PNAS RESEARCH ARTICLE | PLANT BIOLOGY

Dynamics of *cis*-regulatory sequences and transcriptional divergence of duplicated genes in soybean

Chao Fang[a], Mingyu Yang[b,c], Yuecheng Tang[b,c], Ling Zhang[d], Hainan Zhao[a], Hejia Ni[c], Qingshan Chen[c], Fanli Meng[b,1], and Jiming Jiang[a,e,f,1]

Edited by Douglas Soltis, University of Florida, Gainesville, FL; received March 7, 2023; accepted September 19, 2023

Transcriptional divergence of duplicated genes after whole genome duplication (WGD) has been described in many plant lineages and is often associated with subgenome dominance, a genome-wide mechanism. However, it is unknown what underlies the transcriptional divergence of duplicated genes in polyploid species that lack subgenome dominance. Soybean is a paleotetraploid with a WGD that occurred 5 to 13 Mya. Approximately 50% of the duplicated genes retained from this WGD exhibit transcriptional divergence. We developed accessible chromatin region (ACR) datasets from leaf, flower, and seed tissues using MNase-hypersensitivity sequencing. We validated enhancer function of several ACRs associated with known genes using CRISPR/Cas9-mediated genome editing. The ACR datasets were used to examine and correlate the transcriptional patterns of 17,111 pairs of duplicated genes in different tissues. We demonstrate that ACR dynamics are correlated with divergence of both expression level and tissue specificity of individual gene pairs. Gain or loss of flanking ACRs and mutation of *cis*-regulatory elements (CREs) within the ACRs can change the balance of the expression level and/ or tissue specificity of the duplicated genes. Analysis of DNA sequences associated with ACRs revealed that the extensive sequence rearrangement after the WGD reshaped the CRE landscape, which appears to play a key role in the transcriptional divergence of duplicated genes in soybean. This may represent a general mechanism for transcriptional divergence of duplicated genes in polyploids that lack subgenome dominance.

gene duplication | transcriptional divergence | accessible chromatin region | MNase hypersensitivity | CRISPR/Cas

Significance

We analyzed the transcriptional regulation of 17,111 pairs of duplicated genes in soybean, which were derived from a whole genome duplication (WGD) that occurred about 5 to 13 Mya. We demonstrate that gain or loss of flanking regulatory sequences and mutation of *cis*-regulatory elements within the sequences can change the balance of the expression level and/or tissue specificity of duplicated genes. These results support our hypothesis that dynamics of the *cis*-regulatory sequences after the recent WGD event has played an important role in transcriptional divergence of duplicated genes in soybean, which may represent a general mechanism for divergence of duplicated genes in polyploids that lack subgenome dominance.

Whole genome duplication (WGD) is a universal phenomenon associated with the evolution of higher eukaryotes and has been reported widely in plants (1–5). WGDs, also known as polypoid events, double the number of genes in eukaryotes and therefore provide abundant genetic materials to adaptation and evolution for the new species (1, 3, 6). WGD events may generate classical polyploids, including both autopolyploids and allopolyploids (2). Due to functional redundancy, duplicated genes often fractionate (are lost) and/or diverge after a WGD event. Duplicated gene pairs may reach several different evolutionary fates: 1) one copy loses its function (pseudogenization); 2) one copy acquires a new function (neofunctionalization); 3) two copies partition functions of their ancestor (subfunctionalization); and 4) both copies retain functions of their ancestor (5, 7–9).

Divergence of duplicated genes often occurs at the transcriptional level, which may create a condition favorable for the retention of both copies of the gene (8, 10, 11). Transcriptional divergence of duplicated genes in polyploids can be resulted from "subgenome dominance," in which genes from one of the parental genomes (subgenomes) are preferentially retained or gain higher levels of expression than those from other subgenomes (12). Subgenome dominance has been documented in diverse polyploids (12–20). For example, maize is an ancient tetraploid and experienced a WGD 5 to 12 Mya (21). The maize genome has undergone uneven loss of genes from the two ancestral subgenomes (22) and exhibits overexpression of genes from the subgenome that has experienced less gene loss (15, 20). This subgenome dominance can be attributed to subgenome-specific DNA methylation and distribution of transposable elements (TEs) (12, 23) or response to pathogen infection (24, 25). Nevertheless, not all polyploids exhibit subgenome dominance (12), and the lack of subgenome dominance was proposed to be skewed toward autopolyploids (26). It remains unknown what causes the transcriptional divergence of duplicated genes in autopolyploids or allopolyploids with highly similar subgenomes that lack subgenome dominance.

Soybean (*Glycine max*) is one of the most important crops in the world and is a major legume source of oil and protein. Soybean has gone through two rounds of WGDs that occurred at ~59 and 5 to 13 Mya, respectively (27). Nearly 75% of soybean genes have more than one copy in the genome (27). In addition, approximately 50% of the duplicated

Author contributions: C.F., F.M., and J.J. designed research; C.F., M.Y., Y.T., L.Z., and H.N. performed research; Q.C. contributed new reagents/analytic tools; C.F., H.Z., F.M., and J.J. analyzed data; and C.F. and J.J. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: mengfanli@iga.ac.cn or jiangjm@msu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2303836120/-/DCSupplemental.

Published October 23, 2023.

genes from the last WGD exhibit transcriptional divergence (28). Unlike maize, subgenome dominance was not detected in soybean (26), which was attributed to the fact that the two subgenomes were likely genetically highly similar prior to the polyploidization event (29). Thus, soybean provides a useful model to investigate the mechanisms underlying transcriptional divergence of highly similar duplicated genes. We hypothesized that dynamic mutation and rearrangement of the *cis*-regulatory sequences after WGD is a potential force for the transcriptional divergence of duplicated genes in soybean. We used MNase-hypersensitivity sequencing (MH-seq) to develop genome-wide datasets for accessible chromatin regions (ACRs) in the soybean genome. ACR datasets were then used to examine and correlate the transcription patterns of duplicated gene pairs in different soybean tissues. We demonstrate that dynamics of *cis*-regulatory elements (CREs) within the ACRs, including CRE mutations and gain/loss of ACRs, play an important role in the transcriptional divergence of the duplicated genes in soybean.

Results

Development of MH-Seq Datasets in Soybean. We performed MH-seq (30) to detect ACRs in the soybean genome. We developed two MH-seq libraries from each of three tissues (leaves, flowers, and seeds) from soybean cultivar Williams 82. A total of 256 million MH-seq sequence reads were obtained and mapped to the Williams 82 reference genome (27). The MH-seq reads were clearly enriched toward the distal ends of all soybean chromosomes that are more enriched with genes than the pericentromeric regions (27) (*SI Appendix*, Fig. S1). In addition, a strong correlation (r = 0.83 to 0.93) was found between the two biological replicates of each tissue. Therefore, data from two biological replicates were combined for identification of MNase hypersensitive site (MHS) using F-seq (31). We identified 97,993, 87,407, and 105,381 MHSs in leaf, flower, and seed tissues, respectively.

We grouped the MHSs into four categories: 1) "upstream MHSs," which are located within 1 kb upstream of the transcriptional start site (TSS); 2) "genic MHSs," which are located within the 5′UTR, exon, intron, and 3′UTR; 3) "downstream MHSs," which are located within 1 kb downstream of the transcriptional terminal site (TTS); and 4) the remaining MHSs were designated as "intergenic MHSs." MHSs located in genic and ±1 kb flanking regions accounted for 52 to 56% of the total MHSs (*SI Appendix*, Fig. S2).

Comparative Analysis of Soybean ACRs Detected by MH-Seq and ATAC-Seq. We recently found in *Arabidopsis thaliana* that MH-seq has superior resolution and reveals ACRs that cannot be identified by the ATAC-seq technique (32). To investigate a potentially similar phenomenon in soybean, we downloaded ~470 million sequence reads of two ATAC-seq libraries generated from Williams 82 (33). We compared the ACRs identified by MH-seq and ATAC-seq, respectively, since leaf tissues from Williams 82 at the same developmental stage (V2) were used in both MH-seq and ATAC-seq experiments. We found that ~87% of ATAC-seq peaks were covered by MHSs. Strikingly, only 23% of MHSs were covered by ATAC-seq peaks (Fig. 1*A*), which are referred to as "common MHS" (cMHSs). Nearly 77% (75,581) of the MHSs were not covered by ATAC-seq peaks, which were referred to as "specific MHS" (sMHSs). Huang et al. (2022) recently reported ATAC-seq datasets from six soybean tissues (34). The leaf, flower, and seed tissues used in the study were at the similar developmental stages as the same tissues used in our study. Again, we found that a high proportion of ATAC-seq peaks (leaf: 96%; flower: 87%; seed: 84%) were covered by MHSs, but only a low proportion of MHSs (leaf: 32%; flower: 28%; seed: 33%) were covered by ATAC-seq peaks.

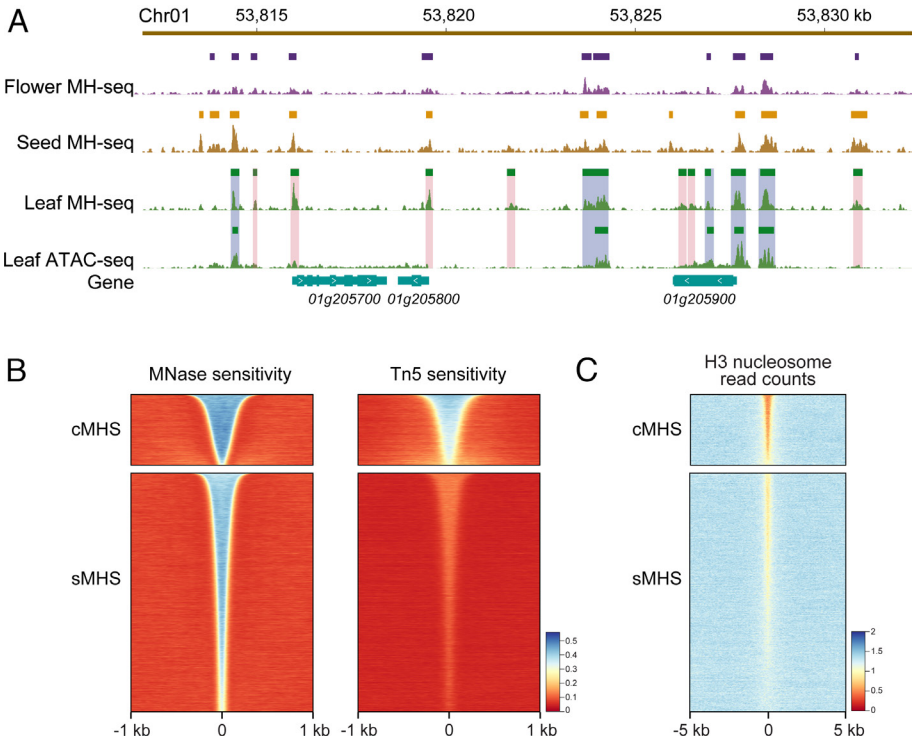

Fig. 1. Identification of MHSs in the soybean genome. (*A*) A selected genomic region of soybean chromosome 1 showing MHSs identified in leaf, flower, and seed tissues and ATAC-seq peaks in leaf. ATAC-seq data derived from Lu et al. (33). Pink boxes mark sMHSs identified by MH-seq only. Blue boxes mark cMHSs identified by both MH-seq and ATAC-seq. (*B*) Sensitivity to MNase (*Left*) or Tn5 (*Right*) of genomic regions associated with cMHSs and sMHSs, respectively. Note: genomic regions associated with sMHSs showed a very low level of Tn5 sensitivity. (*C*) Nucleosome occupancy of the genomic regions associated with cMHSs and sMHSs, respectively.

To further confirm the lack of ATAC-seq sequence enrichment at the sMHSs, we mapped the ATAC-seq reads (33) to the Williams 82 reference genome and identified ~250 million uniquely mapped reads. The numbers of MH-seq reads and ATAC-seq reads associated with each cMHS and sMHS peak were compared after being normalized by the total number of uniquely mapped reads. This comparison revealed that the Tn5 sensitivity associated with sMHSs is significantly lower than the MNase sensitivity with sMHSs (Fig. 1*B*). We also analyzed the nucleosome occupancy of sMHSs using histone H3 chromatin immunoprecipitation-sequencing (ChIP-seq) dataset from the same study (33). A clear eviction of nucleosomes was associated with both sMHSs and cMHSs (Fig. 1*C*), confirming that genomic regions containing the sMHSs indeed lack nucleosomes.

To further analyze the soybean ACRs that were not detected by ATAC-seq, we examined gene expression using the RNA-seq data provided by the same ATAC-seq study using leaf tissue from Williams 82 (33). We identified 40,269 expressed genes (FPKM > 0) and classified these active genes into three groups: 1) Expressed genes associated with ATAC-seq peaks, which were named as "ATAC-seq genes" (n = 13,926) (*SI Appendix*, Fig. S3*A*). The ATAC-seq genes showed an average FPKM of 21.73; 2) Expressed genes that were not associated with ATAC-seq peaks, but were associated with sMHSs, which were named as "sMHS genes" (n = 19,648). The sMHS genes showed an average FPKM of 13.72; 3) The remaining genes are not associated with either ATAC-seq peaks or sMHSs and were named "Peak-free genes" (n = 6,695). The peak-free genes showed an average FPKM of 5.94. Thus, the expression level of sMHS genes is substantial, although lower than the ATAC-seq genes (*SI Appendix*, Fig. S3*A*). We also examined the MNase/Tn5 sensitivity levels around 5′ regions of the genes. The ATAC-seq genes showed a high sensitivity level for both Tn5 and MNase (*SI Appendix*, Fig. S3*B*). However, an appreciable level of MNase sensitivity, but not Tn5 sensitivity, was detected at the 5′ of the sMHS genes (*SI Appendix*, Fig. S3*B*).

Collectively, the comparative MH-seq and ATAC-seq analysis supported our previous conclusion in *A. thaliana* that MNase is capable of accessing chromatin regions that are not accessible to Tn5 (32). Notably, only ~22% of MHSs were not covered by ATAC-seq peaks in *A. thaliana* (32), whereas nearly 67 to 77% of the MHSs were not covered by ATAC-seq peaks in soybean, suggesting that ATAC-seq is particularly less effective in identifying ACRs in soybean. This could be due to a reduced Tn5 activity in soybean cells and/or unique characteristics associated with soybean chromatin.

Functional Validation of MHSs Related to Flowering and Seed Development. The in vivo function of MHSs can be validated by deletion analysis using CRISPR/Cas-based genome editing (35, 36). We attempted to validate the function of a few selected MHSs associated with known soybean genes. *E1* is a soybean-specific maturity gene and encodes a B3-like protein. *E1* has the largest effect on flowering time in soybean (37, 38). Transcription of *E1* is activated in leaves under long day-length (LD) conditions and inhibits flowering by down-regulating *FT2a* and *FT5a*, which are the orthologues of the *A. thaliana FLOWERING LOCUS T* (*FT*) gene (37, 39). Although the expression of *E1* has been profiled extensively, the transcriptional regulation of this gene is largely unknown.

We identified four MHSs associated with *E1* from the leaf MH-seq dataset, two at the 5′ and two at the 3′ of *E1*, respectively (Fig. 2*A*). One MHS (123 bp, an sMHS) is located ~3.6 kb downstream of *E1* and was not detected by the ATAC-seq datasets. This MHS contains a GGGACCAC motif related to the TEOSINTE BRANCHED 1, CYCLOIDEA, and PCF (TCP) transcription factors (TFs) (Fig. 2*A*), which are known for their roles in regulating flowering in *A. thaliana* (40, 41). We developed three homozygous deletion lines using CRISPR/Cas9. The deletions span 163 to 330 bp, including the TCP-binding motif (Fig. 2*A*). We grew the deletion lines under LD conditions and collected leaf tissues at 35 days after germination (DAG) for gene expression analysis. *E1* expression decreased by 29 to 53% in the deletion lines compared to the wild type (Fig. 2*B*). Concordantly, the expression of *FT2a* and *FT5a* was significantly higher ($P < 0.05$, *t* test) in the deletion lines than in the wild type (Fig. 2*B*). We counted the number of days from sowing to flowering for all plants. The deletion lines flowered 2 to 4 d earlier than the wild type (Fig. 2 *C*–*E*). These results showed that this target MHS plays a role in regulating *E1* expression under LD condition, which is likely controlled by the TCP TFs in soybean.

Glyma.03g229700 encodes an amino acid transporter protein (42) and is highly expressed in seed tissues (*SI Appendix*, Fig. S4*B*). A single MHS at the putative promoter region was detected in leaf and seed tissues. However, five additional MHSs around and within the gene were detected only in the seed MH-seq dataset (*SI Appendix*, Fig. S4*A*), suggesting that these MHSs are likely responsible for the seed-specific expression of this gene. We selected a 144-bp MHS (an sMHS), which is located upstream of the putative promoter and was not detected in ATAC-seq datasets, for CRISPR/Cas experiment. We developed three homozygous deletion lines, including T22 (270 bp, loss of all 144 bp of the MHS),

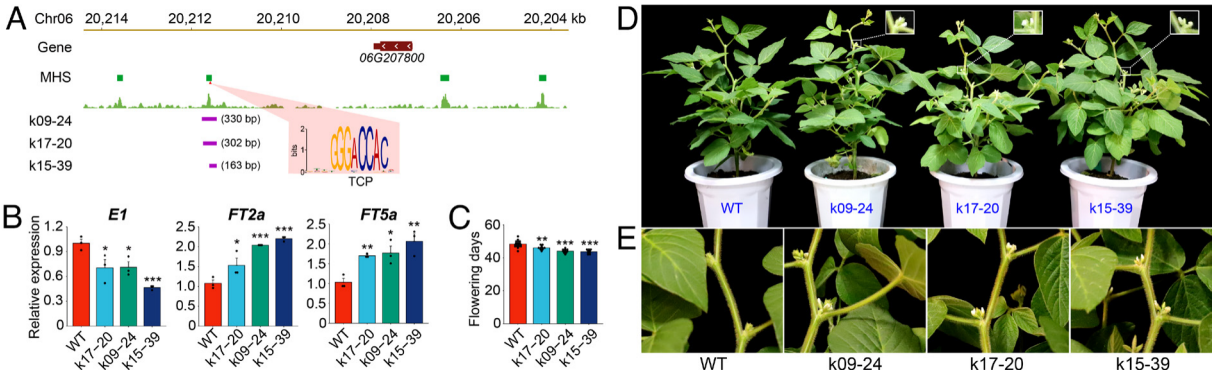

Fig. 2. Functional validation of an MHS associated with the *E1* gene. (*A*) Development of three homozygous CRISPR/Cas deletion lines (purple rectangles) spanning the target MHS. The red arrowhead indicates the position of a TCP motif within the MHS. (*B*) Expressions of *E1*, *FT2a*, and *FT5a* in leaf tissue of wild type (n = 3) and three deletion lines (n = 3) at 35 DAG. (*C*) Numbers of flowering days of wild type (n = 18) and three deletion lines (n = 9 to 10) at 35 DAG. **P* < 0.05, ***P* < 0.01, and ****P* < 0.001, *t* test. (*D*) Plant phenotype of wild type and three deletion lines under LD condition at 45 DAG. One raceme from each of the deletion lines is exemplified. (*E*) Phenotype of a part of soybean plants containing two racemes at 45 DAG.

T2 (175 bp, loss of 92 bp of the MHS), and T27 (117 bp, loss of 54 bp of the MHS) (*SI Appendix*, Fig. S4*A*). R4-stage seeds from the deletion lines were collected for gene expression analysis. The expression levels of *Glyma.03g229700* decreased by 10 to 63% in the deletion lines compared to the wild type (*SI Appendix*, Fig. S4*C*). These results validated the contribution of this MHS to the seed-specific expression of *Glyma.03g229700*.

Tissue-Specific MHSs and Gene Expression. We further explored the role of MHSs in tissue-specific gene expression. Tissue-specific MHSs are defined as MHSs exhibiting significantly higher levels of MNase sensitivity in one tissue than in both of the two other tissues (*Materials and Methods*). We identified 1,036, 459, and 2,086 MHSs that are specific to the leaf, flower, and seed tissue, respectively. We then compared the expression levels of genes associated with the tissue-specific MHSs. Genes associated with tissue-specific MHSs showed significantly high levels of expression in the respective tissue, while the genes associated with MHSs shared by three tissues showed similar expression levels in different tissues (*SI Appendix*, Fig. S5*A*). To test whether both cMHSs and sMHSs play a similar role in tissue-specific gene expression, we separated the leaf-specific MHSs into leaf-specific cMHSs and leaf-specific sMHSs, respectively. We then examined the impact of these two types of MHSs on gene expression. Genes associated with either leaf-specific cMHSs or leaf-specific sMHSs showed higher expression levels in leaves than in flowers and seeds (*SI Appendix*, Fig. S6*A*).

We used the MEME tool (43) to search for enriched DNA sequence motifs in the tissue-specific MHSs using the shared MHSs as a control. A total of seven enriched DNA motifs were identified from the leaf-specific MHSs. Four of these motifs are related to regulation of leaf development or flowering time (*SI Appendix*, Fig. S5*B*). For example, one of the enriched motifs, TGGTCC (P = 4.3e-2), is associated with TCP TFs, which are known to be involved in leaf development (44). We identified a motif CTRGCT (*SI Appendix*, Fig. S5*B*). Genes, which are associated with CTRGCT-containing MHSs and are differentially expressed in the leaf tissue, were found to be associated with leaf development, photosynthesis, and photomorphogenesis (*SI Appendix*, Fig. S5*C* and Table S1). For example, *Glyma.17g055200* encodes a GATA TF and plays an important role in regulating chlorophyll biosynthesis in soybean (45). A leaf-specific MHS containing a CTRGCT motif was identified in the 1.8 kb upstream of this gene (*SI Appendix*, Fig. S5*D*). This gene is highly expressed in leaf tissue and shows only minimum transcription in flowers and most seed tissues (*SI Appendix*, Fig. S5*E*). Similarly, significantly enriched DNA motifs were also identified in flower- and seed-specific MHSs, and these motifs were related to TFs associated with flower/seed development (*SI Appendix*, Fig. S7).

Chromatin Accessibility and Transcriptional Divergence of Duplicated Genes. To investigate the impact of ACR variation on transcriptional divergence of duplicate genes, we first identified 17,111 pairs of soybean genes derived from the recent WGD. These duplicated genes were classified into "divergent" and "nondivergent" pairs based on the differential expression levels between each pair of genes. For example, we identified 5,254 pairs of genes that show ≥2 differential expression levels in the leaf tissue. The 5′ region of the highly expressed copy of each divergent pair showed a higher chromatin accessibility level than the lower expressed copy (Fig. 3*A*). In contrast, gene pairs with nondivergent expression exhibited a similar chromatin accessibility level in the 5′ regions (Fig. 3*A*). Similar results were also observed for duplicated genes expressed in flower and seed tissues (*SI Appendix*, Fig. S8).

To further examine the ACR dynamics between duplicated genes, we identified all MHSs located within ±1 kb regions of each pair of genes and performed sequence alignment of the MHSs. MHS pairs with highly similar sequences (e-value < 0.01) were considered as "syntenic MHSs". We calculated the ratio of syntenic MHSs to all MHSs and used it as the proxy for the "similarity level" of the MHSs associated with each pair of genes (*SI Appendix*, Fig. S9*A*). The duplicate pairs were classified into five subgroups according to the similarity levels of their associated MHSs. Only 21 to 28% of the duplicated genes showed identical orthologous MHSs (Fig. 3*B* and *SI Appendix*, Fig. S8 *C* and *F*). The MHS similarity of duplicate genes is also correlated with the level of transcriptional similarity between the duplicated genes (Fig. 3*C*). Furthermore, we identified 5,757 duplicated genes associated with cMHSs. We used these cMHSs to calculate the MHS similarity between these duplicated genes. Similarly, we identified 13,562 duplicated genes associated with sMHSs, and these sMHSs were used to calculate the MHS similarity between these duplicated genes. The MHS similarity levels derived from either cMHSs or sMHSs were correlated with the level of transcriptional similarity of the duplicated genes (*SI Appendix*, Fig. S6*B*).

Collectively, these results showed that ACR divergence is potentially an important force for the expression divergence of duplicated genes derived from the most recent WGD in soybean.

CRE Mutations and ACR Divergence between Duplicated Genes. Besides the syntenic MHSs, the remaining MHSs do not have a homologous MHS associated with the corresponding duplicated gene. These MHSs were considered as "solo-MHSs." To investigate the origin of the solo-MHSs, we aligned the sequences of the solo-MHSs with the soybean genome to identify their paralogous regions. We found that the best-matched sequences (BMSs) of ~20% solo-MHSs are not located in the regions around the corresponding duplicated genes (*SI Appendix*, Fig. S10*A*). The BMSs of ~65% of the solo-MHSs are located around the corresponding duplicated genes; however, these BMSs were not identified as ACRs by MH-seq (*SI Appendix*, Fig. S10*A*).

We were intrigued by how the BMSs of solo-MHSs lost their capacity as ACRs. We hypothesized that mutations of the CREs within the BMSs may have abolished or altered the binding of regulatory proteins and changed the chromatin accessibility of the regions. To test this hypothesis, we selected 1,432 solo-MHSs (leaf data) and the BMSs associated with their corresponding duplicated genes. We also selected 1,432 pairs of syntenic MHSs (leaf data) as a control. FIMO (46) was used to identify CREs within the sequences. A significantly higher number of motifs (P = 1.9e−10, Mann–Whitney U test) was identified in the solo-MHSs than in their BMSs (Fig. 4*A*). In contrast, a similar number of CREs was identified in the 1,432 pairs of syntenic MHSs (Fig. 4*A*). Similar results were also obtained from the solo-MHSs and their BMSs based on MH-seq datasets from flower and seed tissues (*SI Appendix*, Fig. S10*B*). Furthermore, we separated the solo-MHSs (leaf data) into solo-cMHSs and solo-sMHSs, respectively. We obtained similar results from comparisons of the motif numbers between the two groups of solo-MHSs and their BMSs (*SI Appendix*, Fig. S6*C*). These results suggest that CRE mutations within the BMSs play a role in losing their capacity as ACRs.

To further investigate the CRE dynamics within solo-MHSs and their BMSs in duplicated genes, we analyzed the enrichment of the TCP-binding motif GGACC, which is highly enriched in leaf-specific MHSs (*SI Appendix*, Fig. S5*B*). We used a random sequence GAGCC (not a known motif) as a negative control in the analysis. The number of GGACC motifs was significantly higher (P = 3e−6, Mann–Whitney U test) in the solo-MHSs than in their BMSs (Fig. 4*B*). In

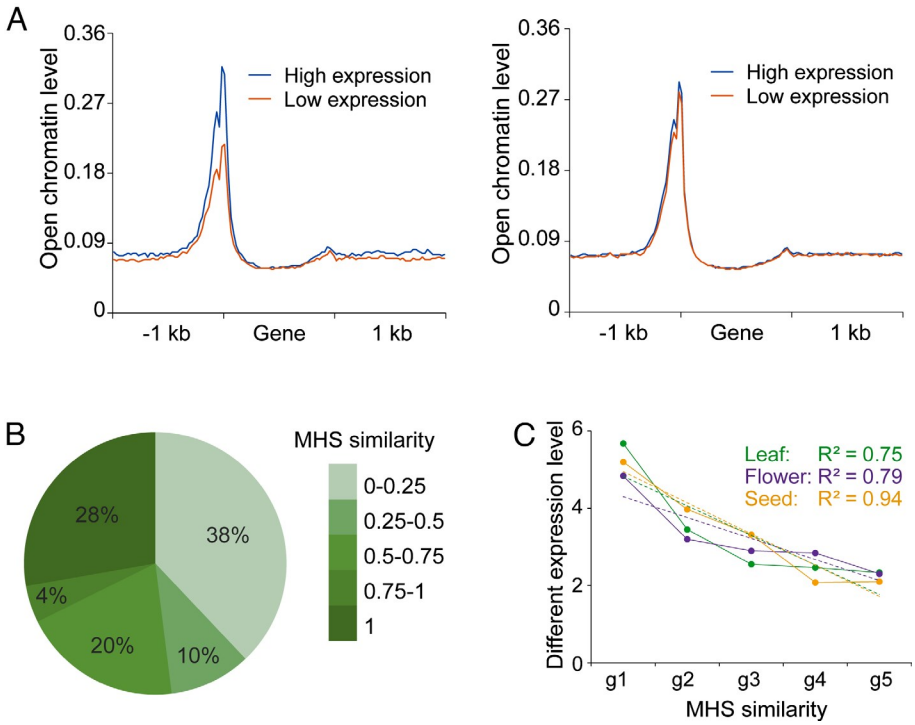

Fig. 3. MNase sensitivity and expression of duplicated genes in leaf tissue. (*A*) Profiles of MNase sensitivity of duplicated genes with divergent (*Left*) and similar (*Right*) expression levels. (*B*) Proportions of duplicated genes with different MHS similarity levels. The duplicated gene pairs were classified into five groups based on the MHS similarity level between each pair, ranging from 0 (group 1) to 1 (group 5). (*C*) The relationship between the MHS similarity levels and expression level of duplicated genes. The duplicated gene pairs were classified into the same five groups as (*B*). The dotted lines indicate linear trendlines.

contrast, we identified a similar number of GGACC motifs between the syntenic MHS pairs (Fig. 4*B*). In addition, we identified similar numbers of the GAGCC control sequence between solo-MHSs and their BMSs, and between the syntenic MHSs (Fig. 4*B*). The numbers of GGACC motifs were also significantly higher in both solo-cMHSs and solo-sMHSs than in their BMSs (*SI Appendix*, Fig. S6*D*).

Single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) (1 to 32 bp) were found to contribute to the

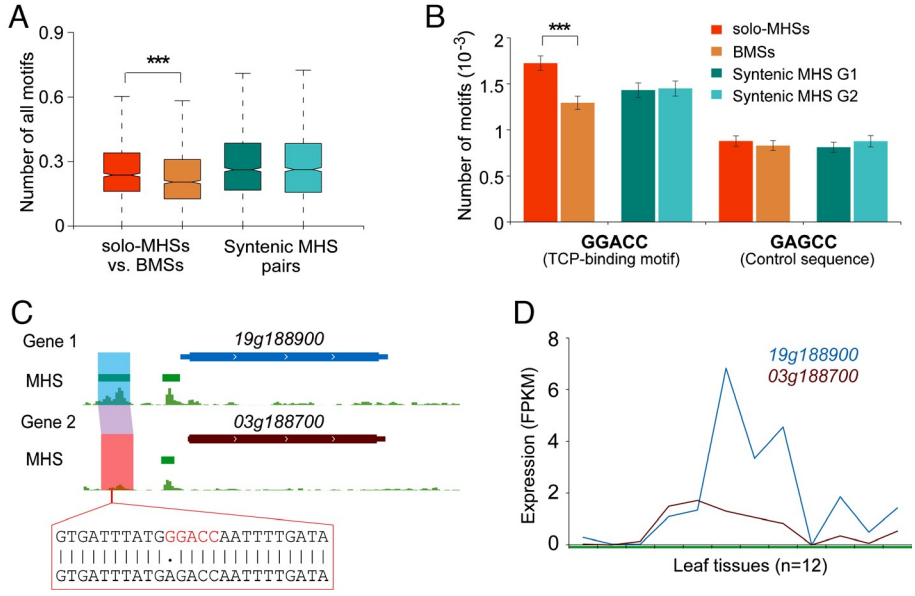

Fig. 4. Mutations of CREs and ACR divergence between duplicated genes. (*A*) Comparison of the numbers of DNA motifs identified within solo-MHSs vs. their best-matched sequences in the corresponding duplicated genes or between syntenic MHS pairs in leaf tissue. The numbers of motifs were counted for each MHS or best-matched sequence and normalized by its sequence length. ***$P < 0.001$, Mann–Whitney *U* test. (*B*) Comparison of the number of GGACC and GAGCC sequence motifs between solo-MHSs and their best-matched sequences in duplicated genes or between syntenic MHS pairs. ***$P < 0.001$, Mann–Whitney *U* test. (*C*) Mutation of CREs and ACR dynamics associated with a duplicated pair of genes *Glyma.19g188900* and *Glyma.03g188700*. Two MHSs were identified in the 5′ region of *Glyma.19g188900*, and only one MHS was associated with *Glyma.03g188700*. One TCP-binding motif GGACC was identified in the solo-MHS associated with *Glyma.19g188900*. A G/A SNP was detected in this motif in the matched sequence associated with *Glyma.03g188700*, which may impact the sensitivity to MNase in this region. (*D*) Expression of *Glyma.19g188900* and *Glyma.03g188700* in 12 different leaf tissues from young (*Left*) to old (*Right*).

mutation of the TCP-binding motifs in the BMSs. For example, we identified three MHSs (leaf data) associated with one pair of duplicate genes: *Glyma.19g188900* and *Glyma.03g188700*. These two genes encode protein kinase family proteins and share highly similar protein sequences (94% identity), but are associated with different MHS profiles (Fig. 4*C*). One solo-MHS (317 bp) was identified 419 bp upstream of *Glyma.19g188900*. We identified the BMS of this solo-MHS in the 5′ end of *Glyma.03g188700*, which is 318 bp long and 564 bp upstream of the TSS (Fig. 4*C*). We identified a single GAGCC motif in the solo-MHS of *Glyma.19g188900*. However, a G/A SNP within the motif was identified in the BMS of *Glyma.03g188700* (Fig. 4*C*). The expression of *Glyma.19g188900* was found to be either equal to or significantly higher than *Glyma.03g188700* in 12 different RNA-seq datasets derived from leaf tissue (Fig. 4*D*). Thus, the solo-MHS likely serves as a distal enhancer to *Glyma.19g188900* in leaf tissue.

Deletions and ACR Divergence between Duplicated Genes. We were not able to identify a BMS (e-value < 1e−5) in the Williams 82 genome for approximately 14% (7,388) of solo-MHSs, which were named as "orphan solo-MHSs" (orMHSs) (*SI Appendix*, Fig. S10*A*). We hypothesized that the origin of orMHSs is resulted from deletions of their BMSs, which occurred during sequence rearrangements after WGD. We predicted that some deletion events could be validated by comparing with the common bean (*Phaseolus vulgaris*) sequences. The common bean diverged from soybean ~19.2 Mya, which predated the last soybean WGD (5 to 13 Mya) (47). We were able to identify orthologous sequences for 18% (1,315/7,388) of the orMHSs in the reference genome of *P. vulgaris* (48). We further validated the hypothesized deletion events by examining the reference pan-genomes from 26 soybean lines, including 3 wild accessions of *Glycine soja*, 9 landraces, and 14 cultivars of *G. max* (49). We identified a BMS for only 328 (4.4%) of the 7,388 orMHSs in at least one of the 26 genomes. These results supported our hypothesis on the origin of orMHSs due to sequence deletions.

We next investigated the impact of the orMHSs on gene expression. The 7,388 orMHSs consist of 2,281 leaf orMHSs, 2,327 flower orMHSs, and 2,780 seed orMHSs. The 2,281 leaf orMHSs are associated with 2,117 genes. Nearly 28% of these genes showed twofold higher expression levels than their duplicated genes. We randomly selected 2,117 genes that are expressed in leaf tissue but are not associated with an orMHS. Only 14% of the 2,117 genes showed 2-fold higher expression levels than their duplicated genes, which is significantly lower ($P < 2.2e{-}16$, chi-square test) than the comparison of gene pairs with orMHSs. Similar results were found for the genes associated with flower orMHSs (31% vs. 16%) and seed orMHSs (23% vs. 12%). Collectively, these results showed that deletions of MHS-related sequences have impacted the transcriptional divergence of a large number of duplicated genes in soybean.

For example, we identified a total of eight orMHSs in three different tissues between the 3′ end of gene *Glyma.13g356400* and the 5′ end of gene *Glyma.13g356500* on chromosome 13 (Fig. 5*A*). The duplicated region of these two genes was found on chromosome 15 in the Williams 82 genome, including genes *Glyma.15g017600* and *Glyma.15g017700* (Fig. 5*A*). The 5′UTR of *Glyma.15g017600* overlaps with 3'UTR *of Glyma.15g017700* due to a ~2.5-kb deletion of the intergenic sequence between the two genes. This deletion resulted in loss of the homologous sequences of the eight orMHSs (Fig. 5*A*). Interestingly, this deletion was found in eight landraces and 13 cultivars of soybean. In contrast, it was not identified in the corresponding chromosome 15 region of the three wild soybean accessions (Fig. 5*B*). The two genes on chromosome 15 (*Soyw01.15g017900* and *Soyw01.15g017800*) and two genes on chromosome 13 (*Soyw01.13g348400* and *Soyw01.13g348500*) shared similar intergenic sequences in the Soyw01 genome. These results confirmed the predicted deletion in chromosome 15 in Williams 82.

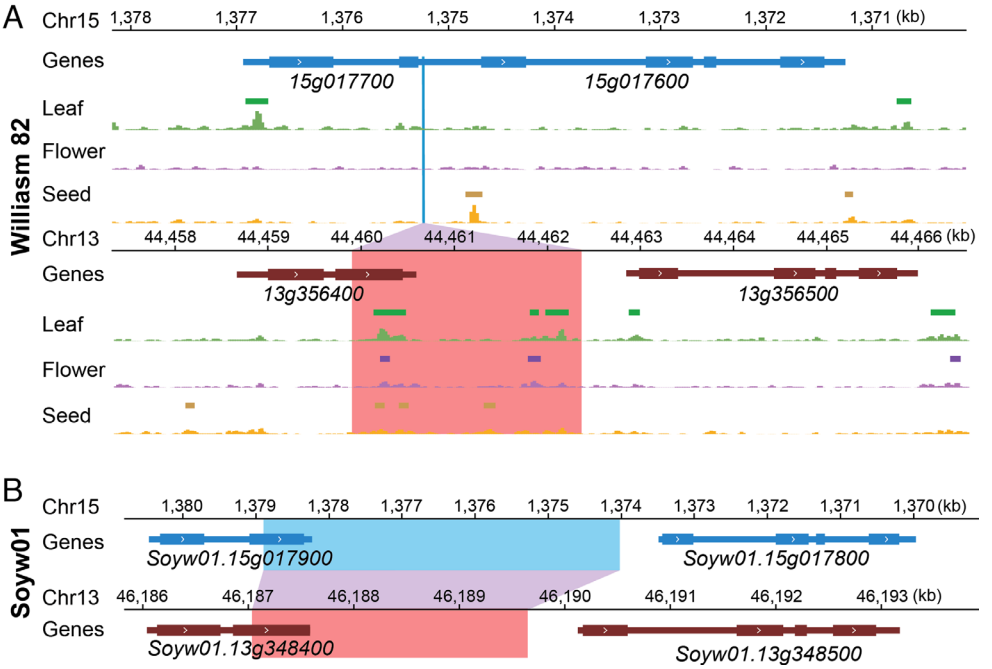

Fig. 5. Deletions and MHS dynamics between duplicated genes. (*A*) Identification of MHSs associated with two pairs of duplicated genes on chromosomes 13 and 15 in Williams 82. A total of eight orMHSs were identified in three different tissues between the 3′ of gene *Glyma.13g356400* and the 5′ of gene *Glyma.13g356500* on chromosome 13. No MHS was detected in the duplicated region on chromosome 15, including genes *Glyma.15g017600* and *Glyma.15g017700*. (*B*) Schematic presentation of the homologous regions in Soyw01 genome. The red box indicates the homologous region of the sequence between *Glyma.13g356400* and *Glyma.13g356500* in (*A*). The blue box indicates the orthologous region, which was deleted in the Williams 82 genome.

We next analyzed the expression patterns of the two pairs of duplicated genes in different tissues of Williams 82. Both pairs of genes (*Glyma.13g356400* and *Glyma.15g017700*; *Glyma.13g356500* and *Glyma.15g017600*) showed differential expression patterns in leaf, flower, and seed tissues (*SI Appendix*, Fig. S11). Compared to *Glyma.15g017700*, its duplicated gene *Glyma.13g356400* showed higher expression levels in most flower and seed tissues but a lower level in leaf tissue (*SI Appendix*, Fig. S11*A*). *Glyma.13g356500* and *Glyma.15g017600* showed similar expression levels in leaf tissues, but they showed different expression patterns in flower and seed tissues (*SI Appendix*, Fig. S11*C*). Although only a limited number of transcriptional datasets were available in wild soybean accession Soyw01, both pairs of duplicated genes showed relatively similar transcriptional patterns (*SI Appendix*, Fig. S11 *B* and *D*). For example, *Glyma.13g356500* and *Glyma.15g017600* showed 1.6-fold transcriptional difference in Williams 82 leaf tissues. In contrast, their corresponding pair of genes, *Soyw01.13g348500* and *Soyw01.15g017800*, showed only 0.98-fold difference in leaf tissues in Soyw01 (*SI Appendix*, Fig. S11*E*). These results support that the loss of the eight orMHSs significantly altered the expression patterns of both duplicated pairs of genes in Williams 82.

Identification of Neo-MHSs Associated with Duplicated Genes. The BMSs of 11,065 (~20%) solo-MHSs were located near unrelated genes rather than the corresponding duplicated genes (*SI Appendix*, Fig. S10*A*). We speculated that some of these solo-MHSs were transferred from a donor gene to one copy of the duplicated genes. Thus, these solo-MHSs are potentially acquired MHSs, or "neo-MHSs," for duplicated genes (*SI Appendix*, Fig. S9*B*). We developed a computational pipeline to further validate the neo-MHS nature of these solo-MHSs, which resulted in a final set of 2,138 neo-MHSs (*SI Appendix*, Fig. S9*B*). These neo-MHSs originated from two possible modes: 1) The neo-MHSs were originally associated with donor genes and were transferred to the duplicated genes. Thus, these are true neo-MHSs (*SI Appendix*, Fig. S9*B*); 2) the syntenic MHSs (or BMSs) of the neo-MHSs were originally associated with the duplicate genes but were transferred to other genomic regions, which resulted in the loss of MHS/BMS of a duplicated gene, rather than gaining of a neo-MHS.

We next aligned the 2,138 neo-MHSs to the common bean genome to identify orthologous sequences and their flanking genes. If the flanking gene is orthologous to a putative donor gene in the soybean genome, it would suggest that the solo-MHS was truly associated with the donor gene and was transferred to the duplicated gene as a neo-MHS. This comparative analysis confirmed a total of 159 neo-MHSs (*SI Appendix*, Tables S2–S4). We next investigated the impact of these neo-MHSs on expression of the orthologous gene pairs. The 159 neo-MHSs consist of 63 leaf MHSs, 44 flower MHSs, and 52 seed MHSs. The 63 leaf neo-MHSs are associated with 62 genes. We analyzed the expression patterns of these gene pairs in 12 RNA-seq samples derived from leaf tissues. We found that 25 gene pairs show twofold differential expression in at least one of the 12 RNA-seq datasets. Similarly, 12 and 22 duplicated gene pairs associated with flower and seed neo-MHSs showed diverged gene expression in flower and seed tissues, respectively.

Functional Validation of a Neo-MHS Associated with Soybean Nodulation. *Glyma.05g029800* on chromosome 5 and *Glyma.17g097000* on chromosome 17 are a pair of duplicated genes. We identified two MHSs (α, 237 bp; β, 231 bp) at the 5′ of *Glyma.05g029800* (Fig. 6*A*). A sequence homologous to the α MHS (α′) was identified in the upstream of *Glyma.17g097000* (Fig. 6*A*). However, a sequence homologous to the β MHS (β′) was identified in the upstream of *Glyma.17g096900*, a nonhomologous gene flanking *Glyma.17g097000* (Fig. 6*A*). In common bean, the homologous sequence of the β MHS (β″) is located at the 5′ of *003g194400*, which is orthologous to *Glyma.17g096900* (Fig. 6*A*). Therefore, the β MHS is a neo-MHS for *Glyma.05g029800*, possibly resulting from deletion of a gene on chromosome 5, which is homologous to *Glyma.17g096900* (Fig. 6*A*). This sequence arrangement was detected in all 3 wild accessions, 9 landraces, and 14 cultivars of soybean (49), indicating that this arrangement was generated immediately after the last WGD.

We next investigated whether acquisition of the β MHS (an sMHS) to *Glyma.05g029800* has impacted its expression pattern. *Glyma.05g029800* showed high levels of expression in leaf tissues and low levels of expression in seed tissues compared to its duplicate *Glyma.17g097000* (Fig. 6*B*). Interestingly, their orthologous gene in common bean (*003g192700*) showed high expression in seeds and low expression in leaf tissue (Fig. 6*C*), which resembles *Glyma.17g097000*. In addition, gene *Glyma.17g096900* and its orthologous gene in common bean (*003g194400*) exhibited a higher expression in leaf tissue than in flower and seed tissues (*SI Appendix*, Fig. S12*A*), a pattern resembling *Glyma.05g029800*. These results suggested that the acquisition of the β MHS may have contributed to the increased expression of *Glyma.05g029800* in the leaf tissues. To validate its *cis*-regulatory function in leaf tissue, the β MHS was cloned into vector pCAMBIA-CRE-LUC (50), which contains the minimal cauliflower mosaic virus (CaMV) 35S (mini35S) promoter (−50 to −2 bp) followed by a firefly luciferase reporter gene. The construct was infiltrated into *Nicotiana benthamiana* leaves. The construct containing the β MHS showed significantly higher bioluminescence signals ($P = 1.7e{-}4$, *t* test) than the construct containing only the mini35S promoter (*SI Appendix*, Fig. S12 *B* and *C*), confirming the potential regulatory activity of the β MHS in the leaf tissue.

Glyma.05g029800 belongs to the SPFH/B and 7/PHB domain-containing membrane-associated protein family. A previous study showed that it is highly expressed in nodules and potentially interacts with a known nodulation gene *GmFWL1* (51). *Glyma.05g029800* showed a significantly higher level of expression in nodules than both *Glyma.17g097000* and *Glyma.17g096900* based on the RNA-seq datasets developed from nodules at different developmental stages (52) (Fig. 7*A*). We identified a sequence motif (tctcgctgctgtaaa) within the β MHS that is associated with the NODULE INCEPTION-Like PROTEIN (NLP) TF (Fig. 7*B*). The NLP TFs are involved in nodulation by negatively regulating nodule numbers (53). Interestingly, this motif is not located within the β' MHS associated with *Glyma.17g096900* (*SI Appendix*, Fig. S12*D*).

We predicted that the β MHS may contribute to the higher expression of *Glyma.05g029800* in nodules. To test this hypothesis, we developed two deletion lines using CRISPR/Cas9. The two deletions, Q2 and Q30, spanned 216 bp and 625 bp, respectively, including the NLP-binding motif (Fig. 7*B*). Seedlings of the deletion lines were inoculated with rhizobia and grown in vermiculite with 2 mM KNO_3. Leaf tissues and nodules were collected from the deletion lines for gene expression analysis and phenotyping at 14 dpi (day postinoculation). The expression level of *Glyma.05g029800* in the deletion lines decreased by 26 to 27% in leaves and dramatically reduced by 81 to 91% in nodules compared to the wild type (Fig. 7*C*). In addition, the deletion lines had significantly higher numbers of nodules per plant ($P < 0.05$, *t* test) (Fig. 7 *D* and *E*). This phenotype is correlated with the loss of NLP-binding motif in the deletion lines. However, the deletion lines had a similar total weight of all nodules per plant compared to the wild type.

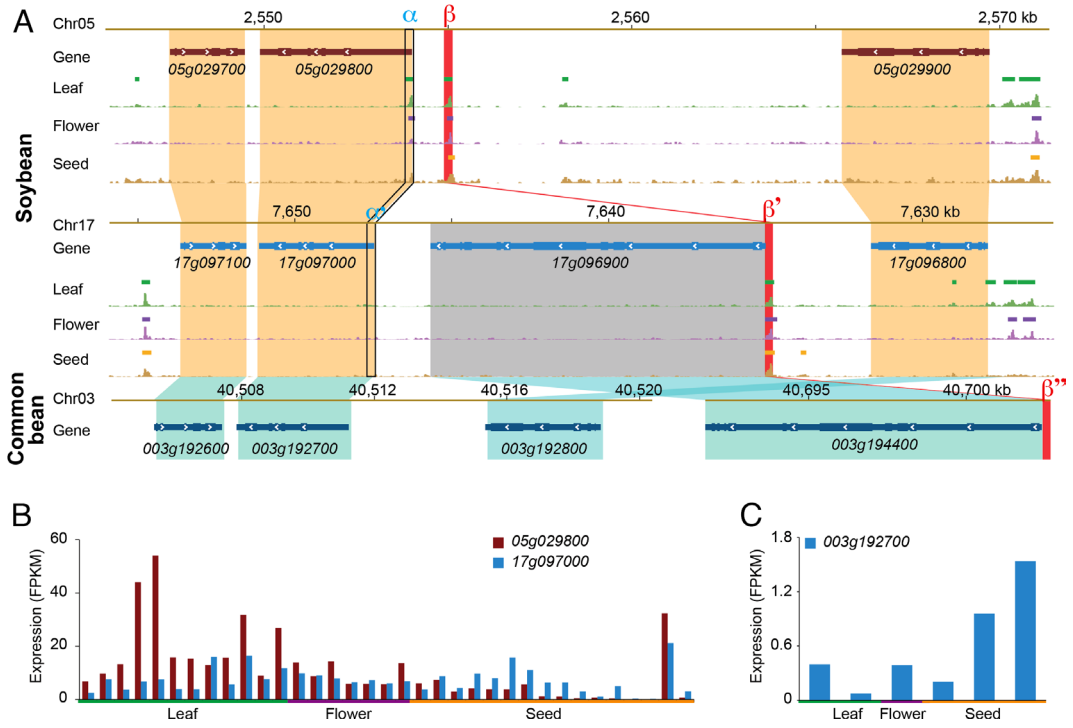

Fig. 6. Identification of a neo-MHS and its impact on gene expression. (*A*) Identification of a neo-MHS (β) associated with gene *Glyma.05g029800*. Brown boxes indicate duplicated pairs of genes. Blue boxes indicate orthologous genes between soybean and common bean. Black open boxes indicate an MHS (α) associated with gene *Glyma.05g029800* and its best matching sequence at the 5′ of gene *Glyma.17g097000*. The β MHS was likely originally associated with a gene homologous to *17g096900* (gray box) in chromosome 17. Deletion of this gene caused the transfer of the β MHS to the 5′ of *Glyma.05g029800*. (*B*) Expression patterns of the duplicated genes *Glyma.05g029800* and *Glyma.17g097000* in different tissues of Williams 82. (*C*) Expression patterns of gene *003g192700* in common bean.

Collectively, our data suggest that the acquisition of the β MHS changed the expression pattern of *Glyma.05g029800*. Most notably, the expression level of *Glyma.05g029800* is substantially higher than *Glyma.17g096900* in nodules (Fig. 7*A*). The β and β′ MHSs span 295 bp and 231 bp, respectively. Sequence homology was detected only within ~150 bp in the middle of the two sequences (*SI Appendix*, Fig. S12*D*). Most importantly, the NLP-binding motif is not located within the β′ MHS. Therefore, the NLP-binding motif was likely recruited to be part of the ACR after the transfer of β MHS to *Glyma.05g029800* and became a key CRE for this gene.

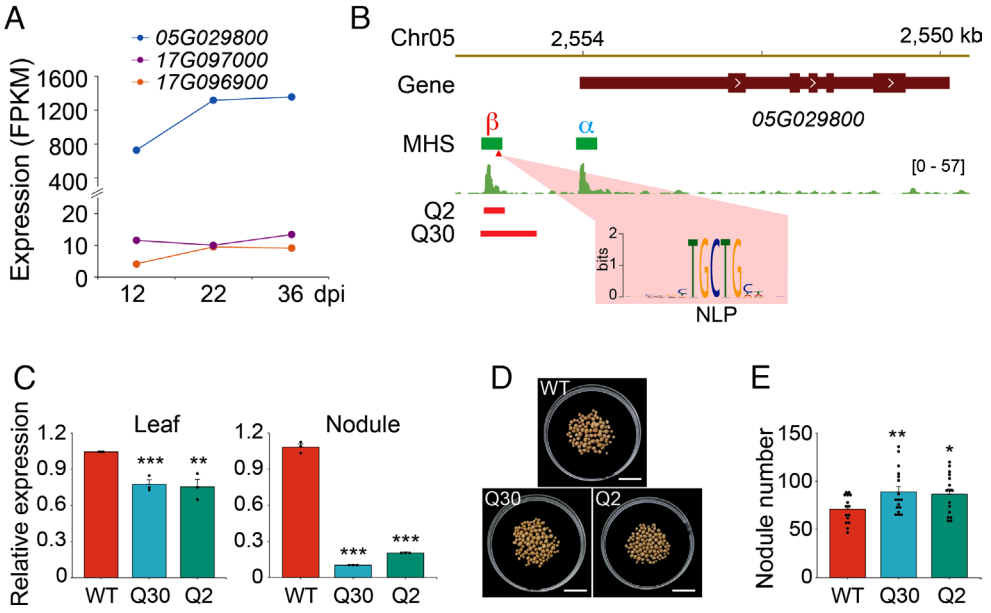

Fig. 7. Functional validation of the β MHS associated with gene *Glyma.05g029800*. (*A*) Comparison of expression levels of *Glyma.05g029800, Glyma.17g097000,* and *Glyma.17g096900* in nodules at three different development stages. "dpi" indicates day postinoculation. (*B*) Development of two CRISPR/Cas deletion lines spanning the β MHS. The red arrowhead indicates a motif related to the NLP TF. The red bars indicate the positions of the deleted regions. (*C*) Comparison of the expression levels of *Glyma.05g029800* in leaf (*Left*) and nodules (*Right*) between the wild type (n = 3) and deletion lines (n = 3) at 14 dpi. (*D*) Phenotype of nodules from the deletion lines and wild type at 14 dpi. Each image shows all nodules from a representative plant. Bars = 15 mm. (*E*) Comparison of nodule number per plant between the deletion lines and wild type at 14 dpi (nWT=15, nDel=16-17). **P* < 0.05, ***P* < 0.01, and ****P* < 0.001, *t* test.

Discussion

Transcriptional divergence of duplicated genes after WGD has been identified in many polyploid species (11, 19, 28, 54–58). However, the contribution of CRE dynamics to this transcriptional divergence was investigated in few studies (11, 59–62). A classic example is the *liguleless related sequence1* (*lrs1*) and *liguleless2* (*lg2*) genes, which are a duplicated pair of maize genes. The proteins encoded by *lg2* and *lrs1* are nearly identical. However, only *lg2*, but not *lrs1*, exhibits a role in ligule development (59, 63). Comparative sequence analysis of *lrs1* and *lg2* together with their rice ortholog revealed 30 conserved noncoding sequences (CNSs) surrounding this duplicated gene pair (59). Fractionation of some of these CNSs was proposed to explain the functional divergence of *lg2* and *lrs1* (59). CNSs were found to be enriched with known TFs or other *cis*-acting binding sites in plants (64). However, less than 50% of the CNSs are associated with ACRs in rice (65). Thus, a high false-positive rate would be expected if a functional CRE is predicted exclusively based on CNSs.

A. thaliana experienced at least three ancient WGD events with the most recent one occurring ~24 to 40 Mya (66–69). More than two-thirds of the duplicated genes in *A. thaliana* showed transcriptional divergence (11). A systematic analysis of the promoter sequences of the duplicated genes indicated that the transcriptional neo- and subfunctionalization is restricted to only a fraction of the CREs associated with the duplicated gene pairs (11). Analysis of duplicated genes related to stress responses revealed that the duplicated genes losing most of their stress responses are those that also lost more of the putative CREs in the promoter regions (60). These early studies showed the correlation between promoter sequence variation and the transcriptional divergence of duplicated genes. These studies relied on the traditional assumption that functional CREs are located mostly in the promoter regions, typically within 1 kb upstream of the TSSs of genes. However, mapping of ACRs in several different plant species showed that the promoter regions contain only a fraction of the CREs surrounding genes (65, 70–72). In soybean, ACRs located in the promoter regions account only for ~20% of total ACRs (*SI Appendix*, Fig. S2). It has been demonstrated in *A. thaliana* that >70% of the ACRs located in introns and intergenic regions show enhancer activities (35, 73). Thus, analysis of promoter sequence alone will understate the role of CREs in transcriptional divergence of duplicated genes.

We demonstrate that the dynamics of ACRs is well correlated with divergence of both expression level and the tissue specificity of duplicated soybean genes (Fig. 3 and *SI Appendix*, Fig. S5). We found that nearly 45% of the MHSs (39% in leaf, 47% in flower, and 48% in seed) associated with 17,111 pairs of duplicated genes are solo-MHSs. Analysis of the solo-MHSs revealed several different types of ACR dynamics: 1) An ACR associated with one copy of a duplicated gene pair may lose its function due to mutation of the CREs located within the ACR (Fig. 4); 2) Deletion of ACRs associated with one copy of duplicated genes (Fig. 5); and 3) Acquisition of a neo-ACR to one copy of a duplicated gene pair (Fig. 6). These three classes of ACR dynamics can be used to explain the transcriptional divergence of 16% (16% for leaf; 18% for flower; 15% for seed) pairs of duplicated genes. However, this number can be significantly underestimated because the solo-MHSs were identified based on a computational threshold. Leaf, flower, and seed tissues used in this study contain multiple cell types and cells at different developmental stages. ACRs associated with a single cell type may not be detected using bulk tissue-based techniques (35). There are additional classes of ACR/CRE dynamics that may not be found by the current methodology. For example, it has recently been demonstrated in *A. thaliana* that each ACR often contains multiple CREs (35). Thus, mutation of a single CRE may not alter the chromatin accessibility, but the mutation may alter the expression pattern of the associated gene.

We found that nearly 14% (7,388 total) of solo-MHSs associated with duplicated genes do not have the BMSs in the Williams 82 reference genome. In addition, only 328 (4.4%) of these solo-MHSs were found to have a BMS in at least one of the 26 available soybean genome sequences. These results indicate that the majority of these solo-MHSs may have been deleted in one copy of the duplicated genes after the recent WGD. Manual examination of several of the solo-MHSs confirmed the presence of their homologous sequences in wild soybean and/or in common bean (Figs. 5 and 6). These results showed that extensive sequence rearrangements may have occurred after the recent WGD and contributed to the divergence of ACRs associated with a large number of duplicated genes in soybean.

In summary, our results collectively support the hypothesis that reshaping of the CRE landscape after the recent WGD is an important force for the transcriptional divergence of duplicated genes in soybean. This may represent a general mechanism for transcriptional divergence of duplicated genes in polyploids that lack subgenome dominance.

Materials and Methods

Construction of MH-Seq Libraries. Soybean plants of Williams 82 were grown in a greenhouse and experimental station of Michigan State University. Trifoliate leaves were collected from seedlings at stage V2 (74). The seeds were collected from plants at stage R4 (74). Soybean plants were grown in a greenhouse with 12 h light (28 °C) and 12 h dark (25 °C). Flower tissues were collected from the R1 stage plants grown in the experimental station. Collected materials were ground into fine powder in liquid nitrogen. The powder (0.5 g for leaf and 1 g for flower and seed) was fixed with 1% of final concentration of formaldehyde for 10 min and followed by neutralization with 0.125 M glycine for 5 min. Nuclei were then prepared following published protocols (75). Purified nuclei were suspended in 0.8 mL MNase digestion buffer (MNB, 10% sucrose, 50 mM Tris-HCl, pH 7.5, 4 mM MgCl2, and 1 mM CaCl2) and divided into four 1.5-mL Eppendorf tubes. The aliquoted nuclei were digested at 37 °C for 10 min using a series of MNase (N3755-50UN, Sigma) in 0, 0.01, 0.03, and 0.1 unit, respectively. After terminating digestion with 10 mM EDTA, the MNase-treated nuclei with 200 mM NaCl were incubated at 65 °C for 4 h to reverse the cross-linking. DNA was extracted by adding 400 µL CTAB and incubated at 65 °C for 15 min following the CTAB method (76). The extracted DNA was then separated by running 2% agarose gel in 1× TAE buffer. DNA fragments <100 bp were excised from the agarose gel to prepare MH-seq library by following standard Illumina library preparation procedures. Two biological replicates were performed for each tissue.

MHS Analyses and Development of CRISPR/Cas Lines. Detailed procedures for MHS analyses and CRISPR/Cas experiments are included in *SI Appendix, Supplemental Methods*.

Data, Materials, and Software Availability. MH-seq datasets and MHSs were deposited in NCBI Gene Expression Omnibus (GEO) under accession number GSE167578 (77). ATAC-seq data and respective RNA-seq data as well as H3 ChIP-seq data were downloaded from NCBI GEO under accession number GSE128434 (78) and BioProject accession number PRJNA751745 (79). RNA-seq data from tissues at similar stages with samples used for MH-seq were downloaded from NCBI SRA under the following accession numbers: SRR1174229 (leaf), SRR1174220 (flower), and SRR1174208 (seed) (80). The remaining RNA-seq datasets of leaf, flower, and seed tissues were summarized in *SI Appendix*, Table S5. The 26 soybean genome datasets were downloaded from the National Genomics Data Center under PRJCA002030 (https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA002030) (81).

ACKNOWLEDGMENTS. We are grateful to Drs. Zhixi Tian, Charles An, and Pat Edgar for their valuable discussion and comments for the development of this manuscript. We thank Dr. Bob Stupar for providing the Williams 82 seeds,

Dr. Dechun Wang for growing plant materials in the field, Dr. Yuefeng Guan for providing the soybean multiplex CRISPR pGES401vector, and Luke Strickland for editing the manuscript. This research is supported by the National Key R&D Program of China (2021YFF1001202) and the National Natural Science Foundation of China (32172032) to F.M., the National Project of China (CARS-04) to Q.C., and NSF grants MCB-1412948 and ISO-2029959 and MSU startup funds to J.J.

Author afiliations: [a]Department of Plant Biology, Michigan State University, East Lansing, MI 48824; [b]Northeast Institute of Geography and Agroecology, Key Laboratory of Soybean Molecular Design Breeding, Chinese Academy of Sciences, Harbin 150081, China; [c]Key Laboratory of Soybean Biology in Chinese Ministry of Education, Northeast Agricultural University, Harbin 150030, China; [d]Agro-Biotechnology Research Institute, Jilin Academy of Agricultural Sciences, Changchun 130033, China; [e]Department of Horticulture, Michigan State University, East Lansing, MI 48824; and [f]Michigan State University AgBioResearch, East Lansing, MI 48824

1. K. L. Adams, J. F. Wendel, Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8, 135–141 (2005).
2. L. Comai, The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846 (2005).
3. Y. N. Jiao *et al.*, Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100 (2011).
4. K. Vanneste, G. Baele, S. Maere, Y. Van de Peer, Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* 24, 1334–1347 (2014).
5. N. Panchy, M. Lehti-Shiu, S. H. Shiu, Evolution of gene duplication in plants. *Plant Physiol.* 171, 2294–2316 (2016).
6. M. Semon, K. H. Wolfe, Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512 (2007).
7. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, New York, 1970).
8. A. Force *et al.*, Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545 (1999).
9. H. Innan, F. Kondrashov, The evolution of gene duplications: Classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108 (2010).
10. Z. L. Gu *et al.*, Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66 (2003).
11. G. Haberer, T. Hindemitt, B. C. Meyers, K. F. X. Mayer, Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of Arabidopsis. *Plant Physiol.* 136, 3009–3022 (2004).
12. E. I. Alger, P. P. Edger, One subgenome to rule them all: Underlying mechanisms of subgenome dominance. *Curr. Opin. Plant Biol.* 54, 108–113 (2020).
13. B. C. Thomas, B. Pedersen, M. Freeling, Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946 (2006).
14. L. Flagel, J. Udall, D. Nettleton, J. Wendel, Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC Biol.* 6, 16 (2008).
15. J. C. Schnable, N. M. Springer, M. Freeling, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4069–4074 (2011).
16. X. T. Wang *et al.*, Transcriptome asymmetry in synthetic and natural allotetraploid wheats, revealed by RNA-sequencing. *New Phytol.* 209, 1264–1277 (2016).
17. K. A. Bird, R. VanBuren, J. R. Puzey, P. P. Edger, The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol.* 220, 87–93 (2018).
18. K. A. Bird *et al.*, Replaying the evolutionary tape to investigate subgenome dominance in allopolyploid *Brassica napus*. *New Phytol.* 230, 354–371 (2021).
19. M. D. Li, W. Q. Sun, F. Wang, X. M. Wu, J. B. Wang, Asymmetric epigenetic modification and homoeolog expression bias in the establishment and evolution of allopolyploid *Brassica napus*. *New Phytol.* 232, 898–913 (2021).
20. L. W. Yin, G. Xu, J. L. Yang, M. X. Zhao, The heterogeneity in the landscape of gene dominance in maize is accompanied by unique chromatin environments. *Mol. Biol. Evol.* 39, msac198 (2022).
21. Z. Swigonova *et al.*, Close split of sorghum and maize genome progenitors. *Genome Res.* 14, 1916–1923 (2004).
22. M. R. Woodhouse *et al.*, Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLoS. Biol.* 8, e1000409 (2010).
23. Y. Y. Zhang *et al.*, Transposable elements orchestrate subgenome-convergent and -divergent transcription in common wheat. *Nat. Commun.* 13, 6940 (2022).
24. J. J. Powell *et al.*, The defence-associated transcriptome of hexaploid wheat displays homoeolog expression and induction bias. *Plant Biotechnol. J.* 15, 533–543 (2017).
25. G. W. de Jong, K. L. Adams, Subgenome-dominant expression and alternative splicing in response to *Sclerotinia* infection in polyploid *Brassica napus* and progenitors. *Plant J.* 114, 142–158 (2023).
26. O. Garsmeur *et al.*, Two evolutionarily distinct classes of paleopolyploidy. *Mol. Biol. Evol.* 31, 448–454 (2014).
27. J. Schmutz *et al.*, Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183 (2010).
28. A. Roulin *et al.*, The fate of duplicated genes in a polyploid plant genome. *Plant J.* 73, 143–153 (2013).
29. M. X. Zhao, B. A. Zhang, D. Lisch, J. X. Ma, Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. *Plant Cell* 29, 2974–2994 (2017).
30. A. Zhang, X. Li, H. Zhao, J. Jiang, W. Zhang, Genome-wide identification of open chromatin in plants using MH-seq. *Methods Mol. Biol.* 2594, 29–43 (2023).
31. A. P. Boyle, J. Guinney, G. E. Crawford, T. S. Furey, F-Seq: A feature density estimator for high-throughput sequence tags. *Bioinformatics* 24, 2537–2538 (2008).
32. H. N. Zhao *et al.*, Genome-wide MNase hypersensitivity assay unveils distinct classes of open chromatin associated with H3K27me3 and DNA methylation in *Arabidopsis thaliana*. *Genome Biol.* 21, 24 (2020).
33. Z. F. Lu *et al.*, The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nat. Plants* 5, 1250–1259 (2019).
34. M. K. Huang *et al.*, Identification of the accessible chromatin regions in six tissues in the soybean. *Genomics* 114, 110364 (2022).
35. F. L. Meng *et al.*, Genomic editing of intronic enhancers unveils their role in fine-tuning tissue-specific gene expression in *Arabidopsis thaliana*. *Plant Cell* 33, 1997–2014 (2021).
36. H. N. Zhao *et al.*, Identification and functional validation of super-enhancers in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U.S.A.* 119, e2215328119 (2022).
37. Z. J. Xia *et al.*, Positional cloning and characterization reveal the molecular basis for soybean maturity locus *E1* that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci. U.S.A.* 109, E2155–E2164 (2012).
38. Y. Tsubokura *et al.*, Natural variation in the genes responsible for maturity loci E1, E2, E3 and E4 in soybean. *Ann. Bot-London* 113, 429–441 (2014).
39. M. L. Xu *et al.*, The soybean-specific maturity gene E1 family of floral repressors controls night-break responses through down-regulation of FLOWERING LOCUS T orthologs. *Plant Physiol.* 168, 1735–1746 (2015).
40. E. Balsemao-Pires, L. R. Andrade, G. Sachetto-Martins, Functional study of TCP23 in *Arabidopsis thaliana* during plant development. *Plant Physiol. Bioch.* 67, 120–125 (2013).
41. L. E. Lucero, P. A. Manavella, D. E. Gras, F. D. Ariel, D. H. Gonzalez, Class I and class II TCP transcription factors modulate SOC1-dependent flowering at multiple levels. *Mol. Plant* 10, 1571–1574 (2017).
42. P. I. L. Joaquim *et al.*, Nitrogen compounds transporters: Candidates to increase the protein content in soybean seeds. *J. Plant Interactions* 17, 309–318 (2022).
43. T. L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36 (1994).
44. T. Koyama, N. Mitsuda, M. Seki, K. Shinozaki, M. Ohme-Takagi, TCP transcription factors regulate the activities of ASYMMETRIC LEAVES1 and miR164, as well as the auxin response, during differentiation of leaves in Arabidopsis. *Plant Cell* 22, 3574–3588 (2010).
45. C. J. Zhang *et al.*, A GATA transcription factor from soybean (*Glycine max*) regulates chlorophyll biosynthesis and suppresses growth in the transgenic *Arabidopsis thaliana*. *Plants-Basel* 9, 1036 (2020).
46. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: Scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018 (2011).
47. M. Lavin, P. S. Herendeen, M. F. Wojciechowski, Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary. *Syst. Biol.* 54, 575–594 (2005).
48. J. Schmutz *et al.*, A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46, 707–713 (2014).
49. Y. C. Liu *et al.*, Pan-genome of wild and cultivated soybeans. *Cell* 182, 162–176 (2020).
50. Y. Lin, F. L. Meng, C. Fang, B. Zhu, J. M. Jiang, Rapid validation of transcriptional enhancers using agrobacterium-mediated transient assay. *Plant Methods* 15, 21 (2019).
51. Z. Z. Qiao *et al.*, The GmFWL1 (FW2-2-like) nodulation gene encodes a plasma membrane microdomain-associated protein. *Plant Cell Environ.* 40, 1442–1455 (2017).
52. D. Niyikiza *et al.*, Interactions of gene expression, alternative splicing, and DNA methylation in determining nodule identity. *Plant J.* 103, 1744–1766 (2020).
53. H. Nishida *et al.*, A NIN-LIKE PROTEIN mediates nitrate-induced control of root nodule symbiosis in *Lotus japonicus*. *Nat. Commun.* 9, 499 (2018).
54. G. Blanc, K. H. Wolfe, Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. *Plant Cell* 16, 1679–1691 (2004).
55. B. Chaudhary *et al.*, Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium). *Genetics* 182, 503–517 (2009).
56. R. J. Buggs *et al.*, Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid *Tragopogon mirus*. *New Phytol.* 186, 175–183 (2010).
57. M. Groszmann, T. Paicu, J. P. Alvarez, S. M. Swain, D. R. Smyth, SPATULA and ALCATRAZ, are partially redundant, functionally diverging bHLH genes required for Arabidopsis gynoecium and fruit development. *Plant J.* 68, 816–829 (2011).
58. S. L. Liu, G. J. Baute, K. L. Adams, Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. *Genome Biol. Evol.* 3, 1419–1436 (2011).
59. R. J. Langham *et al.*, Genomic duplication, fractionation and the origin of regulatory novelty. *Genetics* 166, 935–945 (2004).
60. C. Zou, M. D. Lehti-Shiu, M. Thomashow, S. H. Shiu, Evolution of stress-regulated gene expression in duplicate genes of *Arabidopsis thaliana*. *PLoS Genet.* 5, e1000581 (2009).
61. A. E. Yocca, Z. F. Lu, R. J. Schmitz, M. Freeling, P. P. Edger, Evolution of conserved noncoding sequences in *Arabidopsis thaliana*. *Mol. Biol. Evol.* 38, 2692–2703 (2021).
62. J. L. Han *et al.*, Genome-wide chromatin accessibility analysis unveils open chromatin convergent evolution during polyploidization in cotton. *Proc. Natl. Acad. Sci. U.S.A.* 119, e2209743119 (2022).
63. L. Harper, M. Freeling, Interactions of luguleless1 and liguleless2 function during ligule induction in maize. *Genetics* 144, 1871–1882 (1996).
64. M. Freeling, S. Subramaniam, Conserved noncoding sequences (CNSs) in higher plants. *Curr. Opin. Plant Biol.* 12, 126–132 (2009).
65. W. L. Zhang *et al.*, High-resolution mapping of open chromatin in the rice genome. *Genome Res.* 22, 151–162 (2012).
66. C. Simillion, K. Vandepoele, M. C. Van Montagu, M. Zabeau, Y. Van de Peer, The hidden duplication past of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13627–13632 (2002).
67. G. Blanc, K. Hokamp, K. H. Wolfe, A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.* 13, 137–144 (2003).
68. J. E. Bowers, B. A. Chapman, J. Rong, A. H. Paterson, Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438 (2003).
69. Y. Henry, M. Bedhomme, G. Blanc, History, protohistory and prehistory of the *Arabidopsis thaliana* chromosome complement. *Trends Plant Sci.* 11, 267–273 (2006).
70. W. L. Zhang, T. Zhang, Y. F. Wu, J. M. Jiang, Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *Plant Cell* 24, 2719–2731 (2012).
71. R. Oka *et al.*, Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.* 18, 137 (2017).
72. H. N. Zhao *et al.*, Proliferation of regulatory DNA elements derived from transposable elements in the maize genome. *Plant Physiol.* 176, 2789–2803 (2018).

73. B. Zhu, W. L. Zhang, T. Zhang, B. Liu, J. M. Jiang, Genome-wide prediction and validation of intergenic enhancers in Arabidopsis using open chromatin signatures. *Plant Cell* 27, 2415–2426 (2015).
74. W. R. Fehr, C. E. Caviness, "Stages of soybean development" (*Iowa Agricultural and Home Economics Experiment Station Special Report*, Iowa State University, 1977), pp. 3–11.
75. A. Saleh, R. Alvarez-Venegas, Z. Avramova, An efficient chromatin immunoprecipitation (ChIP) protocol for studying histone modifications in Arabidopsis plants. *Nat. Protoc.* 3, 1018–1025 (2008).
76. M. G. Murray, W. F. Thompson, Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–4325 (1980).
77. C. Fang, H. N. Zhao, F. L. Meng, J. J. Jiang, Genome-wide MNase hypersensitive sites in soybean. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167578. Deposited 25 February 2021.
78. R. J. Schmitz, The prevalence, evolution and chromatin signatures of plant regulatory elements. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128434. Deposited 18 March 2019.
79. M. K. Huang *et al.*, The landscape of accessible chromatin regions in the soybean genome architecture. NCBI Sequence Read Archive. https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA751745. Deposited 3 August 2021.
80. Y. T. Shen *et al.*, Soybean 28 Tissues Collected from the Different Developmental Stages for RNA-seq. NCBI Sequence Read Archive. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA238493. Deposited 18 February 2014.
81. Y. C. Liu *et al.*, IGDB soybean pan-genome project. The Genome Warehouse. https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA002030. Deposited 14 December 2019.