nature genetics

Article

https://doi.org/10.1038/s41588-024-01738-2

Long noncoding RNAs underlie multiple domestication traits and leafhopper resistance in soybean

Received: 2 July 2023

Accepted: 28 March 2024

Published online: 29 April 2024

Check for updates

Weidong Wang[1,2,3,9], Jingbo Duan[1,2,9], Xutong Wang[1,2,7,9], Xingxing Feng[4,9], Liyang Chen[1,2], Chancelor B. Clark[1,2], Stephen A. Swarm[5,8], Jinbin Wang[1,2], Sen Lin[1,2], Randall L. Nelson[5], Blake C. Meyers[6], Xianzhong Feng[4] ✉ & Jianxin Ma[1,2] ✉

The origin and functionality of long noncoding RNA (lncRNA) remain poorly understood. Here, we show that multiple quantitative trait loci modulating distinct domestication traits in soybeans are pleiotropic effects of a locus composed of two tandem lncRNA genes. These lncRNA genes, each containing two inverted repeats, originating from coding sequences of the MYB genes, function in wild soybeans by generating clusters of small RNA (sRNA) species that inhibit the expression of their MYB gene relatives through post-transcriptional regulation. By contrast, the expression of lncRNA genes in cultivated soybeans is severely repressed, and, consequently, the corresponding MYB genes are highly expressed, shaping multiple distinct domestication traits as well as leafhopper resistance. The inverted repeats were formed before the divergence of the *Glycine* genus from the *Phaseolus*–*Vigna* lineage and exhibit strong structure–function constraints. This study exemplifies a type of target for selection during plant domestication and identifies mechanisms of lncRNA formation and action.

The domestication of a crop from its wild relative is a complex process of artificial selection for a suite of favorable traits, which are generally controlled by different genetic loci[1]. Such a process gives rise to a new form of plants, known as domesticates, to meet human needs. Nevertheless, it also leads to drastic reduction in genetic diversity in domesticates, hindering the sustainability of crop improvement[2]. To better understand the dynamic processes of crop domestication and exploit untapped genetic variation in crop wild relatives for enhancement of elite cultivars, it is important to decipher the genetic and molecular basis underlying domestication-related traits (DRTs).

In the past few decades, tremendous work has been done to identify quantitative trait loci (QTL) underlying DRTs in major crops, such as (cultivated) soybean (*Glycine max*), an economically important leguminous crop domesticated from wild soybean (*Glycine soja*)[3,4]. Most wild soybean accessions exhibit a procumbent or climbing growth habit, with long, slender, prolifically branched stems and small leaves that grow with appressed pubescence, whereas the majority of cultivated soybean varieties display a bush-type upright growth habit, with short, scout primary stems and sparse branches and large leaves with semi-appressed or erect pubescence. Here, we report that multiple QTL underlying different DRTs as well as resistance to leafhoppers

[1]Department of Agronomy, Purdue University, West Lafayette, IN, USA. [2]Center for Plant Biology, Purdue University, West Lafayette, IN, USA. [3]College of Agronomy and Biotechnology, China Agricultural University, Beijing, China. [4]Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, China. [5]Department of Crop Sciences, University of Illinois at Urbana–Champaign, Urbana, IL, USA. [6]Genome Center and Department of Plant Sciences, University of California, Davis, Davis, CA, USA. [7]Present address: College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China. [8]Present address: Beck's Hybrids, Atlanta, IN, USA. [9]These authors contributed equally: Weidong Wang, Jingbo Duan, Xutong Wang, Xingxing Feng. ✉e-mail: fengxianzhong@iga.ac.cn; maj@purdue.edu

https://doi.org/10.1038/s41588-024-01738-2

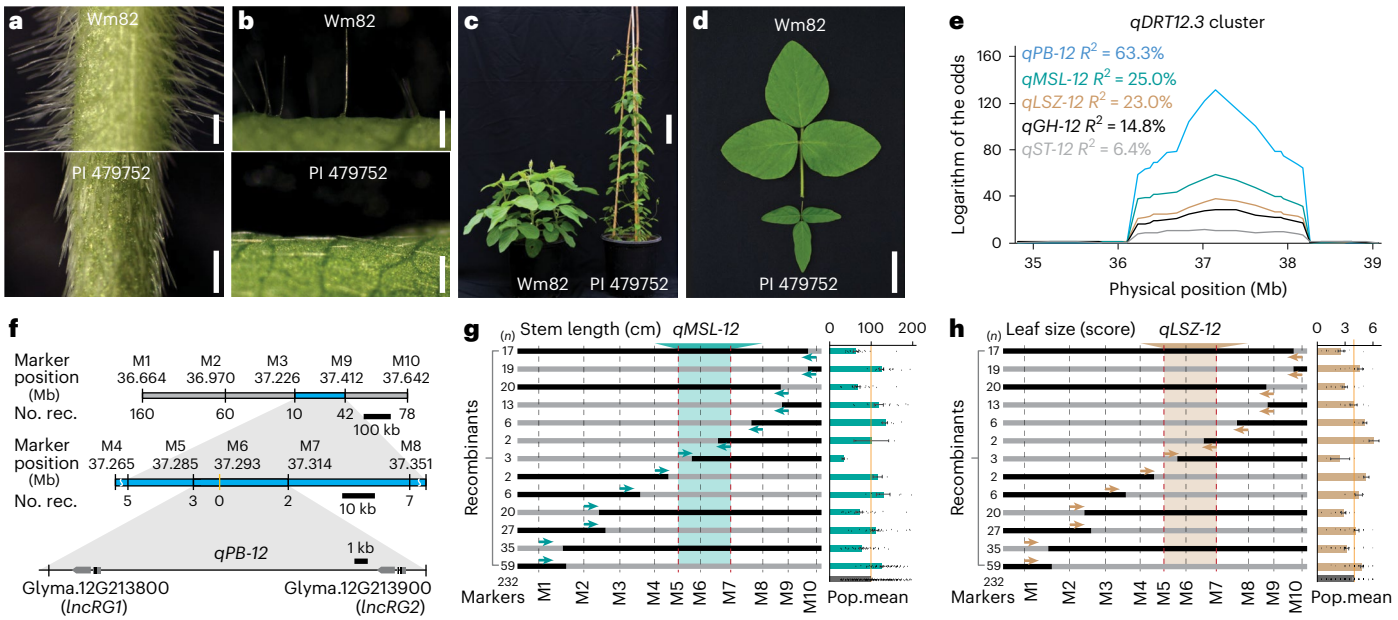


Fig. 1 | Map-based cloning of multiple DRT QTL identifies a single locus with pleiotropic effects. a,**b**, Comparisons of pubescence form on stems (**a**) and leaves (**b**) between Wm82 (*G. max*) and PI 479752 (*G. soja*). Scale bar, 3 mm. **c**, Comparisons of stem height and growth habit between Wm82 and PI 479752. Scale bar, 10 cm. **d**, Comparison of leaf size between Wm82 and PI 479752. Scale bar, 5 cm. **e**, Primary mapping region of *qDRT12.3* on chromosome 12. The *y* axis represents the $\log_{10}$ likelihood ratio, and R^2 values indicate phenotypic variations explained by each QTL. **f**, Fine mapping of *qPB-12*. Rec., recombinants. **g**,**h**, Fine mapping of *qMSL-12* (**g**) and *qLSZ-12* (**h**). Each bar represents the genotype of the recombinants with the same haplotype at all markers. Numbers on the left of the bars indicate number of recombinants sharing the same haplotype, and dots indicate phenotypic values of the recombinants. The black color represents the *G. soja* genotype, and the gray color represents the *G. max* genotype. Arrows indicate the deduced location of the QTL. Green and brown shading highlights the final mapping interval. Pop.mean, population mean. Data are represented as mean ± s.e.m.

in cultivated soybeans are the result of artificial selection of reduced expression of two tandemly duplicated lncRNA genes each carrying MYB gene coding sequence-derived inverted repeats, which have undergone strong purifying selection in the *Glycine* genus.

Results

Map-based cloning pinpoints multi-DRT QTL to the same locus

Using a subset of the 2,287 recombinant inbred lines (RILs) derived from a cross between soybean cultivar Williams 82 (Wm82) and *G. soja* accession PI 479752, we initially mapped >100 QTL associated with various DRTs[4]. Remarkably, many of the QTL regions, which underlie different DRTs, physically overlap. One such region, *qDRT12.3* on chromosome 12, was found to harbor five QTL, *qPB-12*, *qMSL-12*, *qLSZ-12*, *qGH-12* and *qST-12*, which explained 63.3%, 25.0%, 23.0%, 14.8% and 6.4% of the phenotypic variation in pubescence form, main stem length, leaf size, growth habit and stem twining, respectively (Fig. 1a–e).

To determine whether these QTL are attributed to different genes or pleiotropic effects of the same gene, or both, we first conducted fine mapping of three (*qPB-12*, *qMSL-12* and *qLSZ-12*) of the five QTL, independently, using the entire RIL population. Two insertion–deletion markers, M1 and M10, which initially defined the boundaries of the *qDRT12.3* region, were used to genotype all 2,287 RILs, and we identified 238 recombinants between the two markers (Fig. 1f and Supplementary Table 1). These recombinants were then genotyped with eight additional markers within the *qDRT12.3* region and first examined for pubescence form. Combination of the genotypic and phenotypic data delimited *qPB-12* to a 29-kb region between markers M5 and M7 (Fig. 1f). Subsequently, the 238 recombinants were measured for main stem length and leaf size, respectively. Based on the eight markers, these recombinants were divided into 13 haplotypes, and the average phenotypic value of recombinants within each haplotype group was compared to the population mean to calculate the phenotypic scores of individual haplotypes to fine map *qMSL-12* and *qLSZ-12*. Interestingly, these two QTL were also defined to the same 29-kb region (Fig. 1g,h). According to the Wm82 reference genome, this region harbors only two genes, Glyma.12G213800 and Glyma.12G213900, both lncRNA species.

It has been observed that semi-appressed or erect pubescence is linked to reduced defoliation caused by *Cicadellidae* insects[5]. To investigate whether *qPB-12* is responsible for such resistance, we conducted a genome-wide association study (GWAS) on pubescence form using resequencing data from 74 *G. soja* and 594 *G. max* accessions[6] and their phenotypic data from the US Department of Agriculture (USDA) soybean germplasm collection[7] as well as a GWAS on both pubescence form and leafhopper resistance and susceptibility using SNP data and phenotypic data from 784 diverse accessions in the USDA soybean germplasm collection[7] (Supplementary Tables 2–5). We found that these two traits were both primarily modulated by a single major QTL harboring the 29-kb genomic region and that molecular markers within the fine-mapped *qPB-12* region were significantly associated with leafhopper resistance, and no additional regions in the entire genome were found to be associated with leafhopper resistance (Extended Data Fig. 1a–d). These observations, together with the reported association between erect pubescence and resistance to *Cicadellidae* insects[5], suggest that it is very likely that *qPB-12* also underlies leafhopper resistance.

In the set of resequenced diverse *G. soja* and *G. max* accessions[6], only 13.4% of *G. soja* accessions have erect pubescence, whereas 71.3% and 96.7% of the landraces and elite cultivars possess it, respectively (Extended Data Fig. 1e and Supplementary Table 2), suggesting that the QTL for erect pubescence and leafhopper resistance was a target for selection during soybean domestication and improvement. The artificial selection at this QTL was also echoed by the selective sweep surrounding it (Extended Data Fig. 1f), as detected by the resequencing data. Collectively, these observations suggest that Glyma.12G213800 and Glyma.12G213900 are likely the candidate genes regulating pubescence form, main stem length and leaf size as well as leafhopper resistance.

https://doi.org/10.1038/s41588-024-01738-2

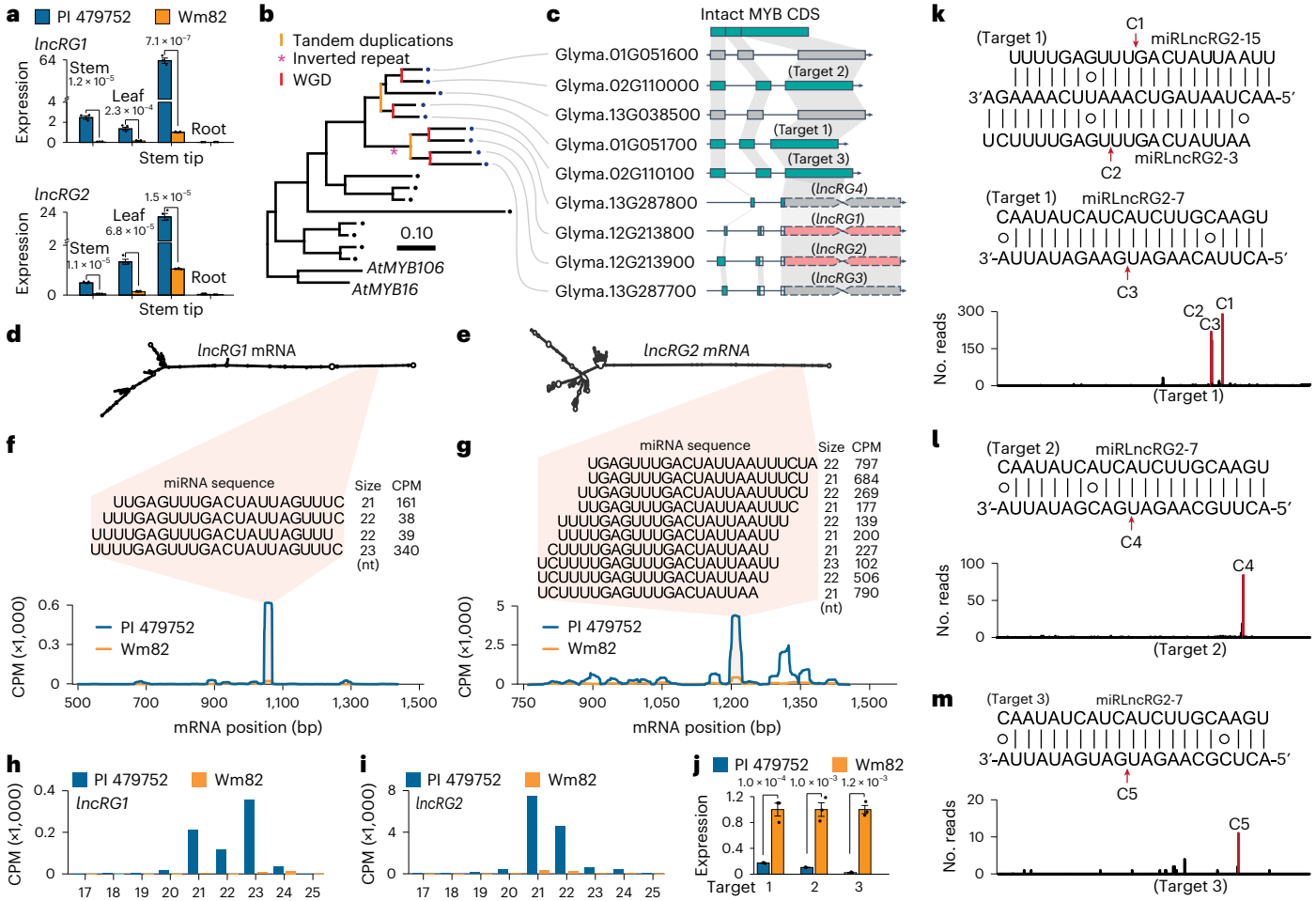


Fig. 2 | *lncRG1* and *lncRG2* harbor inverted repeats and produce abundant sRNA species primarily targeting three closely related MYB genes. a, Expression levels of *lncRG1* and *lncRG2* in different tissues as determined by RT–qPCR, with the Wm82 stem tip set as '1' and the others adjusted accordingly. **b**, Phylogenetic relationships of *lncRG1*, *lncRG2* and their close MYB relatives constructed using the transcript sequences of these genes. Colored lines indicate duplication events. The red asterisk marks the deduced time when the original inverted repeat occurred. **c**, Gene models and transcript sequence alignments of *lncRG1*, *lncRG2* and their close MYB relatives. Green bars represent coding regions, and pink bars represent inverted repeats. Gray shading indicates the syntenic region among the genes. CDS, coding sequence. **d**,**e**, Predicted secondary structures of the *lncRG1* and *lncRG2* transcripts. **f**,**g**, Distribution, abundance and the major cluster of sRNA species produced by *lncRG1* and *lncRG2*. CPM, copies per million reads; nt, nucleotides. **h**,**i**, Abundance of sRNA species in different sizes produced by *lncRG1* and *lncRG2*. **j**, Expression levels of the target genes Glyma.01G051700 (target 1), Glyma.02G110000 (target 2) and Glyma.02G110100 (target 3), as determined by RT–qPCR with Wm82 set as '1' and the others adjusted accordingly. **k**–**m**, The predicted cleavage sites supported by degradome sequencing on the target genes. The letter C represents cleavage sites. In **a**,**j**, dots show values from biologically independent samples ($n = 3$). The numbers above the bars are *P* values determined by a two-sided Student's *t*-test. Data are represented as mean ± s.e.m.

The pleiotropic QTL harbors tandemly duplicated lncRNA genes

The genes Glyma.12G213800 and Glyma.12G213900 in Wm82 produce 1,526-nucleotide and 1,565-nucleotide transcripts, which are predicted to encode 37 and 49 amino acids, respectively. They are defined as lncRNA genes, referred to as *lncRG1* and *lncRG2*. Both *lncRG1* and *lncRG2* are primarily expressed in stems, leaves and stem tips of PI 479752 at the vegetative 1 (V1) developmental stage when the first trifoliate leaflets are fully expanded. However, they are expressed at significantly lower levels in the same tissues of Wm82, as measured by quantitative PCR with reverse transcription (RT–qPCR) (Fig. 2a) and RNA-seq data from these two parental lines (Supplementary Table 6). Moreover, RNA-seq data from nine diverse *G. soja* accessions and 36 diverse *G. max* accessions[8] (Supplementary Table 7) showed significantly higher expression levels of these two genes in *G. soja* accessions than in *G. max* accessions (Extended Data Fig. 1g) as well as a coexpression pattern between *lncRG1* and *lncRG2* (Extended Data Fig. 1h). Therefore, the suppressed expression of *lncRG1* and *lncRG2* is very likely to be responsible for the observed phenotypic changes from wild soybeans to cultivated soybeans.

Comparison of *lncRG1* and *lncRG2* with all other soybean genes in the Wm82 reference genome indicated that not only the putative coding sequences but also large portions of the noncoding sequences of these two lncRNA genes share similarities with typical MYB transcription factor genes (Fig. 2b,c), suggesting that *lncRG1* and *lncRG2* were derived from MYB genes. Further phylogenetic and comparative genomic analyses showed that *lncRG1* and *lncRG2* were tandemly duplicated before the latest whole-genome duplication (WGD) event (Fig. 2b), predicted to have occurred in soybean ~13 million years ago (MYA)[9]. As a result, there are two paralogs of *lncRG1* and *lncRG2*, dubbed *lncRG4* and *lncRG3*, respectively, residing in the WGD-derived region (Fig. 2b and Extended Data Fig. 1i). Nevertheless, *lncRG3* and *lncRG4* are not associated with any of the domestication QTL[4]. Interestingly, all four lncRG genes in soybean possess inverted repeats, each at ~300–400 bp, corresponding to the third exon of their most closely related MYB genes (Fig. 2c).

https://doi.org/10.1038/s41588-024-01738-2

lncRNA genes produce sRNA species targeting related MYB genes

Based on prediction, the inverted repeats within the transcripts of *lncRG1* and *lncRG2* may form double-stranded stem loops at 453 bp and 337 bp, respectively (Fig. 2d,e), which could be processed to generate sRNA, such as microRNA (miRNA), miRNA-like sRNA or small interfering RNA (siRNA). Next, we sequenced sRNA in the V1 stage stem tips of PI 479752 and Wm82, respectively. Abundant, overlapping sRNA species, mainly at 21–23 nucleotides, across the inverted repeats of both *lncRG1* and *lncRG2* were detected in PI 479752, but their relative abundances varied drastically (Fig. 2f,g). The most abundant sRNA species from *lncRG1* were at 23 nucleotides, whereas the most abundant sRNA species from *lncRG2* were at 21 nucleotides (Fig. 2h,i). Overall, *lncRG2* produced ~18 times more sRNA species than *lncRG1* (Fig. 2f–i). This appears to be related to the higher expression level of the former compared to that of the latter (Supplementary Table 6). Consistent abundances and distribution patterns of the sRNA species produced by *lncRG1* and *lncRG2* were observed in a pair of RILs, RIL186 (*qdrt12.3*) and RIL334 (*qDRT12.3*) (Extended Data Fig. 2a–d), suggesting that the abundance of individual sRNA species is tightly regulated and not randomly produced from the inverted repeats.

A total of 163 genes were predicted to be targets of 27 distinct sRNA species from *lncRG1* and *lncRG2*, with a relative abundance of >100 copies per million sRNA reads (Supplementary Tables 8 and 9). Of these putative targets, only Glyma.01G051700, Glyma.02G110000 and Glyma.02G110100 showed significantly reduced levels of expression in PI 479752 compared with Wm82, with at least twofold changes in stem tips, stems and leaves as determined by RNA-seq and RT–qPCR (Fig. 2j and Supplementary Table 10). Degradome sequencing showed that the mRNA of these three genes was predominantly cleaved at the predicted sRNA target sites in PI 479752 (Fig. 2k–m). Interestingly, all three targets are typical MYB genes that are most closely related to *lncRG1* and *lncRG2* based on the phylogenetic relationships established with the transcript sequences (Fig. 2b). Thus, these MYB gene-derived *lncRG1* and *lncRG2* are likely to modulate DRTs by producing plentiful miRNA-like sRNA species to primarily repress their MYB gene relatives by post-transcriptional regulation.

Overexpressing sRNA promotes wild soybean-type phenotypes

To determine whether the sRNA species produced by *lncRG1* and *lncRG2* underlie DRTs, we first generated Wm82 transgenic lines that overexpress the 'stem loop' part of each gene by the cauliflower mosaic virus (CaMV) 35S promoter. The transgenic lines displayed elevated abundance of sRNA from the stem loops (Extended Data Fig. 2e,f) and showed expected phenotypic changes including appressed pubescence form, decreased plant height and smaller leaf size in comparison to Wm82 (Fig. 3a–c and Extended Data Fig. 2g,h). In addition, we constructed two artificial miRNA precursors (aMIR-sRlncRG1-1 and aMIR-sRlncRG2-3) by replacing the miR172a and miR172a* sequences from the soybean miR172a precursor MIR172a with sRlncRG1-1 and its complementary sRlncRG1-1* or with sRlncRG2-3 and its complementary sRlncRG2-3*, respectively. Overexpression of the two artificial miRNA precursors using the 35S promoter in Wm82 resulted in an appressed or semi-appressed pubescence form, reduced plant height and smaller leaf size compared to Wm82 (Fig. 3d–f). As expected, these transgenic lines exhibited increased expression levels of the corresponding artificial sRNA species and decreased expression levels of the three MYB genes as determined by stem loop and regular RT–qPCR, respectively (Fig. 3g,h). The mRNA of the target genes was confirmed to be principally cleaved at the predicted sRlncRG1-1 and sRlncRG2-3 cleavage sites in the transgenic lines, but such cleavages were not detected in the Wm82 control line using the RNA ligase-mediated rapid amplification of 5′ complementary DNA (cDNA) ends (RLM-RACE) technique followed by deep sequencing (Fig. 3i,j). These observations

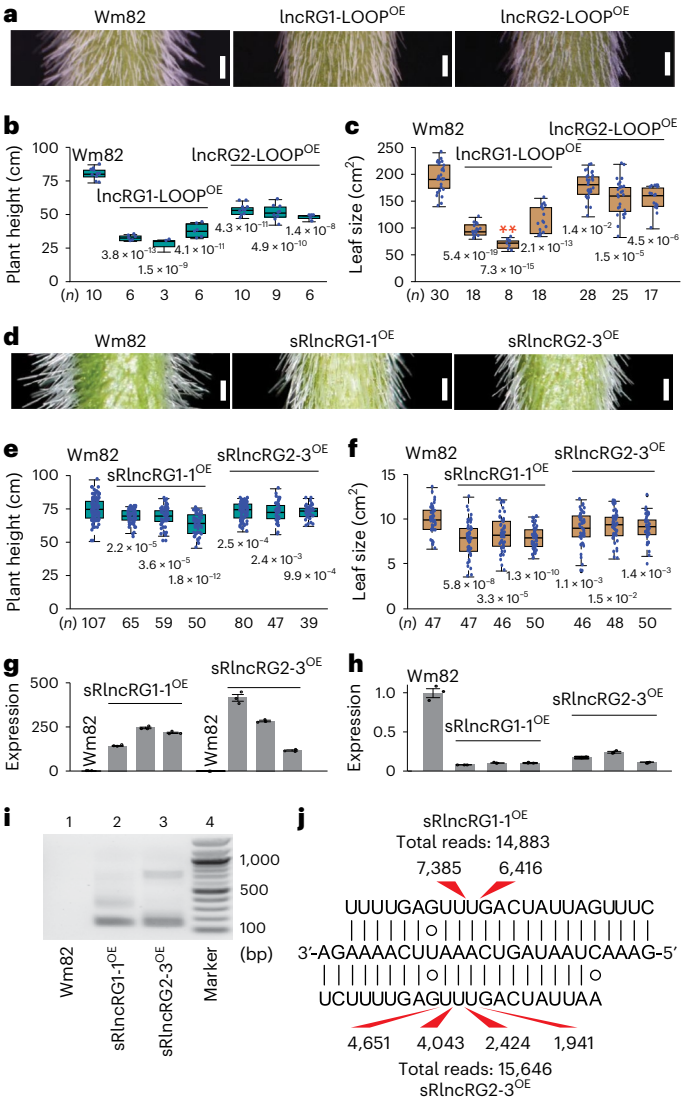


Fig. 3 | Overproduction of sRNA in cultivated soybean promotes wild soybean-type phenotypes. a–c, Phenotypic changes in pubescence form (**a**), plant height (**b**) and leaf size (**c**) of stem loop overexpression (LOOP^{OE}) lines compared with those of Wm82. Scale bars, 3 mm. **d–f**, Phenotypic changes in pubescence form (**d**), plant height (**e**) and leaf size (**f**) of artificial miRNA overexpression lines compared with those of Wm82. Scale bars, 3 mm. **g**,**h**, Expression levels of artificial miRNA (**g**) and the three target MYB genes (**h**) in transgenic lines compared with those in Wm82 as determined by stem loop RT–qPCR and regular RT–qPCR, respectively, with Wm82 set as '1' and the others adjusted accordingly. Dots show values from biologically independent samples ($n = 3$). Data are represented as mean ± s.e.m. **i**, Gel electrophoresis image of RLM-RACE from the transgenic lines and Wm82. The RLM-RACE assay was repeated at least two times. **j**, Cleavage frequencies detected by RLM-RACE followed by deep sequencing. Numbers show the total read number and the read number at each cleavage site. In **b**,**c**,**e**,**f**, horizontal lines indicate the median, and boxes represent the interquartile range (IQR). Whiskers represent the range of 1.5× IQR, and dots beyond the whiskers are outlier values. Numbers at the bottom of the plots indicate the number of independent individuals measured; numbers below the boxes are *P* values determined by a two-sided Student's *t*-test.

confirm that the specific sRNA species produced from *lncRG1* and *lncRG2* affected at least these three DRTs (Fig. 3d–f) and suggest that these sRNA species use a miRNA-like mechanism to repress their targets.

As *lncRG1* and *lncRG2* are predicted to encode two small peptides, we wondered whether the small peptides also contribute to the DRTs.

https://doi.org/10.1038/s41588-024-01738-2

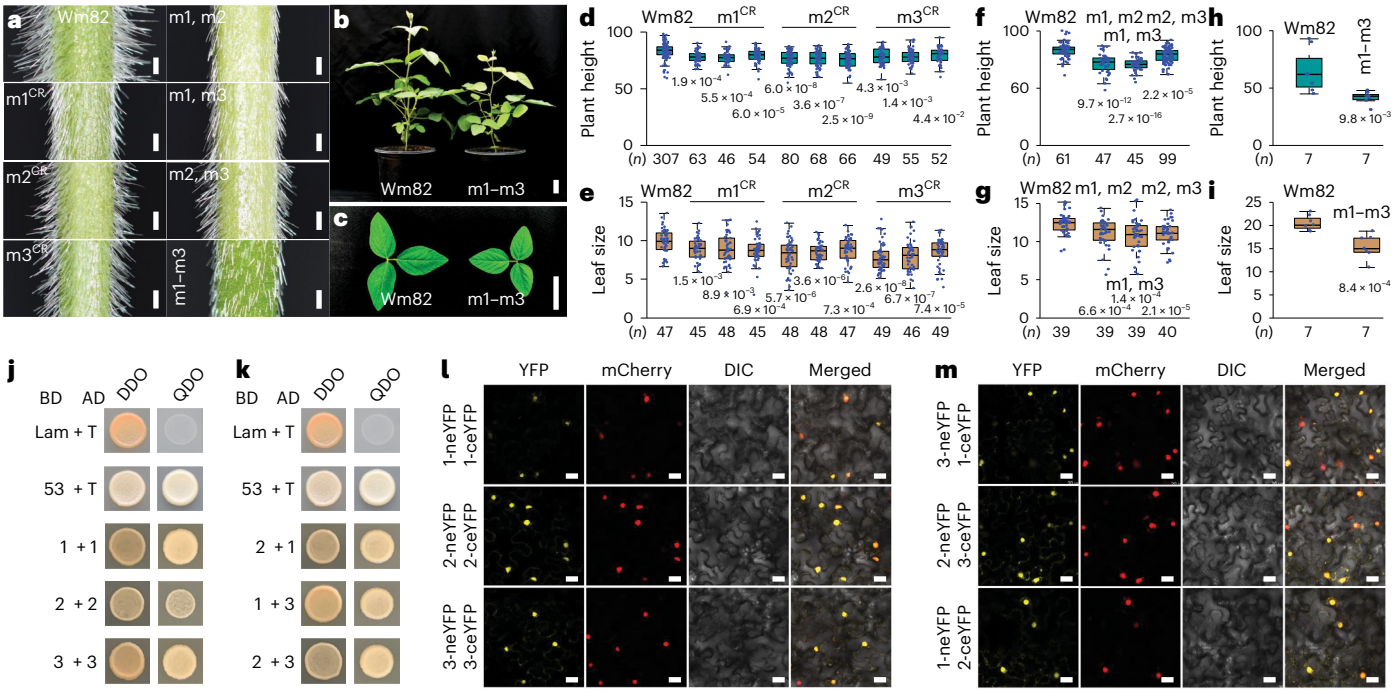


Fig. 4 | Functional redundancy and divergence of the three MYB genes targeted by the sRNA species. a–**c**, Photographic illustration of phenotypic changes in the pubescence form (**a**), plant height (**b**) and leaf size (**c**) of gene-edited mutants compared with Wm82. m1, m2 and m3 are mutants of target 1, target 2 and target 3, respectively. Scale bars, 3 mm in a and 5 cm in **b**,**c**. **d**,**e**, Statistics of plant height (**d**) and leaf size (**e**) of single mutants and Wm82. **f**,**g**, Statistics of plant height (**f**) and leaf size (**g**) of the double mutants and Wm82. CR, CRISPR. **h**,**i**, Statistics of plant height (**h**) and leaf size (**i**) of the triple mutants and Wm82. **j**,**k**, Homo (**j**) and hetero (**k**) protein–protein interactions among the three target genes detected by Y2H assays. AD, activation domain; BD, binding domain; DDO, double dropout; QDO, quadruple dropout; Lam, lamin; T, T-antigen. **l**,**m**, Homo (**l**) and hetero (**m**) protein–protein interactions among the three target genes detected by the BiFC assay. Scale bars, 20 μm. The BiFC assay was repeated two times. eYFP, enhanced yellow fluorescent protein; ceYFP, C-terminal half of eYFP; neYFP, N-terminal half of eYFP; DIC, differential interference contrast image. In **d**–**i**, horizontal lines indicate medians, and boxes represent the IQR. Whiskers represent the range of 1.5× IQR, and dots beyond the whiskers are outlier values. Numbers at the bottom of the plots indicate the number of independent individuals measured; numbers below the boxes are *P* values determined by a two-sided Student's *t*-test.

Therefore, we generated Wm82 transgenic lines that overexpress the predicted coding sequence for the small peptide of each gene by the 35S promoter. No phenotypic differences between any of the transgenic lines and the negative controls were observed, suggesting that the predicted coding sequences are unlikely to modulate DRTs (Extended Data Fig. 2i,j).

MYB targets exhibit functional redundancy and divergence

To gain insights into the mechanism by which the three MYB genes regulate DRTs, we generated Wm82 knockout lines for each of the three MYB genes separately using CRISPR–Cas9 (Extended Data Fig. 3a–c). Knocking out any of the three genes resulted in appressed or semi-appressed pubescence, reduced plant height and smaller leaf size; however, their effects on each DRT slightly varied (Fig. 4a,d,e). We then crossed the knockout lines for different MYB genes to generate double mutants, which were further crossed to create triple mutants. Overall, the double and triple mutants exhibited stronger phenotypic changes than the single mutants (Fig. 4a–c,f–i), suggesting an additive effect of the three MYB genes. As exemplified in Supplementary Videos 1 and 2, appressed pubescence in MYB gene mutants made it easier for leafhoppers to climb than erect pubescence in WM82, which was attributed to leafhopper resistance.

Given that protein dimerization often plays a crucial role in transcription factor activity, we wondered whether the three MYB genes enable homodimerization or heterodimerization. Although these MYB genes are putative transcription factors, none of them showed self-activation activity in yeast two-hybrid (Y2H) assays (Extended Data Fig. 3d); therefore, Y2H assays were suitable and used to test possible homodimerization or heterodimerization involving these MYB genes, followed by validation with bimolecular fluorescence complementation (BiFC) assays in tobacco leaves. Both self and pairwise protein–protein interactions were detected among the three MYB genes (Fig. 4j–m), and, as expected, both homodimers and heterodimers were localized in the nucleus (Fig. 4l,m). Furthermore, the three target MYB genes were shown to be able to interact with their more ancestral MYB genes (Fig. 2b), such as Glyma.07G228600, Glyma.20G032900 and Glyma.04G166900; however, the strengths of the interactions involving each of the three target MYB genes varied (Extended Data Fig. 3e). These observations suggest that the three target MYB genes possess both redundant and divergent functions.

In an attempt to dissect the genetic pathways mediated by the MYB genes modulating the DRTs, we fused the coding sequences of Glyma.01G051700 and Glyma.02G110000 from Wm82 with that for the FLAG epitope, separately, generating transgenic lines that overexpress each of the fused proteins by the 35S promoter. We then conducted chromatin immunoprecipitation followed by sequencing (ChIP–seq) using stem tips collected from transgenic plants. In total, we detected 36,616 and 26,139 peaks, respectively, in both ChIP–seq assays (Supplementary Tables 11 and 12). Among these, 63% were in annotated promoter or gene body regions (Extended Data Fig. 3f,g). As expected, the most frequent binding sites within genic regions were found around the annotated transcription start sites (Extended Data Fig. 3h,i). In total, 8,167 genes were detected as putative targets of the proteins in both assays (Extended Data Fig. 3j). Gene ontology (GO) analysis revealed significant enrichment of genes associated with photosystem and

https://doi.org/10.1038/s41588-024-01738-2

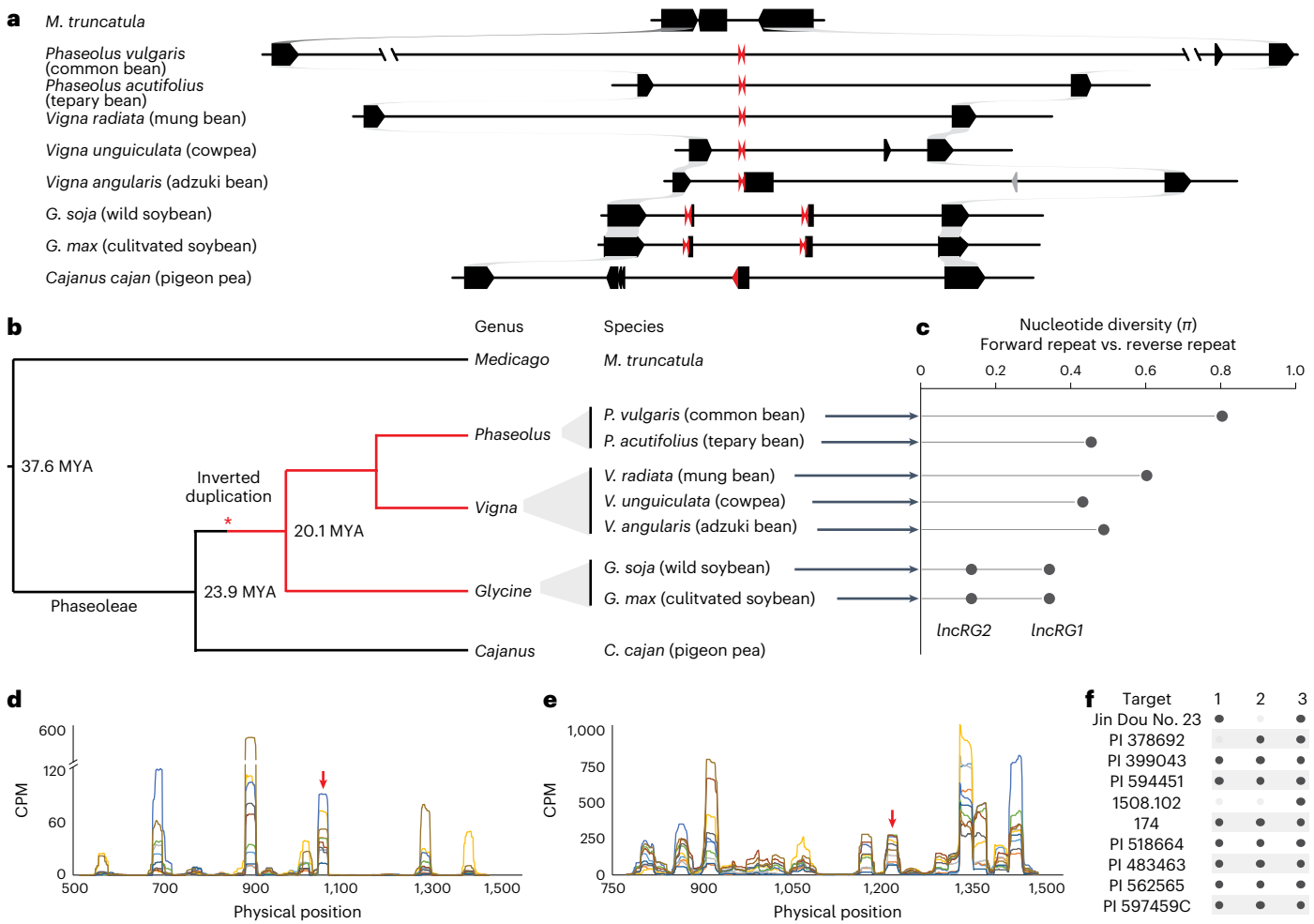


Fig. 5 | The birth and evolutionary consequences of lncRG genes in legumes. **a**, Collinearity analysis of nine legume species at the region harboring the orthologs of *lncRG1* and *lncRG2*. Black boxes present genes, and gray shading connects ortholog genes between species. Red triangles represent inverted repeats. **b**, Phylogenetic relationships of the nine legume species as determined in previous studies[10,11]. Red lines highlight genera that carry the inverted repeats, and the asterisk indicates the deduced time point when the original inverted repeats occurred. **c**, Nucleotide diversity between the forward and reverse repeats in each species. **d**,**e**, Distribution patterns of the sRNA species produced by *lncRG1* (**d**) and *lncRG2* (**e**) in ten diverse soybean accessions as indicated by different colors. Arrows point to the position of major sRNA peaks of PI 479752. **f**, The three MYB genes (targets 1, 2 and 3, as shown in Fig. 2c) predicted to be targeted by the top 20 sRNA species produced by *lncRG1* and *lncRG2* in each of the ten soybean accessions. Black dots indicate predicted targets, while gray dots indicate that they are not predicted to be targets.

photosynthesis and auxin-activated signaling pathways (Extended Data Fig. 3k), which may explain the increased plant height and leaf size through domestication.

Structure–function constraints lead to purifying selection

To track the origin and evolutionary variation of the lncRG genes, we compared the mapped *lncRG1* and *lncRG2* region and its flanking regions of *G. max* and *G. soja* with the corresponding orthologous regions in seven additional leguminous species belonging to the *Phaseolus*, *Vigna* and *Cajanus* genera using *Medicago truncatula* as an outgroup. It appears that the tandem duplication event occurred after the divergence of *Glycine* and *Phaseolus*–*Vigna* from a common ancestor ~20 MYA[10,11] (Fig. 5a,b). However, the inverted repeats were also seen in *Phaseolus* and *Vigna* but not in *Cajanus* and *M. truncatula*, suggesting that the inverted repeats were formed before the divergence of *Glycine* from *Phaseolus* and *Vigna* but after its divergence from *Cajanus* ~20–24 MYA[10,11] (Fig. 5a,b). According to the 26 well-assembled *G. soja* and *G. max* genomes[6], the *lncRG1* and *lncRG2* regions are highly conserved in terms of gene content, without deletion or insertion of genic sequences (Extended Data Fig. 4a). The inverted repeats of *lncRG1* and *lncRG2* in *G. soja* and *G. max* exhibited the lowest level of divergence compared with the inverted repeats in the orthologs of *lncRG1* and *lncRG2* in *Phaseolus* and *Vigna* (Fig. 5c), indicating that the inverted repeats, as functional parts of the *lncRG1* and *lncRG2* gene bodies, have experienced strong 'purifying selection'.

Why were *lncRG3* and *lncRG4* not associated with any of the DRTs modulated by *lncRG1* and *lncRG2*? We found that, different from *lncRG1* and *lncRG2*, which were highly expressed in V1 stage stem tips of PI 479752 to modulate the 'wild' phenotypes, *lncRG3* and *lncRG4* were expressed at very low levels in the same tissue of PI 479752 and produced few sRNA species (Supplementary Table 6 and Extended Data Fig. 4b–d). Although *lncRG3* was expressed at a much higher level in Wm82 (Supplementary Table 6 and Extended Data Fig. 4b), the relative abundance of sRNA species mapping to *lncRG3* was extremely low (Extended Data Fig. 4d). Moreover, *lncRG3* and *lncRG4* in the 26 genomes[6] exhibited higher levels of sequence divergence between respective inverted repeat sequences than *lncRG2* and *lncRG1*, respectively (Extended Data Fig. 4e). Together, these observations suggest that the two pairs of paralogs (*lncRG2* and *lncRG1* versus *lncRG3* and *lncRG4*) in wild soybeans have diverged functionally, perhaps through reduction of *lncRG3* and *lncRG4* expression and/or loss of their capability to produce sRNA species.

sRNA species exhibit diverse distribution patterns in soybeans

The availability of sRNA sequencing data from nine *G. soja* and 36 *G. max* accessions[8] allowed us to compare the distribution and relative abundance of sRNA species generated by *lncRG1* and *lncRG2* at the population level (Fig. 5d,e, Extended Data Fig. 5a,b and Supplementary Table 13). As expected, all nine *G. soja* accessions and a cultivated soybean accession (Jin Dou No. 23) with appressed pubescence produced abundant sRNA species from *lncRG1* and *lncRG2*. By contrast, few sRNA species were produced from *lncRG1* and *lncRG2* in the remaining 35 cultivated soybean accessions with erect pubescence. Remarkably, sRNA distribution patterns varied drastically among the ten accessions with appressed pubescence, and, in most cases, different sRNA species were predicted to target the three MYB genes, and up to 41% of the predicted sRNA targets in one accession were not shared by another accession (Fig. 5f and Supplementary Table 14). As observed in PI 479752 (Fig. 2g), *lncRG2* in each of the ten accessions produced more nonredundant and more abundant sRNA species than *lncRG1*, first with 21-nucleotide and then 22-nucleotide sRNA species as the predominant forms (Extended Data Fig. 5a,b).

Lower lncRNA gene expression is linked to CpG methylation

Transcriptional gene silencing is often associated with promoter methylation in both animals and plants[12]; we thus investigated the distribution of CpG, CHG and CHH DNA methylation along the *lncRG1* and *lncRG2* genomic sequences using whole-genome bisulfite sequencing data from the panel of 45 *G. soja* and *G. max* accessions[8]. We observed that CpG methylation levels in the promoter regions of both *lncRG1* and *lncRG2* in *G. max* accessions were significantly higher than those in *G. soja* accessions (Extended Data Fig. 6a). Moreover, we observed significant negative correlations between CpG methylation levels in the promoter regions of *lncRG1* and *lncRG2* and expression levels of the two genes (Extended Data Fig. 6b,c). These observations may suggest that elevated levels of CpG methylation in the promoter regions of *lncRG1* and *lncRG2* could be responsible for reduced expression levels of the two genes in cultivated soybeans and thus for the underlying changes of the DRTs.

Discussion

lncRNA species are ubiquitously present in eukaryotes and play important roles in regulating gene expression[13]. However, how they originated and execute their functions remains largely unknown. In this study, we demonstrate that two lncRNA tandem duplicates, *lncRG1* and *lncRG2*, were derived from MYB genes and underwent exonic sequence rearrangement to form inverted repeats. Intragenic inverted repeats are typically lost due to their instability and fitness costs[14]; yet the inverted repeats in *lncRG1* and *lncRG2* have been maintained over the course of 20–24 million years of evolution (Fig. 5), likely due to their crucial role in regulating multiple 'wild' adaptive traits in *Glycine*. The inverted repeat structures are still detectable across the *Phaseolus*, *Vigna* and *Glycine* genera, reflecting their functional constraints at variable levels. Given such a great variation in relative abundance and distribution of *lncRG1*- and *lncRG2*-derived sRNA species among different wild soybean accessions, the functional constraints in wild soybeans may be implemented through purifying selection across the entire inverted repeat regions. It would be interesting to explore whether inverted repeats in other legumes have similar functionality and regulate comparable traits and whether inverted repeats were also targeted for selection during domestication of other leguminous crops.

Integration of inverted repeats in the genome can result from processes such as DNA replication repair or transposable elements (TEs)[15]. Inverted repeats derived from TEs are typically processed by enzymes such as DCL3 or DCL3-like protein, resulting in the production of 24-nucleotide sRNA species[16,17]. These sRNA species often have a noticeable impact on expression of nearby genes and phenotypic traits[17–22]. By contrast, the inverted repeats found in *lncRG1* and *lncRG2* predominantly give rise to 21–23-nucleotide sRNA species (Fig. 2h,i). It is interesting to note that *lncRG1* generates 23-nucleotide sRNA species, while *lncRG2* primarily yields 21–22-nucleotide sRNA species (Fig. 2h,i). In concordance, *lncRG1*, which produces 23-nucleotide sRNA species, displays signs of gene body methylation (Extended Data Fig. 6a), as reported for other inverted repeats. This suggests that, in this instance, 23-nucleotide sRNA species might play a role in triggering DNA methylation, potentially aiding in the regulation of this locus between *G. max* and *G. soja*. However, unlike TE-derived inverted repeats[19], we did not observe differential gene expression in the vicinity of *lncRG1* and *lncRG2* when comparing Wm82 and PI 479752 according to the RNA-seq data. This further supports the idea that *lncRG1* and *lncRG2* operate through a distinct mechanism. In this context, the 21-nucleotide and 22-nucleotide sRNA species generated from the conserved *lncRG2* gene, which account for over 80–89% of the total sRNA species produced in these loci in *G. soja* accessions, are likely the major contributors to gene regulation. Consequently, the mechanism responsible for processing inverted repeats in *lncRG1* and *lncRG2* to generate sRNA species is probably distinct from the process seen in TE-derived inverted repeats and more akin to the miRNA pathway, as previously reported for evolutionarily young miRNA species[23].

While sRNA may also repress translation without cleaving mRNA[24], it is unclear whether the remaining 160 predicted sRNA targets, which show no difference in expression levels between Wm82 and PI 479752 (Supplementary Tables 9 and 10), are directly regulated by sRNA from *lncRG1* and *lncRG2* through translational inhibition. Given the fact that the three MYB targets also interact with additional, more divergent copies of MYB genes, that the predominant sizes of sRNA produced from *lncRG1* and *lncRG2* are different and that sRNA species from *lncRG1* and *lncRG2* and their putative targets are highly variable among different accessions, the pleiotropic effects of *lncRG1* and *lncRG2* and the mechanisms by which they execute their full suite of functions are likely to be more extensive than what has been observed.

A few domestication genes have been shown to exhibit pleiotropic effects on multiple traits[2], such as *TEOSINTE BRANCHED1* in maize, which controls branching, inflorescence architecture and plant height[25], and *PROSTRATE GROWTH1* in rice, which controls tiller angle, panicle size and seed shattering[26]. Compared to these genes, the mechanism by which *lncRG1* and *lncRG2* execute their pleiotropic effects is unique and reflective of evolutionary innovation triggered by varied types of duplication events including exonic duplication, genic duplication and WGD. In soybean, approximately 75% of genes exist in multiple copies, which were primarily generated via two rounds of WGD events that occurred 59 and 13 MYA[9]. Consequently, mutations within a single gene can often be 'rescued' by its functionally redundant duplicates. In such a case, phenotypic transition of a DRT during soybean domestication would have involved artificial selection of mutations within two or more duplicated genes. As the sRNA species produced by *lncRG1* and *lncRG2* enable simultaneous repression of multiple duplicated MYB genes and most likely additional genes as well, artificial selection of the DRTs regulated by these genes was achieved simply by selecting reduced expression of *lncRG1* and *lncRG2* within a single locus producing fewer sRNA species. Although this locus possesses pleiotropic effects on multiple morphological traits, the most favorable phenotype targeted by ancient farmers for selection may be insect resistance attributed to erect pubescence, which was modulated mainly by this single major QTL detected in the whole genome.

The causal mutations for reduced expression of *lncRG1* and *lncRG2* in cultivated soybeans remain poorly understood. Genome-wide association analysis with the resequencing data from 74 *G. soja* and 596 *G. max* accessions[6] showed numerous polymorphic sites across the entire mapping region that are highly associated with the phenotypic differences in pubescence form (Extended Data Fig. 1a,b), but no single polymorphic sites in the putative promoters of the two genes or other parts of the region could explain the phenotypic differences better

than the others. This is not unexpected, given that the entire region has undergone a selective sweep (Extended Data Fig. 1f). On the other hand, expression levels of *lncRG1* and *lncRG2* were also found to be associated with differences in CpG methylation in their promoter regions between *G. max* and *G. soja* (Extended Data Fig. 6). Because *lncRG1* and *lncRG2* are coexpressed across different tissues and developmental stages, there is a possibility that these two genes are regulated by the same genetic or epigenetic (or both) regulatory element(s) within the mapped 29-kb region. Under this caveat, extensive functional assays are needed to pinpoint the causal mutation(s) for reduced *lncRG1* and *lncRG2* expression.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01738-2.

References

1. Olsen, K. M. & Wendel, J. F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
2. Doebley, J. F., Gaut, B. S. & Smith, B. D. The molecular genetics of crop domestication. *Cell* **127**, 1309–1321 (2006).
3. Sedivy, E. J., Wu, F. & Hanzawa, Y. Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol.* **214**, 539–553 (2017).
4. Swarm, S. A. et al. Genetic dissection of domestication-related traits in soybean through genotyping-by-sequencing of two interspecific mapping populations. *Theor. Appl. Genet.* **132**, 1195–1209 (2019).
5. Broersma, D., Bernard, R. & Luckmann, W. Some effects of soybean pubescence on populations of the potato leafhopper. *J. Econ. Entomol.* **65**, 78–82 (1972).
6. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176 (2020).
7. Song, Q. et al. Fingerprinting soybean germplasm and its utility in genomic research. *G3* **5**, 1999–2006 (2015).
8. Shen, Y. et al. DNA methylation footprints during soybean domestication and improvement. *Genome Biol.* **19**, 128 (2018).
9. Schmutz, J. et al. Genome sequence of the palaeopolyploid soybean. *Nature* **463**, 178–183 (2010).
10. Choi, H.-K. et al. Estimating genome conservation between crop and model legume species. *Proc. Natl Acad. Sci. USA* **101**, 15289–15294 (2004).
11. Zheng, F. et al. Molecular phylogeny and dynamic evolution of disease resistance genes in the legume family. *BMC Genomics* **17**, 402 (2016).
12. Vaucheret, H. & Fagard, M. Transcriptional gene silencing in plants: targets, inducers and regulators. *Trends Genet.* **17**, 29–35 (2001).
13. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
14. Parniske, M. et al. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. *Cell* **91**, 821–832 (1997).
15. Reams, A. B. & Roth, J. R. Mechanisms of gene duplication and amplification. *Cold Spring Harb. Perspect. Biol.* **7**, a016592 (2015).
16. Cuerda-Gil, D. & Slotkin, R. K. Non-canonical RNA-directed DNA methylation. *Nat. Plants* **2**, 16163 (2016).
17. Gagliardi, D. et al. Dynamic regulation of chromatin topology and transcription by inverted repeat-derived small RNAs in sunflower. *Proc. Natl Acad. Sci. USA* **116**, 17578–17583 (2019).
18. Lu, C. et al. Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in *Oryza sativa*. *Mol. Biol. Evol.* **29**, 1005–1017 (2012).
19. Arce, A. L. et al. Polymorphic inverted repeats near coding genes impact chromatin topology and phenotypic traits in *Arabidopsis thaliana*. *Cell Rep.* **42**, 112029 (2023).
20. Wu, N. et al. A MITE variation-associated heat-inducible isoform of a heat-shock factor confers heat tolerance through regulation of *JASMONATE ZIM-DOMAIN* genes in rice. *New Phytol.* **234**, 1315–1331 (2022).
21. Niu, C. et al. Methylation of a MITE insertion in the *MdRFNR1-1* promoter is positively associated with its allelic expression in apple in response to drought stress. *Plant Cell* **34**, 3983–4006 (2022).
22. Xu, L. et al. Regulation of rice tillering by RNA-directed DNA methylation at miniature inverted-repeat transposable elements. *Mol. Plant* **13**, 851–863 (2020).
23. Bradley, D. et al. Evolution of flower color pattern through selection on regulatory small RNAs. *Science* **358**, 925–928 (2017).
24. Fabian, M. R. & Sonenberg, N. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat. Struct. Mol. Biol.* **19**, 586–593 (2012).
25. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485–488 (1997).
26. Tan, L. et al. Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* **40**, 1360–1364 (2008).

Methods

Plant materials

The entire mapping population consisted of 2,287 $F_{6:7}$ RILs derived from a cross between *G. max* (Wm82) and *G. soja* (PI 479752). The recombinants, as listed in Supplementary Table 1, were identified by screening the entire RIL population using the boundary markers M1 and M10 from primary QTL mapping. The resequenced association mapping population, as listed in Supplementary Table 2, for pubescence form was selected from the soybean pan-genome project[6]. The association mapping population for leafhopper resistance and pubescence form, as listed in Supplementary Table 4, was sourced from the USDA soybean germplasm collection[7] (https://www.ars-grin.gov/). The 45 highly diverse soybean accessions, as listed in Supplementary Table 7, which have RNA-seq, sRNA and WGBS data available, were from a previous study[8]. Wm82 was used for stable transformation and genome editing; *Nicotiana benthamiana* was used for BiFC assays.

Quantitative trait loci and association mapping

A subset of the RIL population (510 RILs) was genotyped using the genotyping-by-sequencing method. Approximately 8,000 SNP markers were used to identify QTL underlying the following DRTs: pubescence form, main stem length, leaf size, growth habit and stem twining. QTL mapping was performed in R (version 4.2.1)[27] using the composite interval mapping method[28] incorporated in R/qtl (version 1.66)[29]. Phenotypic data for association mapping were downloaded from the USDA National Plant Germplasm System (https://npgsweb.ars-grin.gov/), and SoySNP50K data were obtained from a previous study[7]. Resequencing data were from the soybean pan-genome study[6]. Association mapping was performed using TASSEL 5 (ref. 30) with a mixed linear model[31].

Recombinant genotyping and phenotyping

All mapping markers were designed based on resequencing data of PI 479752 from a previous study[32]. DRTs were examined for all recombinants in the field at the Purdue Agronomy Center for Research and Education in 2018. Pubescence form was classified as erect, semi-appressed and appressed; main stem length was measured in cm from the soil surface to the top node of the main stem; growth habit was classified with a visual score on a scale of 1–5 to describe growth tendencies (1, erect *G. max*-like growth type; 5, prolific *G. soja*-like growth); leaf size was determined based on the length of a terminal leaflet from the top third of the canopy; stem twining was determined based on a scale of 1–4 to describe the degree of stem twining (1, no twining; 4, *G. soja*-like twining). All primers used in this study were synthesized by Eurofins Genomics and are listed in Supplementary Table 15.

Transgene constructs

For stem loop overexpression, the stem loops of *lncRG1* and *lncRG2* were amplified from genomic DNA of PI 479752 using primers with 20-bp recombination arms, and nested PCR was used to amplify the stem loops. Meanwhile, the plasmid vector pPTN1171 was digested with the restriction enzymes NcoI (R0193S, New England Biolabs) and XbaI (R0145S, New England Biolabs) at 37 °C for 4 h. PCR products and the linearized vector were purified using the PureLink Quick Gel Extraction Kit (K210012, Thermo Fisher Scientific). Stem loops were inserted into the plasmid vector using the ClonExpress II One Step Cloning Kit (C112, Cellagen Technology). The final constructs were confirmed by Sanger sequencing.

For artificial miRNA overexpression, soybean miR172a was used as the backbone. The miR172a and miR172a* sequences were replaced by sRlncRG1-1 and sRlncRG2-3 and their corresponding reverse complementary sequences. The replaced sequences were synthesized at Integrated DNA Technologies. The forward sequence and the complementary sequence were annealed for 5 min at 95 °C and then cooled to room temperature to form dimers and inserted into pPTN1171. The final constructs were confirmed by Sanger sequencing.

For CRISPR–Cas9 editing, four sgRNA species were designed for each target gene, Glyma.01G051700, Glyma.02G110000 and Glyma.02G110100, using CRISPR-P, a web-based guide RNA design tool[33]. The primer pairs were annealed for 5 min at 95 °C and then cooled to room temperature to form dimers. The dimers were inserted into the pGEL201 vector, linearized by the restriction enzyme BsaI (R0535, New England Biolabs)[34]. The final constructs were confirmed by Sanger sequencing. During transformation, four agrobacteria with different sgRNA species were mixed equally before infection.

For Y2H assays, the full-length coding sequences of Glyma.01G051700, Glyma.02G110000 and Glyma.02G110100 as well as other MYB genes were cloned from the cDNA sample of 'Wm82' and then inserted into the vectors pGBKT7 and pGADT7 using the ClonExpress II One Step Cloning Kit (C112, Cellagen Technology). The final constructs were confirmed by Sanger sequencing.

For the BiFC assay, the full-length coding sequences of Glyma.01G051700, Glym.02G110000 and Glyma.02G110100 were amplified and cloned into plasmids pCNHP-neYFP-C and pCNHP-ceYFP-C, which express fusion proteins with either neYFP or ceYFP at their N terminus, using the ClonExpress II One Step Cloning Kit (C112, Cellagen Technology), respectively. The final constructs were confirmed by Sanger sequencing.

Soybean transformation

Mature seeds from soybean cultivar 'Wm82' were disinfected using chlorine gas for 12 h. The disinfected seeds were soaked in distilled water for 12 h at room temperature in the dark. Half seeds were soaked in resuspended agrobacterium liquid co-cultivation medium (OD_{650} = 0.6, 3.21 g l^{-1} Gamborg B-5 Basal Medium, 30 g l^{-1} sucrose, 3.9 g l^{-1} MES, 0.4 g l^{-1} L-cystine, 0.1542 g l^{-1} DTT, 0.25 mg l^{-1} GA3, 1.67 mg l^{-1} 6-BA and 0.3924 g l^{-1} acetosyringone, pH 5.4) for 30 min. After infection, explants were transferred to solid co-cultivation medium. The plates were sealed with Micropore tape (1530-0, 3M) and incubated in the dark at 21 °C for 4 d. After co-cultivation, explants were inserted into a plate with shoot-induction medium (3.21 g l^{-1} Gamborg B-5 Basal Medium, 30 g l^{-1} sucrose, 0.59 g l^{-1} MES, 0.25 g l^{-1} timentin, 0.1 g l^{-1} cefradine, 1.67 mg 6-BA, 2.5 mg l^{-1} glufosinate, pH 5.7, 2 g l^{-1} gellan gum powder). Shoot induction was carried out at 26 °C with a photoperiod of 18 h and a light intensity of 40–70 μE m^{-2} s^{-1}. After 4 weeks, the inducted shoots were cut from cotyledons and transferred to shoot-elongation medium (4.43 g l^{-1} Murashige & Skoog modified medium with Gamborg vitamins, 30 g l^{-1} sucrose, 0.59 g l^{-1} MES, 0.25 g l^{-1} timentin, 0.1 g l^{-1} cefradine, 0.05 g l^{-1} asparagine, 0.05 g l^{-1} glutamine, 0.5 mg l^{-1} GA3, 0.1 mg l^{-1} IAA, 1 mg l^{-1} zeatin, 5 mg l^{-1} glufosinate, pH 5.7, 2 g l^{-1} gellan gum powder) under the same temperature and photoperiod conditions. After 2–4 weeks in shoot-elongation medium, the glufosinate-resistant shoots were cut and transferred to rooting medium (4.43 g l^{-1} Murashige & Skoog modified medium with Gamborg vitamins, 30 g l^{-1} sucrose, 0.59 g l^{-1} MES, 0.05 g l^{-1} asparagine, 0.05 g l^{-1} glutamine, 0.1 mg l^{-1} IBA, pH 5.7, 3 g l^{-1} gellan gum) for further shoot and root elongation. After roots grew longer than 1 cm, plants were transferred to moistened Berger BM2 soil (Berger) and kept enclosed in a clear plastic tray in a growth chamber at 26 °C with a 16-h photoperiod at 250–350 μE m^{-2} s^{-1}.

Genotyping and phenotyping transgenic and edited lines

Genomic DNA was extracted from leaf samples of T_0, T_1 and T_2 plants. The presence of transgenes in the transgenic plants was confirmed by PCR with primers specific to the vector and the corresponding transgene. Expression of the transgene was monitored by RT–qPCR for mRNA or stem loop RT–qPCR for sRNA. For genome-editing lines, target genes were amplified and sequenced to confirm the presence of the frameshift mutation.

The pubescent forms of both transgenic and genome-editing lines were assessed during the V1 stage in the greenhouse. Plant height measurements for both transgenic and genome-editing lines were taken at

the R7 stage (beginning of maturity) in the field. For transgenic lines that overexpress stem loops, leaf area (LA) was determined following the general equation LA = 2.0185 × (length × width)[35]. For transgenic lines that overexpress an sRNA and genome-editing lines, leaf sizes were measured by scanning the unifoliolate leaf at the VC stage (cotyledons expanded) and analyzing LAs using ImageJ (version 1.53k)[36].

RNA extraction, regular RT–qPCR and stem loop RT–PCR

Tissues collected from plants were immediately frozen in liquid nitrogen. Samples were stored at −70 °C before RNA extraction. Total RNA was extracted using the TRIzol reagent (15596018, Invitrogen) following the manufacturer's protocol. RNA concentration and purity were evaluated using the NanoDrop 2000 spectrophotometer (ND-2000, Thermo Fisher Scientific). Twelve micrograms of total RNA was treated with the Invitrogen TURBO DNA-free Kit (AM1907, Invitrogen) following the user manual. Two micrograms of DNA-free RNA was used to synthesize cDNA with Promega M-MLV Reverse Transcriptase (M1701, Promega) following the user manual. RT–qPCR was performed using Applied Biosystems Power SYBR Green PCR Master Mix (4368577, Applied Biosystems) on an Applied Biosystems StepOnePlus Real-Time PCR System (4376600, Applied Biosystems). For stem loop RT–PCR, miRNA-specific stem loop RT primers bind to the 3′ portion of the miRNA molecules, and reverse transcription occurs with M-MLV Reverse Transcriptase (M1701, Promega). Next, the RT product was quantified using Applied Biosystems Power SYBR Green PCR Master Mix (4368577, Applied Biosystems) plus miRNA-specific forward primers and common reverse primers[37]. Relative gene expression levels were calculated by the $2^{-\Delta\Delta Ct}$ method.

mRNA, sRNA and degradome sequencing and data analysis

RNA samples were prepared in accordance with Novogene's sample preparation instructions. RNA-seq, sRNA and degradome libraries were constructed by Novogene. Cleaned data were obtained after sequencing. To improve mapping quality of the nonreference accession, SNP-corrected references were made for *G. soja* PI 479752 as well as all other accessions used in this study. SNP-corrected references were made by taking the Wm82 reference fasta file and replacing the nucleotides where a SNP was present between Wm82 and other accessions.

Each sequencing data file was aligned to its respective corrected reference. RNA-seq reads were mapped to the respective genomes using STAR (version 2.5.4b)[38] with the following parameters: '–outFilterMultimapNmax 1–alignIntronMin 20–alignIntronMax 10000'. Expression levels (FPKM) were calculated using the cuffnorm function in cufflinks (version 2.2.1)[39]. sRNA species shorter than 17 nucleotides or longer than 25 nucleotides were excluded in the study. sRNA and degradome reads were mapped to the respective genomes using Bowtie 2 (version 2.5.1)[40], with only unique mapped reads kept and no mismatches allowed (-v 0 -a -m 1). The potential target genes of miRNA produced by *lncRG1* and *lncRG2* were analyzed using CleaveLand (version 4.5)[41] with the following parameters: -r 0.6 and -c 2. WGBS data were collected and extracted from the NCBI database using sra-toolkit (https://www.ncbi.nlm.nih.gov/sra). The reads were uniquely mapped to each corrected pseudo-reference genome by Bismark (version 0.23.1)[42]. After filtering the duplicate reads, the methylation information for each cytosine site was extracted. The average methylation levels of 300-bp sliding windows with 50-bp steps were calculated.

Chromatin immunoprecipitation followed by sequencing and data analysis

Chromatin immunoprecipitation was performed using stem tips of Glyma.01G051700-FLAG- and Glyma.02G110000-FLAG-overexpression lines at the V1 stage. In brief, 3 g of stem tips was fixed with 1% formaldehyde for 20 min. Subsequently, nuclei were isolated, and the chromatin solution was subjected to 30 min of sonication to fragment the DNA into sizes ranging from 200 to 500 bp. For immunoprecipitation, we employed ANTI-FLAG M2 Magnetic Beads (Sigma, M8823-1ML) targeting the FLAG epitope and the IgG control (Sigma-Aldrich, I5006). Immune complexes were captured using protein G agarose (Millipore, 16-266), and DNA purification was carried out using the QIAquick PCR Purification Kit (Qiagen, 28106). The purified DNA samples were then forwarded to Novogene for sequencing.

ChIP–seq reads were processed by aligning them to the soybean reference genome with the Burrows–Wheeler Aligner program (version 0.7.15)[43]. Peak identification was accomplished with model-based analysis of ChIP–seq (MACS2) (version 2.1.0)[44]. To identify over-represented GO terms among the MYB target genes, we conducted GO enrichment analysis using clusterProfiler (version 4.0)[45]. Significance (*P* value) was adjusted for false discovery rate. GO terms with *a q* value < 0.05 were considered significantly enriched. Finally, network visualization was executed using BiNGO (version 3.0.3)[46].

RNA ligase-mediated 5′ rapid amplification of cDNA ends

Total RNA was extracted using the TRIzol reagent (15596018, Invitrogen) according to the manufacturer's protocol. RNA concentration and purity were evaluated using the NanoDrop 2000 spectrophotometer (ND-2000, Thermo Fisher Scientific). Twelve micrograms of total RNA was treated with the Invitrogen TURBO DNA-free Kit (AM1907, Invitrogen). mRNA was then ligated with 5′ RACE oligonucleotide adaptors for reverse transcription using the GeneRacer Kit (L150202, Thermo Fisher Scientific), followed by nested PCR. The purified PCR products were sequenced using the WideSeq method (https://www.purdue.edu/hla/sites/genomics/wideseq-2/).

Phylogenetic analysis and nucleotide diversity calculation

Sequence alignments of the MYB genes in Fig. 2c and construction of the phylogenetic tree in Fig. 2b were performed using the maximum likelihood method[47] in MEGA7 (ref. 48) using transcript (nucleotide) sequences of the MYB genes. Nucleotide diversity was calculated using VCFtools (version 0.1.16)[49].

RNA secondary structure prediction

The secondary structures of *lncRG1* and *lncRG2* were predicted using the RNAfold server incorporated in ViennaRNA Web Services (http://rna.tbi.univie.ac.at/) and using the transcript sequences of *lncRG1* and *lncRG2* as input.

MicroRNA target prediction

Potential targets of the miRNA species from *lncRG1* and *lncRG2* were predicted using the online tool psRNATarget (https://www.zhaolab.org/psRNATarget/, Schema V2 2017 release) with the expectation cutoff set to 2.5 (ref. 50).

Yeast two-hybrid assays

Y2H assays were performed using the Matchmaker Gold Yeast Two-Hybrid System Kit (630489, Takara Bio). Different combinations of the constructs were cotransformed into the yeast strain Y2H Gold by following the manufacturer's protocol. Transformed yeast cells were spread on SD (−Trp, −Leu) medium. The plates were incubated at 30 °C for 3–5 d. Five to ten colonies were picked from each plate and resuspended in 0.9% (wt/vol) NaCl solution. Next, the yeast cells were spotted on SD (−Trp, −Leu, −Ade, −His) selection medium. Plates were incubated at 30 °C for 3 d to observe yeast growth. pGADT7-T + pGBKT7-53 was used as the positive control; pGADT7-T + pGBKT7-Lam was used as the negative control.

Bimolecular fluorescence complementation

Different constructs were transformed into *Agrobacterium tumefaciens* strain EHA105. Single colonies for each construct were picked and cultured at 28 °C in 3 ml LB medium supplemented with 50 mg l^{-1} rifampicin and 50 mg l^{-1} kanamycin to an OD_{600} of about 2.0. Bacterial

cultures were pelleted, washed with 10 mM $MgCl_2$ and MES (pH 5.7) solution containing 200 µM acetosyringone and incubated in the same solution for an additional 2 h at room temperature. Before infiltration, cultures were mixed to reach a final OD_{600} of 0.6 for each of the constructs used.

The agrobacterium suspension was injected into the abaxial surface of 4–6-week-old *N. benthamiana* leaves with a needleless syringe. Plasmid used to express mCherry-labeled Petunia hybrida's histone H1-3 (acting as the nuclear marker) was co-infiltrated with the expression construct of each target gene. Seventy-two hours after infiltration, fluorescent signals in detached leaves were imaged using a Zeiss LSM 880 laser scanning confocal microscope (Zeiss). The excitation wavelength and emission bandwidth recorded for each fluorescent protein were optimized by the default presets in ZEN 2.6 software (Zeiss) and were as follows: eYFP (excitation, 514 nm; emission, 519–583 nm), mCherry (excitation, 561 nm; emission, 580–651 nm).

Statistical analysis

P values and sample sizes are provided in the individual figures and/or figure legends. Statistical differences between two groups were assessed using a two-sided Student's *t*-test in Excel. Two-sided Pearson correlation coefficients and their corresponding *P* values were calculated using R (version 4.2.1). *P* values from GWAS were determined using the *F*-test for each marker in TASSEL (version 5.0). *P* values from peak enrichment in the ChIP–seq analysis were determined using the Poisson test in MACS2 (version 2.1.0). *P* values for GO enrichment were determined using Fisher's exact test in clusterProfiler (version 4.0).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data are available in the main text, the Supplementary Information, public databases or referenced studies. All raw sequence data generated in this study have been deposited in the NCBI database under BioProject PRJNA876203. Genotypic data from the USDA soybean germplasm collection used for the GWAS on pubescence form and leafhopper resistance in Extended Data Fig. 1c,d were downloaded from the SoyBase database (https://soybase.org/snps/download.php). Genotypic data of the resequenced soybean accessions used for the GWAS on pubescence form in Extended Data Fig. 1a,b were downloaded from the Genome Variation Map database in BIG Data Center (http://bigd.big.ac.cn/gvm/getProjectDetail?project=GVM000063). RNA-seq, sRNA and WGBS data of the 45 highly diverse soybean accessions were download from the Sequence Read Archive database at NCBI under accession number PRJNA432760 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA432760). Source data are provided with this paper.

Code availability

All software used in this study is publicly available as described in the Methods and the Reporting summary. Detailed parameters used for analyzing each type of sequencing data have been described in the Methods. An in-house Perl scrip used for creating SNP-corrected genomes is available at Zenodo (https://doi.org/10.5281/zenodo.10801184)[51].

References

27. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (2013).
28. Zeng, Z.-B. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468 (1994).
29. Broman, K. W., Wu, H., Sen, Ś. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).
30. Bradbury, P. J. et al. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).
31. Yu, J. et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* **38**, 203–208 (2006).
32. Zhou, Z. et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).
33. Lei, Y. et al. CRISPR-P: a web tool for synthetic single-guide RNA design of CRISPR-system in plants. *Mol. Plant* **7**, 1494–1496 (2014).
34. Bai, M. et al. Generation of a multiplex mutagenesis population via pooled CRISPR–Cas9 in soya bean. *Plant Biotechnol. J.* **18**, 721–731 (2020).
35. Richter, G. L. et al. Estimating leaf area of modern soybean cultivars by a non-destructive method. *Bragantia* **73**, 416–425 (2014).
36. Abràmoff, M. D., Magalhães, P. J. & Ram, S. J. Image processing with ImageJ. *Biophotonics Int.* **11**, 36–42 (2004).
37. Chen, C. et al. Real-time quantification of microRNAs by stem-loop RT–PCR. *Nucleic Acids Res.* **33**, e179 (2005).
38. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
39. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
40. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
41. Addo-Quaye, C., Miller, W. & Axtell, M. J. CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics* **25**, 130–131 (2009).
42. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
44. Zhang, Y. et al. Model-based analysis of ChIP–seq (MACS). *Genome Biol.* **9**, R137 (2008).
45. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
46. Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**, 3448–3449 (2005).
47. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
48. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870–1874 (2016).
49. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
50. Dai, X., Zhuang, Z. & Zhao, P. X. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* **46**, W49–W54 (2018).
51. Wang, X. An in-house Perl script used for creating SNP-corrected references. *Zenodo* https://doi.org/10.5281/zenodo.10801184 (2024).

Acknowledgements

We thank X. Chen, D. Lisch and R. Schmitz for constructive comments on this work. This work was mainly supported by the Agriculture and Food Research Initiative of the USDA National Institute of Food and Agriculture (grants 2018-67013-27425, 2021-67013-33722 and 2022-67013-37037) and partially supported by the United Soybean Board, the North Central Soybean Research Program, the Indiana Soybean Alliance and Ag Alumni Seed.

Author contributions

J.M. and Xianzhong Feng designed the research. W.W., J.D., Xingxing Feng, X.W., L.C., C.B.C., S.A.S., R.L.N., S.L. and J.W. performed experiments. W.W., X.W., B.C.M. and J.M. analyzed data. W.W. and J.M. wrote the manuscript, and B.C.M. edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

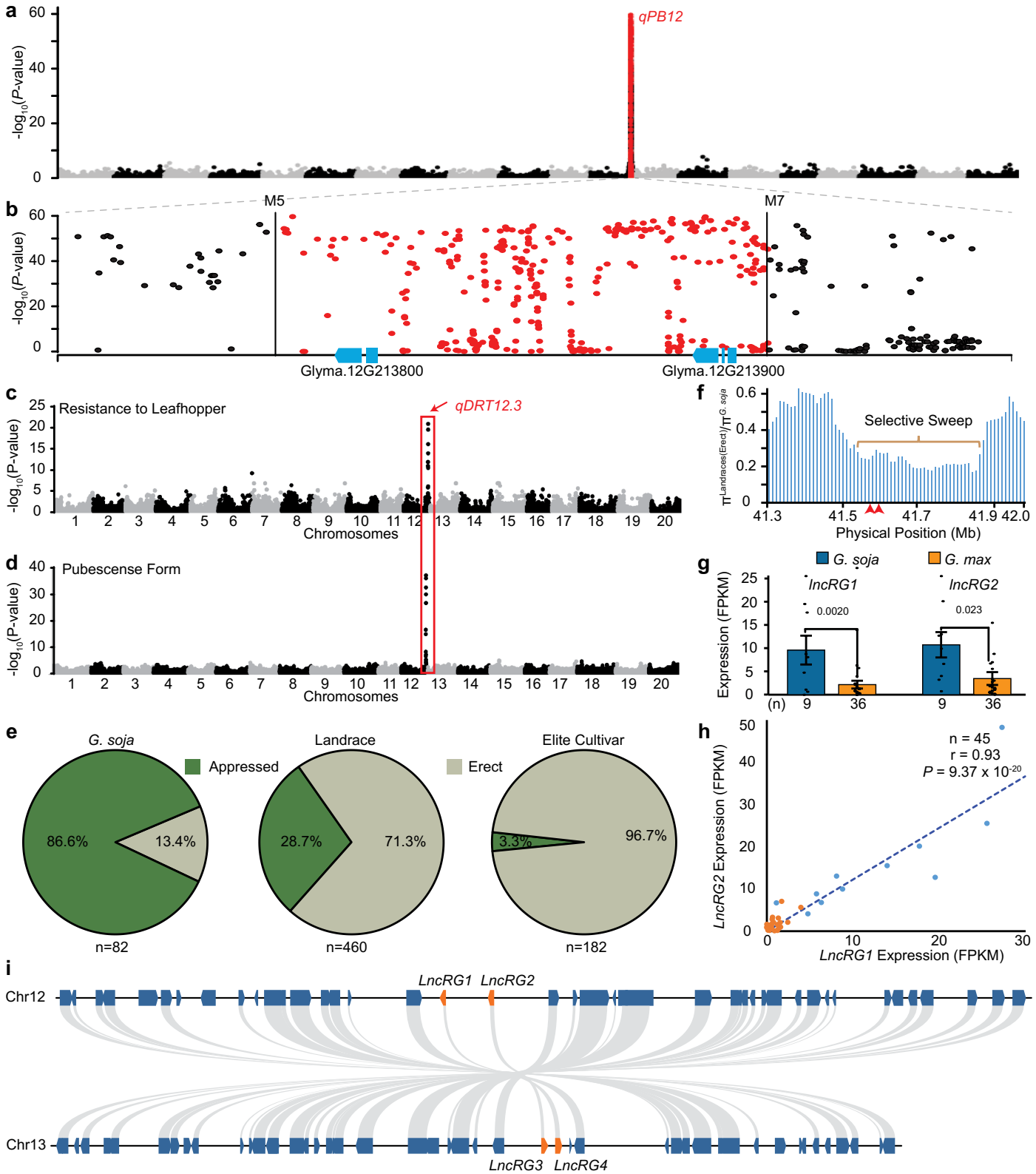
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01738-2.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01738-2.

Correspondence and requests for materials should be addressed to Xianzhong Feng or Jianxin Ma.

Peer review information *Nature Genetics* thanks Yong-Qiang An and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-024-01738-2

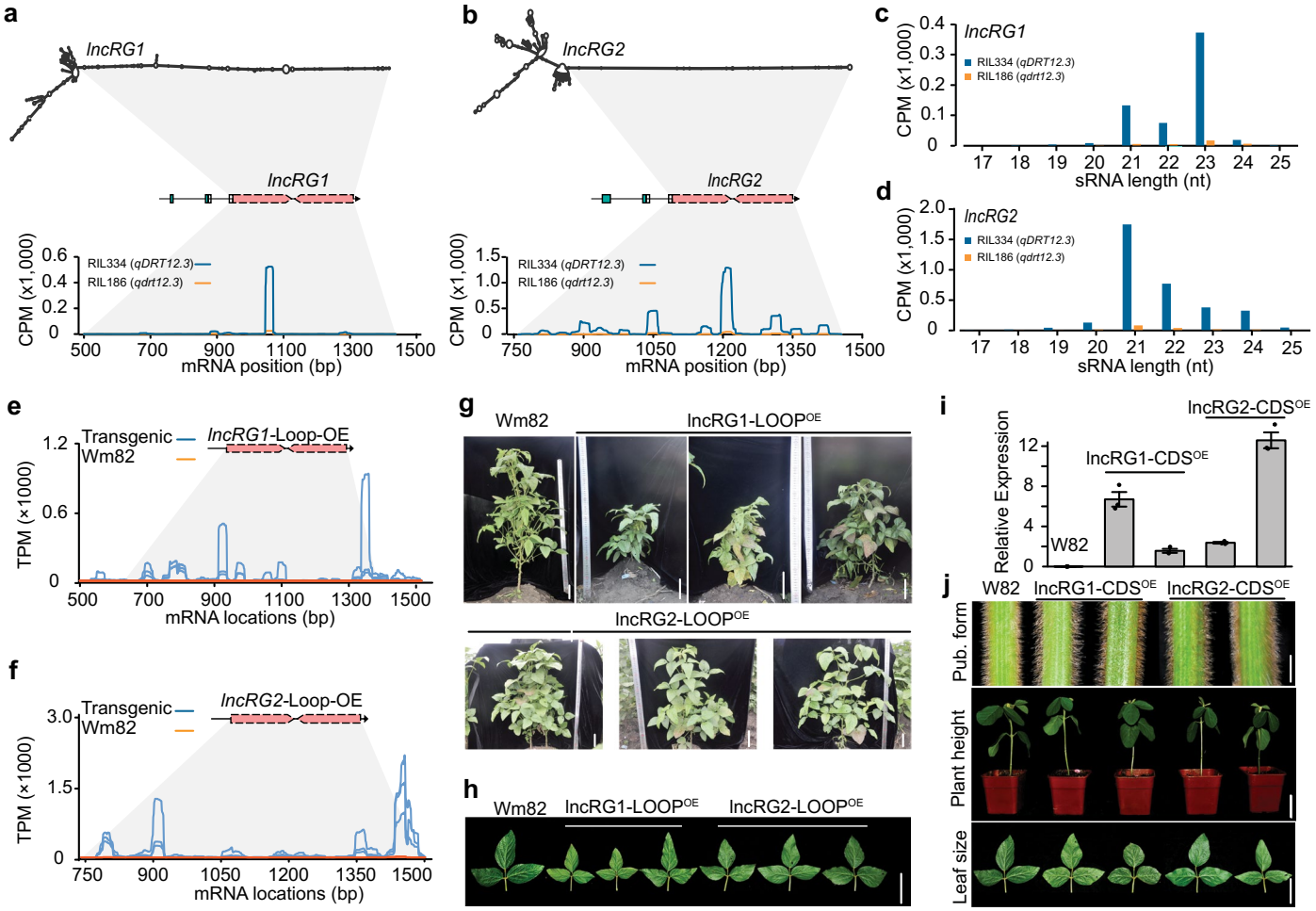


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41588-024-01738-2

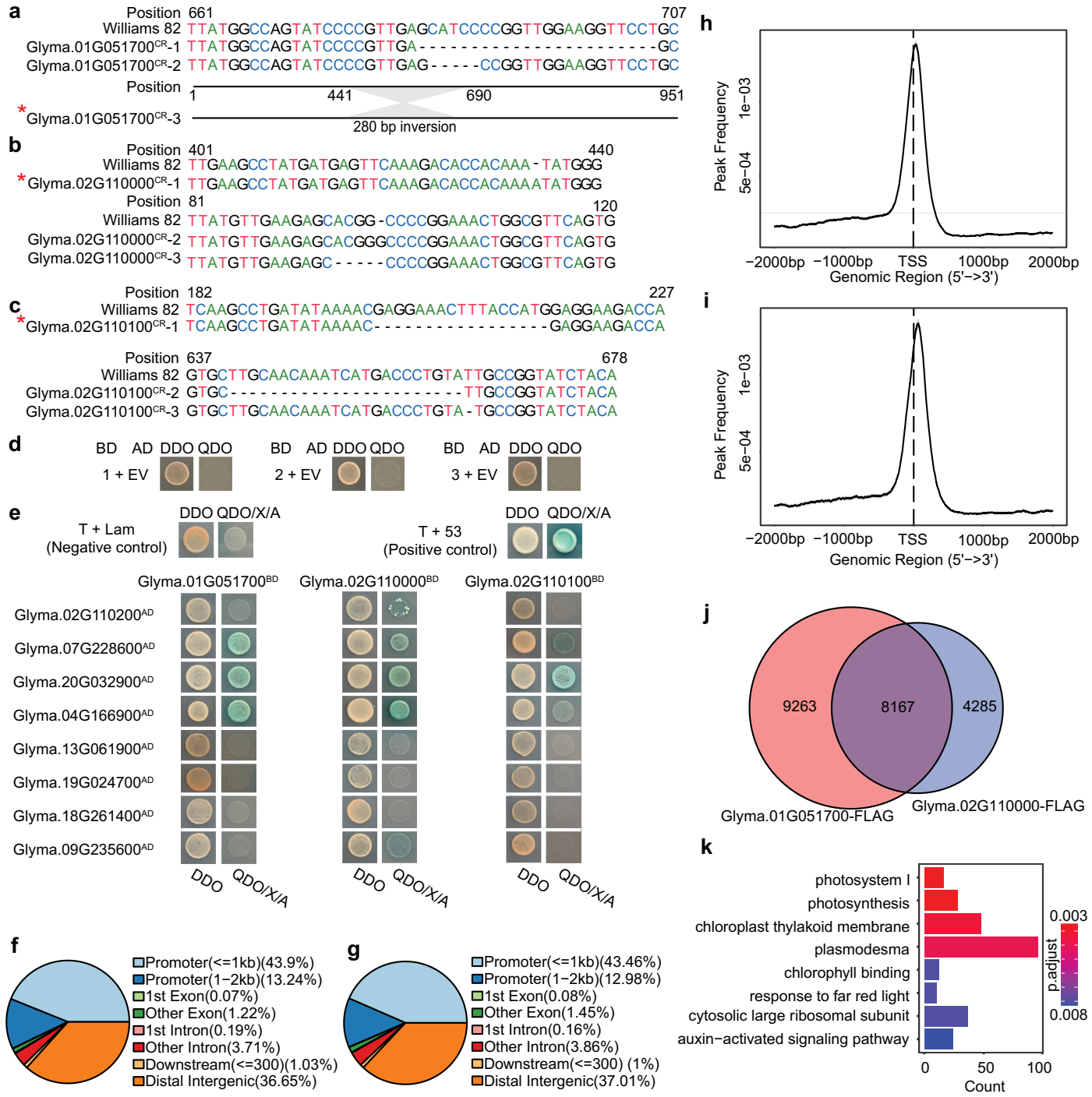
Extended Data Fig. 1 | Association studies, selection analyses and expression analyses. a-**b**, Genome-wide association study (GWAS) on pubescence form using the re-sequencing data from 74 *G. soja* and 594 *G. max* accessions[6] and corresponding phenotypic data from the USDA soybean germplasm database (Supplementary Table 2). The red color highlights markers within the fine-mapped *qDRT12.3* region. **c**-**d**, GWAS on leafhopper resistance (c) and pubescence form (d) using the genotypic data from 784 soybean accession[7] and corresponding phenotypic data from the USDA database (Supplementary Table 4). The rectangle highlights the *qDRT12.3* locus. In (a-d), the *P* values were determined by the F-test for each marker. **e**, Frequencies of erect and appressed pubescence form in *G. soja*, landrace and elite cultivar sub-populations[6]. n indicates the number of soybean accessions in each sub-population. **f**, Selective sweep surrounding the *qDRT12.3* region. The *y*-axis is the ratio of nucleotide diversity (π) of landraces (n = 328) with erect pubescence over *G. soja* (n = 103)[6] calculated for every 100-kb window with 10-kb sliding steps. Each vertical bar represents the value at the middle point of each sliding window. The red arrows pinpoint the positions of *lncRG1* and *lncRG2*. The *x*-axis presents the physical positions based on the Zhonghuang 13 (v2) genome assembly. **g**, Expression levels of *lncRG1* and *lncRG2* in the V1-stage stem tips of *G. soja* (n = 9) and *G. max* (n = 36) (Supplementary Table 7). The expression levels were measured with RNA-seq data[8] and represented as mean ± SEM. FPKM, fragments per kilobase of transcript per million mapped reads. The dots indicate the values from biologically independent samples (n = 3). The numbers above the bars are *P* values determined by a two-sided Student's *t*-test. **h**, Co-expression between *lncRG1* and *lncRG2* in the V1-stage stem tips. The expression levels of *lncRG1* and *lncRG2* were measured with the RNA-seq data[8]. Each dot represents a single soybean accession, with blue dots for *G. soja* haplotype (n = 11) and orange dots for *G. max* haplotype (n = 34). Dashed line is the trend line. The *P* value is obtained by a two-sided Pearson's correlation test. **i**, Collinearity between the l*ncRG1*-*lncRG2* region and the *lncRG3*-*lncRG4* region. Boxes represent genes and grey shades connect WGD pairs.

https://doi.org/10.1038/s41588-024-01738-2



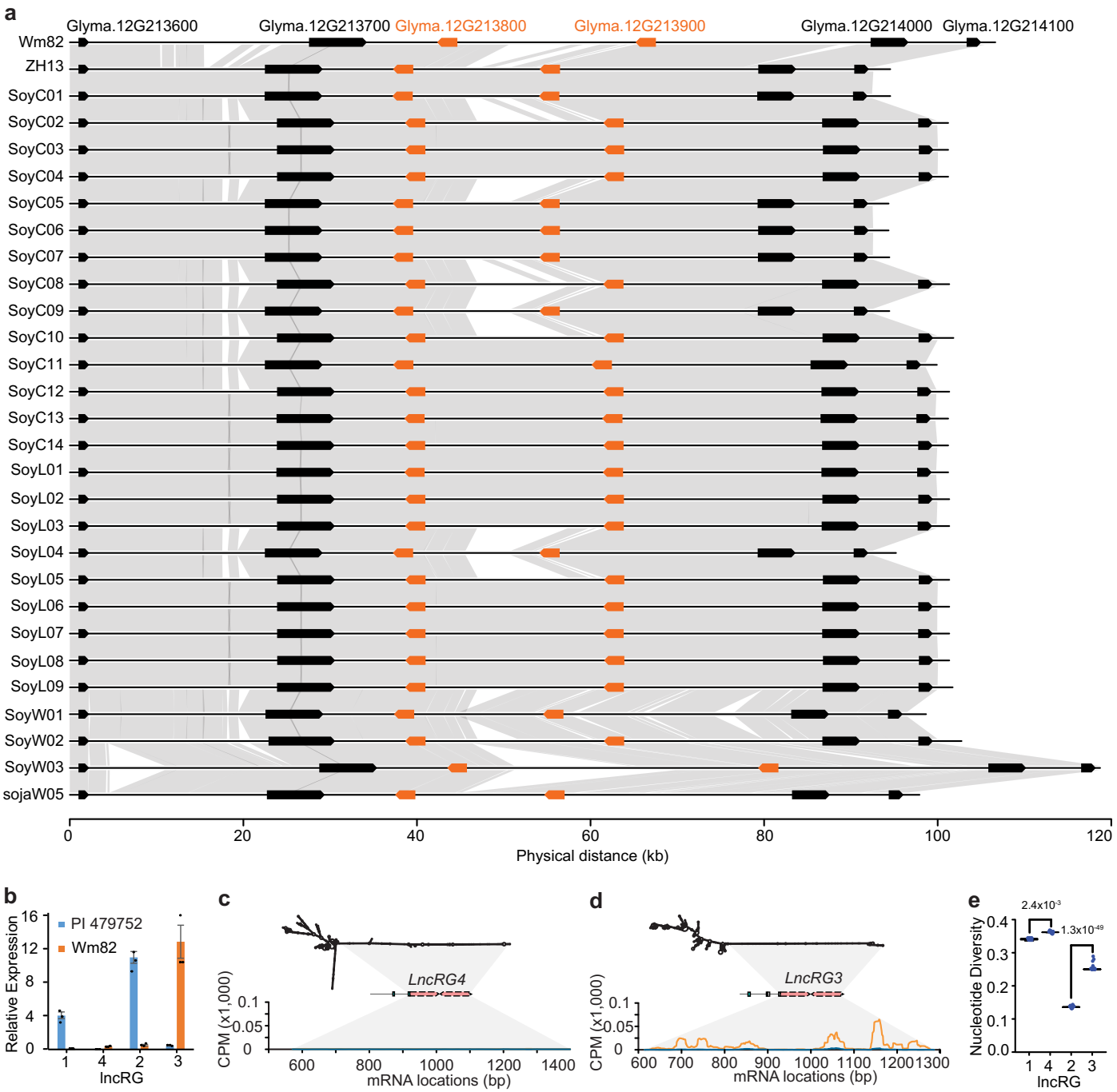
Extended Data Fig. 2 | Abundance and distribution of sRNAs produced by *lncRG1* and *lncRG2* in a pair of RILs and the transgenic lines, and images of transgenic lines. a, Abundance and distribution of sRNAs produced by *lncRG1* in RIL186 (*qdrt12.3*) and RIL334 (*qDRT12.3*). The *x*-axis shows the position on the *lncRG1* transcript, and the *y*-axis is the abundance in copy per million reads (CPM). **b**, Abundance and distribution of sRNAs produced by *lncRG2* in RIL186 (*qdrt12.3*) and RIL334 (*qDRT12.3*). The *x*-axis shows the position on the *lncRG2* transcript, and the *y*-axis is abundance in copy per million reads (CPM). **c**, Frequencies of sRNA from *lncRG1* at different sizes from 17nt to 25nt in RIL186 (*qdrt12.3*) and RIL334 (*qDRT12.3*). **d**, Frequencies of sRNA from *lncRG2* at different sizes 17nt to 25nt in RIL186 (*qdrt12.3*) and RIL334 (*qDRT12.3*). **e**, Abundance and distribution of sRNAs along the transcript of *lncRG1* in the lncRG1-LOOP^{OE} transgenic lines. The *x*-axis shows the position on the *lncRG1* transcript, and the *y*-axis is the abundance in copy per million reads (CPM). **f**, Abundance and distribution of sRNAs along the transcript of *lncRG2* in the lncRG2-LOOP^{OE} transgenic lines. The *x*-axis shows the position on the *lncRG2* transcript, and the *y*-axis is the abundance in copy per million reads (CPM). **g**, Plant images of the transgenic lines that overexpress the inverted repeats of *lncRG1* and *lncRG2*. Bars = 10 cm. **h**, Leaf images of the transgenic lines that overexpress the inverted repeats of *lncRG1* and *lncRG2*. Bars=5 cm. **i**, Relative expression levels of the predicted CDS of *lncRG1* and *lncRG2* in the transgenic lines that overexpress the predicted CDS, as determined by qRT-PCR with Wm82 set as "1" and the others adjusted accordingly. The dots show the values from biologically independent samples (n = 3). Data are represented as mean ± SEM. **j**, images of the transgenic plants that overexpress the predicted CDS of *lncRG1* and *lncRG2*, Bars = 5 mm, 5 cm, 5 cm in top, middle and bottom, respectively.

https://doi.org/10.1038/s41588-024-01738-2



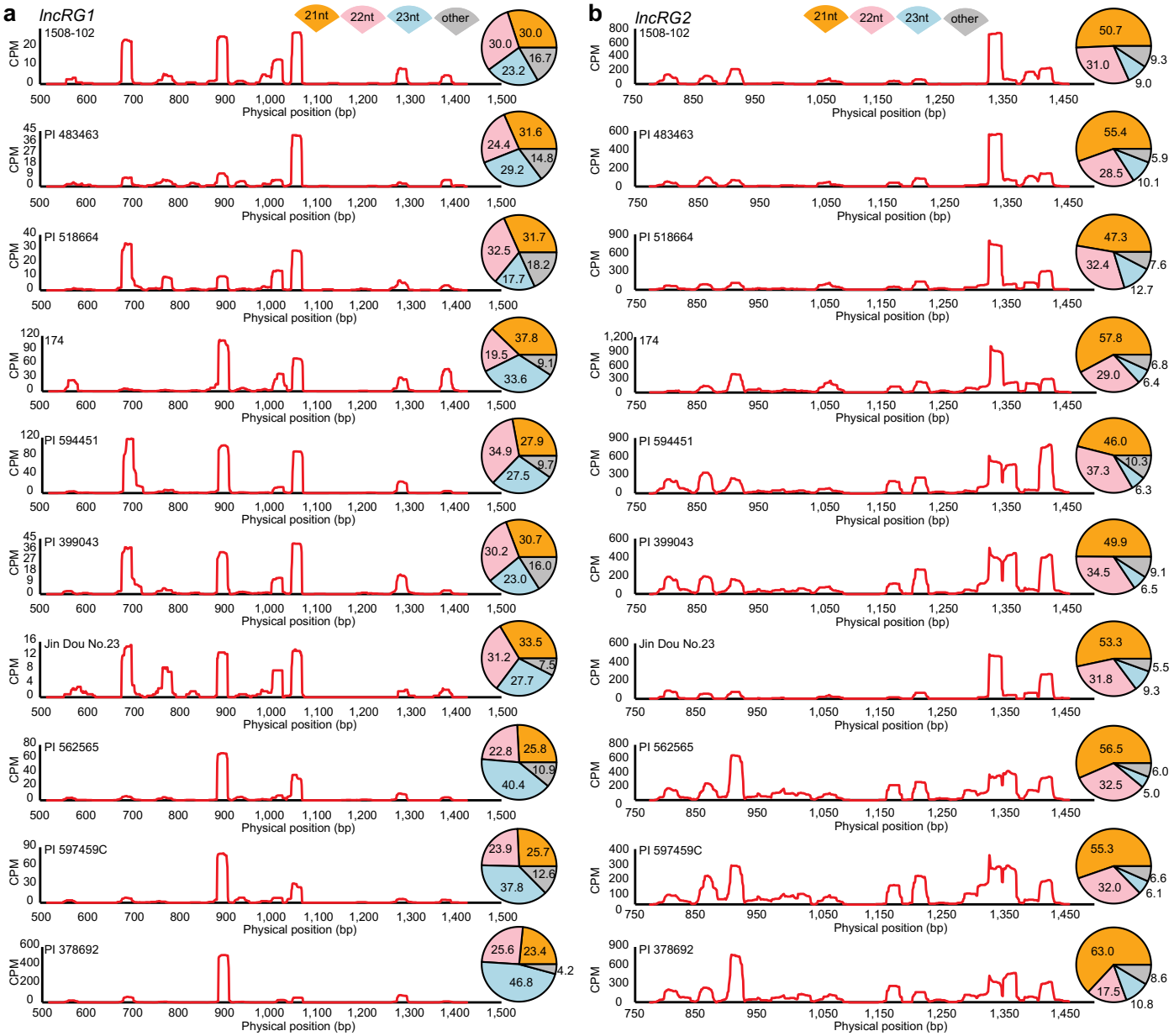
Extended Data Fig. 3 | Mutations created by CRISPR-Cas9, protein-protein interaction as detected by Y2H and ChIP-seq analysis. a-c, Frameshift mutants created by CRISPR-Cas9 for each of the three MYB genes, Glyma.01G051700 (a), Glyma.02G110000 (b) and Glyma.02G110100 (c). The top sequence shows the Wm82 sequence and the position of each base pair in Wm82. - represent deletions in the editing lines. Red asterisk indicates the lines selected for crossing to make double editing lines. **d**, Primary Y2H tests to confirm whether the MYB target genes can active the reporter gene. EV represents empty vector. **e**, Protein-protein interactions among MYB transcription factors as detected by the yeast two hybrid (Y2H) system. Colonies on DDO plate indicate the successful transformation of the construct in yeast cells. Blue colonies on QDO/X/A plates indicate positive protein-protein interactions. AD, activation domain; BD, binding domain; DDO, double dropout; QDO, quadruple dropout. X, X-alpha-Gal; A, Aureobasidin A. **f-g**, Distribution of the locations of the ChIP-seq peaks relative to target genes detected in the Glyma.01G051700-FLAG and Glyma.02G110000-FLAG transgenic lines, respectively. **h-i**, Frequency of the ChIP-Seq peaks surrounding the transcription start sites (SST) detected in the Glyma.01G051700-FLAG and Glyma.02G110000-FLAG transgenic lines, respectively. **j**, Number of potential downstream genes identified by ChIP-seq in the Glyma.01G051700-FLAG and Glyma.02G110000-FLAG transgenic lines. **k**, Gene ontology (GO) classification for the genes detected in both the Glyma.01G051700-FLAG and Glyma.02G110000-FLAG transgenic lines. The *P* value was determined by Fisher's exact test adjusted for false discovery rate.

https://doi.org/10.1038/s41588-024-01738-2



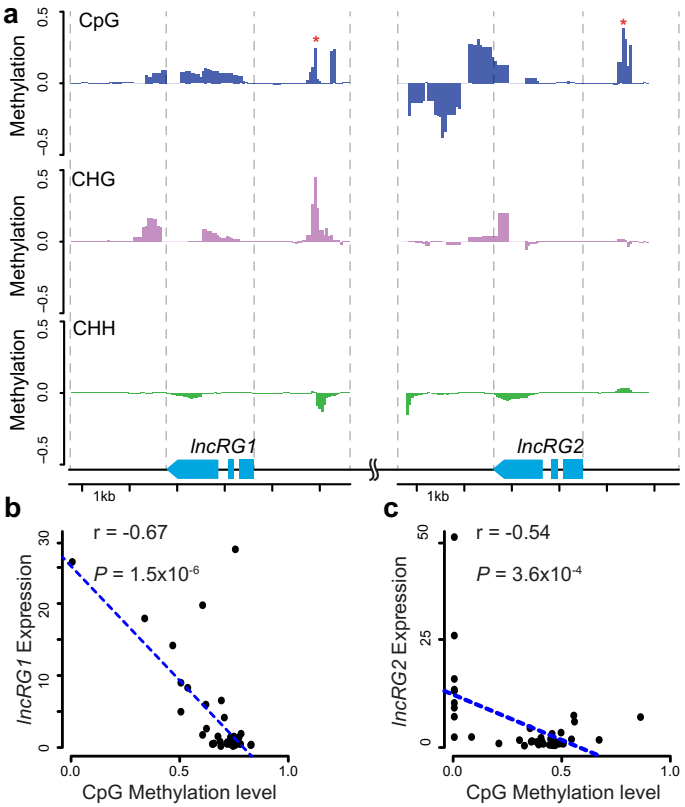
Extended Data Fig. 4 | Copy number conservation of *lncRG1* and *lncRG2* in the soybean pan-genome and evolution of *lncRG3* and *lncRG4*. **a**, Genomic sequence and gene alignments among the soybean pan-genome accessions at the lncRG1-lncRG2 region, including flanking genes. Boxes represent genes and grey color indicate syntenic blocks among genomes. **b**, Relative expression levels of *lncRG1*, *lncRG2*, *lncRG3* and *lncRG4* in the stem tips of Wm82 and PI 479752, as determined by qRT-PCR. The dots show the values from biologically independent samples (n = 3). Data are represented as mean ± SEM. **c**-**d**, Secondary structures of *lncRG3* and *lncRG4* and the sRNAs mapped to their inverted repeats. **e**, nucleotide diversity within the inverted repeats of *lncRG1*, *lncRG2*, *lncRG3* and *lncRG4*. The dots show the values of nucleotide diversity calculated from different soybean pan-genome accessions (n = 27). The horizontal lines indicate the medians, and the boxes represent the interquartile range (IQR). The whiskers represent the range of 1.5 times IQR and dots beyond the whiskers are outlier values. The numbers above the boxes are *P* values determined by a two-sided Student's *t*-test.

https://doi.org/10.1038/s41588-024-01738-2



Extended Data Fig. 5 | Distribution of the sRNAs produced by *lncRG1* and *lncRG2* in ten diverse soybean accessions. The *x*-axis shows the position on the *lncRG1* (a) or *lncRG2* (b) transcripts, and the *y*-axis is abundance in copy per million reads (CPM). The relative abundances of sRNAs of different sizes detected in individual accessions (Supplementary Table 7) are shown in percentage (%) in individual pies.

https://doi.org/10.1038/s41588-024-01738-2



Extended Data Fig. 6 | Association between epigenetic variations and expression levels of *lncRG1* and *lncRG2*. **a**, Differences of CpG, CHG and CHH DNA methylation between *the G. max* haplotype (n = 29) and the *G. soja* haplotype (n = 10) surrounding *lncRG1* and *lncRG2* (Supplementary Table 7). Each vertical bar represents the average methylation level difference within a 300 bp window between the two haplotypes with sliding step=50 bp. The purple color highlights the differences in the promoter regions of the two genes. The red asterisk indicates the window used for correlation analysis in (b) and (c). **b**-**c**, Correlations between the CpG methylation differences in the promoter regions of *lncRG1* and *lncRG2* with their expression levels as measured by Pearson's correlation coefficient (n = 41). The *P* values are obtained by a two-sided Pearson's correlation test. Dashed lines are the trend lines.

nature portfolio

Corresponding author(s): Jianxin Ma, Xianzhong Feng

Last updated by author(s): 03/21/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	sra-toolkit (v2.11.0) was used to collect public sequencing data from NCBI database
Data analysis	All softwares or online tools including R (v4.2.1), r/qtl (v1.66), TASSEL 5 (https://tassel.bitbucket.io/), CRISPR-P (http://crispr.hzau.edu.cn/CRISPR2/), STAR (v2.5.4b), cufflinks (v2.2.1), CleaveLand (v4.5), MEGA (7.0), vcftools (v0.1.16), RNAfold server (http://rna.tbi.univie.ac.at/) and psRNATarget (Schema V2 2017 release), bowtie2 (v2.5.1),Bismark (v0.23.1), Wheeler Aligner program (v0.7.15), Model-based Analysis of ChIP-Seq (MACS2), BiNGO (v3.0.3), ImageJ (v1.53k), clusterProfiler (v4.0) used for data analysis in the manuscript have been cited/refereneced.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data are available in the main text, supplemental materials, public databases, or referenced studies. All the raw sequence data generated in this study have been deposited in NCBI database under the BioProject PRJNA876203. The genotypic data of the USDA soybean germplasm collection used for GWAS on pubescence form and leafhopper resistance in Extended Data Fig. 1c-d were downloaded from the SoyBase database (https://soybase.org/snps/download.php). The genotypic data of the re-sequenced soybean accessions used for GWAs on pubescence form in Extended Data Fig. 1a-b were downloaded from the Genome Variation Map (GVM) database in BIG Data Center (http://bigd.big.ac.cn/gvm/getProjectDetail?project=GVM000063). The RNA-seq, sRNA and WGBS data of the 45 highly diverse soybean accessions were download from the Sequence Read Archive (SRA) database in NCBI under Accession Number PRJNA432760 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA432760). Source data are provided with this paper.
All software used in this study is publicly available as described in the Methods and Reporting summary. Detailed parameters used for analyzing each type of sequencing data have been described in the Method. An in-house Perl scrip used for creating SNP-corrected genomes is available at Zenodo (https://doi.org/10.5281/zenodo.10801184).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes are indicated in the manuscript, figures or legends. No statistical method was employed to determine sample size; instead, sample sizes for each experiment were optimized based on practical feasibility and the availability of samples. The size of the RIL mapping population is 2,287. The sample sizes of association studies are 668 and 784 for pubescence form and leafhopper resistance, respectively. The sample size for the qRT-PCR experiments is 3 biologically independent samples. 20-30 plants were planted in each plot for each transgenic line or gene editing line (single and double mutants). 7 homozygous triple mutants were obtained by crossing the double mutants. Small RNA data from 45 soybean accessions were used to analyze the sRNA diversity at population level.
Data exclusions	small RNAs less than 17-nt or greater than 25-nt were excluded.
Replication	Three biologically independent samples (with three technical replicates for each sample) were performed in qRT-PCR experiments; Three biologically independent transgenic events and gene editing lines were created for functional validations. All attempts at replication for the qRT-PCR and functional validation were successful. Small RNAs were obtained from the parental lines and a pair of RILs as biologically independent replications. The RLM-RACE assay was repeated at least two times; The Bimolecular fluorescence complementation (BiFC) assay was repeated two times. All attempts at replication for the RLM-RACE and BiFC were successful.
Randomization	The plots of recombinants, transgenic lines, gene edited lines as well as the control lines were planted randomly in the field for phenotyping. qRT-PCR and RLM-RACE were performed to identify differences between genotypes, therefore, sample allocation is not relevant to these

	experiments. BiFC and Y2H were performed to detect interaction among specific genes, therefore, sample allocation is not relevant to these experiments.
Blinding	Blinding was performed when phenotyping the recombinants, transgenic plants and gene edited lines. Genotype of the plant was not known when the phenotypic data collection and analysis. qRT-PCR and RLM-RACE were performed to identify differences between genotypes, therefore, blinding is not relevant to these experiments. BiFC and Y2H were performed to detect interaction among specific genes, therefore, blinding is not relevant to these experiments.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	ANTI-FLAG® M2 Magnetic Beads (Sigma, M8823-1ML)
Validation	ANTI-FLAG® M2 (sigma): The antibody has been validated by the manufacturer, https://www.sigmaaldrich.com/catalog/product/sigma/fl804

Plants

Seed stocks	The soybean accession, PI 518671 (Williams 82), used for creating the transgenic lines were originally requested from the USDA soybean germplasm collection (https://www.ars-grin.gov/).
Novel plant genotypes	Transgenic lines were generated for the ChIP-seq experiments by using the soybean transformation protocol described in the Method section. We fused the coding sequences of Glyma.01G051700 and Glyma.02G110000 from Wm82 with the FLAG epitope, separately, generating transgenic lines that overexpress each of the fused proteins by the 35S promoter.
Authentication	Over-expression of the trangenes were validated by qRT-PCR.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	All the raw sequence data generated from ChIP-seq have been deposited in NCBI database under the BioProject PRJNA876203 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA876203). All the called peaks from .BED files were provided in supplementary table 11 and 12.
Files in database submission	BioSample: SAMN40263196; Sample name: Glyma.02G110000-Flag-input; SRA: SRS20659442 BioSample: SAMN40263195; Sample name: Glyma.02G110000-Flag-ChIP; SRA: SRS20659441 BioSample: SAMN40263194; Sample name: Glyma.01G051700-Flag-input; SRA: SRS20659440 BioSample: SAMN40263193; Sample name: Glyma.01G051700-Flag-ChIP; SRA: SRS20659439
Genome browser session (e.g. UCSC)	no longer applicable

Methodology

Replicates	1 replicate for Glyma.01G051700-Flag-ChIP and Glyma.02G110000-Flag-ChIP, only the peaks identified by both experiments were further analyzed.

Sequencing depth	Glyma.02G110000-Flag-input, depth: 8.1X, total reads: 29862028, uniquely mapped reads: 22600135,150bp paired-end Glyma.02G110000-Flag-ChIP, depth: 9.7X, total reads: 35745189, uniquely mapped reads: 27479899,150bp paired-end Glyma.01G051700-Flag-input, depth: 12.3X, total reads: 44938763, uniquely mapped reads: 33425686,150bp paired-end Glyma.01G051700-Flag-ChIP, depth: 11.0X, total reads: 40319731, uniquely mapped reads: 30302209,150bp paired-end
Antibodies	ANTI-FLAG® M2 Magnetic Beads (Sigma, M8823-1ML)
Peak calling parameters	macs2: callpeak -t Glyma.01G051700-Flag-ChIP.bam -c Glyma.01G051700-input-ChIP.bam -f BAM -g 1100000000 -n Glyma.01G051700-peak -B -q 0.01 macs2: callpeak -t Glyma.02G110000-Flag-ChIP.bam -c Glyma.02G110000-input-ChIP.bam -f BAM -g 1100000000 -n Glyma.02G110000-peak -B -q 0.01
Data quality	sequencing quality control were performed by Novogene Corporation Inc. (Sacramento, CA). 8727 and 9732 peaks with enrichment greater than 5 were identified from Glyma.01G051700-Flag-ChIP and Glyma.02G110000-Flag-ChIP, respectively.
Software	Burrows Wheeler Aligner program (v0.7.15) Model-based Analysis of ChIP-Seq (MACS2) (v2.1.0)