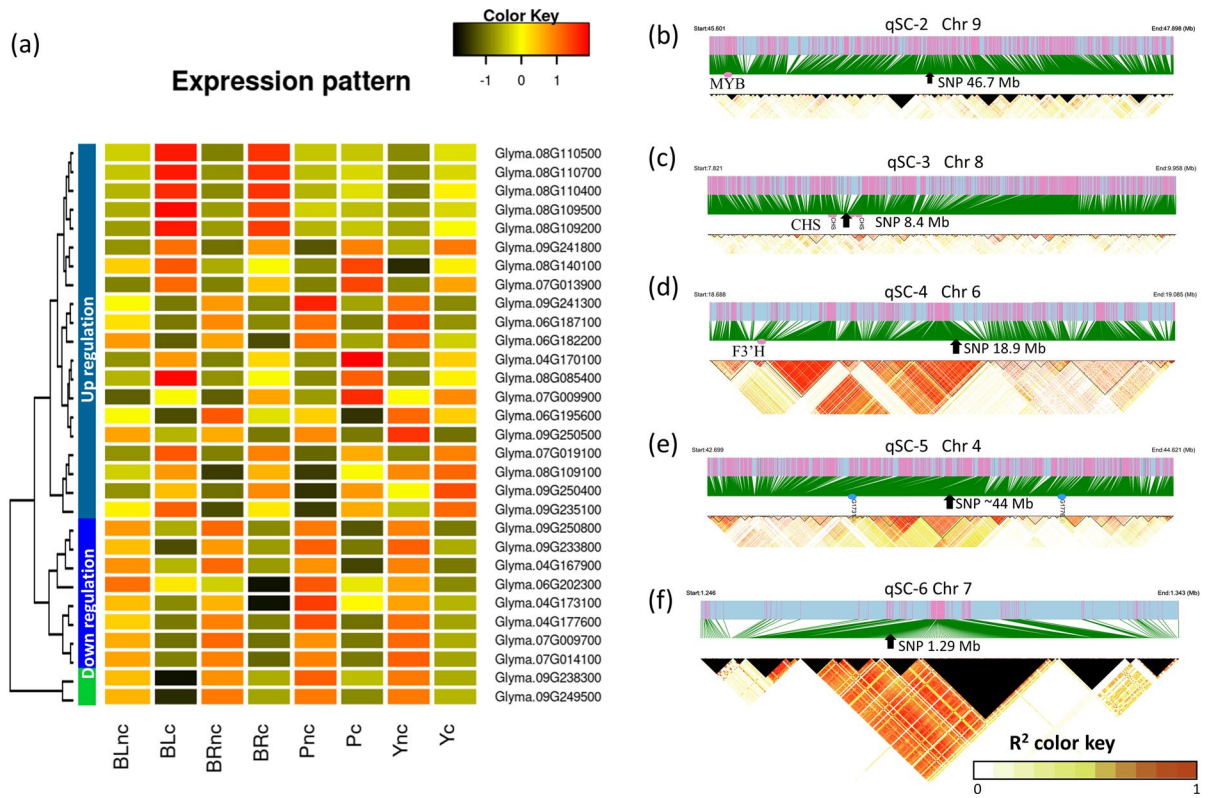


**Fig. 5** Key gene expressional regulation in mapped loci for seed color accumulation. **a** The heatmap showing the expression level of key differentially expressed genes near associated SNPs during pigment accumulation in the seed coat. The average gene expression levels are the *Z*-score of average values of fragments per kilobase of transcript per million mapped reads from three biological experiments. The gene names were listed on the right of each row. The X-axis shows the sample names with different seed coat colors, where BL, BR, P, and Y represent black, brown, purple, and yellow, respectively. The tree on the left side of the heatmap shows the grouped expressional patterns of genes across samples. The suffix nc and c in the sample name represent the control before and after color formation, respectively. **b**, **c**, **d**, **e**, **f** The linkage disequilibrium (LD) blocks of the associated locus on regions *qSC-2*, *-3*, *-4*, *-5*, and *-6* on chromosomes, respectively. The threshold to estimate LD blocks was set to *R*-squared 0.8. The arrow pointed position indicates the most significant SNP in the genome-wide association study. The red dots denote the position of a known gene. The blue dots represent the position of newly mapped two genes on chromosome 4. The significant single nucleotide polymorphism (SNP) position on chromosome 4 is the average value of two adjacent SNPs. The black triangle areas represent the *R*-squared values that were not available

In the *qSC-6* region on chromosome 7 (Fig. 5f), 29 DEGs were identified for the color trait formation. Five DEGs (*Glyma.07G009700*, *Glyma.07G009900*, *Glyma.07G013900*, *Glyma.07G014100*, and *Glyma.07G019100*) were upregulated and shared by all four colors (Fig. 5a). These DEGs were located together between 0.72M and 1.51M on chromosome 7 which were close to the GWAS associating a single SNP site at 1.29 Mbp. The estimated largest LD distance of this associated SNP was ~40 Kbp (Fig. 5f).

Of the QTLs*1~7* and *G* locus, which were detected by BSA mapping and association analysis, only 273 genes with FST (Landraces vs Wild and Improved vs Wild) > 0.6 in the coding sequence (CDS) region were identified, suggesting subjection to domestication selection (Stable 7). Among them, the *Glyma.01G198500*(*G* locus), *Glyma.08G109100*, and *Glyma.08G109500*(*I* locus) were selected. This information may be useful in the selection of candidate genes.

## Characterization of functions of DEG's role for the seed coat color in pathways

Further pathway analysis of DEGs in the overlapped regions against KEGG (accessed by Oct 9th, 2022) showed that 68 DEGs were located in KEGG pathways,

orthologs, or networks. These DEGs mainly were mapped into the pathways of the metabolism (36 DEGs), environmental information processing (5 DEGs), cellular processes (2 DEGs), organismal systems (10 DEGs), and others (Fig. 6). The top DEG-involved pathway was the metabolism, suggesting gene expressional changes in the metabolism could largely affect the seed coat color (Stable 7). These DEGs affected the carbohydrate metabolism, energy metabolism, lipid metabolism, amino acid metabolism, metabolism of other amino acids, glycan biosynthesis and metabolism, metabolism of terpenoids and polyketides, and biosynthesis of other secondary metabolites. We found that some common DEGs function as catalyzing enzymes for compound synthesis. DEGs *Glyma.08G109200*, *Glyma.08G109500*, *Glyma.08G110400*, *Glyma.08G110500*, and *Glyma.08G110700* encode CHS (KEGG id K00660) in flavonoid biosynthesis pathway in glycan biosynthesis and metabolism. These identified CHS were consistent with the genes for color accumulation in previous studies (Cho et al. 2013; Senda et al. 2004; Senda et al. 2013; Todd 1996; Wang et al. 2018; Zhou et al. 2015; Tuteja et al. 2009).

Besides, we found many novel DEGs responsible for the color accumulation. The *qSC-3* commonly associated with DEG *Glyma.08G109100* encodes UDP-glucuronate 4-epimerase (KEGG id K08679) in the pectin biosynthesis (Gu and Bar-Peled 2004). The *qSC-5* associated with common DEG *Glyma.04G167900* encodes the light-harvesting complex I chlorophyll a/b binding protein 4 (LHCA4, KEGG id K08910) in the energy metabolism. In the *qSC-6* region, the common DEG *Glyma.07G009900* functions as a fatty acid omega-hydroxy dehydrogenase (KEGG id K15403) in the lipid metabolism; *Glyma.07G013900* encodes an inositol oxygenase (EGG id K00469) for D-glucuronate biosynthesis in the ascorbate and aldarate metabolism, while *Glyma.07G019100* encodes malonyl-CoA/methylmalonyl-CoA synthetase (KEGG id K18660) in lipid metabolism and amino acid metabolism. The DEG *Glyma.08G085400*, identified in purple seed only, encodes gibberellin 2 beta-dioxygenase (GA2ox, KEGG id K04125) in the metabolism of terpenoids and polyketides. Seven transcription factors were also annotated as DEGs for color accumulation, including one reported previously. Among those six were novel DEGs, including *Glyma.08G140100* for transcription factor TGA (KEGG id K14431) in salicylic acid induced disease resistance signal pathway, *Glyma.09G241800* for homeobox-leucine zipper protein (HD-ZIP) (KEGG id K09338), *Glyma.09G233800* for ethylene response factor (ERF) (Sakuma et al. 2002) (KEGG id K09286) and *Glyma.09G250500* for WRKY transcription factor 2 (Jiang and Yu 2009) (WRKY2, KEGG id K18835), *Glyma.04G170100*, *Glyma.06G195600*, and *Glyma.09G235100* for transcription factor MYB (KEGG id K09422). The gene *Glyma.09G235100* in the reported *R* locus was reported to affect seed color formation (Zabala et al. 2014; Gillman et al. 2011). Functions and GOs of more DEGs were summarized (Stable 7).

## Discussion

### Increasingly improved discovery of genetic loci for seed coat color

A visible trait makes breeding selection much easy. The seed coat color, often yellow, brown, or black, is the breeder's favorite marker in the soybean improvement process (Dixon and Sumner 2003; Koes et al. 2005). To date, the reported six genetic loci (*I*, *T*, *R*, *G*, *W1*, and *O*) mainly associate the flavonoid and

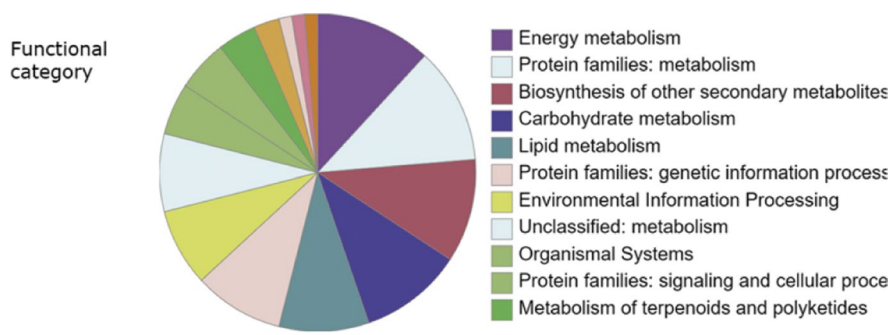


**Fig. 6** Distribution of functional pathways involved by differentially expressed genes for seed coat color. Distribution of functional categories of differentially expressed genes during seed color formation in soybean. The functional categories were classified at KEGG (https://www.kegg.jp/, accessed by Oct 9th, 2022)

anthocyanin pathways in soybean (Song et al. 2016; Yang et al. 2010; Zabala and Vodkin 2007; Cho et al. 2017; Xie et al. 2003). However, the seed coat color is a complex trait. Here we discovered more genes and loci for seed coat color using an omics strategy from 763 accessions, which is different from all previous studies that used a much small soybean population (Kim et al. 2020; Song et al. 2016; McClean et al. 2018). The detected four genetic loci (*I*, *T*, *R*, *G*) in this study were consistent with those in existing reports, suggesting a robust association. Besides, more novel loci with gene candidates were found here.

Color control in plant tissue is mainly through regulating pigment production and modifying existing pigments. Color in other plants is regulated by genes in multiple pathways, such as the anthocyanin biosynthesis for rice pericarp (Yang et al. 2019), chlorophyll catabolism, carotenoids, and flavonoid pathways (Wang et al. 2019; Sun et al. 2022). Here, we also identified related genes for controlling seed coat color from these pathways in soybean. Our discoveries were well supported by omics results from BSA-seq, GWAS, and gene expressional profiles, suggesting robust multiple genes' control on seed coat color in soybean. The omics strategy demonstrated an increased power of gene discovery for gene trait mapping, which may be extendable to other crops. The results of QTL and gene candidates provide a genetic basis for further investigation of the molecular mechanism of gene regulation and help to guide improvement of the soybean breeding.

## Molecular regulatory networks of color trait

For a better understanding, we summarized the molecular regulatory networks and interpreted gene candidates with knowledge from the literature and our novel findings. The regulatory networks comprise at least six major regulatory sections (Fig. 7). The first one is the flavonoid and anthocyanins biosynthesis, in which most regulatory loci have been identified by now, including CHS (*I* locus), F3′H (*T* locus), F3′5′H (*W1* locus), and MYB transcription factor (*R* locus) regulation. Many metabolites in this biosynthesis, e.g., flavonoids and anthocyanins (Khoo et al. 2017), are colorful compounds that may explain the most causal genes in this section. The second section involves multiple transcription factors, including several MYBs, HD-ZIP, WRKY2, TGA, and EREBP-like factors (Sakuma et al. 2002; Chen et al. 2022a; Han et al. 2022; Zhang et al. 2022). WRKY2 is known to mediate seed germination (Jiang and Yu 2009) and was found, here, to regulate seed coat color, indicating that seed coat color affects seed germination via WRKY2. The MYB *R* locus has been reported to regulate seed coat color in soybean (Zabala et al. 2014; Gillman et al. 2011; Gao et al. 2021). Here, we found more MYB transcription

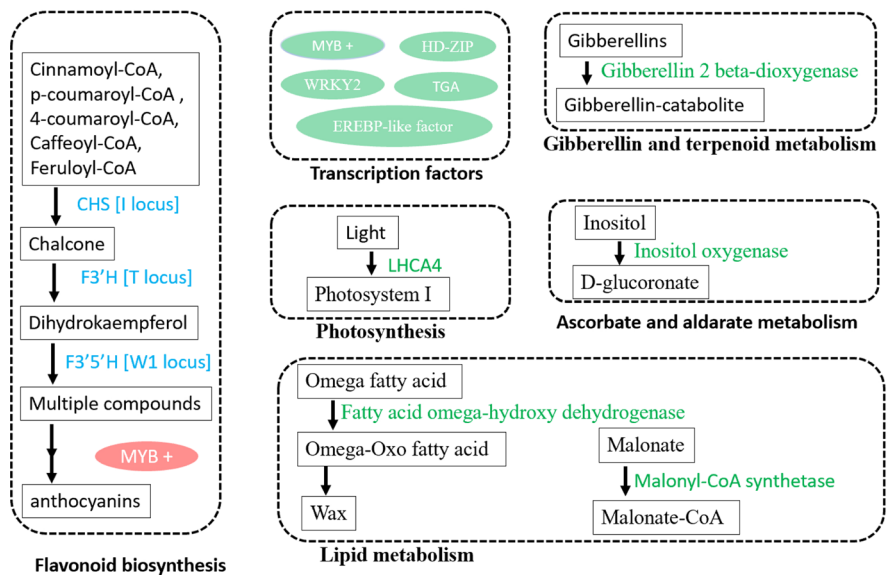


**Fig. 7** An overview of genes and pathways regulating seed coat color in soybean. The gene names in light blue and green color present known and novel genes identified in this study, respectively. LHCA4 for light-harvesting complex I chlorophyll a/b binding protein 4, CHS for chalone synthase, F3′H for flavonoid 3′-hydroxylase, F3′5′H for flavonoid 3′5′-hydroxylase, HD-ZIP for homeobox-leucine zipper protein, and ERF for ethylene response factor

factors involved in forming seed coat colors. We propose that multiple transcription factors may co-act to regulate genes for the seed coat color formation, which could be the interest in future studies. The third section is the gibberellin and terpenoid metabolism, such as the gibberellin 2 beta-dioxygenase gene locus identified in this study. Gibberellin, as a plant hormone, plays a role in cell growth and responses to environmental stresses like drought or cold (Hedden 2020). Terpenoids are essential for the biosynthesis of pigments, including carotenoids for color, toxins against herbivores, and other compounds for environmental stresses (Yazaki et al. 2017; Wang et al. 2021). Therefore, the gibberellin and terpenoid metabolism may confer the roles of seed coat color. The fourth one is the pigment light-harvesting complex I chlorophyll a/b binding protein 4 (LHCA4) in photosynthesis, which harvests the light and reflects the color we can see. The fifth part is the ascorbate and aldarate metabolism (vitamin C metabolism), which protects the plant cells from damage by reactive oxygen species (Smirnoff 2018). Some colors could result from an oxidized state, while this vitamin C metabolism may regulate the oxidization of pigment in the seed coat. The identified inositol oxygenase gene in vitamin C metabolism also is involved in the biosynthesis of plant hormones abscisic acid and jasmonic acid, which regulates many processes in plant development. The sixth network is lipid metabolism, where we identified fatty acid omega-hydroxy dehygrogenase and malonyl synthetase genes for seed coat color. The soybean seed is famous for its enriched oil. Carotenoids, colorful pigments for yellow, orange, and red, could be derived from the lipid metabolism (Maoka 2020; Sun et al. 2022). These two genes in lipid metabolism may involve in carotenoid regulation in the seed coat.

Overall, many biosynthesis, metabolisms, transcription factors, hormones, and photosynthesis are involved in regulating seed coat color. The regulation network of seed coat color is complex, but increasing loci or gene candidates have been identified (Fig. 7). Further validation and characterization of these genes are needed. We want to point out that other regulations may also affect the seed coat color, e.g., an interaction between identified genes or locus. In addition, gene-by-environment interactions may affect the regulation (Li and Lam 2022).

### Common and specific regulatory genes for each color

In this study, we mainly focused on interpreting common regulatory genes and loci during the seed coat color formation (Fig. 5). These genes could be the key effective genes for seed coat color formation in soybean. Here we found five tandem repeated genes *Glyma.08G109200*, *Glyma.08G109500*, *Glyma.08G110400*, *Glyma.08G110500*, and *Glyma.08G110700* encoding CHS, which may indicate an enhanced natural selection for seed coat phenotype. However, some DEGs or loci were also identified to be specifically involved in formatting one color or a color attribute, e.g., the locus for Hue on chromosome 7. Another example is that one DEG may regulate some colors, but not all colors. For example, *Glyma.09G250500*, encoding WRKY transcription factor 2, was only significantly regulated in yellow and purple colors (Stable 7). The specific DEG could be particular to that color. A study showed that a single pigment gene (*P*) could affect the white seed phenotype in common bean (*Phaseolus vulgaris*) (McClean et al. 2018). Further functional validation of the particular color-specific gene is worthy, e.g., by over-expression and silence of this gene. But it may be challenging because of the complexity of multiple effects. The particular gene may interact with other common DEGs to regulate the final visible color in the seed coat. These specific genes may have possible rare DNA variants that only interfere with the regulation involved in the particular color.

## Conclusion

The seed coat color of soybean is a complex quantitative trait and is genetically co-regulated at the expressional level. Robust QTL and candidates were discovered to be responsible for seed coat colors in this omics study with 763 germplasms, including genes in flavonoid biosynthesis, transcription factors, gibberellin and terpenoid metabolism, photosynthesis, ascorbate and aldarate metabolism, and lipid metabolism. The candidates include more genes than genes/loci ever reported. The color formation was regulated by shared DEGs among colors or maybe by color-specific DEGs. The highly consistent results provide a valuable guide for further molecular investigation and the breeding process for better cultivars with colorful seeds.

**Author contributions** All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Jian Song, Qingyuan Guo, Ruixin Xu, and Xuewen Wang. The first draft of the manuscript was written by Jian Song and Xuewen Wang. Caiyu Wu helped on the experimental treatment, and Yinghui Li conceived the mutant and provided data support. Li-Juan Qiu supervised the project and reviewed the manuscript. Jun Wang revised the manuscript. All authors read and approved the final manuscript.

**Funding** This work was supported by the National Natural Scientific Foundation of China (Grant No.: 32072016).

**Data availability** The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

**Declarations**

**Competing interests** The authors declare no competing interests.

## References

Ballaré C (2003) Stress under the sun: spotlight on ultraviolet-B responses. Plant Physiol 132(4):1725–1727. https://doi.org/10.1104/pp.103.027672

Chen X, Xue H, Zhu L, Wang H, Long H, Zhao J, Meng F, Liu Y, Ye Y, Luo X, Liu Z, Xiao G, Zhu S (2022a) ERF49 mediates brassinosteroid regulation of heat stress tolerance in Arabidopsis thaliana. BMC Biol 20(1):254. https://doi.org/10.1186/s12915-022-01455-4

Chen Y, Xiong Y, Hong H, Li G, Gao J, Guo Q, Sun R, Ren H, Zhang F, Wang J, Song J, Qiu L (2022b) Genetic dissection of and genomic selection for seed weight, pod length, and pod width in soybean. Crop J. https://doi.org/10.1016/j.cj.2022.11.006

Cho YB, Jones SI, Vodkin L (2013) The transition from primary siRNAs to amplified secondary siRNAs that regulate chalcone synthase during development of Glycine max seed coats. PLoS One 8(10):1–10. https://doi.org/10.1371/journal.pone.0076954

Cho YB, Jones SI, Vodkin LO (2017) Mutations in argonaute5 illuminate epistatic interactions of the K1 and I loci leading to saddle seed color patterns in Glycine max. Plant Cell 29(4):708–725. https://doi.org/10.1105/tpc.17.00162

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Darrigues A, Hall J, Knaap EVD, Francis DM, Gray S (2008) Tomato analyzer-color test: a new tool for efficient digital phenotyping. J Am Soc Hortic 133(4):579–586. https://doi.org/10.21273/JASHS.133.4.579

Dixon RA, Sumner LW (2003) Legume natural products: understanding and manipulating complex pathways for human and animal health. Plant Physiol 131(3):878–885. https://doi.org/10.1104/pp.102.017319

Dobbels AA, Michno JM, Campbell BW, Virdi KS, Stec AO, Muehlbauer GJ, Naeve SL, Stupar RM (2017) An induced chromosomal translocation in soybean disrupts a KASI ortholog and is associated with a high-sucrose and low-oil seed phenotype. G3 Genesgenetics 7(4):1215–1223. https://doi.org/10.1534/g3.116.038596

Dong SS, He WM, Ji JJ, Zhang C, Guo Y, Yang TL (2020) LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. Brief Bioinform 22(4). https://doi.org/10.1093/bib/bbaa227

Fang C, Li C, Li W, Wang Z, Zhou ZK, Shen YT, Wu M, Wu YS, Li GQ, Kong LA, Liu CM, Jackson SA, Tian Z (2014) Concerted evolution of D1 and D2 to regulate chlorophyll degradation in soybean. Plant J 77(5):700–712. https://doi.org/10.1111/tpj.12419

Gao R, Han T, Xun H, Zeng X, Li P, Li Y, Wang Y, Shao Y, Cheng X, Feng X, Zhao J, Wang L, Gao X (2021) MYB transcription factors GmMYBA2 and GmMYBR function in a feedback loop to control pigmentation of seed coat in soybean. J Exp Bot 72(12):4401–4418. https://doi.org/10.1093/jxb/erab152

Gillman JD, Tetlow A, Lee JD, Shannon JG, Bilyeu K (2011) Loss-of-function mutations affecting a specific Glycine max R2R3 MYB transcription factor result in brown hilum and brown seed coats. BMC Plant Biol 11(1):155–155. https://doi.org/10.1186/1471-2229-11-155

Gu X, Bar-Peled M (2004) The biosynthesis of UDP-galacturonic acid in plants. Functional cloning and characterization of Arabidopsis UDP-d-glucuronic acid 4-epimerase. Plant Physiol 136(4):4256–4264. https://doi.org/10.1104/pp.104.052365

Han J, Xie X, Zhang Y, Yu X, He G, Li Y, Yang G (2022) Evolution of the dehydration-responsive element-binding protein subfamily in green plants. Plant Physiol 190(1):421–440. https://doi.org/10.1093/plphys/kiac286

Hedden P (2020) The current status of research on gibberellin biosynthesis. Plant Cell Physiol 61(11):1832–1849. https://doi.org/10.1093/pcp/pcaa092

Hill JT, Demarest BL, Bisgrove BW, Gorsi B, Yost HJ (2013) MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. Genome Res 23(4):687–697. https://doi.org/10.1101/gr.146936.112

Jiang W, Yu D (2009) Arabidopsis WRKY2 transcription factor mediates seed germination and postgermination arrest of development by abscisic acid. BMC Plant Biol 9(1):96. https://doi.org/10.1186/1471-2229-9-96

Khoo HE, Azlan A, Tang ST, Lim SM (2017) Anthocyanidins and anthocyanins: colored pigments as food, pharmaceutical ingredients, and the potential health benefits. Food Nutr Res 61(1):1361779. https://doi.org/10.1080/16546628.2017.1361779

Kim JH, Park JS, Lee CY, Jeong MG, Xu JL, Choi Y, Jung HW, Choi HK (2020) Dissecting seed pigmentation-associated genomic loci and genes by employing dual approaches of reference-based and k-mer-based GWAS with 438 Glycine

accessions. PLoS One 15(12):e0243085. https://doi.org/10.1371/journal.pone.0243085
Koes R, Verweij W, Quattrocchio F (2005) Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. Trends Plant Sci 10(5):236–242. https://doi.org/10.1016/j.tplants.2005.03.002
Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. https://doi.org/10.1038/nmeth.1923
Li MW, Lam HM (2022) Genomic studies of plant-environment interactions. Int J Mol Sci 23(11):13943. https://doi.org/10.3390/ijms23115871
Li YH, Qin C, Wang L, Jiao CZ, Hl H, Tian Y, Li YF, Xing GN, Wang J, Gu YZ, Gao XP, Li DL, Li HY, Liu ZX, Jing X, Feng BB, Zhao T, Guan RX, Guo Y et al (2022) Genome-wide signatures of the geographic expansion and breeding of soybean. Sci China Life Sci 454:1–16. https://doi.org/10.1007/s11427-022-2158-7
Maoka T (2020) Carotenoids as natural functional pigments. J Nat Med 74(1):1–16. https://doi.org/10.1007/s11418-019-01364-x
McClean PE, Bett KE, Stonehouse R, Lee R, Pflieger S, Moghaddam SM, Geffroy V, Miklas P, Mamidi S (2018) White seed color in common bean (Phaseolus vulgaris) results from convergent evolution in the P (pigment) gene. New Phytol 219(3):1112–1123. https://doi.org/10.1111/nph.15259
Mckenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303. https://doi.org/10.1101/gr.107524.110
Nakano M, Yamada T, Masuda Y, Sato Y, Kobayashi H, Ueda H, Morita R, Nishimura M, Kitamura K, Kusaba M (2014) A green-cotyledon/stay-green mutant exemplifies the ancient whole-genome duplications in soybean. Plant Cell Physiol 55(10):1763–1771. https://doi.org/10.1093/pcp/pcu107
Palmer RG, Pfeiffer TW, Buss GR, Kilen TC (2004) Qualitative genetics in soybeans: improvement, produnction, and uses, 3rd edn. ASA, CSSA, AND SSSA, Madison(WI), pp 137–233
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909. https://doi.org/10.1038/ng1847
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira M, Bender D, Maller J, Sklar P, Bakker P, Daly MJ (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. https://doi.org/10.1086/519795
Rodríguez GR, Moyseenko JB, Robbins MD, Huarachi Morejón N, Francis DM, Esther VDK (2010) Tomato analyzer: a useful software application to collect accurate and detailed morphological and colorimetric data from two-dimensional objects. J Vis Exp 37(37):e1856. https://doi.org/10.3791/1856
Sadohara R, Long Y, Izquierdo P, Urrea CA, Morris D, Cichy K (2021) Seed coat color genetics and genotype × environment effects in yellow beans via machine-learning and genome-wide association. Plant Genome 15:e20173. https://doi.org/10.1002/tpg2.20173
Saghai-Maroof MA (1985) Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics. P Natl Acad Sci 81(24):8014–8018. https://doi.org/10.1073/pnas.81.24.8014
Sakuma Y, Liu Q, Dubouzet JG, Abe H, Shinozaki K, Yamaguchi-Shinozaki K (2002) DNA-binding specificity of the ERF/AP2 domain of Arabidopsis DREBs, transcription factors involved in dehydration- and cold-inducible gene expression. Biochem Biophys Res Commun 290(3):998–1009. https://doi.org/10.1006/bbrc.2001.6299
Santos-Buelga C, Mateus N, Freitas VD (2014) Anthocyanins. plant pigments and beyond. J Agric Food Chem 62(29):6879–6884. https://doi.org/10.1021/jf501950s
Senda M, Nishimura S, Kasai A, Yumoto S, Takada Y, Tanaka Y, Ohnishi S, Kuroda T (2013) Comparative analysis of the inverted repeat of a chalcone synthase pseudogene between yellow soybean and seed coat pigmented mutants. Breed Sci 63(4):384–392. https://doi.org/10.1270/jsbbs.63.384
Senda M, Masuta C, Ohnishi S, Goto K, Kasai A (2004) Patterning of virus-infected Glycine max seed coat is associated with suppression of endogenous silencing of chalcone synthase genes. Plant Cell 16(4):807–818. https://doi.org/10.1105/tpc.019885
Smirnoff N (2018) Ascorbic acid metabolism and functions: a comparison of plants and mammals. Free Radic Biol Med 122:116–129. https://doi.org/10.1016/j.freeradbiomed.2018.03.033
Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2014) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J 13(2):211–221. https://doi.org/10.1111/pbi.12249
Song J, Li Z, Liu Z, Guo Y, Qiu L-J (2017) Next-generation sequencing from bulked-segregant analysis accelerates the simultaneous identification of two qualitative genes in soybean. Front Plant Sci 8:919. https://doi.org/10.3389/fpls.2017.00919
Song J, Liu Z, Hong H, Ma Y, Tian L, Li X, Li Y-H, Guan R, Guo Y, Qiu L-J (2016) Identification and validation of loci governing seed coat color by combining association mapping and bulk segregation analysis in soybean. PLoS One 11(7):e0159064. https://doi.org/10.1371/journal.pone.0159064
Sun L, Miao Z, Cai C, Zhang D, Zhao M, Wu Y, Zhang X, Swarm SA, Zhou L, Zhang ZJ (2015) GmHs1-1, encoding a calcineurin-like protein, controls hard-seededness in soybean. Nat Genet 47(8):939. https://doi.org/10.1038/ng.3339
Sun T, Rao S, Zhou X, Li L (2022) Plant carotenoids: recent advances and future perspectives. Mol Horticulture 2(1):3. https://doi.org/10.1186/s43897-022-00023-2
Toda K, Yang D, Yamanaka N, Watanabe S, Harada K, Takahashi R (2002) A single-base deletion in soybean flavonoid 3′-hydroxylase gene is associated with gray pubescence color. Plant Mol Biol 50(2):187–196
Todd JJ (1996) Duplications that suppress and deletions that restore expression from a chalcone synthase multigene

family. Plant Cell 8(4):687–699. https://doi.org/10.1105/tpc.8.4.687

Tuteja JH, Zabala G, Varala K, Hudson M, Vodkin LO (2009) Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in Glycine max seed coats. Plant Cell 21(10):3063–3077. https://doi.org/10.1105/tpc.109.069856

Wallace TC, Giusti MM (2015) Anthocyanins. Adv Nutr 6(5):620–622. https://doi.org/10.3945/an.115.009233

Wang J, Wang H, Fu Y, Huang T, Liu Y, Wang X (2021) Genetic variance and transcriptional regulation modulate terpenoid biosynthesis in trichomes of Nicotiana tabacum under drought. Ind Crop Prod 167:113501. https://doi.org/10.1016/j.indcrop.2021.113501

Wang M, Li W, Fang C, Xu F, Liu Y, Wang Z, Yang R, Zhang M, Liu S, Lu S (2018) Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat Genet 50(10):1435–1441. https://doi.org/10.1038/s41588-018-0229-2

Wang X, Liu B-y, Zhao Q, Sun X, Li Y, Duan Z, Miao X, Luo S, Li J (2019) Genomic variance and transcriptional comparisons reveal the mechanisms of leaf color affecting palatability and stressed defense in tea plant. Genes 10(11):929

Xie DY, Sharma SB, Paiva NL, Ferreira D, Dixon RA (2003) Role of anthocyanidin reductase, encoded by BANYULS in plant flavonoid biosynthesis. Science 299:396–399. https://doi.org/10.1126/science.1078540

Xie M, Chung YL, Li MW, Wong FL, Wang X, Liu A, Wang Z, Leung KY, Wong TH, Tong SW (2019) A reference-grade wild soybean genome. Nat Commun 10(1):1216. https://doi.org/10.1038/s41467-019-09142-9

Yang K, Jeong N, Moon JK, Lee YH, Lee SH, Kim HM, Hwang CH, Back K, Palmer RG, Jeong SC (2010) Genetic analysis of genes controlling natural variation of seed coat and flower colors in soybean. J Hered 101(6):757–768. https://doi.org/10.1093/jhered/esq078

Yang X, Xia X, Zhang Z, Nong B, Zeng Y, Wu YY, Xiong F, Zhang YX, Liang HF, Pan YH, Dai GX, Deng GF, Li D (2019) Identification of anthocyanin biosynthesis genes in rice pericarp using PCAMP. Plant Biotechnol J 17(9):1700–1702. https://doi.org/10.1111/pbi.13133

Yazaki K, Arimura G-i, Ohnishi T (2017) 'Hidden' terpenoids in plants: their biosynthesis, localization and ecological roles. Plant Cell Physiol 58(10):1615–1621. https://doi.org/10.1093/pcp/pcx123

Yuan B, Yuan C, Wang Y, Liu X, Qi G, Wang Y, Dong L, Zhao H, Li Y, Dong Y (2022) Identification of genetic loci conferring seed coat color based on a high-density map in soybean. Front Plant Sci 13:968618. https://doi.org/10.3389/fpls.2022.968618

Zabala G, Vodkin L (2003) Cloning of the pleiotropic T locus in soybean and two recessive alleles that differentially affect structure and expression of the encoded flavonoid 3' hydroxylase. Genetics 163(1):295–309

Zabala G, Vodkin LO (2007) A rearrangement resulting in small tandem repeats in the F3′5′H gene of white flower genotypes is associated with the soybean locus. Crop Sci 47(S2):S113–S124. https://doi.org/10.2135/cropsci2006.12.0838tpg

Zabala G, Vodkin LO, Cui Z (2014) Methylation affects transposition and splicing of a large CACTA transposon from a MYB transcription factor regulating anthocyanin synthase genes in soybean seed coats. PLoS One 9(11):e111959. https://doi.org/10.1371/journal.pone.0111959

Zhang S, Du H, Ma Y, Li H, Kan G, Yu D (2021) Linkage and association study discovered loci and candidate genes for glycinin and β-conglycinin in soybean (Glycine max L. Merr.). Theor Appl Genet 134(3):1201–1215. https://doi.org/10.1007/s00122-021-03766-6

Zhang Y, Guo C, Deng M, Li S, Chen Y, Gu X, Tang G, Lin Y, Wang Y, He W, Li M, Zhang Y, Luo Y, Wang X, Chen Q, Tang H (2022) Genome-wide analysis of the ERF family and identification of potential genes involved in fruit ripening in octoploid strawberry. Int J Mol Sci 23(18). https://doi.org/10.3390/ijms231810550

Zhou Z, Yu J, Zheng W, Gou Z, Tian Z (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33(4):408–414. https://doi.org/10.1038/nbt.3096

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.