New Phytologist

Research

Dual-function C2H2-type zinc-finger transcription factor GmZFP7 contributes to isoflavone accumulation in soybean

Yue Feng[1,2]*, Shengrui Zhang[1]*, Jing Li[1]*, Ruili Pei[1], Ling Tian[1], Jie Qi[1], Muhammad Azam[1], Kwadwo Gyapong Agyenim-Boateng[1], Abdulwahab S. Shaibu[1], Yitian Liu[1], Zuofeng Zhu[2], Bin Li[1] and Junming Sun[1]

[1]The National Engineering Research Center for Crop Molecular Breeding, MARA Key Laboratory of Soybean Biology (Beijing), Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, 12 Zhongguancun South Street, Beijing 100081, China; [2]State Key Laboratory of Plant Physiology and Biochemistry, China Agricultural University, 2 Yuanmingyuan West Road, Beijing 100193, China

Authors for correspondence:
Bin Li
Email: libin02@caas.cn

Junming Sun
Email: sunjunming@caas.cn

Received: *28 May 2022*
Accepted: *5 November 2022*

New Phytologist (2022)
doi: 10.1111/nph.18610

Key words: C2H2 zinc-finger transcription factor, dual function, *GmZFP7*, haplotype, isoflavones, soybean [*Glycine max* (L.) Merrill].

Summary

- Isoflavones are a class of secondary metabolites produced by legumes and play important roles in human health and plant stress tolerance. The C2H2 zinc-finger transcription factor (TF) functions in plant stress tolerance, but little is known about its function in isoflavone regulation in soybean (*Glycine max*).
- Here, we report a C2H2 zinc-finger TF gene, *GmZFP7*, which regulates isoflavone accumulation in soybean. Overexpressing *GmZFP7* increased the isoflavone concentration in both transgenic hairy roots and plants. By contrast, silencing *GmZFP7* expression significantly reduced isoflavone levels. Metabolomic and qRT-PCR analysis revealed that GmZFP7 can increase the flux of the phenylpropanoid pathway. Furthermore, dual-luciferase and electrophoretic mobility shift assays showed that GmZFP7 regulates isoflavone accumulation by influencing the expression of *Isoflavone synthase 2* (*GmIFS2*) and *Flavanone 3 β-hydroxylase 1* (*GmF3H1*).
- In this study, we demonstrate that GmZFP7 contributes to isoflavone accumulation by regulating the expression of the gateway enzymes (GmIFS2 and GmF3H1) of competing phenylpropanoid pathway branches to direct the metabolic flux into isoflavone. A haplotype analysis indicated that important natural variations were present in *GmZFP7* promoters, with P-Hap1 and P-Hap3 being the elite haplotypes.
- Our findings provide insight into how GmZFP7 regulates the phenylpropanoid pathway and enhances soybean isoflavone content.

Introduction

Isoflavones are important plant secondary metabolites, mainly biosynthesized by legumes. Plants use isoflavones as signaling molecules to induce the formation of nitrogen-fixing root nodules (Zanetti *et al.*, 2010), in addition to their important roles in the defense response against pathogens and the tolerance of some abiotic stresses (Dixon *et al.*, 2002). When consumed by humans, isoflavones also play important roles in the prevention and treatment of cancer (Li *et al.*, 2005), cardiovascular disease (Hsieh *et al.*, 2020), and gynecological diseases (Ollberding *et al.*, 2012), with isoflavones acting as natural selective estrogen receptor modulators (Yang *et al.*, 2013). Due to their essential functions, isoflavones have attracted increasing research attention.

Three main isoflavone aglycones in soybean [*Glycine max* (L.) Merrill] are genistein (4′,5,7-trihydroxyisoflavone), daidzein (4′,7-dihydroxyisoflavone), and glycitein (4′,7-dihydroxy-6-methoxyisoflavone; Kudou *et al.*, 1991), which are biosynthesized in two branches of the phenylpropane isoflavone pathway (Yu *et al.*, 2003). In this pathway, phenylalanine is first converted into naringenin chalcone by the action of phenylalanine ammonia-lyase (PAL; Zhang & Liu, 2014), cinnamic acid-4-hydroxylase (C4H), 4-coumaroyl CoA ligase (4CL; Fraser & Chapple, 2011), and chalcone synthase (CHS; Schröder & Schröder, 1990; Pandith *et al.*, 2019). Next, naringenin chalcone is biosynthesized into liquiritigenin and naringenin by chalcone reductase (CHR) and chalcone isomerase (CHI), respectively (van Tunen *et al.*, 1991; Ralston *et al.*, 2005). Liquiritigenin is further catalyzed by isoflavone synthase (IFS) and 2-hydroxyisoflavanone dehydratase to synthesize daidzein and glycitein (Jung *et al.*, 2000; Yu *et al.*, 2000). Naringenin is a common substrate for the biosynthesis of isoflavones and flavonoids. It is converted into genistein by the enzymes IFS and 2-hydroxyisoflavanone dehydratase (Yu *et al.*, 2000). Also, naringenin is converted into flavonols and flavones under the catalysis of the enzymes, flavanone 3 hydrolase (F3H), flavonoid 3′-hydroxylase, flavonoid 3′5′-hydroxylase, and flavonol synthase for flavonols, and flavonoid 3′-hydroxylase and

*These authors contributed equally to this work.

flavone synthase for flavones, all in the phenylpropanoid pathway (Dong & Lin, 2021).

In addition to the biosynthesis of isoflavones, the phenylpropanoid pathway is also involved in the biosynthesis of lignin, stilbene, procyanidins, and anthocyanins. Due to the complexity of isoflavone biosynthesis pathway, researchers are increasingly focused on the overall regulation including isoflavone biosynthesis, transport, and accumulation. The gene *GmMPK1*, encoding a mitogen-activated protein kinase, was mapped in a genome-wide association study and reported to increase isoflavone content and resistance to *Phytophthora sojae* (Wu *et al.*, 2020). The multidrug and toxic compound extrusion transporter GmMATE1, and possibly GmMATE2, function as isoflavone transporters that promote isoflavone accumulation in soybean seeds (Ng *et al.*, 2021). GmMYB176, an R1-type MYB transcription factor (TF), can combine with the bZIP TF GmbZIP5 to affect the biosynthesis of isoflavones and isoflavonoid phytoalexins (Yi *et al.*, 2010; Li *et al.*, 2012; Vadivel *et al.*, 2021). The CCA1-like R1 MYB TF GmMYB133 and the R2R3-type MYB TF GmMYB29 both influence isoflavone biosynthesis by regulating the expression of *GmCHS8* and *GmIFS2* (Chu *et al.*, 2017; Bian *et al.*, 2018), while the R2R3-MYB TFs GmMYB53 and GmMYB205 also regulate the expression of *GmIFS2* (Han *et al.*, 2017). In addition, GmMYB39 and GmMYB100 have been shown to negatively regulate isoflavone biosynthesis by inhibiting the expression of several structural biosynthesis genes (Liu *et al.*, 2013; Yan *et al.*, 2015).

Isoflavone biosynthesis is under the control of complex regulatory networks. At present, 297 quantitative trait loci (QTL) related to isoflavones have been identified in soybean, according to the SoyBase database (http://www.soybase.org), but the regulatory mechanisms underlying their activity remain unclear. In previous studies, we developed a high-density genetic map based on a recombinant inbred line (RIL) population and successfully fine-mapped a major QTL for isoflavone content within a genomic region of around 200 kbp on chromosome Gm20 (Li *et al.*, 2014; Pei *et al.*, 2018). In the current study, we further identified the candidate genes within this genomic region and performed functional analyses to discover the key gene underlying isoflavone content. As a result, we found that a zinc-finger-type TF gene, *Zinc-Finger Protein 7* (*GmZFP7*, *Glyma.20G012700*), contributes to isoflavone content in soybean. Moreover, the potential regulatory mechanisms of this gene were proposed and discussed. We provide direct evidence that GmZFP7 is a positive regulator of isoflavone accumulation by regulating the expression levels of *GmIFS2* and *GmF3H1*. These findings also facilitate the future genetic engineering and improvement of isoflavone content in soybean.

Materials and Methods

Plant materials and growth conditions

The soybean (*G. max* (L.) Merrill) cultivars Luheidou2 (LHD; high total isoflavone content of 5406.853 $\mu g\,g^{-1}$), Nanhuizao (NHZ; low total isoflavone content of 1912.903 $\mu g\,g^{-1}$), and Williams 82 (total isoflavone content: 3413.955 $\mu g\,g^{-1}$) were used in this study. The field experiments were performed at Nankou Experimental Station, Beijing, China (40°13′N and 116°12′E). The seeds were planted in 1.5-m rows with 0.1- and 0.4-m intra- and inter-row spacing, respectively. Standard agronomic practices were followed. In the glasshouse, soybean plants were grown in pots containing nutritional soil under a 12 h : 12 h, light : dark photoperiod at 30°C. All leaves and seeds were harvested on the same day and from the same position from plants grown under the same conditions.

For the haplotype analysis, 1557 soybean accessions were provided by the soybean genetic resource research group of the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS), including 126 wild accessions, 927 landraces, and 504 cultivars. All accessions were planted at Sanya, Hainan (18°24′N and 109°5′E) and Changping, Beijing, China (40°13′N and 116°12′E) in 2017 and 2018, and at Hefei, Anhui, China (33°61′N and 117°E) in 2017. The experiment was laid out in a randomized incomplete block design, with planting locations as replications. All details of the field experiments have been reported previously (Azam *et al.*, 2020). The mean total isoflavone content in all accessions ranged from 745.1 to 5688.2 $\mu g\,g^{-1}$ (Azam *et al.*, 2020).

Resequencing the two parents, LHD and NHZ

Fifteen days after germination, 10 fresh leaves of LHD and NHZ were randomly selected. The genomic DNA of these samples was extracted using the CTAB method (Kanegae & Wada, 1998). DNA integrity and purity were determined using agarose gel electrophoresis, and DNA purity and concentration were detected using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The DNA library was completed by fragmentation, terminal repair, polyA and sequencing splices, purification, and PCR enrichment for subsequent sequencing, and then sent to BioMarker Co. (Beijing, China) for sequencing. A total of 95.78 Gbp of clean data were obtained, and the Q30 was 90.30%. The average sequence similarity between the sample and the reference genome was 98.94%, the average coverage depth was 43×, and the genomic coverage was 98.16% (at least one base cover). SNP calling for LHD and NHZ was conducted using GATK software (https://www.broadinstitute.org/gatk/guide/best-practices.php), and the obtained SNPs were annotated using SNPEFF software (v.4.3; Cingolani *et al.*, 2012).

GmZFP7 cloning

Total RNA was extracted from the leaves of LHD and NHZ soybean cultivars using the TransZol Up reagent (TransGen Biotech, Beijing, China). The quality and concentration of genomic RNA were determined using agarose gel electrophoresis and a NanoDrop 2000 spectrophotometer, respectively. First-strand cDNAs were synthesized with TransScript® One-Step gDNA Removal and cDNA Synthesis SuperMix (TransGen Biotech). The full-length cDNA of *GmZFP7* (*Glyma.20G012700.1*) was amplified using specific primers designed according to the predicted cDNA sequences in PHYTOZOME v.12.1 (https://

phytozome.jgi.doe.gov/pz/portal.html#) and cloned into a pEASY®-T3 Cloning Vector (TransGen Biotech) for sequencing confirmation. The *GmZFP7* sequence was further confirmed using the BLAST tool in SoyBase (http://soybase.org/).

Isoflavone extraction and quantification

The extraction and determination of isoflavones were performed according to the protocol described by Sun *et al.* (2011). A cyclone mill (A10 basic; IKA, Staufen, Germany) was used to grind *c.* 20 g of soybean tissues (seeds, leaves, pods, stems, and roots) into a fine powder. From this, 0.1 g of powder was weighed into a 10-ml centrifuge tube, to which 5 ml of a solution containing 70% (v/v) ethanol and 0.1% (v/v) acetic acid was added. The samples were shaken and mixed for 12 h to be fully homogenized. After centrifugation at 2700 ***g*** for 10 min at 4°C, the supernatant was filtered through a YMC Duo filter (YMC, Kyoto, Japan) with a pore size of 0.2 µm. An Agilent 1260 HPLC System (Agilent Technologies, Santa Clara, CA, USA) was used to determine the isoflavone contents. A YMC ODS AM-303 column (250 mm × 4.6 mm ID, S-5 µm, 120 Å) was used for the quantitative analysis. The mobile phases A and B consisted of 0.1% acetic acid in distilled water and acetonitrile, respectively. The solvent flow rate was 1.0 ml min^{-1}, the injection volume was 10 µl, and a 70-min linear gradient of 13–30% acetonitrile (v/v) was used. The wavelength of the UV detector was set to 260 nm, and the column temperature was set to 35°C. Six isoflavone standards, daidzin, daidzein, genistin, malonyldaidzin, malonylglycitin, and malonylgenistin, were used for the identification and quantification of each isoflavone component. The standards were provided by Dr Akio Kikuchi (National Agricultural Research Center of Tohoku Region, Japan). The concentration of isoflavone was calculated using the formula described by Sun *et al.* (2011). The total isoflavone concentration was specified as the sum of all isoflavone component contents.

GmZFP7 expression pattern

RNA was isolated from the roots, stems, and leaves of the soybean plants during the full-blossom period; the pod walls on the 10^{th} day after flowering; the seeds at 20, 30, 40, and 50 d after flowering; and the mature seeds. Quantitative RT-PCR (qRT-PCR) was used to detect the expression of *GmZFP7* in these tissues in LHD and NHZ. The qRT-PCR primers are summarized in Supporting Information Table S1.

Phylogenetic tree analysis

The typical C2H2 zinc-finger protein sequences in *Arabidopsis thaliana*, *Zea mays*, *G. max*, *Solanum lycopersicum*, *Oryza sativa japonica*, *Triticum aestivum*, and *Petunia axillaris* were obtained from the TF database Plant TFDB (http://planttfdb.gao-lab.org/) and used for a phylogenetic analysis. The phylogenetic tree was inferred using the neighbor-joining method and the software MEGA7 (www.megasoftware.net). A bootstrap analysis with 1000 replicates was performed.

Subcellular localization

The open reading frame (ORF) region of *GmZFP7* (without the stop codon) was amplified from the LHD and NHZ cDNA and then inserted them into the *Xba*I site of the PTF101-GFP plasmid to create a GFP fusion protein, CaMV35S:GmZFP7(LHD/NHZ)-GFP. The resulting vectors, PTF101-GmZFP7(LHD/NHZ)-GFP, were confirmed by sequencing and used for the subcellular localization analysis. The CaMV35S:GFP plasmid was used as control. Briefly, 100 g of mixed plasmid was coated in 1-m-diameter gold dust in 0.5 mM spermidine, precipitated by 0.1 mM $CaCl_2$, and washed three times in ethanol. DNA-coated gold dust was delivered into onion epidermal cells precultured for 4 h by Gene Gun (BioRad). The fusion and nucleus marker (with RFP) plasmids were transformed into the *Agrobacterium tumefaciens* EHA105 strain using the freeze–thaw method, and the cells were injected into tobacco leaves and cultivated for 2–3 d. The GFP signal was examined using confocal fluorescence microscopy. The basic sequence of the nuclear localization marker was CaMV35S:AT1G07790:RFP.

RNA extraction, reverse transcription, and quantitative PCR

Total RNA was extracted using the Plant rapid RNA Extraction Kit (GeneBetter, Beijing, China). TransScript® One-Step gDNA Removal and cDNA Synthesis Supermix (TransGene Biotech) were used to obtain cDNA. qRT-PCR was performed using an ABI7900 system (Applied Biosystems, Foster City, CA, USA), and a PCR mixture containing 1 µl of 1 : 5 diluted cDNA, 0.4 µl gene-specific primers (10 ng μl^{-1}), 5 µl PowerUp SYBR Green Master Mix (Applied Biosystems), and water to a final volume of 10 µl was used. The PCR conditions were as follows: 95°C for 3 min, followed by 45 cycles of 95°C for 5 s and 60°C for 30 s. Three biological replicates were performed for each reaction. The qRT-PCR primers are summarized in Table S1.

Soybean hairy root transformation

The ORF regions of *GmZFP7* cloned from LHD and NHZ were inserted into the pGUSGFPplus (pGGP) plasmid with the CaMV35S promoter and CaMV35S:GFP to produce the pGGP-GmZFP7-OE overexpression vector. A 288-bp (from 408 to 695 bp) fragment of *GmZFP7* was cloned to construct the pGGP-RNAi plasmid. A soybean hairy root transformation was performed using this accession with the pGGP-GmZFP7-OE overexpression vector and the pGGP-GmZFP7-RNAi vector, while the pGGP empty vector was used as the control. The vector was transferred into *Agrobacterium rhizogenes* using the click transformation method, and then *A. rhizogenes* was used to infect soybean cotyledon nodes to induce the formation of hairy roots (Chen *et al.*, 2018a). A portable hand-held blue-green lamp (Luyor-3260, Shanghai, China) was used to screen for positively transformed hairy roots (Fig. S1). The positive hairy roots were harvested from several independent transgenic lines and used for gene expression or isoflavone content analysis.

Transient transcription dual-LUC reporter assays

Approximately 2-kbp promoter regions (including the 5′ UTR) of important genes in the isoflavonoid pathway were identified in PHYTOZOME v.13 and cloned. *GmIFS1*, *GmIFS2*, and *GmF3H1-pro*:*LUC* fusions containing the promoters of the genes and the *CaMV35S*:*GmZFP7*:*GFP* fusions were cloned and used as reporters and effectors, respectively. *CaMV35S*:*GFP* was used as the control. The reporter plasmids were introduced into the EHA105 (pSoup) strain. In the transient assays, each bacterial mixture was made with 350 µl (OD_{600} = 0.8) of EHA105 (pSoup) containing the reporter (pGreenII-0800-Promoter-LUC) and 350 µl (OD_{600} = 0.8) of EHA105 (pSoup) containing the effector (PTF101-GmZFP7(L/N)-GFP or PTF101-GFP), and 300 µl of the P19 microbial culture (OD_{600} = 0.6). Tobacco plants were grown in the glasshouse with a 16 h : 8 h, light : dark cycle at 22°C. After agro-infiltration with the mixture followed by 3-d incubation, 2-cm pieces were harvested from tobacco leaves and stored at −70°C, and the protein was extracted using a protein extraction kit (CoWin Biosciences, Jiangsu, China). The firefly luciferase (LUC) and *Renilla* LUC were assayed using the Dual-Glo Luciferase Assay System (Promega, Madison, WI, USA). The enzyme label assay from SYNERGY H1 was used to detect the activity of LUC. The *Renilla* LUC reporter was used as an internal control for normalization, and the microbial solution of PTF101-GFP with pGreenII-0800-Promoter-LUC was the control (Chu *et al.*, 2017). The relative LUC activities normalized to REN activity are shown (LUC/REN). Each sample was tested three times at 5-min intervals to ensure a complete response.

In vitro electrophoretic mobility shift assays

Based on the results of the dual-LUC complementation assay results, the GmZFP7 target probe was designed using the online tool JASPAR (http://jaspar.genereg.net): GmIFS2: 5′-CCCAG ATGCAGTGACAATCGAAGGAAAAGACAAAACCCAAAT ATATGTTTTTTA-3′, GmF3H1: 5′-AAGCATTGCATTC TGCTATTTAATTCCACTACGTACACGCACATTCTCCT CA AA-3′. The *GmZFP7(L/N)* sequences were cloned into the expression vector PGEX-4T containing the GST protein and transferred into competent cells of the *Escherichia coli* strain Rosetta. The expression of the TFs was induced with IPTG, and the proteins were purified using protein purification resins and eluted with glutathione elution buffer. The eluted proteins and biotin end-labelled duplex DNA probes were mixed with electrophoretic mobility shift assay (EMSA) binding buffer (LightShift box; Thermo Fisher Scientific) and then resolved using native gel electrophoresis with 0.5 TBE. The bound protein–DNA sample was transferred from the native gel to a positive nylon membrane. After UV cross-linking for 10 min, the Chemiluminescent Detection Module kit (Thermo Fisher Scientific) was used to detect and image biotin luminescence.

CRISPR/Cas9-mediated mutagenesis of *GmZFP7*

The CRISPR/Cas9 system was employed to knock out *GmZFP7*. The target site was selected using the web tool CRISPR-P v.2.0 (Liu *et al.*, 2017) based on the GC content and putative off-target sites. The 20-bp target sequence (5′-TCTGGTTCCAGATTCAAG TTTGG-3′) was located in the *GmZFP7* exon. Nested PCR was used to link the gRNA with the *U6* promoter and integrate them into the CRISPR-Cas9 vector. The vector was then transformed into *A. tumefaciens* strain EHA105 via electroporation. The soybean cultivar Williams82 was used for the transformation, as described previously (Chen *et al.*, 2018b). DNA was extracted from the leaf tissue and used to examine the CRISPR/Cas9-induced mutations at the target sites using a PCR and sequencing analysis. Transgenic plants were grown in a glasshouse. The primers used for sequencing were as follows: CR-ZF20-F: TAAAAGGAAGTGGAGTTA-CATTG; CR-ZF20-R: TTAAAGCCTCAGAGTGAGATCAG.

Transgenic overexpression of *GmZFP7* in soybean

The *GmZFP7* sequence cloned from the cDNA of LHD leaves was ligated to the *Xba*I site of the PF101-GFP vector to form a CaMV35S:GmZFP7(LHD):GFP overexpression vector. The final overexpression vector was introduced into *A. tumefaciens* strain EHA101 and ultimately transformed into Williams82 using the cotyledon-node method (Chen *et al.*, 2018b). The transgenic plants were identified by daubing leaves with 160 mg l^{-1} glufosinate and detecting the PAT proteins using Liberty Link strips. All transgenic and WT plants were grown in glasshouse conditions of 12 h : 12 h, light : dark, 30°C, and 70% relative humidity. The positive plants were detected using *bar* gene strip and glufosinate-ammonium leaf spray (Fig. S2).

Metabolome analysis

For metabolomics analysis, four samples were selected for each transgenic line. Each sample had three biological replicates. Metabolomics analysis was conducted using an ultra-performance liquid chromatography–mass spectrometry system performed on a Qtrap 5500+ mass spectrophotometer (AB Sciex, Framingham, MA, USA) equipped with an electrospray ionization source and a Shimadzu Nexera UHPLC LC-30A system, and an injection volume of 5 µl. Mobile phases were 0.1% formic acid–water and acetonitrile. Electrospray ionization was conducted in positive mode. The extraction and detection were assisted by Shanghai Luming Biological Technology Co. Ltd (Shanghai, China).

Haplotype analysis of *GmZFP7*

GmZFP7 gene sequences for 1557 soybean accessions were compiled from SoyFGB (https://sfgb.rmbreeding.cn/index) and used for the genetic diversity analysis of *GmZFP7* in soybean. All the SNPs located in the promoter and coding regions of *GmZFP7* were identified based on the genome gff3 annotation. The soybean accessions were divided into three populations: wild, landrace, and cultivar. The association test of the isoflavone content and the SNP was performed using SPSS v.16.0 (IBM, Armonk, NY, USA). The haplotype analysis was performed using Perl scripts. The numbers of haplotypes and the haplotype diversity levels were identified and distinguished using DNASP v.5.0

(Rozas *et al.*, 2017). The haplotype network was displayed using PopART v.1.7 (Leigh & Bryant, 2015).

Results

Discovery of candidate genes for soybean isoflavone content in *qIF20-2* QTL

In a previous study, we developed a RIL population of 200 lines by crossing two soybean varieties with significantly different isoflavone contents, LHD (high-isoflavone content) and NHZ (low isoflavone content). In that study, we used next-generation sequencing (NGS) to construct a high-density genetic linkage map using 110 of the 200 lines and identified a major stable QTL on Gm20, *qIF20-2*, which could explain *c.* 20% of phenotypic variances in isoflavone content across multiple environments (Li *et al.*, 2014). To fine-map this locus, we sequenced the remaining 90 lines and constructed an integrated genetic linkage map using all 200 lines (Li *et al.*, 2017). The genetic distance of this locus was successfully reduced from 4.60 to 0.53 cM, with the corresponding genomic region reduced from 575.95 to 243.72 kbp (Pei *et al.*, 2018).

To identify the candidate genes responsible for isoflavone content at this locus, the genomes of the two RIL parents were resequenced using NGS in the present study. Illumina high-throughput sequencing resulted in 40 and 56 Gb of clean reads for LHD and NHZ, respectively (Table S2). Based on the Q20 and Q30 values, the sequencing quality was considered sufficient, and the data were used for further analysis. The clean reads of the two cultivars were aligned to the soybean reference genome (Wm82.a2.v1). The mapping rates of the sequencing data were 99.01% for LHD and 98.87% for NHZ, with an average depth of 36 and 51 for LHD and NHZ, respectively (Table S3). In the current study, we focused on the genes within the genomic region of *qIF20-2*. Within the 243.72-kbp region, 33 genes were annotated, and 17 of them contained nonsynonymous variations between LHD and NHZ (Table S4), although none had previously been reported to be involved in isoflavone biosynthesis. By combining the sequence differences and the developing seed transcriptome expression profiles of the genes within *qIF20-2* (Table S4; Fig. S3), we noticed that two TF-encoding genes, *Glyma.20g011700* (encoding a MYB-like TF) with a nonsynonymous SNP and *Glyma.20g012700* (encoding a zinc-finger TF related to plant tolerance; Table S4) with two nonsynonymous SNPs, respectively, were observed between the parents. Previous studies have shown that TFs can affect isoflavone accumulation in soybean; therefore, these two genes were considered important candidate genes, and their functions were explored.

Characterization of the candidate genes

The full ORF of *Gm20MYB* (*Glyma.20G011700*) contains 891 bp and encodes a protein 296 amino acids in length, with a calculated mass of 31.51 kDa and a pI of 6.26. There is a Phe-Cys variation at the 221st amino acid in Gm20MYB proteins of LHD and NHZ (Fig. S4a); however, in the subsequent hairy root experiment, the isoflavone content was not significantly altered by the overexpression or silencing of *Gm20MYB* in transgenic hairy roots (Fig. S5), suggesting that *Gm20MYB* was not the key gene responsible for the regulation of isoflavone content by *qIF20-2*. We also analyzed another candidate, *GmZFP7* (*Glyma.20G012700.1*; Fig. 1a), which contains a 717-bp ORF encoding a protein of 238 amino acids in length with a predicted molecular mass of 26.23 kDa and a pI of 6.56. The GmZFP7 protein belongs to the C2H2-type zinc-finger subfamily and contains a conserved plant-specific 'QALGGH' motif found in other proteins involved in plant stress tolerance (Fig. S4b; Zhang *et al.*, 2016; Wang *et al.*, 2018). Two amino acid differences were observed in the GmZFP7 proteins of LHD and NHZ. The first was a His-Arg difference at the 114th amino acid, while the second caused a premature termination in the NHZ protein at the 227th amino acid compared with LHD (Fig. S4c). Phylogenetic analysis of C2H2-type zinc-finger proteins from various plant species revealed that GmZFP7 was grouped in the same cluster as genes regulating plant organ development and hormone control (e.g. KNU, GIS, RBE, SUPMAN, and AtYY1; Fig. 1b; Payne *et al.*, 2004; Takeda *et al.*, 2004; An *et al.*, 2012; Li *et al.*, 2016).

GmZFP7 contributes to isoflavone accumulation in soybean hairy roots

An expression pattern analysis revealed that *GmZFP7* was detected in all tissues and that its expression pattern was consistent with the accumulation of isoflavone in the seeds, with both increasing during embryonic development. The expression level of *GmZFP7* in LHD was significantly higher than in NHZ at 40–50 d after flowering, during which time isoflavone was rapidly accumulating in both lines, but to a higher level in LHD (Fig. 1c,d). A subcellular localization analysis showed that GmZFP7 was distributed in the nucleus and cytoplasm in both transgenic onion (*Allium cepa*) epidermal cells and tobacco (*Nicotiana benthamiana*) leaves (Fig. 1e,f).

To further investigate the function of GmZFP7 in isoflavone accumulation, we constructed two vectors, pGGP-GmZFP7L/N (L for LHD-type and N for NHZ-type) and pGGP-GmZFP7-Si, for generating *GmZFP7*-overexpressing and *GmZFP7*-RNAi hairy roots. In the LHD background, the transcript levels of *GmZFP7* were significantly increased by 2.0- to 2421.3-fold in the *GmZFP7L*-overexpressing transgenic hairy roots and 684.2- to 1374.2-fold in the *GmZFP7N*-overexpressing transgenic hairy roots, while their relative expression levels were decreased by around 90% in the corresponding silenced hairy roots (Fig. 2a). Consistent with the transcription level, the relative isoflavone contents were significantly increased by 1.15- to 1.29-fold in the *GmZFP7L*-overexpressing transgenic hairy roots and 1.24- to 1.37-fold in the *GmZFP7N*-overexpressing transgenic hairy roots, while their relative isoflavone contents were decreased by 28% in the silenced hairy roots (Fig. 2b). In the NHZ background, the expression level of *GmZFP7* was significantly increased by 7.2- to 216.9-fold in the *GmZFP7L*-overexpressing transgenic hairy roots and 66.2- to 201.9-fold in the *GmZFP7N*-overexpressing transgenic hairy roots, while the expression decreased by 99% in the *GmZFP7*-silenced

New Phytologist

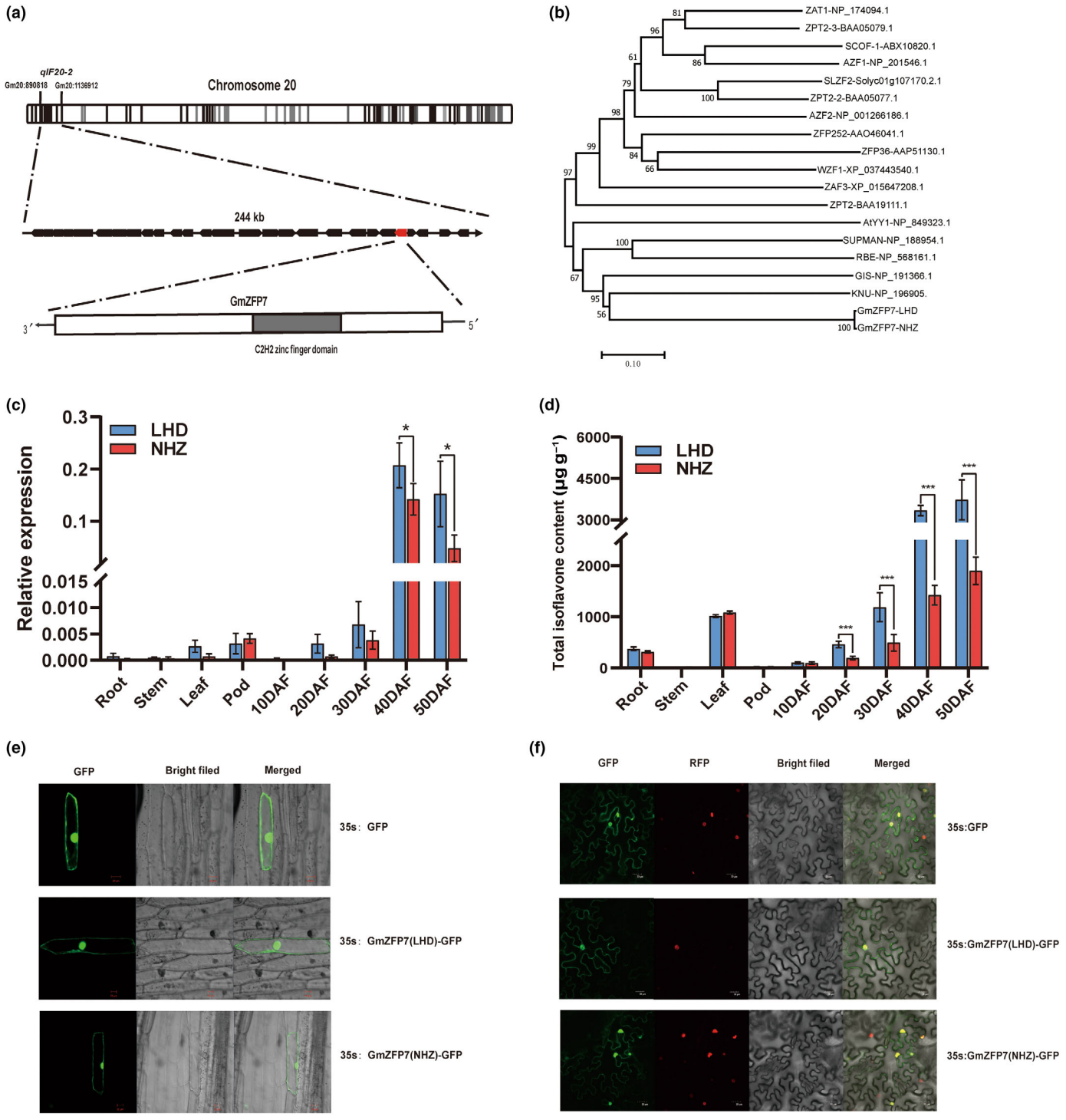


Fig. 1 Characteristics of the candidate gene *GmZFP7*. (a) Position of *GmZFP7* on Gm20 in soybean. (b) Phylogenetic tree of C2H2-type zinc-finger transcription factors. Branch numbers represent a percentage of the bootstrap values in 1000 sampling replicates, and the scale bar indicates branch lengths. (c) Expression of *GmZFP7* in different soybean tissues of Luheidou2 (LHD) and Nanhuizao (NHZ). Bars represent the standard error of three technical replicates in at least three biological replicates. (d) Accumulation of total isoflavones in different soybean tissues of LHD and NHZ. Bars represent the standard error of three technical replicates in at least three biological replicates. The error bars represent the SE. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Different lowercase letters indicate statistically significant differences at the $P < 0.05$ level, as determined using a one-way ANOVA. (e) Subcellular localization analysis of GmZFP7 in (e) onion epidermal cells and (f) tobacco leaves.

hairy roots (Fig. 2a). As a result, the relative isoflavone contents were significantly increased by 2.27- to 3.79-fold in the *GmZFP7L*-overexpressing transgenic hairy roots and 1.31- to 2.06-fold in the *GmZFP7N*-overexpressing transgenic hairy roots, whereas a 27% decrease in isoflavone contents was observed in the silenced hairy roots (Fig. 2b).

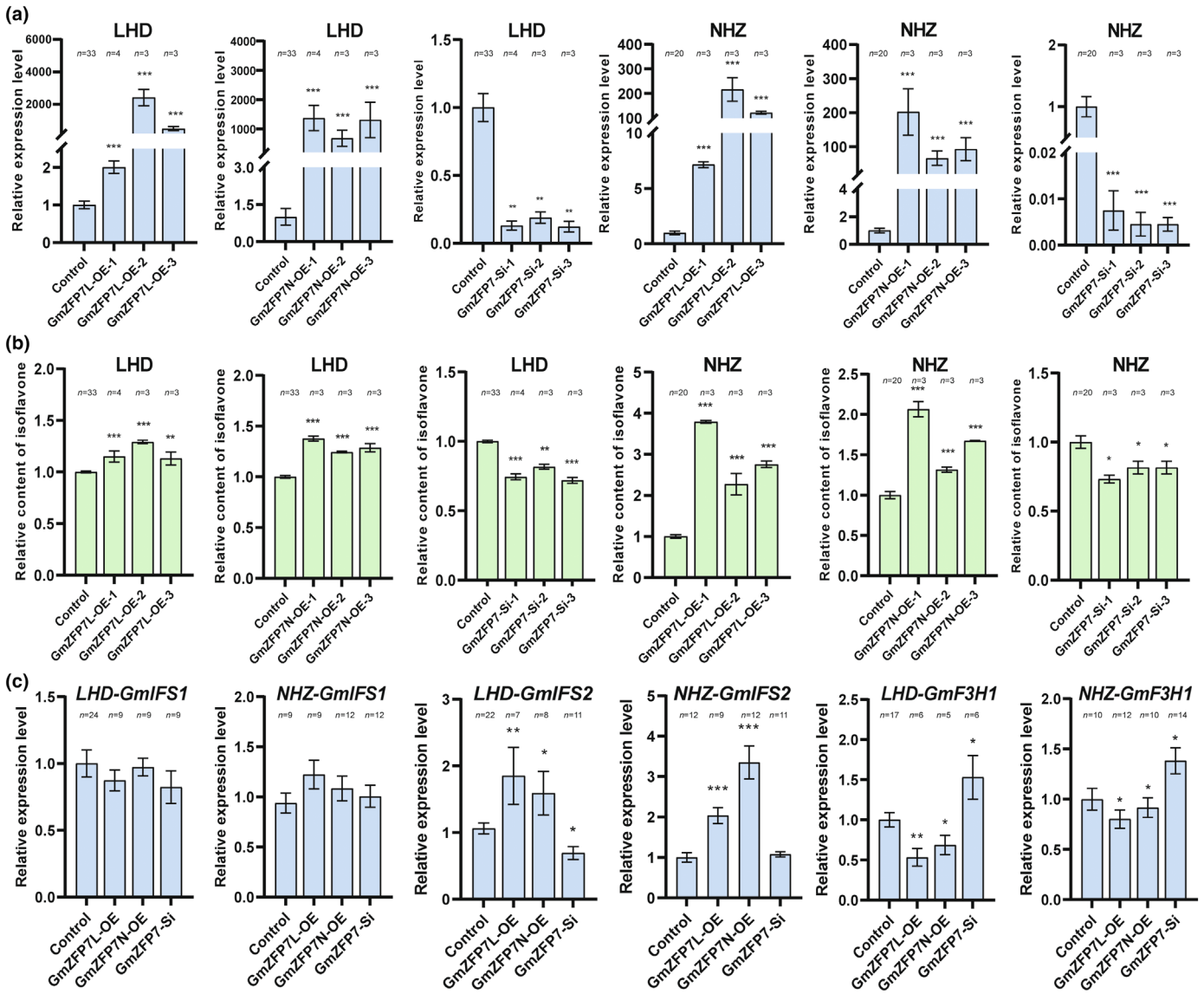


Fig. 2 Gene expression and isoflavonoid content of *GmZFP7*-overexpressing and RNAi-silenced soybean hairy roots. (a) The relative expression level of *GmZFP7* in the hairy roots. The hairy roots were transformed using the pGGP vector as the control, and all results were normalized against the housekeeping gene *GmACTIN*. (b) The relative total isoflavone content in *GmZFP7*-overexpressing and *GmZFP7*-silenced hairy roots. (c) The relative expression level of the isoflavonoid-related gene *GmIFS* and *GmF3H1* in the hairy roots. GmZFP7L-OE and GmZFP7N-OE represent the *GmZFP7*-overexpressing roots in the Luheidou2 (LHD) and Nanhuizao (NHZ) backgrounds, respectively. GmZFP7-Si represents RNAi-silenced hairy roots. The data for each of the three independent overexpression and silencing lines or the control represent the means of three replicates, with error bars indicating the SE. *, $P<0.05$; **, $P<0.01$; ***, $P<0.001$, as determined using a one-way ANOVA..

To explore the role of GmZFP7 in regulating the isoflavone biosynthesis pathway, we performed a qRT-PCR on three gateway enzyme genes involved in isoflavone biosynthesis, *GmIFS1* (*Glyma.07G202300*), *GmIFS2* (*Glyma.13G173500*), and *GmF3H1* (*Glyma.02G048400*) in the transgenic hairy roots. No significant change in *GmIFS1* expression was observed in any of the hairy roots. The expression level of *GmIFS2* was significantly increased in the *GmZFP7L/N*-overexpressing hairy roots, but only decreased in the *GmZFP7*-silenced LHD hairy roots. By contrast, the expression level of *GmF3H1* was significantly decreased in the *GmZFP7L/N*-overexpressing hairy roots and significantly increased in the *GmZFP7*-silenced hairy roots (Fig. 2c).

These results indicated that the overexpression or silencing of *GmZFP7* could significantly enhance or reduce the accumulation of isoflavone in hairy roots by altering the expression of isoflavone biosynthesis gateway genes. It is noteworthy that both *GmZFP7L* and *GmZFP7N* could contribute to isoflavone accumulation in soybean hairy roots, although an abovementioned premature termination SNP was observed in NHZ.

GmZFP7 is a dual-function TF in the isoflavone pathway

To explore whether GmZFP7 directly binds to the promoters of genes in the isoflavone pathway to regulate their transcription, we

determined the activities of GmZFP7 on the expression of three key structural genes, *GmIFS1* (*Glyma.07G202300*), *GmIFS2* (*Glyma.13G173500*), and *GmF3H1* (*Glyma.02G048400*), using dual-luciferase (LUC) reporter assays (Fig. 3a). The overexpression of *GmZFP7L* and *GmZFP7N* significantly increased *GmIFS2* expression and inhibit *GmF3H1* expression, but had no effect on *GmIFS1* (Figs 3b, S6). These results suggested that GmZFP7 plays a dual role in the transcriptional regulation of key genes (*GmIFS2* and *GmF3H1*) in the soybean isoflavone biosynthesis pathway.

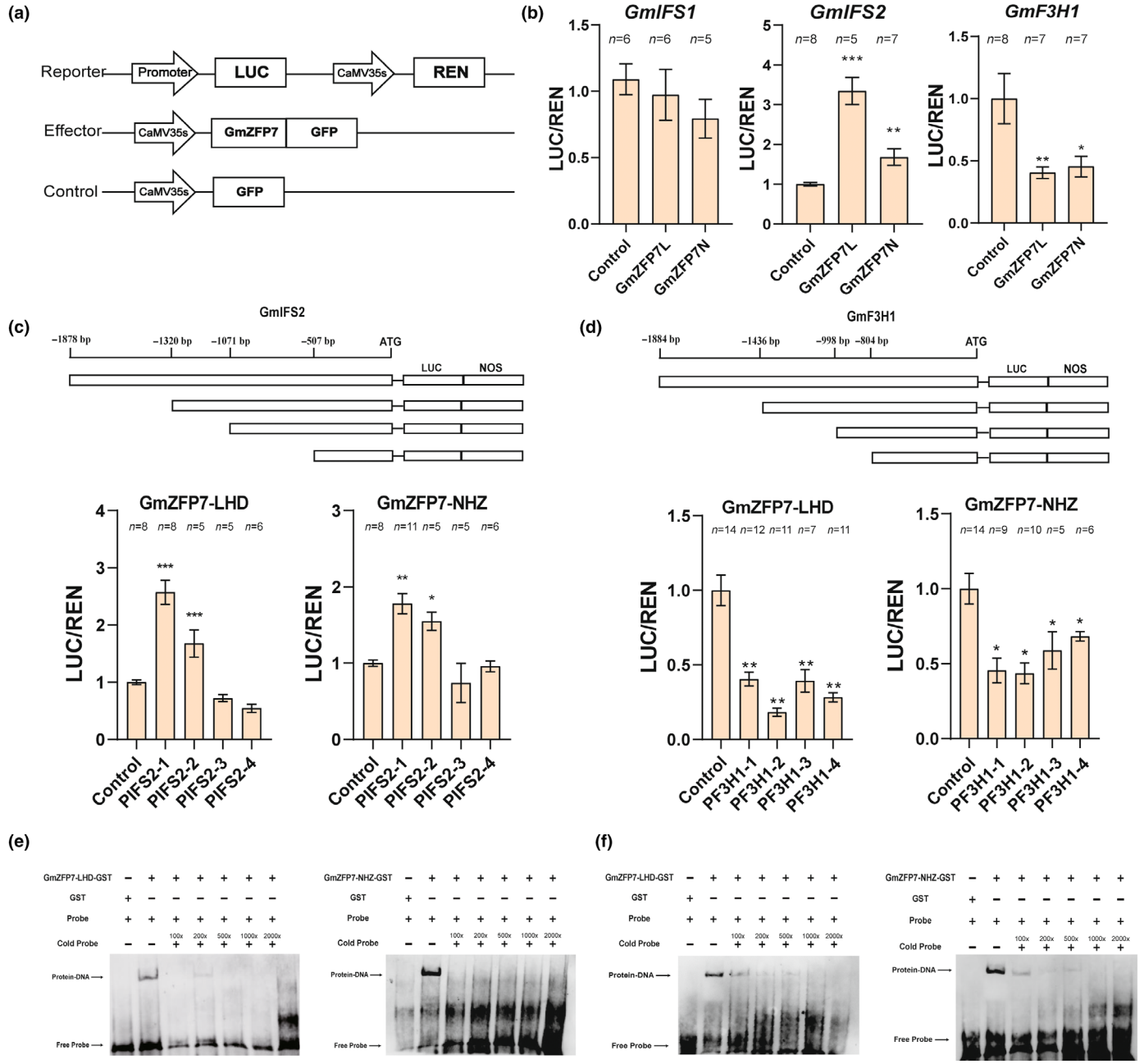


Fig. 3 Dual-luciferase (LUC) complementation and electrophoretic mobility shift assay to verify the interaction between GmZFP7 and the *GmIFS2* and *GmF3H1* promoters. (a) Schematic representation of the reporter, effector, and internal control constructs used in the transient expression assays. (b) GmZFP7-induced promoter activity of *GmIFS1*, *GmIFS2*, and *GmF3H1*. (c) GmZFP7 binding to different regions of the *GmIFS2* promoter. (d) GmZFP7 binding to different regions of the *GmF3H1* promoter. The GmZFP7L/GmZFP7N column represents the relative LUC activity (Firefly/*Renilla*) of the promoter plus the GmZFP7 factor relative to the control (without the GmZFP7 factor). The *PIFS2-1*, *PIFS2-4*, *PF3H1-1*, and *PF3H1-4* columns represent the relative LUC activity (Firefly/*Renilla*) of the different promoters co-expressed with the GmZFP7 factor relative to the control. The ratio of LUC to REN luminescence was measured, and the ratio of the control was set at a value of 1. Each value represents the mean of at least three replicates, and the error bar represents the SE. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$, as determined using a one-way ANOVA. (e) Electrophoretic mobility shift assay (EMSA) was used to verify the interaction between GmZFP7 and *GmIFS2*. (f) EMSA was used to verify the interaction between GmZFP7 and *GmF3H1*.

To identify the GmZFP7 recognition region in the *GmIFS2* and *GmF3H1* promoters, four fragments, PIFS2-1 (0 to −1878 bp), PIFS2-2 (0 to −1320 bp), PIFS2-3 (0 to −1071 bp), PIFS2-4 (0 to −507 bp) and PF3H1-1 (0 to −1884 bp), PF3H1-2 (0 to −1436 bp), PF3H1-3 (0 to −998 bp), PF3H1-4 (0 to −804 bp), were generated by gradual 5′ deletions of the promoter and were then used in dual-LUC reporter assays. Our results showed that the LUC/REN enzyme activity linked to PIFS2-1 and PIFS2-2 yielded similar LUC activities, which were significantly higher than the control. By contrast, the LUC activity in PIFS2-3 and PIFS2-4 was greatly decreased with no significant differences between them and the control. This indicated that the 249-bp region from position −1071 to −1320 in the *GmIFS2* promoter contained the motif required for binding GmZFP7 (Fig. 3c). As for *GmF3H1*, the LUC/REN enzyme activities for all four parts of the *GmF3H1* promoter (PF3H-1 to PF3H-4) were greatly decreased, suggesting that the 804-bp region from position 0 to −804 in the *GmF3H1* promoter contained the motif required for binding GmZFP7 (Fig. 3d). To validate these results, we used online tools to predict binding sites in this region and conducted an EMSA. The results showed that both GmZFP7L and GmZFP7N could physically interact with promoter fragments of *GmIFS2* and *GmF3H1 in vitro*. These results also confirmed that both the LHD- and NHZ-type GmZFP7 proteins could bind to the *GmIFS2* (Fig. 3e) and *GmF3H1* (Fig. 3f) promoters. This demonstrates that the GmZFP7 TF can bind the promotors of both *GmIFS2* and *GmF3H1* and regulate their expression levels.

CRISPR/Cas9-induced gene editing and overexpression of *GmZFP7* in soybean

A CRISPR/Cas9 system was used to perform the targeted mutagenesis of *GmZFP7* in the cultivar Williams82 to validate the function of GmZFP7 in soybean isoflavone accumulation. We conducted a total of five transformation events in this gene-editing experiment and screened out four homozygous genome-edited lines with the single-base insertion (*Gmzfp7-1*) and 5-bp base deletion (*Gmzfp7-2*) mutation in *GmZFP7* (Fig. 4a). The isoflavone contents of these *Gmzpf7* knockout mutants were significantly decreased in the leaves at the full bloom stage (R2 stage) and mature seeds (R8 stage; Fig. 4b,c). Our qRT-PCR results showed that the expression level of *GmZFP7* was not significantly different in *Gmzfp-1*, but significantly decreased in *Gmzfp-2* (Fig. 4d). Furthermore, the expression profiles of genes in isoflavone pathway showed that the expression of *GmC4H*, *GmCHS7*, *GmCHS8*, *GmCHR1*, *GmCHR3*, and *GmIFS2* decreased, while the expression of *GmF3H1* increased in mutant leaves (Figs 4e, S7a).

Constitutive *GmZFP7*-overexpressing lines driven by CaMV35S transgenic plants were also developed. We found that the expression level of *GmZFP7* (Fig. 4h) and the isoflavone content (Fig. 4f,g) were significantly increased in the leaves and seeds of these lines. Moreover, the isoflavone contents of *GmZFP7*-overexpressing lines were dependent on the transcript levels of *GmZFP7* (Fig. 4f–h). The qRT-PCR analysis revealed that the expression level of *GmC4H*, *Gm4CL*, *GmCHS7*, *GmCHS8*, *GmCHR1*, *GmCHR4*, and *GmIFS2* was significantly increased, while that of *GmF3H1* was significantly decreased in the *GmZFP7*-overexpressing leaves (Figs 4i, S7b).

To further investigate the effect of GmZFP7 on phenylpropanoid metabolism, targeted metabolomics analysis was performed in both knockout mutants and overexpression seeds. The metabolome was tested for 129 phenolic metabolites. A total of 57 metabolites were detected in transformed seeds, of which 24 metabolites were different (Table S5). The results showed that after overexpressing *GmZFP7*, upstream products of the isoflavone pathway, such as trans-cinnamic acid, aesculin, isoliquiritigenin, phlorizin in phenylpropane pathway, cosmosiin, luteolin, apigenin in flavonoid pathway, and naringin and naringenin in flavanone pathway, were significantly increased. For the flavonol pathway, the contents of rutin, isorhamnetin-3-O-glucoside, and prunin were significantly increased, but the contents of quercitrin and nicotiflorin were significantly decreased (Fig. S8a; Table S5). In GmZFP7 knocked out mutants, the contents of trans-cinnamic acid, isoliquiritigenin, phlorizin, luteolin, apigenin, naringin, naringenin, and quercitrin were significantly decreased (Fig. S8b; Table S5).

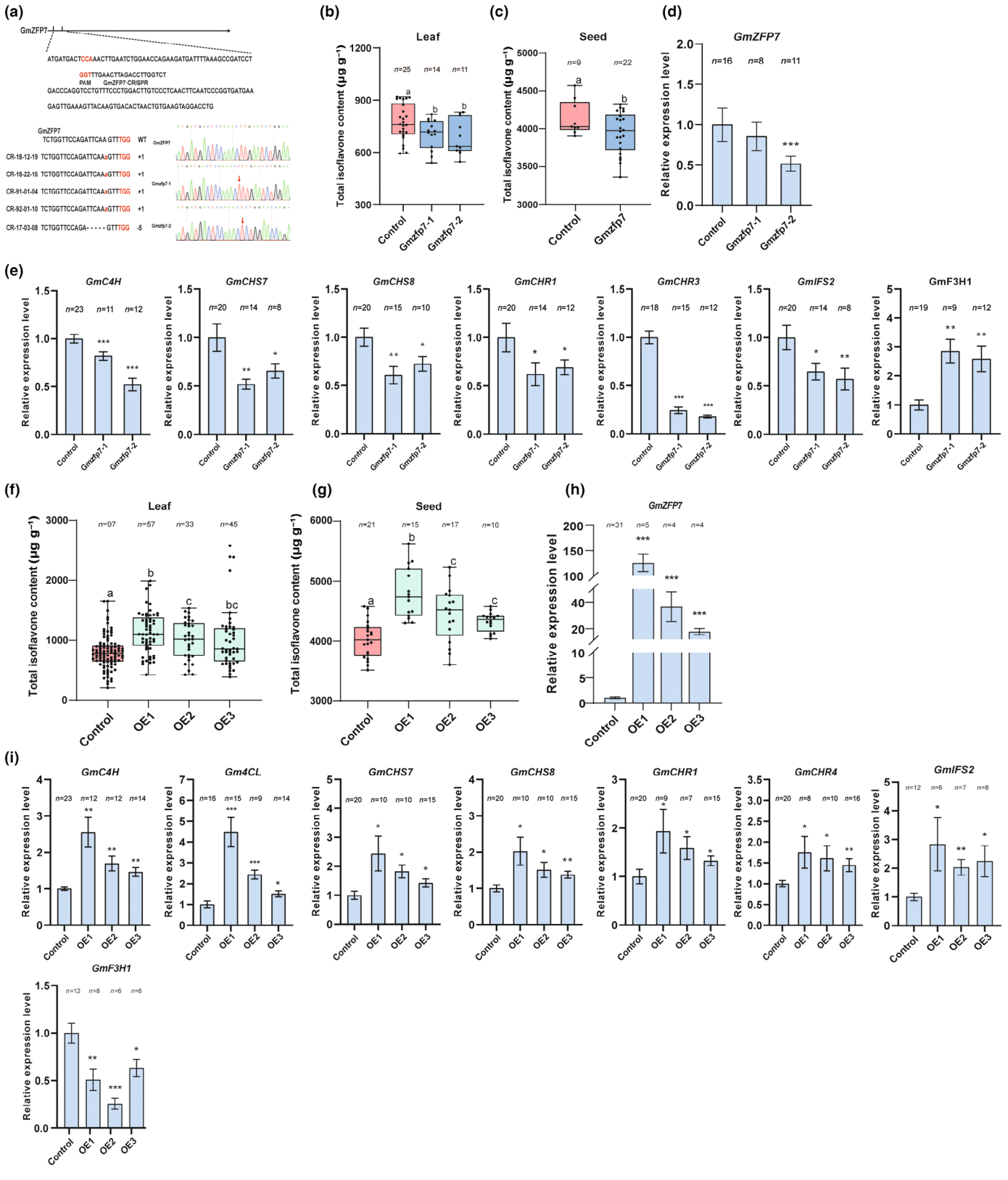
Taken together, the results obtained from *GmZFP7* knockout mutants and overexpressing plants suggest that GmZFP7 can increase the efficiency of the phenylpropanoid pathway by affecting the expression of multiple enzyme genes and participating in the regulation of soybean isoflavone content through its pronounced influence on the expression of *GmIFS2* and *GmF3H1*.

Haplotype analysis of *GmZFP7*

A haplotype analysis was conducted on 1557 soybean accessions to determine the association between the *GmZFP7* allele variation and isoflavone content. Based on an association analysis using a general linear model and mixed linear model, *GmZFP7* was further confirmed to be associated with isoflavone content (Fig. 5a). Five SNPs were identified in the coding regions of the *GmZFP7* sequences from all the accessions, including three nonsynonymous SNPs (Gm20_1070377, Gm20_1070658, and Gm20_1070712), one synonymous SNP (Gm20_1070549), and one termination codon SNP (Gm20_1070374). Eight haplotypes (designated C-Haps) were identified according to these sequence differences (Fig. 5b); however, association analysis with the isoflavone phenotypes revealed that these SNPs and haplotypes did not group the population well (Figs 5c, S9).

To further analyze the effect of the variations in *GmZFP7* on isoflavone content, we also analyzed the SNP diversity in the promoter region and identified 37 haplotypes (designated P-Haps) based on the 41 SNPs identified in the 2000-bp upstream sequence (Fig. S10; Table S6). The promoter sequence presented more genetic diversity in both the cultivated and wild soybeans. Sixteen haplotypes were identified in the wild soybean lines, and the cultivars may have been selected from wild individuals carrying the P-Hap18, P-Hap20, P-Hap24, and P-Hap28 haplotypes (Fig. 5d,e). Ten haplotypes found in more than nine varieties were selected and divided into two major groups. Group A comprises five haplotypes (P-Hap1, P-Hap3, P-Hap55, P-Hap61,

New Phytologist



and P-Hap2) with a higher proportion of landraces (not < 38.4%), while group B consists of P-Hap4, P-Hap11, P-Hap38, P-Hap5, and P-Hap41 and has a higher proportion of cultivars (not < 35.4%; Fig. 5f). In group A, 65% of the accessions had isoflavone contents > 2300 μg g^{-1}, and 88% of the accessions of P-Hap_1 and P-Hap_3 had isoflavone contents

Fig. 4 Gene editing and overexpression of *GmZFP7* in soybean. (a) Gene structure of *GmZFP7* with sites targeted using CRISPR/Cas9. Red letters represent the exon of *GmZFP7*. The red sequence represents the protospacer adjacent motif (PAM) site. The red arrow indicates the location of the mutation. (b) Isoflavone content in the leaves of the CRISPR/Cas9-mediated *GmZFP7* mutant (*GmZFP7*) and Williams82 (Control). (c) Isoflavone content in the seeds of the *Gmzfp7* mutant and Williams82. Points in the figure represent the isoflavone contents of each individual transgenic plant. (d) The relative expression level of *GmZFP7* in the Williams82 and *Gmzfp7* mutant leaves. (e) The relative expression levels of isoflavone synthesis-related genes in the Williams82 and *Gmzfp7* mutant leaves. (f) The total isoflavone contents in the leaves of the *GmZFP7*-overexpressing transgenic plants. (g) The total isoflavone content in the seeds of the *GmZFP7*-overexpressing transgenic plants. The points in the figure represent the isoflavone contents of each individual transgenic plant. (h) The relative expression level of *GmZFP7* in *GmZFP7*-overexpressing leaves. Each column represents the relative expression level in the leaves of one transgenic line. (i) The relative expression level of isoflavone synthesis-related genes in *GmZFP7*-overexpressing leaves. The control column presents the result of plants transformed with the PTF101 empty plasmid. All quantitative RT-PCR (qRT-PCR) results were normalized against the housekeeping gene *GmACTIN*. For boxplots, box represents interquartile range, the horizontal line inside the box represents the median, whisker extend to extremes, and individual values are shown. The error bars represent the SE. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. Different lowercase letters indicate statistically significant differences at the $P < 0.05$ level, as determined using a one-way ANOVA.

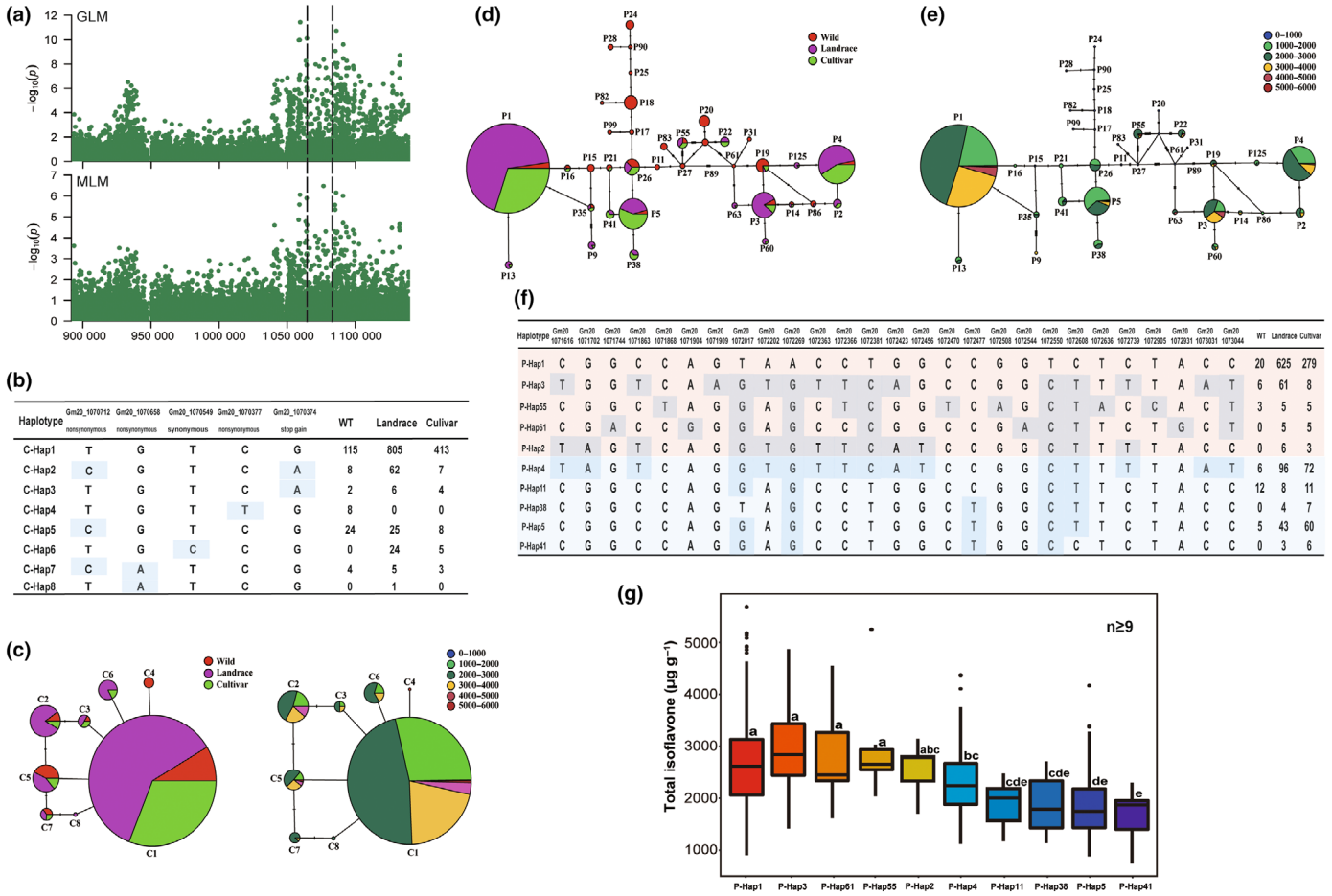


Fig. 5 *GmZFP7* haplotype analysis in soybean accessions. (a) Manhattan plots of the correlation analysis of the total isoflavone content and the SNPs, performed using a green linear model (GLM) and a mixed linear model (MLM). Dashed black lines indicate the *GmZFP7* region. (b) Haplotype analysis based on the *GmZFP7* coding region. (c) Median-joining network of the haplotype distribution and isoflavone analyses based on the *GmZFP7* coding region. Circle size is proportional to the sample quantity within a given haplotype. C, C-Hap. (d) Median-joining network of the haplotype distribution based on the *GmZFP7* promoter region. Circle size is proportional to the sample quantity within a given haplotype. (e) Median-joining network of the haplotype–isoflavone analyses based on the *GmZFP7* promoter region. Circle size is proportional to the sample quantity within a given haplotype. P, P-Hap. (f) Major haplotypes (haplotypes carried by more than nine accessions) based on the *GmZFP7* promoter region. (g) A box plot showing the isoflavone content of the major haplotypes (more than nine accessions) in the promoter region of *GmZFP7*. For boxplots, box represents interquartile range, the horizontal line inside the box represents the median, whisker extend to extremes. The isoflavone contents of haplotypes P-Hap18 and P-Hap20 are not shown due to limitations in the number of assays. P-Hap represents the haplotype according to SNPs in the *GmZFP7* promoter region. Different lowercase letters indicate statistically significant differences at the $P < 0.05$ level, Tukey's multiple comparisons test were used.

higher than 3000 μg g^{-1}. In the other group, 65% of the accessions had isoflavone contents lower than 2300 μg g^{-1} (Fig. 5g; Table S7). This finding suggested that SNPs at positions −564, −811, −1311, −1314, −1329, −1371, −1425, and −1687 of the *GmZFP7* promoter region clearly differentiated between high isoflavone accessions and low isoflavone accessions.

Discussion

Discovery of candidate genes involved in isoflavone biosynthesis

A large number of isoflavone-associated QTL have been reported; however, few have been further verified or fine-mapped (Wang *et al.*, 2015; Pei *et al.*, 2018). Previously, we fine-mapped the major QTL *qIF20-2* to within a genomic region of around 200 kbp (Pei *et al.*, 2018). In the current study, we cloned the gene *GmZFP7* for this locus, containing a zinc-finger motif and a conserved 'QALGGH' sequence found in other proteins involved in plant stress tolerance (Wang *et al.*, 2018). We therefore speculated that GmZFP7 might be involved in plant stress tolerance, which is consistent with the role of isoflavones in plants. Here, we showed that *GmZFP7* is expressed in all tissues and that its expression pattern in seeds is consistent with the accumulation pattern of isoflavones. A subcellular localization showed that GmZFP7 is present in the nucleus and cytoplasm (Fig. 1e,f), which is consistent with the localization of the related TF GmZFP3 known to be involved in drought resistance (Zhang *et al.*, 2016). We therefore inferred that GmZFP7 may be involved in the regulation of isoflavone biosynthesis and plays a role in the soybean response to stress.

GmZFP7 is involved in regulating isoflavone biosynthesis

The C2H2 zinc-finger TFs play important roles in plant interactions with microbes and stress responses (Wang *et al.*, 2018); however, there are few reports of zinc-finger TFs in the phenylpropanoid pathway. In the current study, we verified the function of GmZFP7 in the isoflavone pathway. We first performed *GmZFP7* overexpression, RNAi, and gene-editing experiments in the transgenic hairy roots, leaves, and seeds of soybean. The isoflavone content was significantly increased in the *GmZFP7*-overexpressing hairy roots and seeds and was significantly decreased in the RNAi-silenced hairy roots and mutant leaves and seeds. These results demonstrated that GmZFP7 is a new type of C2H2 zinc-finger TF, which acts as a positive regulator in the production of isoflavones. In previous studies, several MYB TFs (e.g. MYB176, MYB29, MYB58, MYB205, MYB39, and MYB100) were found to be involved in the regulation of isoflavone accumulation in soybean (Yi *et al.*, 2010; Li *et al.*, 2012; Liu *et al.*, 2013; Yan *et al.*, 2015; Chu *et al.*, 2017; Han *et al.*, 2017; Vadivel *et al.*, 2021); herein, we report a zinc-finger-type TF could also regulate isoflavone accumulation in soybean.

GmZFP7 acts as a dual-function TF in isoflavone regulation

Isoflavone biosynthesis is a branch of the phenylpropanoid pathway and shares common substrates with flavone, flavonol, and anthocyanin biosynthesis. Moreover, the gateway enzyme in isoflavone biosynthesis, IFS, competes with F3H and FNS for the common substrate naringenin to direct the metabolic flux into isoflavone rather than flavone, flavonol, and anthocyanin (Liu *et al.*, 2002; Dong & Lin, 2021). The ectopic expression of *GmIFS* in the *Arabidopsis f3h1dfr* mutant and the mutant of *Gmf3h1*, *Gmf3h2*, and *GmfnsII* in soybean both resulted in significantly higher isoflavone content (Liu *et al.*, 2002; Zhang *et al.*, 2020); therefore, this redirection of the metabolic flux could regulate the isoflavone content. In this study, we show that GmZFP7 could significantly increase the overall flux of the phenylpropanoid pathway by increasing the expression levels of *GmC4H*, *Gm4CL*, *GmCHS*, and *GmCHR*. Moreover, GmZFP7 could regulate the accumulation of isoflavones by increasing the expression of *GmIFS2* while inhibiting the expression of *GmF3H1*. It is difficult to assess the effects of increasing *GmIFS2* and inhibiting *GmF3H1* expression in metabolomics analysis due to the overall increase in phenylpropanoid metabolic flux, although some flavonols (e.g. quercitrin and nicotiflorin) mediated by GmF3H were decreased in the *GmZFP7*-overexpressed plants. Few studies have identified dual-function TFs in plants. A C2H2-type zinc-finger TF, Yin Yang1 (AtYY1), in *Arabidopsis* has been reported to have dual regulatory functions, possessing both a transcriptional activation domain and a repressor domain, which can be used as both a repressor in the GAL4 fusion system and an activator in the regulation of GABR1 to coordinate the balance of various systems in the ABA response network (Li *et al.*, 2016). In tumor studies, there are many reports about the complex biological functions of YY1 in humans, which can either activate or repress the expression of downstream genes according to the stimuli received by the cells and the association with other cytokines (Qiang *et al.*, 2011). In addition, ELF3 (variably called ESE-1, ERT, and ESX, an ETS family TF) has been reported to suppress the promoter activity of basal keratin 4 while simultaneously activating the late differentiation–linked SPRR2A promoter in humans (Brembeck *et al.*, 2000). These findings indicate that some TFs may have dual functions in regulatory mechanisms involved in both repression and activation, which can be achieved according to different signal stimuli or interactions with other TFs. However, whether the regulatory function of GmZFP7 changes based on an interaction with plant cytokines or other upstream regulatory signals is yet to be investigated.

Genetic diversity analysis of *GmZFP7* in soybean accessions

In breeding projects, the selection and accumulation of elite alleles is an effective strategy for improving target traits. Although many major genes and TFs involved in the regulation of isoflavone biosynthesis have been identified, little is known about the related genetic polymorphisms and elite haplotypes. The haplotype analysis performed in this study suggested that the variations in the promoter of *GmZFP7* play a more important role in isoflavone accumulation than those in the ORF. This finding is consistent with the similar levels of isoflavone accumulation when *GmZFP7L* or *GmZFP7N* was overexpressed or suppressed in hairy roots, despite *GmZFP7N* bearing two missense SNPs at its ORF region. Moreover, consistent with the haplotype analysis, higher expression levels of *GmZFP7* were observed in LHD than in NHZ during seed development, which may also be attributed to differences in the promoter region. Taken together, we

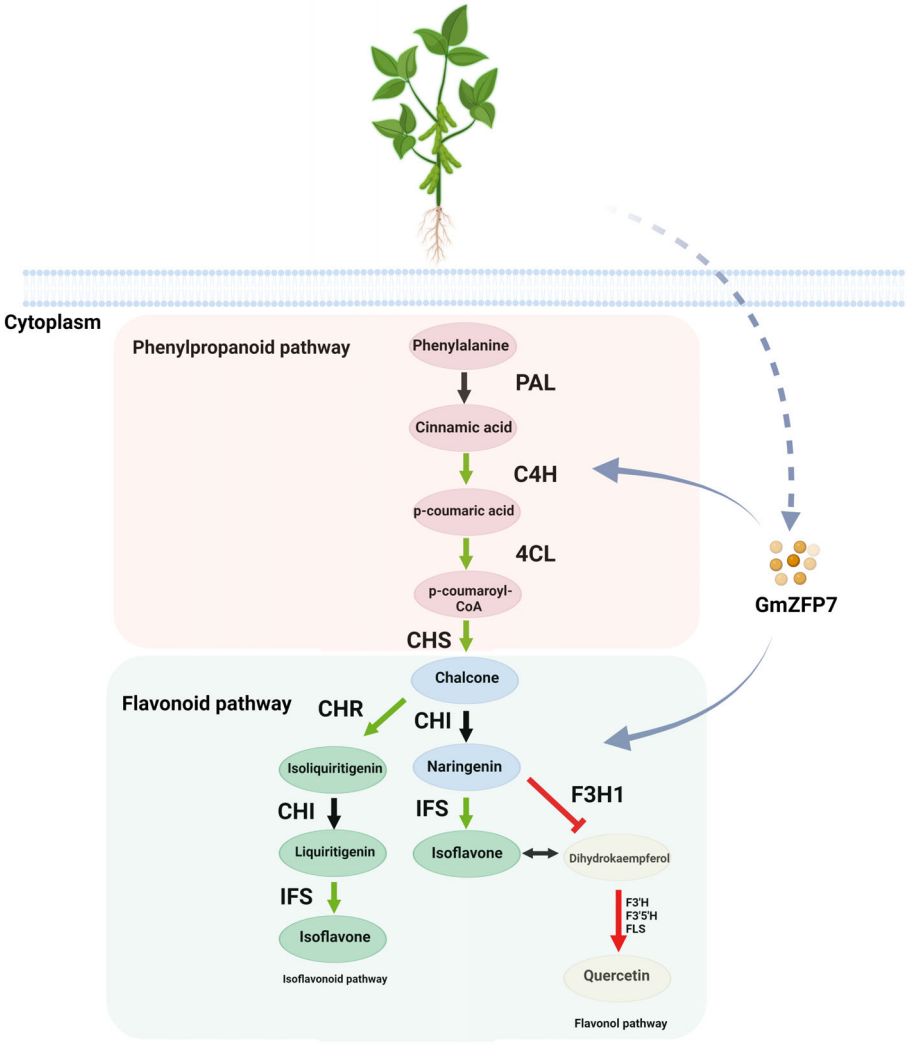


Fig. 6 GmZFP7-mediated regulatory model for soybean isoflavone biosynthesis. Black arrows indicate the relationship between isoflavone-related enzymes. The green arrow indicates a promoting effect. The red 'T' line and red arrow indicate an inhibitory effect. The dashed arrow indicates that the results are yet to be confirmed experimentally. The double arrows indicate competition. 4CL, 4-coumarate-CoA ligase; C4H, cinnamic acid 4-hydroxylase; CHI, chalcone isomerase; CHR, chalcone reductase; CHS, chalcone synthase; F3′5′H, flavonoid 3′5′-hydroxylase; F3′H, flavonoid 3′-hydroxylase; F3H1, flavanone 3-hydroxylase 1; FLS, flavonol synthase; IFS2, isoflavone synthase 2; PAL, phenylalanine ammonia-lyase.

speculate that the SNPs within the *GmZFP7* promoter are responsible for the differences in expression levels, thereby leading to differences in isoflavone accumulation.

Two high isoflavone haplotypes (P-Hap1 and P-Hap3) from the *GmZFP7* promoter region were considered to be elite haplotypes. The *GmZFP7* haplotype group A mainly consists of landrace accessions and has higher isoflavone contents, while group B consists of mainly cultivars and has lower isoflavone contents. These results could be partly explained by the adaptation of landraces to environmental stress conditions as isoflavones confer tolerance to these stresses (Azam *et al.*, 2020). However, breeding for good taste has been an important goal in soybean. So, isoflavones may be selected indirectly for their bitter taste and negative correlation with other important traits such as protein, palmitic acid, stearic acid, and oleic acid contents (Azam *et al.*, 2021). *GmZFP7* was also selected along with the isoflavone content in the process of soybean selection. Regulatory element analysis revealed that many stress-related elements (e.g. ABRE, CuRE, SURE, and WUN) of *GmZFP7* promotors differ among haplotypes (Fig. S11). These elements were reported to be related to the phytohormone signaling pathway and to be involved in multiple stress responses (Fujita *et al.*, 2005; Kropat *et al.*, 2005; Maruyama-Nakashita *et al.*, 2005; Peleg & Blumwald, 2011; Jiang *et al.*, 2014; Shi *et al.*, 2018). Both isoflavone and hormones are known to be involved in plant stress tolerance. Therefore, the differences in stress-related elements may be the reason for the differences in transcript levels of *GmZFP7* and isoflavone content among the haplotypes.

In summary, the C2H2 zinc-finger TF, GmZFP7, has a dual regulatory function, and can redirect metabolic flux to isoflavone by upregulating the expression of *GmC4H*, *Gm4CL*, *GmCHS*, *GmCHR*, and *GmIFS2* while downregulating the expression of *GmF3H1* (Fig. 6). These findings not only provide insights into the functional role of C2H2 zinc-finger TFs in the regulation of the phenylpropanoid pathway, but also provide a theoretical basis for soybean isoflavone improvement.

Acknowledgements

We thank Prof Lijuan Qiu for providing soybean accessions, Prof Wensheng Hou for assistance in *GmZFP7* gene transformation, and Prof Wenxue Li for critical reading of this manuscript

and valuable suggestions. This work was financially supported by the National Natural Science Foundation of China (32272178, 32161143033, 32001574 and 31671716) and the Agricultural Science and Technology Innovation Program of CAAS (2060203-2).

Competing interests

None declared.

Author contributions

BL and JS conceived the research project, designed the experiment, and edited the manuscript. YF, JQ, RP, LT and YL performed the experiments. MA collected the phenotypic data. YF, SZ and JL analyzed the data. YF wrote the manuscript. JL, KGAB, ASS and ZZ edited the manuscript. All authors discussed the data and contributed to the manuscript. YF, SZ and JL contributed equally to this work.

ORCID

Bin Li https://orcid.org/0000-0002-9452-6083
Junming Sun https://orcid.org/0000-0002-5585-0016

Data availability

The resequencing data of soybean cultivars LHD and NHZ have been submitted to the National Genomics Data Center, GEO (accession number: GVM000002).

References

An L, Zhou Z, Su S, Yan A, Gan Y. 2012. *GLABROUS INFLORESCENCE STEMS* (*GIS*) is required for trichome branching through gibberellic acid signaling in Arabidopsis. *Plant & Cell Physiology* **53**: 457–469.

Azam M, Zhang S, Abdelghany AM, Shaibu AS, Feng Y, Li Y, Tian Y, Hong H, Li B, Sun J. 2020. Seed isoflavone profiling of 1168 soybean accessions from major growing ecoregions in China. *Food Research International* **130**: 108957.

Azam M, Zhang S, Qi J, Abdelghany AM, Shaibu AS, Ghosh S, Feng Y, Huai Y, Gebregziabher BS, Li J *et al.* 2021. Profiling and associations of seed nutritional characteristics in Chinese and USA soybean cultivars. *Journal of Food Composition and Analysis* **98**: 103803.

Bian S, Li R, Xia S, Liu Y, Jin D, Xie X, Dhaubhadel S, Zhai L, Wang J, Li X. 2018. Soybean *CCA1-like MYB* transcription factor *GmMYB133* modulates isoflavonoid biosynthesis. *Biochemical and Biophysical Research Communications* **507**: 324–329.

Brembeck FH, Opitz OG, Libermann TA, Rustgi AK. 2000. Dual function of the epithelial specific ets transcription factor, ELF3, in modulating differentiation. *Oncogene* **19**: 1941–1949.

Chen L, Cai Y, Liu X, Guo C, Sun S, Wu C, Jiang B, Han T, Hou W. 2018a. Soybean hairy roots produced in vitro by *Agrobacterium* rhizogenes-mediated transformation. *The Crop Journal* **6**: 162–171.

Chen L, Cai Y, Liu X, Yao W, Guo C, Sun S, Wu C, Jiang B, Han T, Hou W. 2018b. Improvement of soybean *Agrobacterium*-mediated transformation efficiency by adding glutamine and asparagine into the culture media. *International Journal of Molecular Sciences* **19**: 3039.

Chu S, Wang J, Zhu Y, Liu S, Zhou X, Zhang H, Wang C-e, Yang W, Tian Z, Cheng H *et al.* 2017. An R2R3-type MYB transcription factor, GmMYB29, regulates isoflavone biosynthesis in soybean. *PLoS Genetics* **13**: e1006770.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. 2012. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics* **3**: 35.

Dixon RA, Achnine L, Kota P, Liu C-J, Reddy MSS, Wang L. 2002. The phenylpropanoid pathway and plant defence – a genomics perspective. *Molecular Plant Pathology* **3**: 371–390.

Dong NQ, Lin HX. 2021. Contribution of phenylpropanoid metabolism to plant development and plant-environment interactions. *Journal of Integrative Plant Biology* **63**: 180–209.

Fraser CM, Chapple C. 2011. The phenylpropanoid pathway in Arabidopsis. *The Arabidopsis Book* **9**: e0152.

Fujita Y, Fujita M, Satoh R, Maruyama K, Parvez MM, Seki M, Hiratsu K, Ohme-Takagi M, Shinozaki K, Yamaguchi-Shinozaki K. 2005. AREB1 is a transcription activator of novel ABRE-dependent ABA signaling that enhances drought stress tolerance in Arabidopsis. *Plant Cell* **17**: 3470–3488.

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

Hsieh P-L, Liao Y-W, Hsieh C-W, Chen P-N, Yu C-C. 2020. Soy isoflavone genistein impedes cancer stemness and mesenchymal transition in head and neck cancer through activating *miR-34a/RTCB* Axis. *Nutrients* **12**: 1924.

Jiang Y, Duan Y, Yin J, Ye S, Zhu J, Zhang F, Lu W, Fan D, Luo K. 2014. Genome-wide identification and characterization of the Populus WRKY transcription factor family and analysis of their expression in response to biotic and abiotic stresses. *Journal of Experimental Botany* **65**: 6629–6644.

Jung W, Yu O, Lau S-MC, O'Keefe DP, Odell J, Fader G, McGonigle B. 2000. Identification and expression of isoflavone synthase, the key enzyme for biosynthesis of isoflavones in legumes. *Nature Biotechnology* **18**: 208–212.

Kanegae T, Wada M. 1998. Isolation and characterization of homologues of plant blue-light photoreceptor (cryptochrome) genes from the fern *Adiantum capillus-veneris*. *Molecular & General Genetics* **259**: 345–353.

Kropat J, Tottey S, Birkenbihl RP, Depège N, Huijser P, Merchant S. 2005. A regulator of nutritional copper signaling in *Chlamydomonas* is an SBP domain protein that recognizes the GTAC core of copper response element. *Proceedings of the National Academy of Sciences, USA* **102**: 18730–18735.

Kudou S, Fleury Y, Welti D, Magnolato D, Uchida T. 1991. Malonyl isoflavone glycosides in soybean seeds (*Glycine max* Merrill). *Agricultural and Biological Chemistry* **55**: 2227–2233.

Leigh JW, Bryant D. 2015. PopART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution* **6**: 1110–1116.

Li B, Fan S, Yu F, Chen Y, Zhang S, Han F, Yan S, Wang L, Sun J. 2017. High-resolution mapping of QTL for fatty acid composition in soybean using specific-locus amplified fragment sequencing. *Theoretical and Applied Genetics* **130**: 1467–1479.

Li B, Tian L, Zhang J, Huang L, Han F, Yan S, Wang L, Zheng H, Sun J. 2014. Construction of a high-density genetic map based on large-scale markers developed by specific length amplified fragment sequencing (SLAF-seq) and its application to QTL analysis for isoflavone content in *Glycine max*. *BMC Genomics* **15**: 1086.

Li T, Wu X-Y, Li H, Song J-H, Liu J-Y. 2016. A dual-function transcription factor, AtYY1, is a novel negative regulator of the *Arabidopsis* ABA response network. *Molecular Plant* **9**: 650–661.

Li X, Chen L, Dhaubhadel S. 2012. 14-3-3 proteins regulate the intracellular localization of the transcriptional activator GmMYB176 and affect isoflavonoid synthesis in soybean. *The Plant Journal* **71**: 239–250.

Li Y, Ahmed F, Ali S, Philip PA, Kucuk O, Sarkar FH. 2005. Inactivation of nuclear factor κB by soy isoflavone genistein contributes to increased apoptosis induced by chemotherapeutic agents in human cancer cells. *Cancer Research* **65**: 6934–6942.

Liu C-J, Blount JW, Steele CL, Dixon RA. 2002. Bottlenecks for metabolic engineering of isoflavone glycoconjugates in *Arabidopsis*. *Proceedings of the National Academy of Sciences, USA* **99**: 14578–14583.

Liu H, Ding Y, Zhou Y, Jin W, Xie K, Chen L-L. 2017. CRISPR-P 2.0: an improved CRISPR-Cas9 tool for genome editing in plants. *Molecular Plant* **10**: 530–532.

Liu X, Yuan L, Xu L, Xu Z, Huang Y, He X, Ma H, Yi J, Zhang D. 2013. Overexpression of *GmMYB39* leads to an inhibition of the isoflavonoid biosynthesis in soybean (*Glycine max*. L). *Plant Biotechnology Reports* **7**: 445–455.

Maruyama-Nakashita A, Nakamura Y, Watanabe-Takahashi A, Inoue E, Yamaya T, Takahashi H. 2005. Identification of a novel *cis*-acting element conferring sulfur deficiency response in Arabidopsis roots. *The Plant Journal* **42**: 305–314.

Ng M-S, Ku Y-S, Yung W-S, Cheng S-S, Man C-K, Yang L, Song S, Chung G, Lam H-M. 2021. MATE-type proteins are responsible for isoflavone transportation and accumulation in soybean seeds. *International Journal of Molecular Sciences* **22**: 12017.

Ollberding NJ, Lim U, Wilkens LR, Setiawan VW, Shvetsov YB, Henderson BE, Kolonel LN, Goodman MT. 2012. Legume, soy, tofu, and isoflavone intake and endometrial cancer risk in postmenopausal women in the multiethnic cohort study. *Journal of the National Cancer Institute* **104**: 67–76.

Pandith SA, Ramazan S, Khan MI, Reshi ZA, Shah MA. 2019. Chalcone synthases (CHSs): the symbolic type III polyketide synthases. *Planta* **251**: 15.

Payne T, Johnson SD, Koltunow AM. 2004. *KNUCKLES (KNU)* encodes a C2H2 zinc-finger protein that regulates development of basal pattern elements of the *Arabidopsis* gynoecium. *Development* **131**: 3737–3749.

Pei R, Zhang J, Tian L, Zhang S, Han F, Yan S, Wang L, Li B, Sun J. 2018. Identification of novel QTL associated with soybean isoflavone content. *The Crop Journal* **6**: 244–252.

Peleg Z, Blumwald E. 2011. Hormone balance and abiotic stress tolerance in crop plants. *Current Opinion in Plant Biology* **14**: 290–295.

Qiang Z, Stovall DB, Inoue K, Sui G. 2011. The oncogenic role of Yin Yang 1. *Critical Reviews in Oncogenesis* **16**: 163–197.

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

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Molecular Biology and Evolution* **34**: 3299–3302.

Schröder J, Schröder G. 1990. Stilbene and chalcone synthases: related enzymes with key functions in plant-specific pathways. *Journal of Biosciences* **45**: 1–8.

Shi D, Ren A, Tang X, Qi G, Xu Z, Chai G, Hu R, Zhou G, Kong Y. 2018. MYB52 negatively regulates pectin demethylesterification in seed coat mucilage. *Plant Physiology* **176**: 2737–2749.

Sun JM, Sun BL, Han FX, Yan SR, Yang H, Akio K. 2011. Rapid HPLC method for determination of 12 isoflavone components in soybean seeds. *Agricultural Sciences in China* **10**: 70–77.

Takeda S, Matsumoto N, Okada K. 2004. *RABBIT EARS*, encoding a SUPERMAN-like zinc finger protein, regulates petal development in *Arabidopsis thaliana*. *Development* **131**: 425–434.

van Tunen AJ, Mur LA, Recourt K, Gerats AG, Mol JN. 1991. Regulation and manipulation of flavonoid gene expression in anthers of petunia: the molecular basis of the Po mutation. *Plant Cell* **3**: 39–48.

Vadivel AKA, McDowell T, Renaud JB, Dhaubhadel S. 2021. A combinatorial action of GmMYB176 and GmbZIP5 controls isoflavonoid biosynthesis in soybean (*Glycine max*). *Communications Biology* **4**: 356.

Wang K, Ding Y, Cai C, Chen Z, Zhu C. 2018. The role of C2H2 zinc finger proteins in plant responses to abiotic stresses. *Physiologia Plantarum* **165**: 690–700.

Wang Y, Han Y, Zhao X, Li Y, Teng W, Li D, Zhan Y, Li W. 2015. Mapping isoflavone QTL with main, epistatic and QTL × environment effects in recombinant inbred lines of soybean. *PLoS ONE* **10**: e0118447.

Wu D, Li D, Zhao X, Zhan Y, Teng W, Qiu L, Zheng H, Li W, Han Y. 2020. Identification of a candidate gene associated with isoflavone content in soybean seeds using genome-wide association and linkage mapping. *The Plant Journal* **104**: 950–963.

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

Yang G, Shu X-O, Li H-L, Chow W-H, Wen W, Xiang Y-B, Zhang X, Cai H, Ji B-T, Gao Y-T *et al.* 2013. Prediagnosis soy food consumption and lung cancer survival in women. *Journal of Clinical Oncology* **31**: 1548–1553.

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

Yu O, Jung W, Shi J, Croes RA, Fader GM, McGonigle B, Odell JT. 2000. Production of the isoflavones genistein and daidzein in non-legume dicot and monocot tissues. *Plant Physiology* **124**: 781–794.

Yu O, Shi J, Hession AO, Maxwell CA, McGonigle B, Odell JT. 2003. Metabolic engineering to increase isoflavone biosynthesis in soybean seed. *Phytochemistry* **63**: 753–763.

Zanetti ME, Blanco FA, MPı B, Battaglia M, Aguilar OM. 2010. A C subunit of the plant nuclear factor NF-Y required for rhizobial infection and nodule development affects partner selection in the common bean–*Rhizobium etli* symbiosis. *Plant Cell* **22**: 4142–4157.

Zhang D, Tong J, Xu Z, Wei P, Xu L, Wan Q, Huang Y, He X. 2016. Soybean C2H2-type zinc finger protein GmZFP3 with conserved QALGGH motif negatively regulates drought responses in transgenic *Arabidopsis*. *Frontiers in Plant Science* **7**: 325.

Zhang P, Du H, Wang J, Pu Y, Yang C, Yan R, Yang H, Cheng H, Yu D. 2020. Multiplex CRISPR/Cas9-mediated metabolic engineering increases soya bean isoflavone content and resistance to soya bean mosaic virus. *Plant Biotechnology Journal* **18**: 1384–1395.

Zhang X, Liu C-J. 2014. Multifaceted regulations of gateway enzyme phenylalanine ammonia-lyase in the biosynthesis of phenylpropanoids. *Molecular Plant* **8**: 17–27.

Supporting Information

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

Fig. S1 Experimental effect of soybean hairy roots.

Fig. S2 Screening of transgenic *GmZFP7*-overexpressing soybean plants.

Fig. S3 Gene expression profiles within the *qIF20-2* interval in LHD and NHZ during different seed developmental stages.

Fig. S4 Comparison of amino acid sequences of proteins encoded by *Gm20MYB* and *GmZFP7* between LHD and NHZ.

Fig. S5 Gene expression and isoflavonoid content in the *Gm20MYB*-overexpressing and RNAi-silenced soybean hairy roots.

Fig. S6 Fluorescence detection of the transient expression of *GmZFP7* in tobacco leaves.

Fig. S7 The relative expression level of isoflavone biosynthesis-related genes in *GmZFP7* mutant and overexpressing leaves.

Fig. S8 Metabolomic results of *GmZFP7* gene editing and overexpressing soybean seeds.

Fig. S9 Isoflavone content of each *GmZFP7* coding region haplotype.

Fig. S10 Isoflavone content of each *GmZFP7* promoter region haplotype.

Fig. S11 Differential motifs between haplotypes in the *GmZFP7* promoter region.

Table S1 The qRT-PCR primers used in this research.

Table S2 The quality of sequencing data for the two cultivars.

Table S3 Statistical analysis of the sequencing depth and coverage for the two pools.

Table S4 Candidate genes and functional annotation for isoflavone content in the fine-mapping interval (Gm20_49565–Gm20_55277) in soybean.

Table S5 Metabolome analysis results.

Table S6 Haplotype analysis results of the promoter region of *GmZFP7* in soybean accessions.

Table S7 Promoter haplotype information for *GmZFP7* in soybean accessions.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.