Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2024), pp. 1–15 doi: 10.1111/pbi.14541

An R2R3-type MYB transcription factor, GmMYB77, negatively regulates isoflavone accumulation in soybean [*Glycine max* (L.) Merr.]

Yitian Liu†, Shengrui Zhang†, Jing Li†, Azam Muhammad, Yue Feng, Jie Qi, Dan Sha, Yushui Hao, Bin Li* and Junming Sun*

The State Key Laboratory of Crop Gene Resources and Breeding, National Engineering Research Center for Crop Molecular Breeding, Key Laboratory of Soybean Biology (Beijing), Ministry of Agriculture and Rural Affairs, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China

Received 10 August 2024;
revised 4 November 2024;
accepted 17 November 2024.
**Correspondence* (Tel/fax +861082105805; email sunjunming@caas.cn (J.S.) and Tel/fax +861082105805; email libin02@caas.cn (B.L.))
†These authors contributed equally to this work.

Keywords: soybean, malonylglycitin, total isoflavone, *GmMYB77*, haplotype.

Summary

Soybean [*Glycine max* (L.) Merr.] is an exceptionally rich in isoflavones, and these compounds attach to oestrogen receptors in the human body, lessening the risk of breast cancer and effectively alleviating menopausal syndrome symptoms. Uncovering the molecular mechanisms that regulate soybean isoflavone accumulation is crucial for enhancing the production of these compounds. In this study, we combined bulk segregant analysis sequencing (BSA-seq) and a genome-wide association study (GWAS) to discover a novel R2R3-MYB family gene, *GmMYB77*, that regulates isoflavone accumulation in soybean. Using the soybean hairy root transient expression system, we verified that *GmMYB77* inhibits isoflavone accumulation. Furthermore, knocking out *GmMYB77* significantly increased total isoflavone (TIF) content, particularly malonylglycitin, while its overexpression resulted in a notable decrease in contents of malonylglycitin and TIF. We found that GmMYB77 can directly binds the core sequence GGT and suppresses the expression of the key isoflavone biosynthesis genes *Isoflavone synthase 1* (*GmIFS1*), *Isoflavone synthase 2* (*GmIFS2*), *Chalcone synthase 7* (*GmCHS7*) and *Chalcone synthase 8* (*GmCHS8*) by using dual-luciferase assays, electrophoretic mobility shift assays and yeast one-hybrid experiments. Natural variations in the promoter region of *GmMYB77* affect its expression, thereby regulating the malonylglycitin and TIF contents. Hap-P2, an elite haplotype, plays a pivotal role in soybean breeding for substantially enhanced isoflavone content. These findings enhance our understanding of the genes influencing soybean isoflavone content and provide a valuable genetic resource for molecular breeding efforts in the future.

Introduction

Soybean [*Glycine max* (L.) Merr.] is an important source of protein and oil, containing bioactive compounds such as isoflavones (Azam *et al.*, 2020), soluble sugars (Qi *et al.*, 2022), folates (Agyenim-Boateng *et al.*, 2022) and saponins (Guang *et al.*, 2014). Notably, soybean isoflavones are known for their structural similarity to oestrogen, exhibiting oestrogen-like effects that can reduce the risk of breast cancer and osteoporosis while effectively alleviating menopausal syndrome symptoms (Al-Nakkash and Kubinski, 2020; Dixon *et al.*, 2002; Nestel, 2003). In plants, isoflavones serve as signalling molecules that facilitate symbiosis with nitrogen-fixing bacteria, act as phytoalexins for defence against pests and inhibit the spread of pathogenic bacteria (Aoki *et al.*, 2000; Cesco *et al.*, 2012; Lozovaya *et al.*, 2004; Zhang *et al.*, 2012). High isoflavone levels make soybeans the primary source of these compounds in the human diet. Isoflavones have garnered increasing research attention owing to their essential functions. Uncovering the key genes and molecular mechanisms that regulate isoflavone accumulation in soybean seeds is crucial for enhancing their isoflavone content in breeding efforts.

Isoflavones are a group of compounds characterized by a 3-benzopyranone core. In soybean, there are 12 primary isoflavones, including three aglycones (daidzein, genistein and glycitein) and their nine corresponding glycosides (daidzin, glycitin, genistin, acetyldaidzin, acetylglycitin, acetylgenistin, malonyldaidzin, malonylglycitin and malonylgenistin) (Azam *et al.*, 2020; Kudou *et al.*, 1991). The malonyl form is the predominant storage isoflavone in soybean, constituting over 80% of the total isoflavone (TIF) content. The active forms of isoflavones are aglycones, but the functions of different aglycones can vary. Daidzein is beneficial in the treatment of cancer (Laddha and Kulkarni, 2023), while genistein has great potential as an anticancer, antioxidant and antiproliferative agent (Piovesan *et al.*, 2005). Glycitein and its corresponding derivatives confer a major bitter and astringent taste to soybean products, especially malonylglycitin and acetylglycitin, which present the lowest threshold value of bitterness (Kudou *et al.*, 1991). The isoflavone levels within different soybean accessions exhibit

Please cite this article as: Liu, Y., Zhang, S., Li, J., Muhammad, A., Feng, Y., Qi, J., Sha, D., Hao, Y., Li, B. and Sun, J. (2024) An R2R3-type MYB transcription factor, GmMYB77, negatively regulates isoflavone accumulation in soybean [*Glycine max* (L.) Merr.]. *Plant Biotechnol. J.*, https://doi.org/10.1111/pbi.14541.

substantial variation, with the TIF content ranging from 677.25 to 5823.29 μg/g (Azam *et al.*, 2023b).

The isoflavone biosynthesis pathway affects their content, making this an important target for improving the isoflavone composition of soybean. Isoflavone biosynthesis is initiated through the phenylpropanoid pathway, producing compounds that are then converted into conjugated forms (i.e. glucosyl- and malonyl-glucosides) and stored in vacuoles (Yu *et al.*, 2003). Phenylalanine, generated from the shikimate pathway, is used for the biosynthesis of phenylpropanoid metabolites (Dong and Lin, 2021). Phenylalanine is converted into cinnamic acid by phenylalanine lyase (PAL), which is further converted into *p*-coumaric acid and *p*-coumaroyl-CoA by 4-hydroxylase (C4H) and 4-coumarate-CoA ligase (4CL), respectively. *p*-Coumaroyl-CoA is catalysed by chalcone reductase (CHR) and chalcone synthase (CHS) to produce isoliquiritigenin chalcone and naringenin chalcone, which then undergo cyclization catalysed by chalcone isomerase (CHI) to form isoliquiritigenin and naringenin, respectively. (Ralston *et al.*, 2005). The daidzein and genistein aglycones are generated from liquiritigenin and naringenin, respectively, under the catalysis of isoflavone synthase (IFS) and 2-hydroxyisoflavone dehydratase (HID) (Akashi *et al.*, 1999; Dhaubhadel *et al.*, 2003; Jung *et al.*, 2000). Glycitein aglycone is biosynthesized from liquiritigenin by flavonoid 6-hydroxylase (F6H), IFS, HID and *O*-methyltransferase (IOMT) (Uchida *et al.*, 2020). These aglycones are further converted into daidzin, genistin and glycitin by uridine diphosphate glycosyltransferase (UGT), and into their malonyl and acetyl derivatives by malonyl transferase (MT) and acetyl transferase (AT), respectively (Dixon and Steele, 1999; Tohge *et al.*, 2017; Winkel-Shirley, 2001).

Isoflavone contents are complex traits regulated by multiple genes and influenced by environmental factors. Transcription factors (TFs), such as the members of the myeloblastosis (MYB); basic helix–loop–helix (bHLH); WRKYGQK DNA-binding protein (WRKY); MCM1, AGAMOUS, DEFICIENS, SRF box (MADS box); and zinc-finger (ZF) families, play pivotal roles in regulating isoflavonoid biosynthesis. These TFs can either promote or suppress flavonoid biosynthesis by binding specific sequences in the promoters of key genes (LaFountain and Yuan, 2021; Stracke *et al.*, 2001). Several MYB TFs have been implicated in regulating the isoflavone biosynthesis pathway in soybean; for example, GmMYB176 regulates *GmCHS8* expression and isoflavone biosynthesis, and co-immunoprecipitates with the bZIP TF GmbZIP5 (Anguraj Vadivel *et al.*, 2021; Yi *et al.*, 2010). GmMYB29 acts as a positive regulator of isoflavone production, with its overexpression increasing isoflavone accumulation in transgenic soybean hairy roots (Chu *et al.*, 2017). GmMYB133 promotes isoflavonoid biosynthesis in soybean hairy roots by interacting with the GmMYB176 and 14-3-3 proteins (Bian *et al.*, 2018). The overexpression of GmZFP7 in transgenic soybeans enhances isoflavone accumulation in the seeds (Feng *et al.*, 2022). Conversely, GmMYB100 inhibits isoflavonoid production by down-regulating the expression of *CHS*, *CHI* and *IFS* (Yan *et al.*, 2015). In addition to genetic regulation, recent studies have indicated that light signalling also influences isoflavone biosynthesis in soybean; for instance, the blue-light photoreceptors (GmCRY1s, GmCRY2s, GmPHOT1s and GmPHOT2s) and TFs (GmSTF1 and GmSTF2) promote isoflavone accumulation. By contrast, the E3 ubiquitin ligase GmCOP1b inhibits this process (Song *et al.*, 2024). Despite these insights, the gene networks and molecular regulatory mechanisms governing isoflavone accumulation in soybean have remained largely unexplored.

In this study, we identified a major locus regulating the isoflavone biosynthesis pathway using bulk segregant analysis sequencing (BSA-seq) combined with a genome-wide association study (GWAS). We cloned the underlying causal gene, which encodes a MYB TF, GmMYB77. Furthermore, we validated the function of GmMYB77 in hairy roots and transgenic soybean, and identified elite haplotypes associated with seed isoflavone content in the *GmMYB77* sequence. These findings provide valuable insights into the genetic regulation of isoflavone biosynthesis and offer potential targets for improving soybean nutritional quality through molecular breeding.

Results

Identification of the soybean isoflavone accumulation regulator *GmMYB77* through BSA-seq and GWAS

To identify the genes controlling isoflavone content, the seed isoflavone contents of 1551 soybean accessions grown in two locations for 2 years were profiled using high-performance liquid chromatography (HPLC). Two extreme pools with high (4065.10 μg/g) and low (1427.23 μg/g) isoflavone contents were constructed for BSA-seq to identify candidate genes involved in the isoflavonoid biosynthesis pathway (Azam *et al.*, 2023a). The difference in allele frequencies between the two extreme bulks was calculated using the ΔSNP-index. An interval correlated with isoflavone content was discovered on chromosome 4 (physical location: 2.86–2.94 Mb). Within this interval, the candidate gene *Glyma.04g036700*, carrying three mis-sense mutations and five upstream SNPs between the bulks, showed the highest ΔSNP-index peak (Figure 1a; Table S1). This gene encodes a MYB TF, here named *GmMYB77*.

We analysed previous GWAS results to validate the candidate gene identified in the BSA-seq analysis (Azam *et al.*, 2023b). Using 6 149 599 SNPs from 2214 previously sequenced soybean accessions (Li *et al.*, 2023), we identified a strong signal on chromosome 4 (chr4: 2925585 bp, *P*-value: 1.38e-07) associated with the malonylglycitin content. A linkage disequilibrium (LD) block analysis using the 150-kb flanking regions upstream and downstream of each SNP narrowed the candidate region to a smaller 153.4-kb linkage region (2 775 761 bp to 2 929 193 bp) containing 23 genes (Figure 1b,c; Table S2). Combining the results of BSA-seq and GWAS, we identified the MYB TF gene *GmMYB77* as a strong candidate for control of the soybean seed isoflavone content.

GmMYB77 regulates the isoflavone content and expression of isoflavonoid biosynthesis–related genes in soybean hairy roots

The coding sequence (CDS) of *GmMYB77* from soybean variety Tianlong1 (TL1) is 633 bp in length. This sequence encodes a protein consisting of 211 amino acids, with a molecular mass of 23.68 kDa and an isoelectric point (pI) of 7.74 (https://web.expasy.org/protparam/). *GmMYB77* shares the highest identity with other R2R3-MYB TFs with the R2R3 repeat domain in the N-terminal region (Figure S1a). *GmMYB77* was expressed in all soybean organs, with higher expression levels in the roots and lower levels in the seeds (Figure S1b). To investigate the subcellular localization of GmMYB77, recombinant vectors containing *GmMYB77-GFP* [encoding a fusion with green fluorescent protein (GFP)] or *GFP* alone were transformed into tobacco (*Nicotiana benthamiana* Domin) leaves via an *Agrobacterium*-mediated method. The GmMYB77-GFP fusion

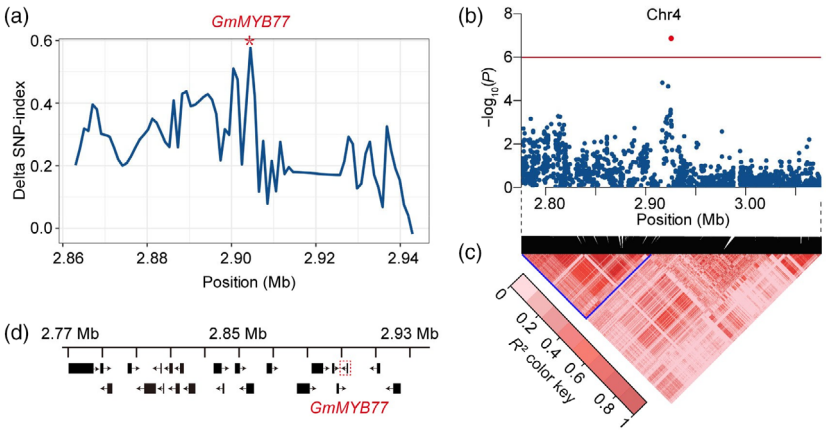


Figure 1 Identification of the candidate gene *GmMYB77* underlying differences in isoflavone content among soybean accessions using bulk segregant analysis sequencing (BSA-seq) and a genome-wide association study (GWAS). (a) Delta index value of BSA-seq in the linkage region of the peak single-nucleotide polymorphism (SNP). The red dot indicates the candidate gene *GmMYB77*. (b) Manhattan plot for GWAS using the malonylglycitin content of soybean seeds collected in Hainan, China, in 2017. The solid red line represents the significant *P*-value threshold corrected using the Bonferroni correction method ($P = 1.0 \times 10^{-6}$). (c) Linkage disequilibrium (LD) analysis of SNPs surrounding the peak SNP on chromosome 4. The colour key (white to red) represents the LD value (r^2) of the accessions. (d) The 2.77–2.93-Mb region of interest on chromosome 4 of soybean (Williams 82 reference genome), which contains 23 predicted genes. Black boxes represent the location of genes, and arrows indicate the gene orientations. The gene in the red box is *GmMYB77*.

protein was localized to the nucleus and cytoplasm, while GFP fluorescence was observed throughout the cells transformed with the *GFP* control vector (Figure S1c).

Compared with the empty vector control, the relative expression level of *GmMYB77* was about 16-fold higher in hairy roots transformed to overexpress this gene (Figure 2a). Importantly, the contents of daidzin, malonylglycitin, malonyldaidzin and TIF were significantly decreased in the *GmMYB77*-overexpressing (*GmMYB77*-OE) transgenic hairy roots, by an average of 18.8%, 20.6%, 20.3% and 23.2%, respectively, compared with the empty vector (Figure 2b–e). Conversely, the *GmMYB77* transcript level was reduced by about 50% in RNA interference (RNAi) hairy roots compared with the empty vector (Figure 2f). The contents of daidzin, malonylglycitin, malonyldaidzin and TIF were increased by 68.8%, 28.8%, 31.6% and 36.9%, respectively, in the *GmMYB77*-RNAi roots compared with the empty vector (Figure 2g–j). The contents of genistin and malonylgenistin in the transgenic hairy roots did not significantly differ from those of the empty vector (Figure S2).

GmMYB77 directly binds to the promoters of *GmIFS1*, *GmIFS2*, *GmCHS7*, and *GmCHS8* and inactivates their expression

We analysed the expression of 22 key genes encoding enzymes in the isoflavone biosynthesis pathway using reverse transcription quantitative PCR (RT-qPCR) and found that the overexpression of *GmMYB77* significantly reduced the expression levels of *GmIFS1*, *GmIFS2*, *GmCHS7*, *GmCHS8*, *GmCHR2* and *GmCHR3* (Figure 2k–n; Figure S3). Conversely, in *GmMYB77*-RNAi hairy roots, the transcription levels of *GmIFS1*, *GmIFS2*, *GmCHS7*, *GmCHS8*, *GmPAL1*, *Gm4CL* and *GmCHR6* were significantly increased, indicating that GmMYB77 negatively regulated these genes. We conducted a dual-luciferase (LUC) assay to determine whether GmMYB77 directly represses *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8*, using the promoters of these genes fused to LUC as reporters, *GmMYB77* driven by the CaMV *35S* promoter as the effector, and the *REN* (encoding Renilla LUC) gene driven by the CaMV *35S* promoter as a control (Figure 3a). The overexpression of *GmMYB77* significantly decreased the activity of the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters by 62.8%, 62.1%, 88.5% and 84.0%, respectively (Figure 3b–e). These results suggest that GmMYB77 plays a critical role in the transcriptional regulation of genes encoding key proteins in the soybean isoflavone biosynthesis pathway.

To further identify the GmMYB77 recognition regions within the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters, 5′ deletions were performed to produce four different fragments of each promoter, with the full-length promoters being approximately 2000 bp each. These fragments were tested in a dual-LUC assay to determine the specific regions required for GmMYB77 binding. The LUC activity of *pIFS1*-2 and *pIFS1*-3 was significantly decreased by GmMYB77, similar to the full-length promoter *pIFS1*-1, while *pIFS1*-4 did not show a significant difference when compared with control (empty vector) samples (Figure 3f), indicating that the crucial binding region lies between the 5′ ends of *pIFS1*-3 and *pIFS1*-4, spanning 492 bp. For *GmIFS2*, the LUC activity of *pIFS2*-1 was significantly decreased by GmMYB77 compared with the control; however, the inhibitory effect disappeared in *pIFS2*-2, *pIFS2*-3 and *pIFS2*-4, indicating that the crucial binding region lies within the 728-bp interval between the 5′ ends of *pIFS2*-1 and *pIFS2*-2 (Figure 3g). Similarly, like the full-length promoter *pCHS7*-1, the LUC activity of *pCHS7*-2 and *pCHS7*-3 was significantly decreased by GmMYB77 compared with the control, while the inhibitory effect disappeared in *pCHS7*-4, suggesting that the crucial binding region is located within the 563 bp between the 5′ ends of *pCHS7*-3 and *pCHS7*-4 (Figure 3h). For *GmCHS8*, the LUC activity of *pCHS8*-2 was significantly decreased by GmMYB77 compared with the control, exhibiting the same inhibitory effect as the full-length promoter *pCHS8*-1, while no inhibitory effect was observed in *pCHS8*-3 and *pCHS8*-4. This indicates that the crucial binding region is located within the 429 bp between the 5′ ends of *pCHS8*-2 and *pCHS8*-3

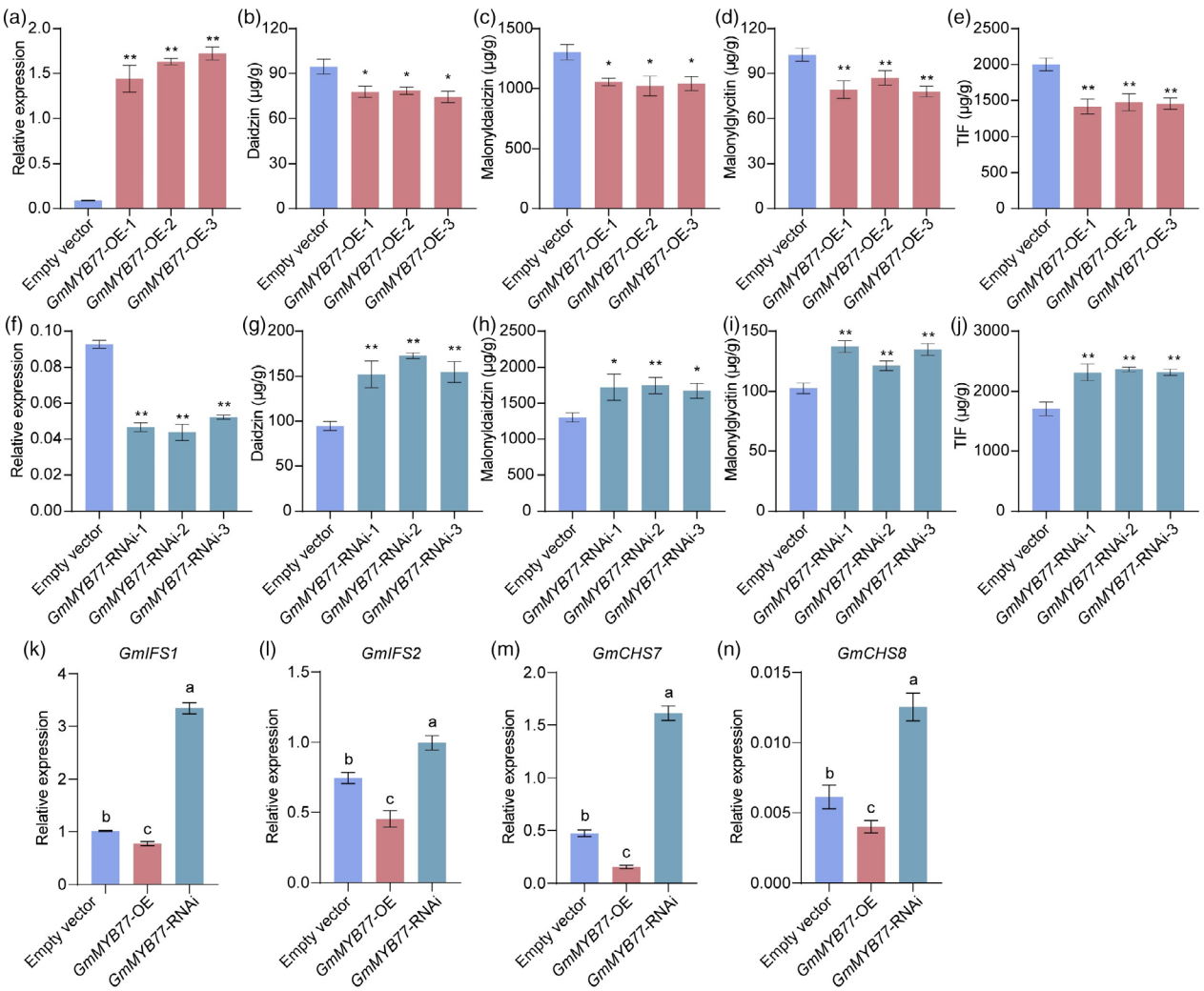


Figure 2 Evaluation of the isoflavone content in *GmMYB77*-overexpressing (*GmMYB77*-OE) and *GmMYB77*-RNAi soybean hairy roots. (a) The relative expression level of *GmMYB77* in the *GmMYB77*-OE hairy roots in soybean. (b–e) Analysis of daidzin (b), malonyldaidzin (c), malonylglycitin (d) and total isoflavone (TIF) contents (e) in *GmMYB77*-overexpressing soybean hairy roots. (f) The relative expression level of *GmMYB77* in the *GmMYB77*-RNAi soybean hairy roots. (g–j) Analysis of daidzin (g), malonyldaidzin (h), malonylglycitin (i) and TIF contents (j) in *GmMYB77*-RNAi soybean hairy roots. Data are shown as means ± SE. Student's *t*-tests were used to identify statistically significant differences between the genotypes (*$P < 0.05$, **$P < 0.01$). (k–n) Transcription levels of the indicated genes in *GmMYB77*-overexpressing (OE) and *GmMYB77*-RNAi soybean hairy roots. Different letters indicate statistically significant differences [analysis of variance (ANOVA), the least significant difference (LSD) method].

(Figure 3i). These results demonstrate that GmMYB77 represses the activity of these genes by binding to specific regions on their promoters.

We performed an electrophoretic mobility shift assay (EMSA) to validate the binding of GmMYB77 to these promoter regions. The 30-bp target probe designed using JASPAR (https://jaspar.elixir.no/) confirmed that GmMYB77 directly and specifically binds to the crucial binding regions of the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters *in vitro*, while this binding was abolished when a competitor was added (Figure 3j). A yeast one-hybrid (Y1H) assay was initially conducted to evaluate the interaction between GmMYB77 and the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters. As shown in Figure 3k, the yeast cells were selected on the SD/−Leu medium containing 200 ng/mL aureobasidin A (AbA). Only the combination of GmMYB77 and the *GmIFS1*, *GmIFS2*, *GmCHS7* or *GmCHS8* promoters survived on the selection medium, consistent with the positive control containing pGADT7-Rec-p53 and p53-pAbAi, whereas the negative control containing empty pGADT7 and promoter-pAbAi did not. These findings provide evidence of the exclusive binding of GmMYB77 to the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters.

To further investigate the binding sequence recognized by the GmMYB77 protein, we analysed the predicted binding sequences of the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters from JASPAR database for GmMYB77 (Figure 4a–d). We found that the predicted binding sequences within the promoters of *GmIFS2*, *GmCHS7* and *GmCHS8* exhibited similar sequences as GGTAGGT, while that of *GmIFS1* is GACGGTT. Therefore, six mutations were generated into each of the two GmMYB77 binding core sequences (GACGGTT for *GmIFS1* and GGTAGGT for *GmCHS8*) and validated the resulting binding interactions using EMSA assays (Figure 4e,f). For the GACGGTT binding sequence, GmMYB77 displayed a binding band with WT-1, which

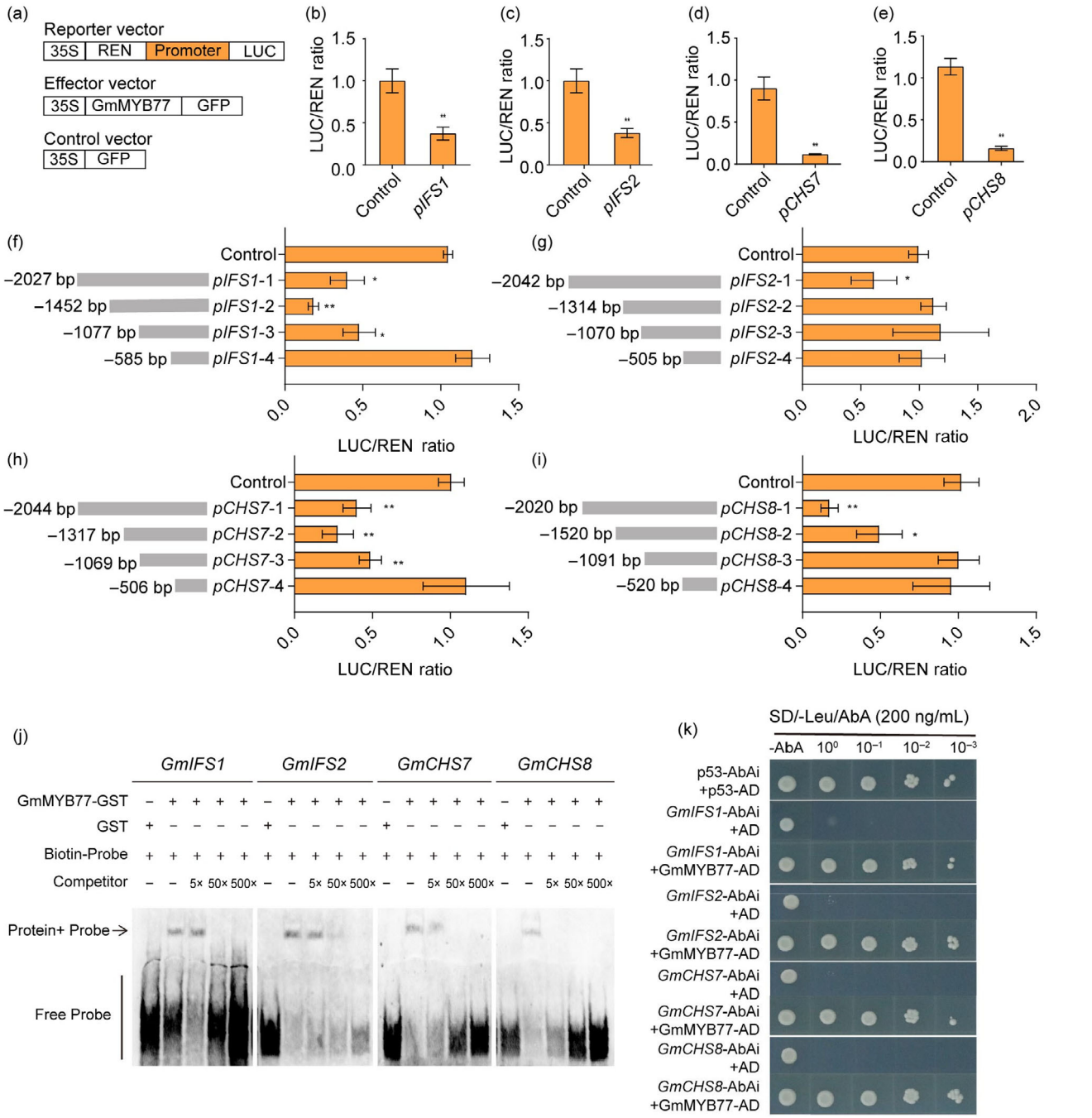


Figure 3 GmMYB77 directly binds the promoters of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* and negatively regulates their expression. (a) Diagrams showing the effector and reporter vectors for the dual-luciferase assay. (b–e) GmMYB77 repressed the promoter activity of *GmIFS1* (b), *GmIFS2* (c), *GmCHS7* (d) and *GmCHS8* (e) in transgenic tobacco leaves. (f–i) GmMYB77 binds to different promoter regions of *GmIFS1* (f), *GmIFS2* (g), *GmCHS7* (h) and *GmCHS8* (i). The numbers on the left of the grey bars represent the distance from the start codon. Data are shown as means ± SE for three biological replicates, with significant differences determined using by Student's *t*-test (*$P < 0.05$; **$P < 0.01$). (j) Electrophoretic mobility shift assay used to verify the interaction between GmMYB77 and the promoter regions of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8*. (k) Interaction between GmMYB77 and the promoter regions of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* in a yeast one-hybrid assay. Empty pGADT7 + promoter-pAbAi and pGADT7-Rec-p53 + p53-AbAi as the negative and positive controls, respectively.

disappeared following deletion or mutation to a uniform sequence, while M4 restored the binding band (Figure 4g). For the GGTAGGT sequence, GmMYB77 demonstrated binding with WT-2, M10 and M12 (Figure 4h). These findings confirm that GmMYB77 directly binds to the GGT core sequence. Taken together, our results demonstrate that GmMYB77 directly binds to GGT core sequence in the promoters of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8*, thereby repressing their expression.

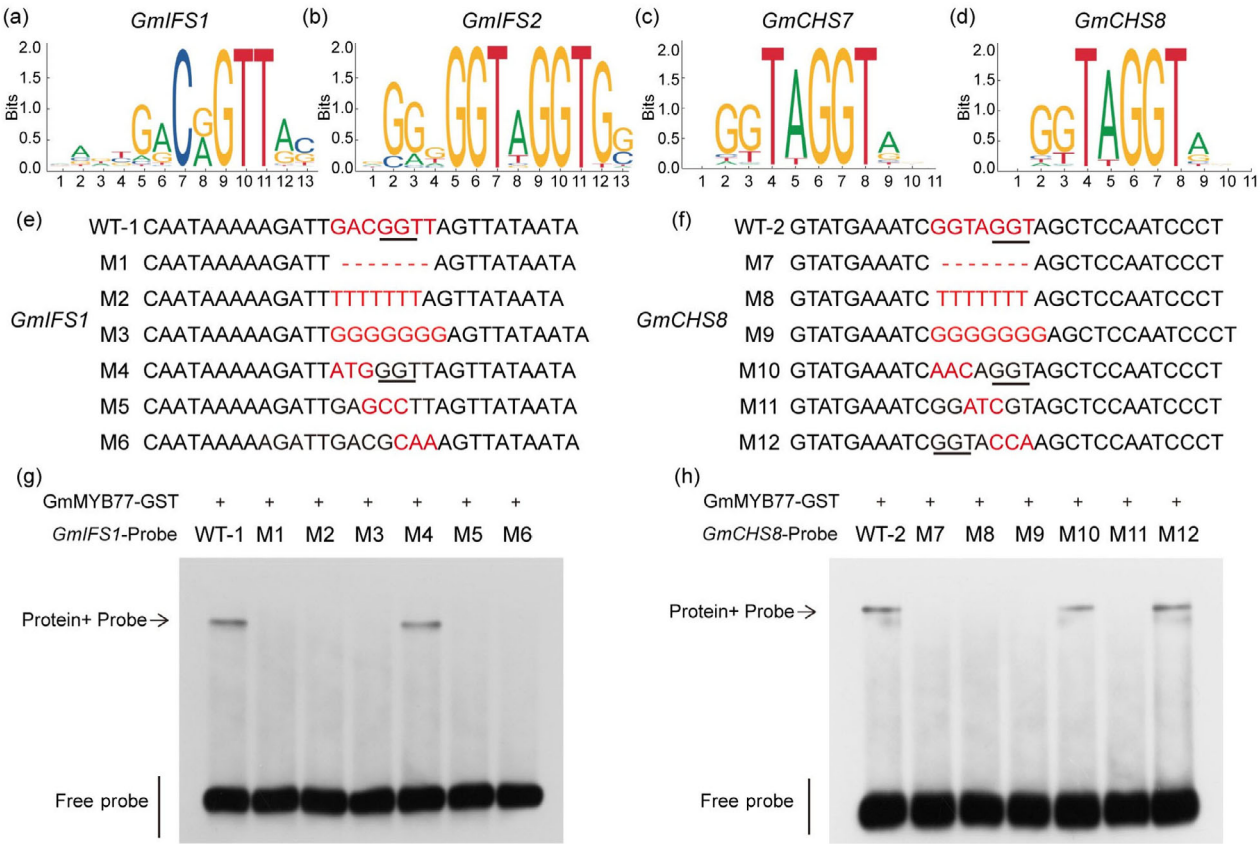


Figure 4 Electrophoretic mobility shift assay (EMSA) showing GmMYB77 binding to the GGT core sequence. Both the wild-type and mutant probes were labelled with biotin. (a–d) Sequence logos of the predicted myeloblastosis (MYB)-binding sequence in the promoter regions of *GmIFS1* (a), *GmIFS2* (b), *GmCHS7* (c) and *GmCHS8* (d) were generated using the JASPAR database. The height of each stack indicates the degree of conservation (bits), and the height of the letters represents the relative frequency of each base. (e–f) Design of the probes for the *GmIFS1* (e) and *GmCHS8* (f) promoter region, with mutated bases shown in red type. (g–h) Binding of the GmMYB77 protein to the *GmIFS1* (g) and *GmCHS8* (h) probes.

GmMYB77 negatively regulates the isoflavone content of soybean seeds

To verify the function of the *GmMYB77* gene, we used clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated nuclease 9 (Cas9) to obtain three *Gmmyb77* mutants lacking the *bar* gene in the soybean TL1 background (*Gmmyb77-1*, *Gmmyb77-2* and *Gmmyb77-3*) (Figure 5a; Figure S4). The relative expression of *GmMYB77* in these mutants was reduced by an average of 86.9% compared with TL1 (Figure 5b). The malonylglycitin content in the T_3 knockout soybean seeds increased by 28.1–58.7% relative to the TL1 seeds (Figure 5c), while the malonylgenistin content was also significantly higher (12.2–16.0% increase; Figure 5d). The *Gmmyb77* knockout seeds showed an 11.0–15.2% increase in TIF content compared with the non-transgenic seeds (Figure 5e). The contents of daidzin, glycitin, genistin and malonyldaidzin did not significantly differ between the TL1 and *Gmmyb77* seeds (Figure S5).

To further investigate the function of *GmMYB77*, we generated constitutive *GmMYB77*-OE lines in TL1 by introducing the CDS region of *GmMYB77* driven by the CaMV *35S* promoter. Three independent transformation events were confirmed via a LibertyLink strip analysis (Figure S6a). A western blot analysis demonstrated the stable accumulation of the recombinant protein in selected leaves (Figure S6b). The overexpression lines demonstrated a significant 4.1- to 7.6-fold increase in *GmMYB77* expression levels (Figure 5f). Seeds from the *GmMYB77*-OE lines had lower TIF content than the non-transgenic TL1 seeds, including reduced malonylglycitin and malonylgenistin levels (Figure 5g–i). The contents of daidzin, glycitin, genistin and malonyldaidzin were not significantly different between the TL1 and *GmMYB77*-OE seeds (Figure S7).

We also analysed the expression levels of key genes encoding enzymes in the phenylalanine metabolic pathway in the transgenic plants (Baldoni *et al.*, 2013; Feng *et al.*, 2022; Lu *et al.*, 2021). *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* were significantly down-regulated in *GmMYB77*-OE plants compared with TL1 (Figure 5j–m). By contrast, the expression levels of *GmF3H1*, *GmFNS1*, *GmFLS2* and *GmCAD* were significantly increased in overexpressing seeds compared with *Gmmyb77* mutants (Figure S8). Relative to TL1, the expression level of *GmIOMT1* increased by 400-fold in the *Gmmyb77* mutant plants, but decreased by 48.7% in the *GmMYB77*-OE plants (Figure 5n). Furthermore, the expression level of the *GmF6H* gene increased by 65.1% in the *Gmmyb77* mutant plants (Figure S8). These findings led us to speculate that GmMYB77 regulates the expression of *GmF6H* and *GmIOMT1*, significantly influencing the accumulation of malonylglycitin in soybean seeds. Taken together, our data show that *GmMYB77* negatively regulates soybean isoflavone accumulation by inhibiting the isoflavone

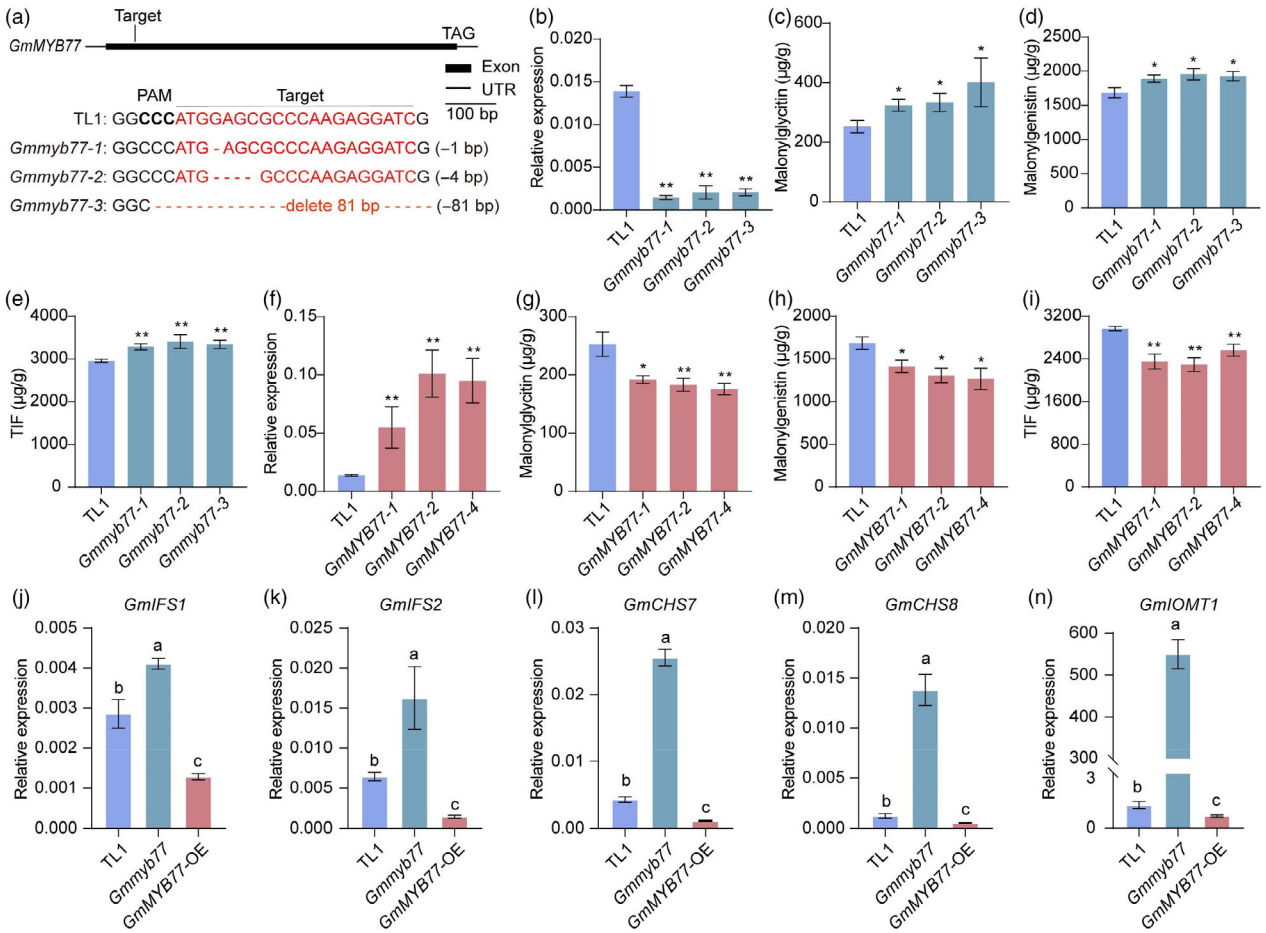


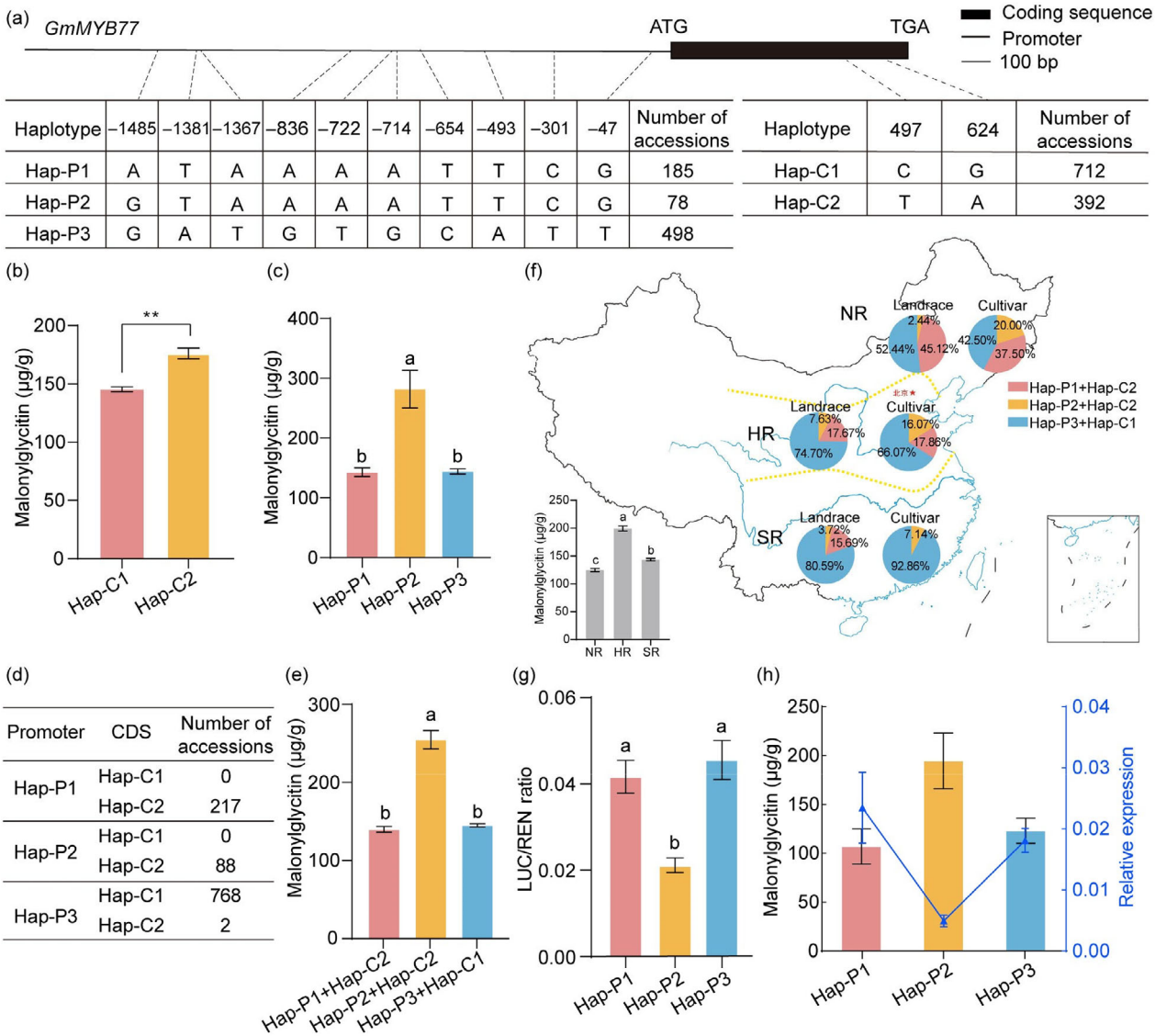
Figure 5 Investigation of isoflavone content in *Gmmyb77* mutant and GmMYB77-overexpressing (OE) soybean seeds. (a) CRISPR/Cas9-generated *Gmmyb77* mutants (TL1 background). Red dashes indicate deleted nucleotides. (b–e) Statistical analysis of relative *GmMYB77* expression level (b), malonylglycitin (c), malonylgenistin (d) and total isoflavone (TIF) contents (e) in the TL1 and *Gmmyb77* mutants. (f–i) Statistical analysis of relative *GmMYB77* expression level (f), malonylglycitin (g), malonylgenistin (h) and TIF contents (i) in the TL1 and *GmMYB77*-OE plants. Student's *t*-tests were used to calculate the *P*-values (**P* < 0.05; ***P* < 0.01). (j–n) Transcription levels of *GmIFS1* (j), *GmIFS2* (k), *GmCHS7* (l), *GmCHS8* (m) and *GmIOMT1* (n) in *GmMYB77*-OE and *GmMYB77*-RNAi hairy roots. Significant differences are indicated by different letters (ANOVA with LSD method).

synthetase genes and promoting the expression of flavonoid pathway synthetase genes. This regulation likely explains the reduced isoflavone accumulation in *GmMYB77*-OE transgenic seeds.

We also compared the key agronomic traits in the transgenic plants, such as plant height, first pod height, node number on the main stem, effective branches, seed weight per plant, seed number per plant and 100-seed weight. No significant differences in these agronomic traits were observed among the *Gmmyb77* mutants, *GmMYB77*-OE and TL1 plants (Figure S9). These results indicate that *GmMYB77* negatively regulates isoflavone content in soybean seeds without changing the agronomic characteristics of the plants.

The Hap-P2 haplotype confers high isoflavone content

Having demonstrated a role for GmMYB77 in the soybean isoflavone content, we analysed the nucleotide polymorphisms in the CDS and promoter regions of *GmMYB77* using 2214 previously resequenced accessions (Li *et al*., 2023). We identified two variations in the CDS region and 10 variations in the promoter region, which revealed two haplotypes (Hap-C1 and Hap-C2) in the coding region, and three main promoter haplotypes (Hap-P1, Hap-P2 and Hap-P3) (Figure 6a, Tables S3, S4). To evaluate the phenotypic effects of the haplotypes, we examined the malonylglycitin and TIF contents of the soybean haplotypes across six environments. Soybean accessions carrying Hap-C2 exhibited significantly higher malonylglycitin and TIF contents (176.20 and 2657.20 μg/g, respectively) than Hap-C1 (145.36 and 2430.71 μg/g, respectively) (Figure 6b, Figure S10). Additionally, soybean accessions with the Hap-P2 haplotype had significantly higher malonylglycitin and TIF contents (250.38 and 3055.95 μg/g, respectively) than Hap-P1 (143.29 and 2524.13 μg/g, respectively) and Hap-P3 (149.00 and 2461.96 μg/g, respectively), while Hap-P1 and Hap-P3 did not differ significantly from each other (Figure 6c; Figure S11a–k). Notably, the frequencies of Hap-C2 and Hap-P2 in the Huang-Huai-Hai region (HR) are higher than in the southern region (SR) and northern region (NR) of China (Figures S12a, S13a). Higher malonylglycitin and TIF contents were observed in the HR compared with the NR and SR, consistent with the finding that Hap-C2 and Hap-P2 contribute to increased malonylglycitin and TIF contents (Figures S12b,c, S13b,c).



Haplotype	−1485	−1381	−1367	−836	−722	−714	−654	−493	−301	−47	Number of accessions
Hap-P1	A	T	A	A	A	A	T	T	C	G	185
Hap-P2	G	T	A	A	A	A	T	T	C	G	78
Hap-P3	G	A	T	G	T	G	C	A	T	T	498

Haplotype	497	624	Number of accessions
Hap-C1	C	G	712
Hap-C2	T	A	392

Promoter	CDS	Number of accessions
Hap-P1	Hap-C1	0
	Hap-C2	217
Hap-P2	Hap-C1	0
	Hap-C2	88
Hap-P3	Hap-C1	768
	Hap-C2	2

Figure 6 Haplotype analysis of *GmMYB77* in soybean accessions. (a) Summary of two coding-region haplotypes and three promoter-region haplotypes for *GmMYB77*. (b, c) Malonylglycitin content variation in the *GmMYB77* haplotypes of the coding region (b) and promoter region (c) in the averaged across five environments. Student's *t*-tests were used to calculate the *P*-values (**$P < 0.01$). Different letters indicate statistically significant differences (ANOVA with LSD method). (d) The distribution of coding-region haplotypes within the promoter-region haplotypes. (e) Malonylglycitin content variation of Hap-P1 + Hap-C2 ($n = 210$ accessions), Hap-P2 + Hap-C2 ($n = 87$) and Hap-P3 + Hap-C1 ($n = 739$) in the averaged across five environments. (f) The distribution of three combined haplotypes among the landraces and cultivars from three different regions of China. NR, HR and SR are the northern, Huang-Huai-Hai and southern regions of China, respectively. Bottom-left box plot represents the malonylglycitin content comparison of accessions in the NR ($n = 162$), HR ($n = 305$) and SR ($n = 403$). (g) Promoter activity analysis of the three promoter-region haplotypes using sequences 2000-bp upstream from the translation initiation site. Data are the means ± SE. Significant differences are indicated by different letters (ANOVA with LSD method). (h) Association between the expression level of *GmMYB77* and the malonylglycitin content among 15 randomly selected Hap-P1, Hap-P2 and Hap-P3 accessions ($n = 5$ biological replicates).

To further explore the allelic effects of *GmMYB77* on the isoflavone content in soybean, we analysed the combined promoter and CDS haplotypes of *GmMYB77*. A total of 305 accessions had the Hap-P1 or Hap-P2 haplotypes in the promoter region, all of which carried the Hap-C2 CDS haplotype. The 770 accessions with the Hap-P3 haplotype predominantly carried the Hap-C1 haplotype, with only two exceptions exhibiting Hap-C2 (Figure 6d). Owing to the low frequency of the Hap-P3 + Hap-C2 combined haplotype, we focused on three combined haplotypes: Hap-P1 + Hap-C2, Hap-P2 + Hap-C2 and Hap-P3 + Hap-C1. Accessions with the Hap-P2 + Hap-C2 combined haplotype (254.82 and 3084.42 μg/g, respectively) had higher malonylglycitin and TIF contents than those with Hap-P1 + Hap-C2 (139.96 and 2476.62 μg/g, respectively) and Hap-P3 + Hap-C1 (145.13 and 2426.04 μg/g, respectively) across all environments (Figure 6e; Figure S14a–k). The geographic distributions of the three combined haplotypes were further examined in the landraces and cultivars. The elite haplotype Hap-P2 + Hap-C2 was more frequent in cultivated varieties than in landraces in all three regions (NR, 20.0% and 2.4%, respectively; HR, 16.1%

and 7.6%, respectively; SR, 7.1% and 3.7%, respectively; Figure 6f; Figure S14l), suggesting that Hap-P2 + Hap-C2 may have been selected during soybean improvement, leading to its accumulation in cultivars. Given that Hap-P2 was only ever found in combination with the Hap-C2 haplotype and that the combination of Hap-P1 with Hap-C2 did not correlate with elevated levels of malonylglycitin and TIF, it can be inferred that Hap-P2 plays a more significant role than Hap-C2 in the enhancement of these compounds.

To investigated whether natural variations in the CDS region of *GmMYB77* resulted in functional differences. First, subcellular localization experiments were conducted for two CDS haplotypes. Both CDS haplotype proteins were localized to the nucleus and cytoplasm, indicating no changes in their subcellular location (Figure S15a). Next, we introduced Hap-C1 and Hap-C2 into soybean using hairy root transformation driven by the CaMV *35S* promoter, the transgenic soybean hairy roots carrying both CDS haplotypes exhibited reduced contents of daidzin, malonylglycitin, malonyldaidzin and TIF compared with the empty vector, while there were no significant differences in the levels of genistin and malonylgenistin (Figure S15b–h). There were no significant differences between the two CDS region haplotypes with respect to the reduction in isoflavone contents. Additionally, dual-LUC assays demonstrated that both Hap-C1 and Hap-C2 inhibited the transcriptional activity of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8*, with no significant differences observed between them (Figure S15i–l). These findings indicate that natural variations in the CDS region of *GmMYB77* do not produce functional differences.

We further investigated the activity of three *GmMYB77* promoter haplotypes. Transient transcription activity assays demonstrated that the *GmMYB77* promoter sequences of Hap-P1 and Hap-P3 had significantly higher transcriptional activity than that of Hap-P2 (Figure 6g). To explore the relationship between these haplotypes and GmMYB77 expression levels, we selected five accessions for each of the three promoter haplotypes (Table S5), which were analysed to elucidate the correlation between *GmMYB77* expression and the malonylglycitin and TIF contents. Our results identified a significant negative correlation between promoter activity and isoflavone content. Accessions containing Hap-P2 exhibited the highest malonylglycitin and TIF contents, along with the lowest *GmMYB77* expression, and the highest expression levels of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* (Figure 6h; Figure S11l,m, Table S5). These findings reaffirm the pivotal role of the Hap-P2 haplotype in elevating isoflavone levels, offering an innovative direction for the genetic improvement of soybean quality traits.

Discussion

Soybean isoflavones have attracted significant interest owing to their beneficial effects on plant and human health. The isoflavone content is a quantitative trait controlled by multiple genes. In this study, we identified a novel R2R3-MYB gene, *GmMYB77*, which plays a negative role in regulating the accumulation of isoflavone content in soybean hairy roots and seeds. We further discovered that GmMYB77 directly binds to the GGT core sequence and significantly suppresses the expression of key downstream genes in the isoflavone biosynthesis pathway, including *GmIFS1*, *GmIFS2*, *GmCHS7*, *GmCHS8* and *GmIOMT1*. The Hap-P2 haplotype in the promoter region of *GmMYB77* is considered elite as it is associated with higher malonylglycitin and TIF contents. These findings deepen our understanding of the genes affecting the isoflavone content of soybean and offer a valuable genetic resource for molecular breeding efforts.

The MYB TF family, one of the largest groups of TFs, plays a crucial role in secondary metabolism in plants (Ambawat *et al*., 2013; Katiyar *et al*., 2012). Examples include SbMYB3, which regulates baicalin biosynthesis in *Scutellaria baicalensis* Georgi (Feng *et al*., 2022); PbMYB12b, which regulates flavonoid biosynthesis in *Pyrus betulifolia* Bunge (Zhai *et al*., 2019); MdMYB22, which regulates anthocyanin biosynthesis in *Malus domestica* (Suckow) Borkh (Wang *et al*., 2017); and GmMYB176, which regulates isoflavone biosynthesis in soybean (Anguraj Vadivel *et al*., 2019). Previous studies have shown that MYB TFs are vital for regulating isoflavone biosynthesis in soybean. The overexpression of MYB genes, such as *GmMYB29*, *GmMYB176*, *GmMYB133*, *GmMYB58* and *GmMYB205*, has been reported to significantly increase the isoflavone content in transgenic soybean hairy roots (Anguraj Vadivel *et al*., 2021; Bian *et al*., 2018; Chu *et al*., 2017; Han *et al*., 2017). Despite the known positive regulators of isoflavone biosynthesis, little is known about the negative regulators of this process within the MYB TF family, although *GmMYB39* and *GmMYB100* were identified as inhibitors of isoflavone accumulation (Liu *et al*., 2013; Yan *et al*., 2015). In our study, we found that GmMYB77 functions as a negative regulator of isoflavone accumulation in soybean. Whether in soybean hairy roots or transgenic soybean seeds, the loss of function of the *GmMYB77* gene (via RNAi or mutation) led to an increase in isoflavone content, while the overexpression of *GmMYB77* decreased the isoflavone content. These comprehensive results highlight the role of GmMYB77 as a negative regulator in the complex network of MYB TFs controlling isoflavone biosynthesis in soybean.

Previously, the functional characterization of the candidate genes regulating soybean isoflavones was primarily validated using transgenic hairy root systems owing to their speed and accuracy; however, the isoflavone composition in hairy roots does not completely match that of soybean seeds. In our study, we observed that silencing *GmMYB77* in soybean hairy roots led to a significant increase in the daidzin, malonyldaidzin, malonylglycitin and TIF contents, while these were decreased by the overexpression of *GmMYB77*. In transgenic soybean seeds, the knockout of *GmMYB77* led to increases in malonylglycitin and TIF contents, similar to the trends observed in hairy roots; however, the accumulation patterns of other specific isoflavone components (daidzin, malonyldaidzin and malonylgenistin) differed between soybean hairy roots and seeds. Additionally, the overexpression of *GmGLY1* in soybean hairy roots significantly increased the levels of genistin, glycitin and TIF contents, while the transgenic *GmGLY1*-OE seeds had increased glycitin and malonylglycitin levels but a decreased TIF content (Zhang *et al*., 2024). Differences in isoflavone accumulation between hairy roots and seeds underscore the importance of validating gene function in whole soybean plants, with transgenic soybean seeds offering the most reliable validation.

In this study, we found that *GmMYB77* not only regulates the TIF content but also significantly impacts the malonylglycitin levels, especially. This significant reduction in malonylglycitin content compared with the overall decrease in TIF suggests that *GmMYB77* plays a crucial role in regulating the biosynthesis of malonylglycitin. Previous studies indicated that MYB TFs regulate the accumulation of isoflavone content by controlling genes for key enzymes in the biosynthetic pathway; for example,

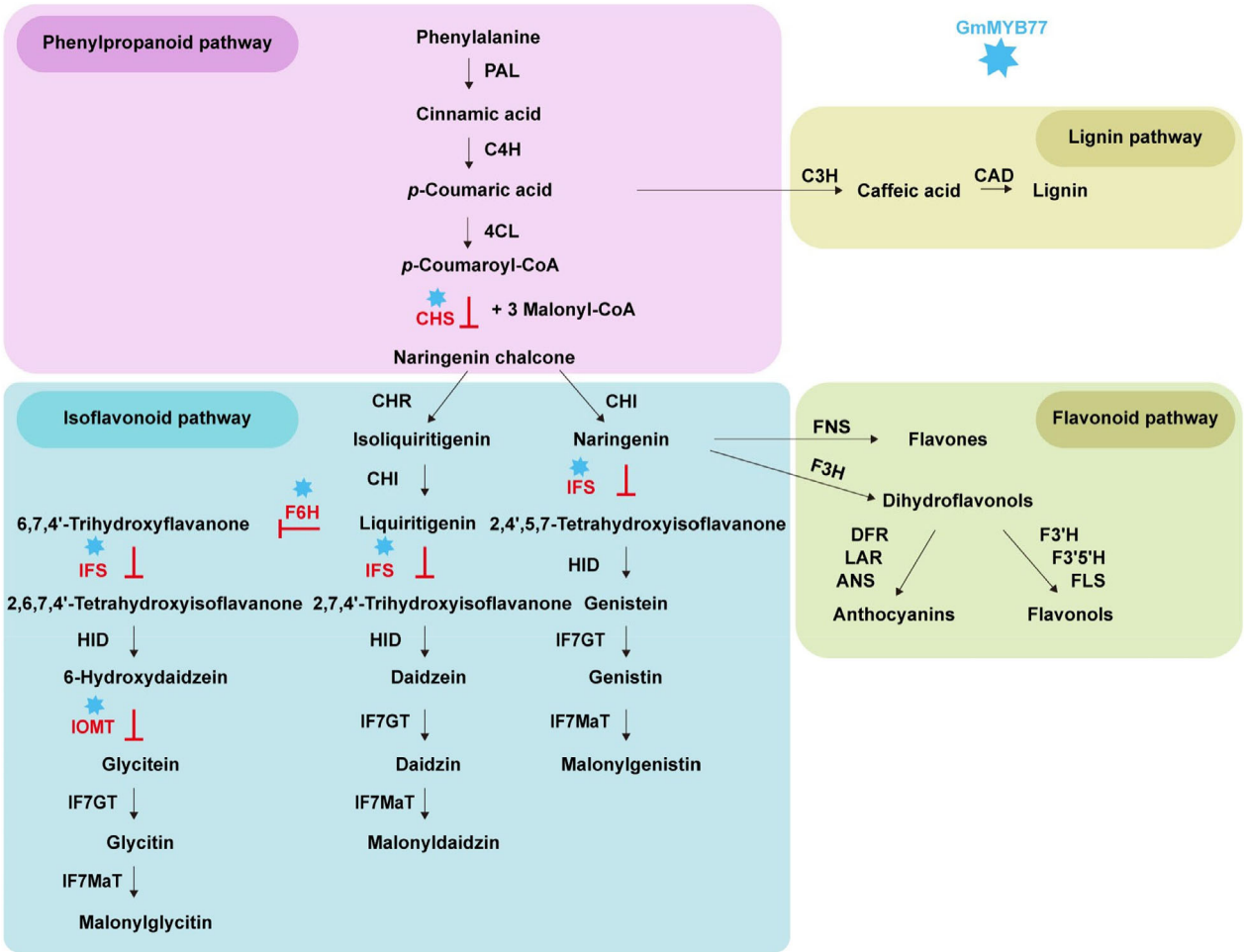


Figure 7 GmMYB77-mediated regulatory model for soybean isoflavone biosynthesis. Black arrows indicate the relationship between isoflavone-related enzymes. The red 'T' lines indicate the inhibitory effect. 4CL, 4-coumarate-CoA ligase; ANR, anthocyanidin reductase; ANS, anthocyanin synthase; C4H, cinnamic acid 4-hydroxylase; CHI, chalcone isomerase; CHR, chalcone reductase; CHS, chalcone synthase; DFR, dihydroflavonol 4-reductase; F3′5′H, flavonoid 3′5′-hydroxylase; F3′H, flavonoid 3′-hydroxylase; F3H, flavanone 3-hydroxylase; F6H, flavonoid 6-hydroxylase; FLS, flavonol synthase; FNS, flavone synthase; HID, 2-hydroxyisoflavanone; IF7GT, isoflavone 7-*O*-glucosyltransferase; IF7MaT, isoflavone 7-*O*-glucoside 6″-*O*-malonyltransferase; IFS, isoflavone synthase; IOMT, isoflavone *O*-methyltransferase; LAR, leucoanthocyanidin reductase; PAL, phenylalanine ammonia-lyase.

GmMYB100 inhibits isoflavonoid production by downregulating *GmCHS*, *GmCHI* and *GmIFS* (Yan *et al.*, 2015), while GmMYB176 activates *GmCHS8* promoter activity (Yi *et al.*, 2010). Previous studies have demonstrated that MYB TFs could bind to the MYB core sequences (*e.g.* ACTGGTAGCTATT, CCGTTG or TTGTTG) in the promoters of target genes (Espley *et al.*, 2009; Lan *et al.*, 2023; Prouse and Campbell, 2012; Zhang *et al.*, 2020). In this study, we discovered that GmMYB77 specifically binds to the GGT sequence of target genes, that is different from the binding sequences of above mentioned MYB TFs. These findings suggest that although MYB-type proteins recognize and bind target genes with their core binding sequences, these sequences vary across different proteins. In addition, we found that GmMYB77 inhibits the expression of *GmIFS1*, *GmIFS2*, *GmCHS7*, *GmCHS8* and *GmF6H1*, with particularly potent inhibition noted for *GmIOMT1* following *GmMYB77* knockout. This causes the metabolic flux of the isoflavone biosynthetic pathway to primarily shift towards malonylglycitin biosynthesis (Figure 7). Malonylglycitin, an important bitterness factor in soybean seeds (Kudou *et al.*, 1991), also plays a crucial role in plant defence mechanisms, particularly in response to pathogen infection, by acting as a key component in the biosynthesis of glycitein and its derivatives (Uchida *et al.*, 2020); therefore, it is essential to regulate malonylglycitin content in soybean purposefully.

Isoflavones play a crucial role in human health as well as in plant defence against pathogens and abiotic stressors. Consequently, identifying and accumulating elite alleles of pivotal genes involved in isoflavone biosynthesis is an important strategy in soybean breeding. Here, the haplotype analysis indicates that although variations in the *GmMYB77* coding region affect isoflavone content, they do not lead to functional differences between the two CDS haplotypes. However, differences in the promoter region have a significantly impact on isoflavone accumulation. Accessions with the Hap-P2 + Hap-C2 allele produced significantly more isoflavones than those with the Hap-P1 + Hap-C2 and Hap-P3 + Hap-C1 alleles. Designing molecular markers based on the elite haplotype of *GmMYB77* and using marker-assisted selection methods could therefore enable the rapid identification of elite germplasm with high malonylglycitin and TIF contents. Additionally, editing *GmMYB77*

using CRISPR/Cas9 can directly produce soybean materials with high malonylglycitin and TIF contents, thereby accelerating the soybean breeding process. These insights into natural variation thus provide valuable genetic resources and molecular markers for improving isoflavone content in soybean seeds through molecular breeding.

In conclusion, we found that GmMYB77 inhibits the expression of key isoflavone biosynthesis enzyme genes by directly binding to the core sequence GGT, thus controlling the isoflavone content of soybean seeds. Eliminating the dominant *GmMYB77* allele effect could accelerate soybean breeding programmes. GmMYB77 is a negative regulator, so the malonylglycitin content of soybean accessions can be increased by removing its effect through gene editing, which would improve the industrial and dietary value of soybeans. Further investigation is needed to explore the molecular function of GmMYB77 in the biosynthesis of flavonoids, anthocyanins and lignin.

Materials and methods

Plant materials and growth conditions

The soybean cultivar Tianlong 1 (TL1) and transgenic plants were grown in the field at Nankou Experimental Station, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (ICS-CAAS), Beijing, China (40°13′ N and 116°5′ E). The seeds were sown in 1.5-m rows with 0.1-m intra-row and 0.5-m inter-row spacings. Recommended agronomic practices were performed, and plants from each line were harvested upon maturity for further analysis. All seeds were stored for 1 month after harvest before their isoflavone content was determined. *Nicotiana benthamiana* plants were grown in a controlled environment at 25 °C with a 10-h light and 14-h dark photoperiod.

GWAS

A natural population comprising 1551 diverse soybean accessions was used in the GWAS analysis (Azam *et al*., 2023b). The soybean accessions were divided into two populations (925 in landraces, 501 in cultivars and 125 of unknown type) and three ecological regions (242 in NR, 437 in HR, 488 in SR and 384 of unknown type). Detailed information about the soybean accessions can be found in a previous study (Azam *et al*., 2023a). Field trials were conducted at three locations: Changping, Beijing (40°13′ N and 116°12′ E) and Sanya, Hainan (18°24′ N and 109°5′ E), in 2017 and 2018, and Hefei, Anhui (33°61′ N and 117° E), in 2017. The accessions were sown in 3-m rows with 0.5-m inter-row and 0.1-m intra-row spacings. A total number of 6 149 599 SNPs with a minor allele frequency (MAF) > 0.01 from previously sequenced soybean accessions were used for the GWAS analysis (Li *et al*., 2023), which was performed using the compressed mixed linear model (cMLM) in the GAPIT program (Azam *et al*., 2023b).

BSA-seq

For the BSA-seq, the average soybean isoflavone contents from four environments (Beijing and Hainan each in 2017 and 2018) were used. Based on the phenotypic results of all accessions, 101 samples constituting two groups with extreme (high and low) isoflavone contents were selected. Genomic DNA from the 50 high isoflavone and 51 low-isoflavone soybean accessions was pooled to create the high- and low-isoflavone pools, respectively. Both pools underwent whole-genome sequencing using the Illumina HiSeq X Ten platform (Illumina, San Diego, CA) with standard paired-end 150-bp sequencing. The annotation data included Glyma (v2.0) identifiers from the soybean genome sequence (Azam *et al*., 2023a).

Isoflavone extraction and determination

Isoflavone was extracted using a previously described method (Sun *et al*., 2011). Approximately 20 g of mature seeds from each accession was ground to a fine powder using a cyclone mill (A10 basic; IKA, Staufen, Germany). The soybean powder (0.1 g) was transferred to a 10-mL centrifuge tube containing 5 mL of 70% (v/v) ethanol and 0.1% (v/v) acetic acid. The samples were homogenized by shaking for 12 h. After centrifugation at 2700 *g* for 10 min, the supernatants were filtered through a YMC Duo-filter (YMC, Kyoto, Japan) with a 0.2-μm pore size and stored at 4 °C until required. The isoflavone content was determined using an Agilent 1260 HPLC System (Agilent Technologies, Santa Clara, CA) with a YMC ODS AM-303 column (250 mm × 4.6 mm I.D., S-5 μm, 120 A; YMC). The mobile phases consisted of 0.1% acetic acid in distilled water (phase A) and acetonitrile (phase B). The solvent flow rate was 1.0 mL/min with a 10-μL injection volume, using a 70-min linear gradient of 13–30% acetonitrile (v/v). The UV detector was set to 260 nm, and the column temperature was maintained at 35 °C. Isoflavone standards included daidzin (D), glycitin (GL), genistin (G), acetyldaidzin (AD), acetylglycitin (AGL), acetylgenistin (AG), daidzein (DE), glycitein (GLE), genistein (GE), malonyldaidzin (MD), malonylglycitin (MGL) and malonylgenistin (MG). In the current study, five isoflavone components (D, MD, MGL, G and MG) were detected in the soybean hairy roots, and seven were detected in the seeds (D, GL, G, MD, MG, MGL and AD). The TIF concentration was calculated as the sum of these components.

RNA extraction and RT-qPCR assays

The full-length sequence of *GmMYB77* was obtained using the NCBI BLAST tool (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The total RNA was isolated from TL1 leaves using the RNA Easy Fast Plant Tissue Kit (DP452; Tiangen Biotech, Beijing, China). Full-length cDNAs were reverse-transcribed using a cDNA synthesis kit (AE311; TransGen Biotech, Beijing, China) and used as templates for RT-qPCR according to the manufacturer's instructions (AQ101; TransGen Biotech). The soybean gene *GmActin6* (GenBank number: NM_001289231) was used as an internal control. The expression levels of each gene were calculated using the $2^{-\Delta\Delta Ct}$ method (Livak and Schmittgen, 2001). All primers used are listed in Table S6. Three biological and three technical replicates were applied to the RT-qPCR assays.

Expression pattern of *GmMYB77*

Total RNA was isolated from the roots, stems, leaves and cotyledons of TL1 plants at the flowering stage, and seeds at 20, 40, 60 and 75 days post-flowering. The samples were flash frozen and stored at −80 °C until use. A RT-qPCR analysis was performed to detect *GmMYB77* expression in these tissues using the primers listed in Table S6.

Subcellular localization of GmMYB77

A *GmMYB77* fragment was separately amplified from the cDNA of TL1 (*GmMYB77*), 16NF239_240 (Hap-C1) and 16NF2299_2300 (Hap-C2) using primers GmMYB77-F and GmMYB77-R (Table S6) and inserted into the pTF101-GFP vector to generate the GFP fusion protein, under the control of the CaMV *35S* promoter. A CaMV *35S::GFP* construct was used as the control. The recombinant plasmid was introduced into *N.*

benthamiana leaf cells via an *A. rhizogenes* infiltration (Liu *et al*., 2014). The GFP signals in the tobacco epidermal cells were observed and photographed using a ZEISS LSM 710 confocal microscope (Carl Zeiss, Oberkochen, Germany), using excitation and emission wavelengths of 488 nm and 495–530 nm, respectively. The nuclear localization marker was AT1G07790-RFP driven by the *35S* promoter.

Generation of transgenic plants

CRISPR/Cas9 target sequences with a G as the first base were designed using Crispr-P v.2.0 (Liu *et al*., 2017). The 20-bp DNA fragment coding for the guide RNA (gRNA) was inserted into the pCas9-AtU6-sgRNA plasmid at the *Xba*I site. A CRISPR/Cas9 expression vector was constructed with Cas9 driven by the *Arabidopsis thaliana* (L.) Heynh. *RPS5A* promoter and gRNA transcribed by the *Arabidopsis* U6 promoter (Tsutsui and Higashiyama, 2016). The resulting vector was introduced into the *A. tumefaciens* strain EHA105 via electroporation and transformed into the soybean cultivar TL1 using the cotyledonary node transformation method (Chen *et al*., 2018b). Mutations were identified by amplifying and sequencing the gRNA target site using *GmMYB77*-specific primers (Table S6).

To investigate the effect of GmMYB77 on isoflavone accumulation, a soybean hairy root system was generated with *GmMYB77* either overexpressing (GmMYB77-OE, Hap-C1 and Hap-C2) or silenced by RNAi (*GmMYB77*-RNAi) using the CaMV *35S* constitutive promoter. To produce transgenic hairy roots, TL1 was transformed following a modified protocol (Chen *et al*., 2018a). The PCR-derived CDS of *GmMYB77* was inserted into the pGUSGFPplus (pGGP) plasmid under the CaMV *35S* promoter to generate the pGGP-*GmMYB77*-OE vector for the overexpression assays. For the RNAi of *GmMYB77*, a 300-bp fragment was inserted into the pGGP plasmid. The vectors were transferred into *A. rhizogenes* using the click transformation method, and soybean cotyledon nodes were transformed, using the pGGP empty vector as a control. The primer sequences used are listed in Table S6. Transgenic hairy roots co-expressed *GFP*, and the roots with strong *GFP* expression were screened using a portable blue-green lamp (Luyor-3260; Luyor, Shanghai, China), harvested and pooled for the gene expression and isoflavone content analyses. Each transgenic hairy root represented an independent transformation event, and the pooled sample set represented a biological replicate.

For the overexpression constructs, the *GmMYB77* CDS was amplified from the TL1 cDNA using the primers listed in Table S6. The T_0 transgenic plants were screened using LibertyLink strip detection (EnviroLogix, Portland, ME).

Western blot analysis

The expression levels of *GmMYB77* in the T_3 transgenic plants were confirmed using a western blot analysis. The total proteins were prepared from young leaves of the T_3 transgenic plants using a plant protein extraction kit (CoWin Biotech, JiangSu, China). The protein samples were separated using a 12% sodium dodecyl sulphate–polyacrylamide gel electrophoresis (SDS-PAGE), and transferred onto a polyvinylidene difluoride (PVDF) membrane (Amersham Hybond; GE Healthcare, Chicago, IL). The membrane was probed with an anti-GFP mouse monoclonal antibody (HT801; TransGen Biotech) diluted 1:5000, followed by a Goat Anti-Mouse IgG (H + L) HRP Conjugate (HS201; TransGen Biotech) at a dilution of 1:3000. The protein–antibody complexes were visualized using the HRP colour development reagent (4-chloro-1-naphthol) (Bio-Rad Laboratories, Hercules, CA).

Dual-LUC reporter assay

To construct the reporter vector, the promoter fragments (approximately −2.0 to 0 kb) of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* were amplified and independently inserted into the pGreen0800-LUC vector. The resulting vectors were transformed into the *A. tumefaciens* strain EHA105 (pSoup). CaMV *35S::GmMYB77-GFP*, CaMV *35S::Hap1* and CaMV *35S::Hap2* were used as the effectors, CaMV *35S::GFP* were used as control. The reporter and effector plasmids were transiently co-transformed into tobacco leaves, as described previously (Feng *et al*., 2022). The LUC and REN activities were quantified using the Dual-Luciferase Reporter Assay System (E1910; Promega, Madison, WI), and the relative promoter activity was calculated as the LUC to REN ratio from three biological replicates.

In vitro EMSA

The *GmMYB77* target probe was designed using JASPAR (http://jaspar.genereg.net/) based on the dual-LUC complementation assay results. The full-length *GmMYB77* CDS was cloned into the pGEX-4T expression vector containing the glutathione-S-transferase (GST) protein and transferred into the *Escherichia coli* strain Rosetta. The fusion protein was induced at 16 °C using 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG), purified using protein purification resins and eluted with glutathione elution buffer. The EMSA primer probes were labelled with biotin at the 5′ end, and unlabeled primers served as competitor probes. The EMSA was carried out following the LightShift Chemiluminescent EMSA Kit manual (20148; Thermo Fisher Scientific, Waltham, MA). All primer sequences used for the EMSA are listed in Table S6.

Y1H assay

To evaluate the interaction between GmMYB77 and the *cis*-elements of the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters, the full-length CDS of *GmMYB77* was amplified and inserted into pGADT7 to construct the effector vector GmMYB77-AD. DNA fragments of the *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* promoters were cloned into the pAbAi vector to construct the baits (*GmIFS1*-pAbAi, *GmIFS2*-pAbAi, *GmCHS7*-pAbAi and *GmCHS8*-pAbAi). The bait-reporter yeast strain was generated through the homologous integration of the bait into the Y1HGold genome (ZK288; Zoman Biotechnology, Beijing, China). The effector was transformed into the bait-reporter strain to determine the DNA–protein interactions. The transformed cells were cultured on SD/-Leu/-Ura medium supplemented with AbA and incubated for 3 days at 30 °C. The yeast cells were diluted (1:10, 1:100 and 1:1000) and plated on SD/-Leu/-Ura medium containing 200 ng/mL AbA. Empty pGADT7 + promoter-pAbAi and pGADT7-Rec-p53 + p53-AbAi as the negative and positive controls, respectively. The primers used in this assay are listed in Table S6.

Haplotype analysis of *GmMYB77*

The *GmMYB77* gene sequences of 2214 soybean accessions were compiled from SoyFGB (https://sfgb.rmbreeding.cn/index) and used to analyse the genetic diversity of this locus. SNPs in the promoter and coding regions were filtered using an MAF > 0.05 cutoff. Owing to the high rate of missing genotypes, a total of 761 soybean accessions were used for the promoter-region

haplotype analysis, and 1104 soybean accessions were used for the coding-region haplotype analysis (Tables S3, S4). In this study, the six environments used to analyse the isoflavone content were Hainan, 2017, Hainan, 2018, Beijing, 2017, Beijing, 2018, Anhui, 2017, and the mean environment of the five environments.

Acknowledgements

This study was financially supported by the National Natural Science Foundation of China (32472193, 32272178, 32161143033 and 32001574) and the Agricultural Science and Technology Innovation Program of CAAS (2060302-2).

Author contributions

Y.L. performed the experiments. Y.L., S.Z. and J.L. wrote the manuscript. B.L. and J.S. designed and supervised this study. Y.L., S.Z., J.L. and Azam. M. performed the GWAS and BSA-seq analysis. Y.L., S.Z., J.L., Y.F., J.Q., D.S. and Y.H. performed the data analysis and visualization. All authors read and approved the final manuscript.

Data availability statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

Agyenim-Boateng, G.K., Zhang, S., Islam, S., Gu, Y., Li, B., Azam, M., Abdelghany, A.M. *et al*. (2022) Profiling of naturally occurring folates in a diverse soybean germplasm by HPLC-MS/MS. *Food Chem.* **384**, 132520.

Akashi, T., Aoki, T. and Ayabe, S.I. (1999) Cloning and functional expression of a cytochrome p450 cDNA encoding 2-hydroxyisoflavanone synthase involved in biosynthesis of the isoflavonoid skeleton in licorice. *Plant Physiol.* **121**, 821–828.

Al-Nakkash, L. and Kubinski, A. (2020) Soy isoflavones and gastrointestinal health. *Curr Nutr Rep.* **9**, 193–201.

Ambawat, S., Sharma, P., Yadav, N.R. and Yadav, R.C. (2013) MYB transcription factor genes as regulators for plant responses: an overview. *Physiol. Mol. Biol. Plants* **19**, 307–321.

Anguraj Vadivel, A.K., McDowell, T., Renaud, J.B. and Dhaubhadel, S. (2021) A combinatorial action of GmMYB176 and GmbZIP5 controls isoflavonoid biosynthesis in soybean (*Glycine max*). *Commun Biol.* **4**, 356.

Anguraj Vadivel, A.K., Renaud, J., Kagale, S. and Dhaubhadel, S. (2019) GmMYB176 regulates multiple steps in isoflavonoid biosynthesis in soybean. *Front. Plant Sci.* **10**, 562.

Aoki, T., Akashi, T. and Ayabe, S.I. (2000) Flavonoids of leguminous plants: structure, biological activity, and biosynthesis. *J. Plant Res.* **113**, 475–488.

Azam, M., Zhang, S., Abdelghany, A.M., Shaibu, A.S., Feng, Y., Li, Y., Tian, Y. *et al*. (2020) Seed isoflavone profiling of 1168 soybean accessions from major growing ecoregions in China. *Food Res. Int.* **130**, 108957.

Azam, M., Zhang, S., Huai, Y., Abdelghany, A.M., Shaibu, A.S., Qi, J., Feng, Y. *et al*. (2023a) Identification of genes for seed isoflavones based on bulk segregant analysis sequencing in soybean natural population. *Theor. Appl. Genet.* **136**, 13.

Azam, M., Zhang, S., Li, J., Ahsan, M., Agyenim-Boateng, K.G., Qi, J., Feng, Y. *et al*. (2023b) Identification of hub genes regulating isoflavone accumulation in soybean seeds *via* GWAS and WGCNA approaches. *Front. Plant Sci.* **14**, 1120498.

Baldoni, A., Von Pinho, E.V., Fernandes, J.S., Abreu, V.M. and Carvalho, M.L. (2013) Gene expression in the lignin biosynthesis pathway during soybean seed development. *Genet. Mol. Res.* **12**, 2618–2624.

Bian, S., Li, R., Xia, S., Liu, Y., Jin, D., Xie, X., Dhaubhadel, S. *et al*. (2018) Soybean *CCA1-like MYB* transcription factor *GmMYB133* modulates isoflavonoid biosynthesis. *Biochem. Biophys. Res. Commun.* **507**, 324–329.

Cesco, S., Mimmo, T., Tonon, G., Tornasi, N., Pinton, R., Terzano, R., Neumann, G. *et al*. (2012) Plant-borne flavonoids released into the rhizosphere: impact on soil bio-activities related to plant nutrition. A review. *Biol Fertil Soils* **48**, 123–149.

Chen, L., Cai, Y., Liu, X., Guo, C., Sun, S., Wu, C., Jiang, B. *et al*. (2018a) Soybean hairy roots produced *in vitro* by *Agrobacterium rhizogenes*-mediated transformation. *Crop J.* **6**, 162–171.

Chen, L., Cai, Y., Liu, X., Yao, W., Guo, C., Sun, S., Wu, C. *et al*. (2018b) Improvement of soybean *Agrobacterium*-mediated transformation efficiency by adding glutamine and asparagine into the culture media. *Int. J. Mol. Sci.* **19**, 3039.

Chu, S., Wang, J., Zhu, Y., Liu, S., Zhou, X., Zhang, H., Wang, C. *et al*. (2017) An R_2R_3-type MYB transcription factor, GmMYB29, regulates isoflavone biosynthesis in soybean. *PLoS Genet.* **13**, e1006770.

Dhaubhadel, S., Mcgarvey, B.D., Williams, R. and Gijzen, M. (2003) Isoflavonoid biosynthesis and accumulation in developing soybean seeds. *Plant Mol. Biol.* **53**, 733–743.

Dixon, R.A., Achnine, L., Kota, P., Liu, C.J., Reddy, M.S. and Wang, L. (2002) The phenylpropanoid pathway and plant defence-a genomics perspective. *Mol. Plant Pathol.* **3**, 371–390.

Dixon, R.A. and Steele, C.L. (1999) Flavonoids and isoflavonoids – a gold mine for metabolic engineering. *Trends Plant Sci.* **4**, 394–400.

Dong, N. and Lin, H. (2021) Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. *J. Integr. Plant Biol.* **63**, 180–209.

Espley, R.V., Brendolise, C., Chagné, D., Kutty-Amma, S., Green, S., Volz, R., Putterill, J. *et al*. (2009) Multiple repeats of a promoter segment causes transcription factor autoregulation in red apples. *Plant Cell* **21**, 168–183.

Feng, Y., Zhang, S., Li, J., Pei, R., Tian, L., Qi, J., Azam, M. *et al*. (2022) Dual-function C2H2-type zinc-finger transcription factor GmZFP7 contributes to isoflavone accumulation in soybean. *New Phytol.* **237**, 1794–1809.

Guang, C., Chen, J., Sang, S. and Cheng, S. (2014) Biological functionality of soyasaponins and soyasapogenols. *J. Agric. Food Chem.* **33**, 8247–8255.

Han, X., Yin, Q., Liu, J., Jiang, W., Di, S. and Pang, Y. (2017) *GmMYB58* and *GmMYB205* are seed-specific activators for isoflavonoid biosynthesis in *Glycine max*. *Plant Cell Rep.* **36**, 1889–1902.

Jung, W., Yu, O., Lau, S.M., O'Keefe, D.P., Odell, J., Fader, G. and McGonigle, B. (2000) Identification and expression of isoflavone synthase, the key enzyme for biosynthesis of isoflavones in legumes. *Nat. Biotechnol.* **18**, 208–212.

Katiyar, A., Smita, S., Lenka, S.K., Rajwanshi, R., Chinnusamy, V. and Bansal, K.C. (2012) Genome-wide classification and expression analysis of MYB transcription factor families in rice and *Arabidopsis*. *BMC Genomics* **13**, 544.

Kudou, S., Fleury, Y., Welti, D., Magnolato, D., Uchida, T., Kitamura, K. and Okubo, K. (1991) Malonyl isoflavone glycosides in soybean seeds (*Glycine max* Merrill). *Agric. Biol. Chem.* **55**, 2227–2233.

Laddha, A.P. and Kulkarni, Y.A. (2023) Pharmacokinetics, pharmacodynamics, toxicity, and formulations of daidzein: an important isoflavone. *Phytother. Res.* **37**, 2578–2604.

LaFountain, A.M. and Yuan, Y.W. (2021) Repressors of anthocyanin biosynthesis. *New Phytol.* **231**, 933–949.

Lan, Y., Zhang, K., Wang, L., Liang, X., Liu, H., Zhang, X., Jiang, N. *et al*. (2023) The R2R3-MYB transcription factor OfMYB21 positively regulates linalool biosynthesis in *Osmanthus fragrans* flowers. *Int. J. Biol. Macromol.* **249**, 126099.

Li, Y., Qin, C., Wang, L., Jiao, C., Hong, H., Tian, Y., Li, Y. *et al*. (2023) Genome-wide signatures of the geographic expansion and breeding of soybean. *Sci. China Life Sci.* **66**, 350–365.

Liu, H., Ding, Y., Zhou, Y., Jin, W., Xie, K. and Chen, L. (2017) CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant* **6**, 530–532.

Liu, T., Song, T., Zhang, X., Yuan, H., Su, L., Li, W., Xu, J. *et al*. (2014) Unconventionally secreted effectors of two filamentous pathogens target plant salicylate biosynthesis. *Nat. Commun.* **5**, 4686.

Liu, X., Yuan, L., Xu, L., Xu, Z., Huang, Y., He, X., Ma, H. *et al*. (2013) Overexpression of *GmMYB39* leads to an inhibition of the isoflavonoid biosynthesis in soybean (*Glycine max*. L). *Plant Biotechnol Rep.* **7**, 445–455.

Livak, K.J. and Schmittgen, T.D. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408.

Lozovaya, V.V., Lygin, A.V., Zernova, O.V., Li, S., Hartman, G.L. and Widholm, J.M. (2004) Isoflavonoid accumulation in soybean hairy roots upon treatment with *Fusarium solani*. *Plant Physiol. Biochem.* **42**, 671–679.

Lu, N., Rao, X., Li, Y., Jun, J.H. and Dixon, R.A. (2021) Dissecting the transcriptional regulation of proanthocyanidin and anthocyanin biosynthesis in soybean (*Glycine max*). *Plant Biotechnol. J.* **19**, 1429–1442.

Nestel, P. (2003) Isoflavones: their effects on cardiovascular risk and functions. *Curr. Opin. Lipidol.* **1262**, 317–319.

Piovesan, A.C., Soares Júnior, J.M., Mosquette, R., Simões, M.D.J., Simões, R.D.S. and Baracat, E.C. (2005) Morphological and molecular effects of isoflavone and estrogens on the rat mammary gland. *Bras. Ginecol.* **27**, 204–209.

Prouse, M.B. and Campbell, M.M. (2012) The interaction between MYB proteins and their target DNA binding sites. *Biochim. Biophys. Acta* **1819**, 67–77.

Qi, J., Zhang, S., Azam, M., Shaibu, A.S., Abdelghany, A.M., Feng, Y., Huai, Y. *et al*. (2022) Profiling seed soluble sugar compositions in 1164 Chinese soybean accessions from major growing ecoregions. *Crop J.* **10**, 1825–1831.

Ralston, L., Subramanian, S., Matsuno, M. and Yu, O. (2005) Partial reconstruction of flavonoid and isoflavonoid biosynthesis in yeast using soybean type I and type II chalcone isomerases. *Plant Physiol.* **137**, 1375–1388.

Song, Z., Zhao, F., Chu, L., Lin, H., Xiao, Y., Fang, Z., Wang, X. *et al*. (2024) GmSTF1/2-GmBBX4 negative feedback loop acts downstream of blue-light photoreceptors to regulate isoflavonoid biosynthesis in soybean. *Plant Commun.* **12**, 100730.

Stracke, R., Werber, M. and Weisshaar, B. (2001) The *R2R3-MYB* gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **4**, 447–456.

Sun, J., Sun, B., Han, F., Yan, S., Yang, H. and Akio, K. (2011) Rapid HPLC method for determination of 12 isoflavone components in soybean seeds. *Agric Sci China.* **10**, 70–77.

Tohge, T., de Souza, L.P. and Fernie, A.R. (2017) Current understanding of the pathways of flavonoid biosynthesis in model and crop plants. *J. Exp. Bot.* **68**, 4013–4028.

Tsutsui, H. and Higashiyama, T. (2016) pKAMA-ITACHI vectors for highly efficient CRISPR/Cas9-mediated gene knockout in *Arabidopsis thaliana*. *Plant Cell Physiol.* **58**, 46–56.

Uchida, K., Sawada, Y., Ochiai, K., Sato, M., Inaba, J. and Hirai, M.Y. (2020) Identification of a unique type of isoflavone *O*-methyltransferase, GmIOMT1, based on multi-omics analysis of soybean under biotic stress. *Plant Cell Physiol.* **61**, 1974–1985.

Wang, N., Xu, H., Jiang, S., Zhang, Z., Lu, N., Qiu, H., Qu, C. *et al*. (2017) MYB12 and MYB22 play essential roles in proanthocyanidin and flavonol synthesis in redfleshed apple (*Malus sieversii* f. *niedzwetzkyana*). *Plant J.* **90**, 276–292.

Winkel-Shirley, B. (2001) Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiol.* **126**, 485–493.

Yan, J., Wang, B., Zhong, Y., Yao, L., Cheng, L. and Wu, T. (2015) The soybean R2R3 MYB transcription factor *GmMYB100* negatively regulates plant flavonoid biosynthesis. *Plant Mol. Biol.* **89**, 35–48.

Yi, J., Derynck, M.R., Li, X., Telmer, P., Marsolais, F. and Dhaubhadel, S. (2010) A single-repeat MYB transcription factor, GmMYB176, regulates *CHS8* gene expression and affects isoflavonoid biosynthesis in soybean. *Plant J.* **62**, 1019–1034.

Yu, O., Shi, J., Hession, A.O., Maxwell, C.A., McGonigle, B. and Odell, J.T. (2003) Metabolic engineering to increase isoflavone biosynthesis in soybean seed. *Phytochemistry* **63**, 753–763.

Zhai, R., Zhao, Y., Wu, M., Yang, J., Li, X., Liu, H., Wu, T. *et al*. (2019) The MYB transcription factor PbMYB12b positively regulates flavonol biosynthesis in pear fruit. *BMC Plant Biol.* **19**, 85.

Zhang, L., Li, L., Jiao, M., Wu, D., Wu, K., Li, X., Zhu, G. *et al*. (2012) Genistein inhibits the stemness properties of prostate cancer cells through targeting Hedgehog-Gli1 pathway. *Cancer Lett.* **323**, 48–57.

Zhang, P., Liu, X., Yu, X., Wang, F., Long, J., Shen, W., Jiang, D. *et al*. (2020) The MYB transcription factor CiMYB42 regulates limonoids biosynthesis in citrus. *BMC Plant Biol.* **20**, 254.

Zhang, P., Yang, C., Wang, J., Jiang, P., Qi, J., Hou, W., Cheng, H. *et al*. (2024) Cytochrome *GmGLY1* is involved in the biosynthesis of glycitein in soybean. *J. Agric. Food Chem.* **72**, 10944–10957.

Supporting information

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

Figure S1 Characteristics of the candidate gene *GmMYB77*. Characteristics of candidate gene *GmMYB77*. (a) Sequence alignment of GmMYB77 and its homologues in other plant species. (b) Tissue-specific expression analysis of *GmMYB77* in cultivar TL1. *GmActin* was used as an internal control with three biological replicates per samples ($n = 3$). (c) Subcellular localization of GmMYB77. The CaMV *35S::GFP* and CaMV *35S::GmMYB77* constructs were transformed into tobacco epidermal cells. A CaMV *35S::GFP* was used as a control. Scale bars, 30 μm.

Figure S2 Isoflavone content analysis in *GmMYB77*-overexpressing (OE) and *GmMYB77*-RNAi soybean hairy roots. Isoflavone content analysis in *GmMYB77*-overexpressing (OE) and *GmMYB77*-RNAi soybean hairy roots. (a, b) Analysis of genistin (a) and malonylgenistin content (b) in *GmMYB77*-OE soybean hairy roots. (c, d) Analysis of genistin (c) and malonylgenistin content (d) in *GmMYB77*-RNAi soybean hairy roots. Data are shown as means ± SE. Student's *t*-tests were used to calculate the *P*-values.

Figure S3 Relative expression levels of isoflavone biosynthesis–related enzyme genes in *GmMYB77*-overexpressing (OE) and *GmMYB77*-RNAi soybean hairy roots. Relative expression levels of isoflavone biosynthesis–related enzyme genes in *GmMYB77*-overexpressing (OE) and *GmMYB77*-RNAi soybean hairy roots. Different letters indicate statistically significant differences (ANOVA with LSD method).

Figure S4 Detection of transgenic soybean plants. Detection of transgenic soybean plants. (a) PCR analysis of *bar* gene in transgenic soybean plants, with wild-type DNA (TL1) as the negative control and our constructed plasmid (+) as the positive control. (b) Detection of the *bar* gene by LibertyLink strip analysis.

Figure S5 Isoflavone content of the *Gmmyb77* mutant in soybean seeds. Isoflavone content of the *Gmmyb77* mutant in soybean seeds. (a–e) Statistical analysis of daidzin (a), glycitin (b), genistin (c), malonyldaidzin (d) and acetyldaidzin content (e) in seeds of TL1 and *Gmmyb77* mutant. Data are shown as means ± SE. Student's *t*-tests were used to calculate the *P*-values.

Figure S6 Overexpression of *GmMYB77* in soybean. Overexpression of *GmMYB77* in soybean. (a) LibertyLink strip analysis results of *GmMYB77*-overexpressing soybean plants. (b) Western bolt analysis of the T_3 transgenic plants. M, protein marker; GFP, pTF101 empty vector.

Figure S7 Isoflavone content of *GmMYB77*-overexpressing (OE) soybean seeds. Isoflavone content of *GmMYB77*-overexpressing (OE) soybean seeds. (a–e) Statistical analysis of daidzin (a), glycitin (b), genistin (c), malonyldaidzin (d) and acetyldaidzin content (e) in TL1 and *GmMYB77*-overexpressing seeds. Data are shown as means ± SE. Student's *t*-tests were used to identify statistically significant differences.

Figure S8 Relative expression levels of isoflavone biosynthesis–related enzyme genes in *Gmmyb77* mutant and *GmMYB77*-

overexpressing (OE) soybean seeds. Relative expression levels of isoflavone biosynthesis–related enzyme genes in *Gmmyb77* mutant and *GmMYB77*-overexpressing (OE) soybean seeds. Data are shown as means ± SE. Different letters indicate statistically significant differences (ANOVA with LSD method).

Figure S9 Evaluation of the main agronomic traits in TL1, *Gmmyb77* mutant and *GmMYB77*-overexpressing (OE) plants. Evaluation of the main agronomic traits in TL1, *Gmmyb77* mutant and *GmMYB77*-overexpressing (OE) plants. (a) Representative photographs of TL1, *Gmmyb77* mutants and *GmMYB77*-OE plants at the maturity stage under natural conditions. The plants were sown in early June and harvested in October in Beijing. Scale bar = 10 cm. (b–h) Statistical analysis of plant height (b), first pod height (c), node number on the main stem (d), effective branches (e), seed weight of per plant (f), seed number of per plant (g) and 100-seed weight (h). Data are shown as means ± SE. ANOVA with LSD method was used to identify statistically significant differences.

Figure S10 Malonylglycitin and total isoflavone contents associated with two *GmMYB77* coding-region haplotypes. Malonylglycitin and total isoflavone contents associated with two *GmMYB77* coding-region haplotypes. (a–e) The malonylglycitin variation in the *GmMYB77* coding-region haplotypes across five environments. (f–k) TIF variation in the *GmMYB77* coding-region haplotypes across six environments. Data are shown as means ± SE. Student's *t*-tests were used to identify statistically significant differences (**$P < 0.01$).

Figure S11 Malonylglycitin and total isoflavone contents associated with three promoter-region *GmMYB77* haplotypes. Malonylglycitin and total isoflavone contents associated with three promoter-region *GmMYB77* haplotypes. (a–e) Malonylglycitin variation in the *GmMYB77* promoter-region haplotypes across five environments. (f–k) TIF variation in the *GmMYB77* promoter-region haplotypes across six environments. (l) Correlation of *GmMYB77* expression levels with TIF in 15 randomly selected Hap-P1, Hap-P2 and Hap-P3 accessions ($n = 5$ biological replicates). (m) Expression levels of *GmIFS1*, *GmIFS2*, *GmCHS7* and *GmCHS8* in 15 randomly selected Hap-P1, Hap-P2 and Hap-P3 accessions ($n = 5$ biological replicates). Data are shown as means ± SE. Significant differences indicated by lowercase letters (ANOVA with LSD method).

Figure S12 Geographic distribution of different *GmMYB77* coding-region haplotypes. Geographic distribution of different *GmMYB77* coding-region haplotypes. (a) Geographic distribution of *GmMYB77* coding-region haplotypes in three different ecoregions. NR, HR and SR are the northern, Huang-Huai-Hai and southern regions of China, respectively. (b, c) The malonylglycitin (b) and total isoflavone (c) contents comparison of accession in NR ($n = 168$), HR ($n = 320$) and SR ($n = 395$). Data are shown as means ± SE. Significant differences indicated by lowercase letters (ANOVA with LSD method).

Figure S13 Geographic distribution of different *GmMYB77* promoter-region haplotypes. Geographic distribution of different *GmMYB77* promoter-region haplotypes. (a) Geographic distribution of *GmMYB77* promoter-region haplotypes in three different ecoregions. NR, HR and SR are the northern, Huang-Huai-Hai and southern regions of China, respectively. (b, c) The malonylglycitin (b) and total isoflavone (c) contents comparison of accession in NR ($n = 65$), HR ($n = 158$) and SR ($n = 128$). Data are shown as means ± SE. Significant differences indicated by lowercase letters (ANOVA with LSD method).

Figure S14 Malonylglycitin and total isoflavone contents associated three combined haplotypes of *GmMYB77*. Malonylglycitin and total isoflavone contents associated with the three combined haplotypes of *GmMYB77*. (a–e) The malonylglycitin variation in the three combined haplotypes of *GmMYB77* across five environments. (f–k) The total isoflavone content variation in the three combined haplotypes of *GmMYB77* across six environments. (l) The total isoflavone content comparison of accessions in the NR ($n = 162$), HR ($n = 305$) and SR ($n = 403$). Data are shown as means ± SE. ANOVA with LSD method was used to calculate significant differences.

Figure S15 Functional analysis of two haplotypes in the coding region of *GmMYB77*. Functional analysis of two *GmMYB77* coding-region haplotypes in soybean. (a) Subcellular localization of two haplotypes in the coding region of GmMYB77. The CaMV *35S::GFP*, CaMV *35S::Hap-C1* and CaMV *35S::Hap-C2* constructs were transformed into tobacco epidermal cells. CaMV *35S::GFP* was used as a control. Scale bars, 20 μm. (b–h) Statistical analysis of relative *GmMYB77* expression level (b), daidzin (c), malonyldaidzin (d), malonylglycitin (e), total isoflavone (f), genistin (g) and malonylgenistin contents (h) in soybean hairy roots expressing Hap-C1 or Hap-C2 haplotypes compared with the empty vector control. (i–l) Both haplotypes of the GmMYB77 coding-region successfully repressed the promoter activities of *GmIFS1* (i), *GmIFS2* (j), *GmCHS7* (k) and *GmCHS8* (l) in transgenic tobacco leaves. Data are shown as means ± SE. Promoter activities were assessed by dual-luciferase assay, and significant differences are indicated by different lowercase letters (ANOVA with LSD method).

Table S1 The candidate genes related to isoflavone content identified using bulk segregant analysis sequencing (BSA-seq).

Table S2 List of candidate genes involved in the regulation of malonylglycitin content located on chromosome 4.

Table S3 The 1104 accessions carrying one of the two *GmMYB77* coding-region haplotypes.

Table S4 The 761 accessions carrying one of the three *GmMYB77* promoter-region haplotypes.

Table S5 Malonylglycitin and total isoflavone contents during the R8 growth stage in different accessions.

Table S6 Primers used in this study.