Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance

Jianxin Ma
maj@purdue.edu
Purdue University https://orcid.org/0000-0002-1474-812X

Weidong Wang
wangwd@cau.edu.cn
China Agricultural University https://orcid.org/0000-0002-7110-5630

Jingbo Duan
duan68@purdue.edu
Purdue University https://orcid.org/0000-0001-6467-9102

Xutong Wang
wangxt881@gmail.com
Purdue University

Xingxing Feng
fengxingxing112@163.com
Chinese Academy of Sciences

Liyang Chen
zjuabcly@gmail.com
Purdue University

Chancelor Clark
clark367@purdue.edu
Purdue University https://orcid.org/0000-0002-2255-2514

Stephen Swarm
stephen.swarm@beckshybrids.com
University of Illinois

Jinbin Wang
wang5549@purdue.edu
Purdue University

Sen Lin
lin1595@purdue.edu
Purdue University

Randall Nelson
rlnelson@illinois.edu
University of Illinois https://orcid.org/0000-0001-9482-1763

Blake Meyers
bmeyers@danforthcenter.org
Donald Danforth Plant Science Center https://orcid.org/0000-0003-3436-6097

Xianzhong Feng
fengxianzhong@iga.ac.cn
Northeast Institute of Geography and Agroecology, CAS https://orcid.org/0000-0002-7129-3731

Article

Keywords:

Posted Date: December 11th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3133590/v1

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Genetics on April 29th, 2024. See the published version at https://doi.org/10.1038/s41588-024-01738-2.

Article:

Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance

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

[1]Department of Agronomy, Purdue University; West Lafayette, IN 47907, USA.

[2]Key Laboratory of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences; Changchun, Jilin 130102, China.

[3]Department of Crop Sciences, University of Illinois at Urbana–Champaign; Urbana, IL 61801, USA.

[4]Donald Danforth Plant Science Center, St. Louis, Missouri, USA.

[5]Division of Plant Science & Technology, University of Missouri-Columbia, Columbia, Missouri, USA.

[6]Present address: College of Agronomy and Biotechnology, China Agricultural University; Beijing 100193, China.

[7]Present address: Hubei Hongshan Laboratory; Wuhan, Hubei 430070, China.

[8]Present address: Beck’s Hybrids; Atlanta, IN 46031, USA.

[9]These authors contributed equally to this work.

*Correspondence: maj@purdue.edu (J.M.), fengxianzhong@iga.ac.cn (X.F.)

Abstract

The origination and functionality of long non-coding RNAs (lncRNAs) 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 (IRs) originated from coding sequences of MYB genes, function by generating clusters of small RNAs in wild soybeans to inhibit the expression of their MYB gene relatives through posttranscriptional regulation. In contrast, the expression of the 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 IRs were formed before the divergence of the Glycine genus from the Phaseolus/Vigna lineage and exhibit strong structure-function constraints. This study exemplifies a new type of targets for selection during plant domestication and uncovers mechanisms of lncRNA formation and action.

Main

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 creates 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]. Most wild soybean accessions exhibit a procumbent or climbing growth habit, with long, slender, prolifically branched stems and small leaves that grow appressed pubescence, whereas 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 in cultivated soybeans are resulted from artificial selection of reduced expression of two tandemly duplicated long non-coding RNA (lncRNA) genes each carrying MYB gene coding sequence-derived inverted repeats (IRs), which have undergone strong purifying selection in the Glycine genus.

Results

Map-based cloning of multiple DRT QTLs identifies a single locus with pleiotropic effects

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% the phenotypic variation in pubescence form, main stem length, leaf size, growth habit, and stem-twining, respectively (Fig. 1a-1e). 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 (InDel) markers, M1 and M10, which initially defined the boundaries of the *qDRT12.3* region, were used to genotype all 2,287 RILs and identified 238 recombinants between the two markers (Fig. 1f). 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), which was echoed by a genome-wide association study using re-sequencing data from 74 *G. soja* and 594 *G. max* accessions[5] (Extended Data Fig. 1a and 1b). 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-1h). According to the Wm82 reference genome, this region harbors only two genes, Glyma.12G213800 and Glyma.12G213900, both lncRNAs.

It has been observed that semi-appressed or erect pubescence is linked to reduced defoliation caused by Cicadellidae insects[6]. To investigate whether *qPB-12* is responsible for such resistance, we conducted a genome-wide association study (GWAS) on leafhopper resistance/susceptibility using the phenotypic and genotypic data from 784 accessions in the USDA soybean germplasm collection[7]. We found that molecular markers within the fine mapped *qPB-12* region were significantly associated with leafhopper resistance (Extended Fig. 1c and 1d) and that erect pubescence indeed contributes to the leafhopper resistance as shown in Supplementary Movies 1 and 2. In a set of re-sequenced diverse *G. soja* and *G. max* accessions chosen from the USDA soybean germplasm collection[5], only 13.4% of the *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), indicating that erect pubescence and its underlying leafhopper resistance was a target for selection during soybean domestication and improvement. Artificial selection at this locus was also echoed by a selective sweep surrounding it (Extended Fig. 1f), as detected by sequencing data from 103 *G. soja* accessions and 328 landraces[5]. Collectively, these observations suggest that Glyma.12G213800 and Glyma.12G213900 are the candidate genes regulating pubescence form, main stem length, leaf size, as well as leafhopper resistance attributed to erect pubescence.

The pleiotropic DRT locus is composed of two tandemly duplicated lncRNA genes, *lncRG1* and *lncRG2*

The genes, Glyma.12G213800 and Glyma.12G213900, in Wm82 produce 1,526-nt and 1,565-nt transcripts, which are predicted to encode 37 and 49 amino acids, respectively. Thus, they are defined as long non-coding RNA (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, but are expressed at significantly lower levels in the same tissues of Wm82, as measured by quantitative reverse transcription-PCR (qRT-PCR) (Fig. 2a). RNA-seq data from 45 highly diverse soybean accessions[8] revealed significantly higher expression levels of these two genes in nine wild soybean accessions than in 36 cultivated soybean accessions (Extended Data Fig. 1g) as well as a pattern of *lncRG1* and *lncRG2* co-expression (Extended Data Fig. 1h). Therefore, the suppressed expression of *lncRG1* and *lncRG2* is most 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 soybean genome reveals that not only the putative coding sequences (CDSs) but the large portions of the non-CDSs of these two lncRNA genes share similarities with typical MYB transcription factor genes (Fig. 2b and 2c). Thus, *lncRG1* and *lncRG2* were derived from MYB genes. Further phylogenetic analysis reveals 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 were two homologs of *lncRG1* and *lncRG2*, dubbed *lncRG3* and *lncRG4,* respectively (Fig. 2c). Nevertheless, *lncRG3* and *lncRG4* are not associated with any of the domestication QTLs[4]. Interestingly, all four lncRGs in soybean possess IRs, each at ~300-400 bp, corresponding to the third exon of their most closely related MYB genes (Fig. 2c).

***LncRG1* and *lncRG2* harbor IRs and produce abundant sRNAs primarily targeting three closely related MYB genes**

Based on prediction, the IRs within the transcripts of *lncRG1* and *lncRG2* may form double-stranded stem loops at 453 bp and 337 bp, respectively (Fig. 2d and 2e), which could be processed to generate small RNAs (sRNAs), such as microRNAs, microRNA (miRNA)-like sRNAs, or short interfering sRNAs (siRNAs). Then, we sequenced sRNAs in the V1-stage stem tips of PI 479752 and WM82, respectively. Abundant, overlapping sRNAs, mainly at 21-23 nucleotides (nt), across the IRs of both *lncRG1* and *lncRG2* were detected in PI 479752, but their relative abundances vary drastically (Fig. 2f and 2g). The most abundant sRNAs from *lncRG1* are at 23nt, whereas the most abundant sRNAs from *lncRG2* are at 21nt (Fig. 2h and 2i). Nevertheless, much more sRNAs were produced from *lncRG2* than *lncRG1*. Consistent abundances and distribution patterns of the sRNAs produced by *lncRG1* and *lncRG2* were observed in a pair of RILs, RIL186 (*qdrt12.3*) and RIL 334 (*qDRT12.3*) (Extended Data Fig. 2a-2d), suggesting that the abundance of individual sRNAs were tightly regulated and not randomly produced from the IRs.

A total of 163 genes were predicted to be targets of 27 distinct sRNAs from *lncRG1* and *lncRG2*, with a relative abundance of >100 copies per million (CPM) sRNA reads (Supplementary Table 1 and 2). 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 2-fold changes in stem tips, stems and leaves as determined by RNA-seq and qRT-PCR (Fig. 2j and Supplementary Table 3). Degradome sequencing revealed that the mRNAs of these three genes were predominantly cleaved at the predicted sRNA target sites in PI 479752 (Fig. 2k-2m). 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 predicted coding sequences (Fig. 2b). Thus, these MYB gene-derived *lncRG1* and *lncRG2* are likely to modulate the DRTs by producing plentiful sRNAs to primarily repress their MYB gene relatives via post-transcriptional regulation.

Overproduction of sRNAs in cultivated soybean promotes the wild soybean-type phenotypes

To determine whether the sRNAs produced by *lncRG1* and *lncRG2* underlie the DRTs, we first generated Williams 82 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 sRNAs from the stem loops (Extended Data Fig. 2e and 2f) and showed expected phenotypic changes including appressed pubescence form, decreased plant height and smaller leaf size in comparison to the Wm82 (Fig. 3a-3c). 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 Williams 82 resulted in appressed/semi-appressed pubescence form, reduced plant height, and smaller leaf size compared to the Wm82 (Fig. 3d and 3e). As expected, these transgenic lines exhibited increased expression levels of the corresponding artificial sRNAs and decreased expression levels of the three MYB genes as determined by stem-loop and regular qRT-PCR, respectively (Fig. 3g and 3h). The mRNAs of the target genes were 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 wild-type control using RNA ligase mediated rapid amplification of the 5’ cDNA ends (RLM-RACE) technique followed by deep sequencing (Fig. 3i and g). These observations indicate that the specific sRNAs produced from *lncRG1* and *lncRG2* are responsible for forming the DRTs and suggest that these sRNAs use miRNA-like mechanism to repress their targets.

Since *lncRG1* and *lncRG2* are predicted to encode two small peptides, we wonder whether the small peptides also contribute to the DRTs. Then, we generated Williams 82 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 the DRTs.

The three MYB genes targeted by the sRNAs exhibit functional redundancy and divergence

To gain insights into the mechanism by which the three MYB genes regulate the DRTs, we generated Wm82 knockout lines for each MYB gene using

CRISPR-Cas9 (Extended Data Fig. 3a-3c). Knocking out any of the three genes resulted in appressed/semi-appressed pubescence, and reduced plant height, and smaller leaf size; however, their effects on each DRT slightly vary (Fig. 4a, 4d and 4e). 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 compared to the single mutants (Fig. 4a-4c and 4f-4i), suggesting an additive effect of the three MYB genes.

Given that protein dimerization often plays a crucial role in transcription factor activity, we wondered whether the three MYB genes enable homo- or hetero-dimerization. Using the yeast two-hybrid (Y2H) system (Fig. 4j and 4k) and the bimolecular fluorescence complementation (BiFC) assay in tobacco leaves (Fig. 4l and 4m), both self- and pairwise-protein-protein interactions were detected among the three MYB genes. As expected, both homo- and hetero-dimers were localized in the nucleus (Fig. 4l and 4m). Furthermore, the three target MYB genes were detected 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 vary (Extended Data Fig. 3d). These observations suggest that the three target MYB genes possess both redundant and diverged functions.

The lncRGs have undergone purifying selection due to the structure-function constraints

To track the origin and evolutionary variation of the lncRGs, 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 and 5b). The IRs were also seen in Phaseolus and Vigna, but not seen in Cajanus and *M. truncatula*, suggesting that the IRs were formed before the divergence of Glycine from Phaseolus/Vigna but after its divergence from Cajanus ~20-24MYA[10,11] (Fig. 5a and 5b). The IRs of *lncRG1* and *lncRG2* in *G. soja* and *G. max* exhibited the lowest level of divergence compared with the IRs in the orthologs of *lncRG1*/*lncRG2* in Phaseolus/Vigna (Fig. 5c), indicating that the IRs, as the functional parts of the *lncRG1* and *lncRG2* gene bodies, have experienced strong “purifying selection”. It is also noticeable that *lncRG2*, which produced more non-redundant and more abundant sRNAs than *lncRG1* in PI 479752, evolved in a slower pace than *lncRG1*.

The lncRG-derived sRNAs exhibit diverse distribution patterns at the population level

The availability of sRNA sequencing data from 45 *G. soja* and *G. max* accessions[8] allowed us to compare the distribution and relative abundance of sRNAs generated by *lncRG1* and *lncRG2* at the population level (Fig. 5d, 5e, Extended Data Fig. 4a, 4b and Supplementary Table 4). As expected, all the nine *G. soja* accessions and a cultivated soybean accession (Jin Dou No. 23) with appressed pubescence produced abundant sRNAs from *lncRG1* and *lncRG2.* In contrast, few sRNAs were produced from *lncRG1* and *lncRG2* in the remaining 35 cultivated soybean accessions, which possess erect pubescence. Remarkably, the sRNA distribution patterns vary drastically among the 10 accessions with depressed pubescence, and in most cases, different sRNAs are predicted to target the three MYB genes, and up to 41% of the predicted sRNA targets in one accession are not shared by another accession (Supplementary Table 5). As observed in PI 479752 (Fig. 2g), *lncRG2* in each of the 10 accessions produced more non-redundant and more abundant sRNAs than *lncRG1*, with 21-nt and then 22-nt sRNAs as the predominant forms (Extended Data Fig. 4a and 4b).

Discussion

LncRNAs are ubiquitously present in eukaryotes and play important roles in regulating gene expression[12]. However, how they are 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 IRs. Intragenic

IRs are typically lost due to their instability and fitness costs[13]; yet the IRs 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. IRs can be induced by DNA replication repair or transposable elements[14], which usually leave specific sequence features surrounding the IR junctions. However, as these features do not generally bring any fitness benefits, they would not be preserved over such a long period of evolutionary time. While the processes leading to the formation of the IRs in *lncRG1* and *lncRG2* remain unclear, it is possible that these are newly emerged or incipient miRNA precursors that hasn’t yet been selected for production of a single, precisely processed duplex and instead are generating siRNAs from the IR precursor[15]. The IR 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 the *lncRG1*- and *lncRG2*-derived sRNAs among different wild soybean accessions, the functional constraints in soybean may be implemented through purifying selection across the entire IR regions. It would be interesting to explore whether the IRs in other legumes have similar functionality and regulate comparable traits, and whether the IRs were also targeted for selection during domestication of other leguminous crops.

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[16], and *PROSTRATE GROWTH1* in rice, which controls tiller angle, panicle size, and seed shattering[17]. 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 duplications events including exonic duplication, genic duplication, and WGD. In soybean, approximately 75% of the genes existing 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 sRNAs 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 the reduced expression of *lncRG1* and *lncRG2* within a single locus producing fewer sRNAs.

The causal mutations for the reduced expression of *lncRG1* and *lncRG2* in cultivated soybeans remain unknown. Genome-wide association analysis with the re-sequencing data from 74 *G. soja* and 596 *G. max* accessions[5] revealed numerous polymorphic sites across the entire mapping region that are highly associated with the phenotypic differences in pubescence form (Extended Data Fig. 1a and 1b), 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 selective sweep (Extended Data Fig. 1f). Because *lncRG1* and *lncRG2* are co-expressed across different tissues and developmental stages, there is a possibility that these two genes are regulated by the same 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.

While sRNAs may also repress translation without cleaving mRNAs[18], it is unclear whether the remaining 160 predicted sRNA targets, which show no difference in expression levels between Wm82 and PI 479752 (Supplementary Table 2 and 3), are directly regulated by the sRNAs from *lncRG1* and *lncRG2* through translation inhibition. Given the fact that the three MYB targets also interacts with additional, more diverged copies of MYB genes, that the predominant sizes of sRNAs produced from lncRG1 and lncRG2 are different, and that the sRNAs 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 detected.

Methods

Plant materials

The mapping population consisted of 2,287 $F_{6:7}$ recombination inbred lines (RIL) derived from a cross between *G. max* (Wm82) and *G. soja* (PI 479752). The association mapping population for leafhopper resistance were sourced from the USDA soybean germplasm collection (https://www.ars-grin.gov/). Wm82 was used for stable transformation and genome editing; *Nicotiana benthamiana* was used for the BiFC assays.

QTL and association mapping

The QTL mapping was performed using composite interval mapping (CIM) method[19] incorporated in the r/qtl package[20]. The phenotypic data for association mapping were downloaded from the USDA National Plant Germplasm System (https://npgsweb.ars-grin.gov/) and the SoySNP50K data were obtained from a previous study[7]. The re-sequencing data were from the soybean pan-genome study[5]. The association mapping was performed using TASSEL 5[21] with a mixed linear model[22].

Recombinants genotyping and phenotyping

All the mapping markers were designed based on the re-sequencing data of PI 479752 from a previous study[23]. The domestication-related traits were examined for all the recombinants in the field at Purdue Agronomy Center for Research & Education in 2018. All the primers used in this study were listed in Supplementary Table 6.

Transgene constructs

For stem loop over-expression, the stem loops of *lncRG1* and *lncRG2* were amplified from genomic DNA of PI 479752. The PCR products and linearized vector were purified using PurLink^{TM} Quick Gel Extraction Kit (K210012, ThermoFisher Scientific). The stem loops were inserted into the plasmid vector, linearized by restriction enzymes Nco I and Xba I, using ClonExpress II One Step Cloning Kit (C112, Cellagen Technology).

For artificial miRNA over-expression, the soybean MIR172a was used as the backbone following a previous protocol[24]. The miR172a/miR172a* sequences were replaced by sRlncRG1-1 and sRlncRG2-3 and their corresponding reverse complementary sequences. The forward sequence and complementary sequence were annealed to form dimmers and inserted into pPTN1171.

For CRISPR-Cas9 editing, 4 sgRNAs were designed for each target gene using CRISPR-P, a web-based guided RNA design tool[25]. The primer pairs were annealed for 5 minutes at 95 °C and then cool down to form dimers. The dimers were inserted into pGEL201, linearized by restriction enzymes Bsa I, vector[26]. During transformation, four agrobacteria with different sgRNA were equally mixed before infection.

For yeast two hybrid assays, the full-length coding sequences of the MYB genes were cloned from the cDNA sample of ‘Wm82’ and then inserted into the vectors pGBKT7 and pGADT7.

For bimolecular fluorescence complementation assay, the full-length coding sequences of the three target genes were amplified and cloned into plasmids pCNHP-neYFP-C and pCNHP-ceYFP-C, which express fusion proteins with either N-terminal half of eYFP (neYFP) or C-terminal half of eYFP (ceYFP) at their N-terminus, respectively.

Soybean transformation

Mature seeds from soybean cultivar ‘Williams 82’ were disinfected using chlorine gas for 12 hours. The disinfected seeds were soaked in distilled water for 12 hours at room temperature at dark. Half-seeds were soaked in resuspended agrobacterium liquid co-cultivation medium (OD650 = 0.6, 3.21g/L Gamborg B-5 basal medium, 30g/L sucrose, 3.9g/L MES, 0.4g/L L-cystine, 0.1542g/L DTT, 0.25mg/L GA3, 1.67mg/L 6-BA and 0.3924g/L acetosyringone, pH = 5.4) for 30 minutes. After infection, the explants were transferred to solid co-cultivation medium. The plates were sealed with Micropore tape (Catelog #1530-0, 3M, St. Paul, MN) and incubated in the dark at 21 ℃ for 4 days. After co-cultivation, explants were inserted into shoot induction medium plate (3.21g/L Gamborg B-5 basal medium, 30g/L sucrose, 0.59g/L MES, 0.25g/L timetin, 0.1 g/L cefradine, 1.67mg 6-BA, 2.5mg/L glufosinate, pH = 5.7, 2g/L gellan gum powder). Shoot induction was carried out at 26 ℃ with a photoperiod of 18 hours and a light intensity of 40-70 μE/m2/s. After 4 weeks, the inducted shoots were cut from cotyledons and transferred to shoot elongation medium (4.43g/L Murashige & Skoog modified medium with Gamborg vitamins, 30g/L sucrose, 0.59g/L MES, 0.25g/L timentin, 0.1g/L cefradine, 0.05g/L asparagine, 0.05g/L glutamine , 0.5mg/L GA3, 0.1mg/L IAA, 1mg/L zeatin, 5mg/L glufosinate, pH = 5.7, 2g/L gellan gum powder) under same temperature and photoperiod. After 2-4 weeks in shoot elongation medium, the glufosinate-resistant shoots were cut and transferred to rooting medium (4.43g/L Murashige & Skoog modified medium with Gamborg vitamins, 30g/L sucrose, 0.59g/L MES, 0.05g/L asparagine, 0.05g/L glutamine, 0.1mg/L IBA, pH = 5.7, 3g/L gellam gum) for further shoot and root elongations. After root grows longer than 1 cm, plants were transferred to moistened Berger BM2 soil (Berger, Saint-Modeste, QC, Canada), and kept enclosed in clear plastic tray in a growth chamber at 26 ℃ with a 16-hour photoperiod at 250 -350 µE/m2/s.

Genotyping the transgenic and genome editing lines

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

RNA extraction, regular qRT-PCR and stem loop RT-PCR

Total RNA was extracted using the TRIzol reagent (Cat. # 15596018, Invitrogen). 2 µg DNA-free RNA was used to synthesize cDNA with the

Promega M-MLV Reverse Transcriptase (Cat. # M1701, Promega). qRT-PCR was performed using Applied Biosystems™ Power SYBR™ Green PCR Master Mix (Cat. # 4368577, Applied Biosystems) on an Applied Biosystems StepOnePlus™ Real-Time PCR Systerm (Cat. # 4376600, Applied Biosystems). Stem-loop RT-PCR was used to examine the expression levels of miRNAs following a previous protocol[27].

mRNA, small RNA and Degradome sequencing

The cleaned RNA-seq reads were mapped to the soybean reference genome[9] using STAR (v2.5.4b) with only unique mapped reads kept[28]. The expression levels (FPKM) were calculated using the cuffnorm function in cufflinks (v2.2.1)[29]. Degradome libraries were constructed and sequenced at Novogene Corporation Inc. (Sacramento, CA). The potential target genes of miRNA produced by *lncRG1* and *lncRG2* were analyzed using CleaveLand with the following parameters, -r 0.6 and -c 2. (v4.5)[30].

RNA ligase-mediated 5′ rapid amplification of cDNA ends (5' RLM-RACE)

RLM-RACE was performed following the protocol described previously[24]. The mRNAs were then ligated with 5′ RACE oligo adaptors for reverse transcription using the GeneRacer kit (Cat. # L150202, ThermoFisher Scientific) followed by nested PCR. The purified PCR products were sequenced by using the WideSeq method (https://www.purdue.edu/hla/sites/genomics/wideseq-2/).

Phylogenetic analysis and nucleotide diversity calculation

Sequence alignments and tree construction were performed using the Maximum Likelihood method[31] in MEGA7[32]. Nucleotide diversity was calculated using vcftools (v0.1.16)[33].

RNA secondary structure prediction

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

miRNA target prediction

Potential targets by the miRNAs 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 as 2.5[34].

Yeast Two-Hybrid (Y2H) assays

Y2H assays were performed using the Matchmaker Gold Yeast Two-Hybrid System (Cat. # 630489, Takara Bio USA Inc). Different combinations of the constructs were co-transformed into the yeast strain Y2H Gold. The transformed yeast cells were spread on SD (–Trp/–Leu) medium. The plates were incubated at 30 °C for 3-5 days. 5-10 colonies were picked from each plate and resuspended in 0.9% (w/v) NaCl solution. Then the yeast cells were spotted on SD (–Trp/–Leu/–Ade/–His) selection medium. Plates were incubated at 30°C for 3 days to observe yeast growth. pGADT7-T + pGBKT7-53 was used as positive control; pGADT7-T + pGBKT7-Lam was used as negative control.

Bimolecular fluorescence complementation (BiFC)

Different constructs were transformed into Agrobacterium tumefaciens strain EHA105. The agrobacterium suspension was injected into the abaxial surface of 4–6-week-old *Nicotiana Benthamiana* leaves with a needleless syringe. Plasmid expressing mCherry-labeled Pentunia hybrida's histone H1-3 (acted as the nuclear marker) was co-infiltrated with the expression construct of each target gene. 72 h after infiltration, the fluorescent signals in detached leaves were imaged using a Zeiss LSM-880 laser-scanning confocal microscope.

Data and code availability

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. This paper does not report original code.

Acknowledgments

We thank Drs. Xuemei Chen, Damon Lisch, and Robert Schmitz for constructive comments on this work. This work was mainly supported by the Agriculture and Food Research Initiative of the U.S. Department of

Agriculture National Institute of Food and Agriculture (grants 2018-67013-27425, 2021-67013-33722, and 2022-67013-37037), and partially supported by United Soybean Board, North Central Soybean Research Program, Indiana Soybean Alliance, and Ag Alumni Seed.

Author contributions

JM and XZF designed the research. WW, JD, XF, XW LC, CBC, SAS, RLN, SL and JW performed the experiments. WW, XW, BCM and JM analyzed the data. WW and JM wrote the manuscript and BCM edited the manuscript.

Declaration of interests

The authors declare no competing interests.

References

1. Olsen, K.M. & Wendel, J.F. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annual review of plant biology* **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 Phytologist* **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. *Theoretical and Applied Genetics* **132**, 1195-1209 (2019).

5. Liu, Y. *et al.* Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162-176. e13 (2020).

6. Broersma, D., Bernard, R. & Luckmann, W. Some effects of soybean pubescence on populations of the potato leafhopper. *Journal of Economic Entomology* **65**, 78-82 (1972).

7. Song, Q. *et al.* Fingerprinting soybean germplasm and its utility in genomic research. *G3: Genes, genomes, genetics* **5**, 1999-2006 (2015).

8. Shen, Y. *et al.* DNA methylation footprints during soybean domestication and improvement. *Genome biology* **19**, 1-14 (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. *Proceedings of the National Academy of Sciences* **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**, 1-13 (2016).

12. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nature reviews Molecular cell biology* **22**, 96-118 (2021).

13. 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).

14. Reams, A.B. & Roth, J.R. Mechanisms of gene duplication and amplification. *Cold Spring Harbor perspectives in biology* **7**, a016592 (2015).

15. Bradley, D. *et al.* Evolution of flower color pattern through selection on regulatory small RNAs. *Science* **358**, 925-928 (2017).

16. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485-488 (1997).

17. Tan, L. *et al.* Control of a key transition from prostrate to erect growth in rice domestication. *Nature genetics* **40**, 1360-1364 (2008).

18. Fabian, M.R. & Sonenberg, N. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nature structural & molecular biology* **19**, 586-593 (2012).

19. Zeng, Z.-B. Precision mapping of quantitative trait loci. *Genetics* **136**, 1457-1468 (1994).

20. Broman, K.W., Wu, H., Sen, Ś. & Churchill, G.A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890 (2003).

21. Bradbury, P.J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633-2635 (2007).

22. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* **38**, 203-208 (2006).

23. Zhou, Z. *et al.* Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature biotechnology* **33**, 408-414 (2015).

24. Ren, B., Wang, X., Duan, J. & Ma, J. Rhizobial tRNA-derived small RNAs are signal molecules regulating plant nodulation. *Science* **365**, 919-922 (2019).

25. Lei, Y. *et al.* CRISPR-P: a web tool for synthetic single-guide RNA design of CRISPR-system in plants. *Molecular plant* **7**, 1494-1496 (2014).

26. Bai, M. *et al.* Generation of a multiplex mutagenesis population via pooled CRISPR‐Cas9 in soya bean. *Plant Biotechnology Journal* **18**, 721-731 (2020).

27. Chen, C. *et al.* Real-time quantification of microRNAs by stem–loop RT–PCR. *Nucleic acids research* **33**, e179-e179 (2005).

28. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

29. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578 (2012).

30. 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).

31. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution* **10**, 512-526 (1993).

32. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution* **33**, 1870-1874 (2016).

33. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156-2158 (2011).

34. Dai, X., Zhuang, Z. & Zhao, P.X. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic acids research* **46**, W49-W54 (2018).

Figure legends

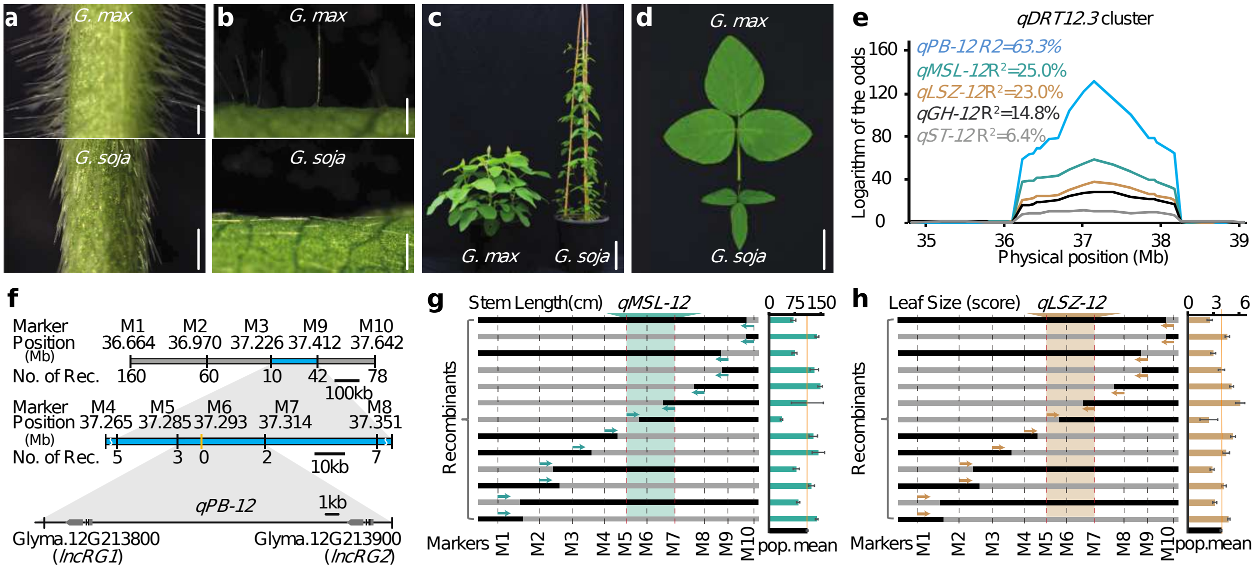
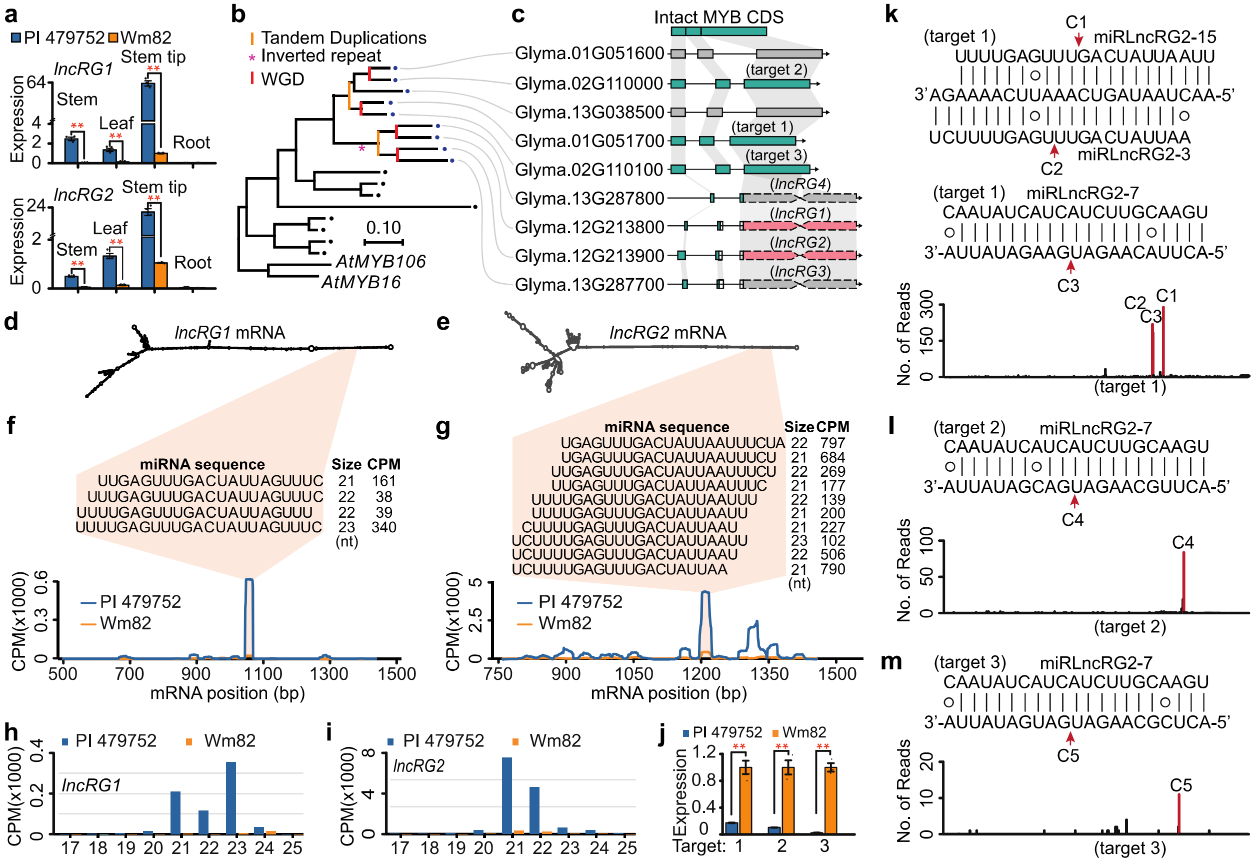


Fig. 1: Map-based cloning of multiple DRT QTLs identifies a single locus with pleiotropic effects. a-b, Comparisons of pubescence form on stems (a) and leaves (b) between *G. max* and *G. soja*. Scale bar = 3 mm. **c**, Comparisons of stem height and growth habit between *G. max* and *G. soja*. Scale bar = 10 cm. **d**, Comparison of leaf size between *G. max* and *G. soja*. Scale bar = 5 cm. **e**, Primary mapping region of *qDRT12.3* on chromosome 12. The *y*-axis represents the log10 likelihood ratio and R^2 values indicate the phenotypic variations explained by each QTL. **f**, Fine mapping of *qPB-12*. **g-h**, Fine mapping of *qMSL12* (g) and *qLSZ12* (h). Each bar represents the genotype of the recombinants with the same haplotype at all markers. The black color represents the *G. soja* genotype and the grey color represents the *G. max* genotype. Arrows indicate the deduced location of the QTL. Green and brown shades highlight the final mapping interval. Data are represented as mean ± SEM.



miRNA sequence	Size	CPM
UUGAGUUUGACUAUUAGUUUC	21	161
UUUGAGUUUGACUAUUAGUUUC	22	38
UUUUGAGUUUGACUAUUAGUUU	22	39
UUUUGAGUUUGACUAUUAGUUUC	23	340
	(nt)	

miRNA sequence	Size	CPM
UGAGUUUGACUAUUAAUUUCUA	22	797
UGAGUUUGACUAUUAAUUUCU	21	684
UUGAGUUUGACUAUUAAUUUCU	22	269
UUGAGUUUGACUAUUAAUUUC	21	177
UUUUGAGUUUGACUAUUAAUUU	22	139
UUUUGAGUUUGACUAUUAAUU	21	200
CUUUUGAGUUUGACUAUUAAU	21	227
UCUUUUGAGUUUGACUAUUAAUU	23	102
UCUUUUGAGUUUGACUAUUAAU	22	506
UCUUUUGAGUUUGACUAUUAA	21	790
	(nt)	

Fig. 2: *lncRG1* and *lncRG2* harbor IRs and produce abundant sRNAs primarily targeting three closely related MYB genes. a, Expression levels of *lncRG1* and *lncRG2* in different tissues as determined by qRT-PCR with Wm82 stem tip set as “1” and the others adjusted accordingly. **b**, Phylogenetic relationships of *lncRG1*, *lncRG2*, and their close MYB relatives. Colored lines indicate duplication events. Red asterisk marks the deduced time when the original IR occurred. **c**, Gene models and alignments of *lncRG1*, *lncRG2*, and their close MYB relatives. Green bars represent coding regions and pink bars represent the inverted repeats (IRs). **d**-**e**, Predicted secondary structures of *lncRG1* and *lncRG2* transcripts. **f**-**g**, Distribution, abundance, and the major cluster of sRNAs produced by *lncRG1* and *lncRG2*. **h**-**i**,

Abundance of sRNAs 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 qRT-PCR with Wm82 set as "1" and the others adjusted accordingly. **k**-**m**, The predicted cleavage sites supported by degradome sequencing on the target genes. Letter Cs represents cleavage sites. In (a) and (j), The dots show the values from different biological replicates (n=3), and the red asterisks indicate the significant level at $P < 0.01$ (Student's *t*-test) and data are represented as mean ± SEM.

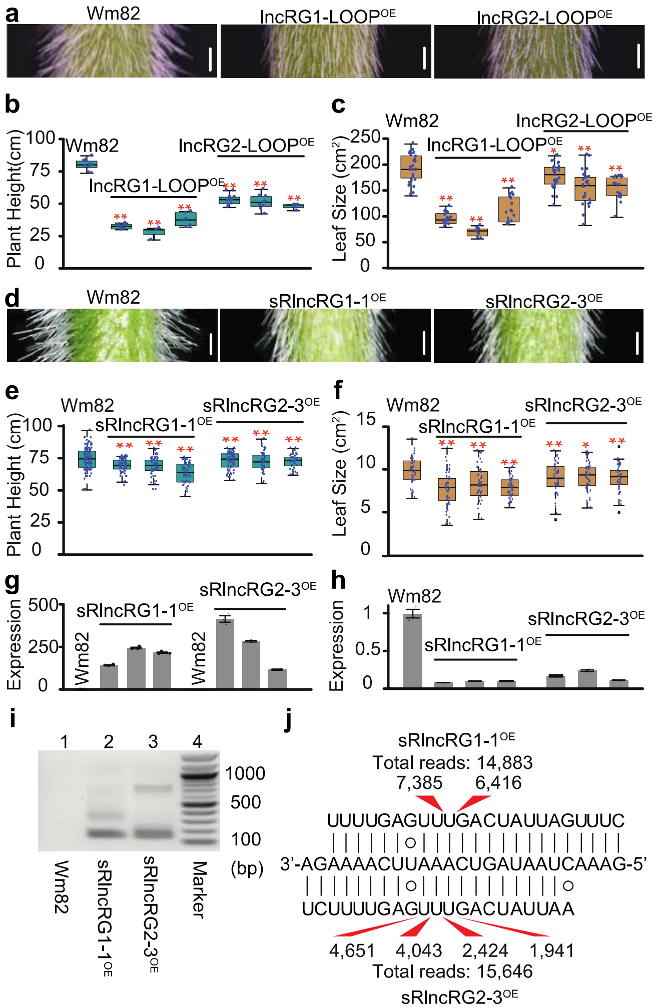


Fig. 3: Overproduction of sRNAs in cultivated soybean promotes the wild soybean-type phenotypes. a-**c**, The phenotypic changes in pubescence form (a), plant height (b) and leaf size (c) of the stem-loop over-expression lines compared with those of Wm82. Scale bars=3 mm. **d**-**e**, The phenotypic changes in pubescence form (d), plant height (e) and leaf size (f) of the artificial miRNA over-expression lines compared with those of Wm82. Scale bars=3 mm. **g**-**h**, The expression levels of the artificial miRNAs (g) and the three target MYB genes (h) in the transgenic lines compared with those inWm82 as determined by stem loop RT-qPCR and regular qRT-PCR,

respectively, with Wm82 set as “1” and the others adjusted accordingly. The dots show the values from different biological replicates (n=3). Data are represented as mean ± SEM. **i**, Gel electrophoresis image of RLM-RACE from the transgenic lines and Wm82. **j**, Cleavage frequencies detected by RLM-RACE followed by deep sequencing. Numbers show the total reads number and the read number at each cleavage site. In (b), (c), (e), and (f), 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. Red asterisks indicate significant levels at P <0.01 or P<0.05 (Student’s *t*-test).

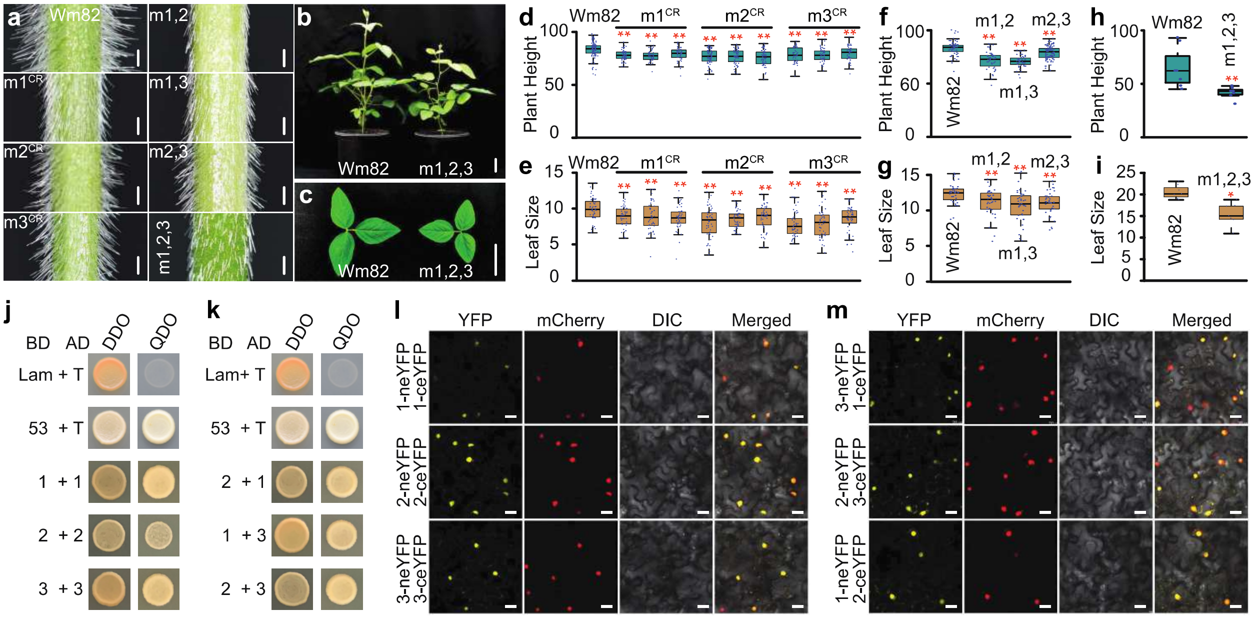


Fig. 4: Functional redundancy and divergence of the three MYB genes targeted by the sRNAs. a-c, Photographic illustration of the phenotypic changes in the pubescence form (a), plant height (b), and leaf size (c) of the gene-edited mutants compare with Wm82. The m1, m2 and m3 are mutants of target 1, target 2 and target 3, respectively. Scale bars=3mm in (a) and 5cm in (b-c). **d-e**, Statistics of the plant height (d) and the leaf size (e) of single mutants and Wm82. **f-g**, Statistics of the plant height (f) and the leaf size (g) of the double mutants and Wm82. **h-i**, Statistics of the plant height (h) and leaf size (i) of the triple mutants and Wm82. **j-k**, Home- (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. **l-m**. Home- (l) and hetero- (m) protein-protein interactions among the three target genes detected by BiFC assay. Scale bars=20μm. In (d-i), 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. Red asterisks indicate significant levels at $P < 0.01$ or $P < 0.05$ (Student's *t*-test).

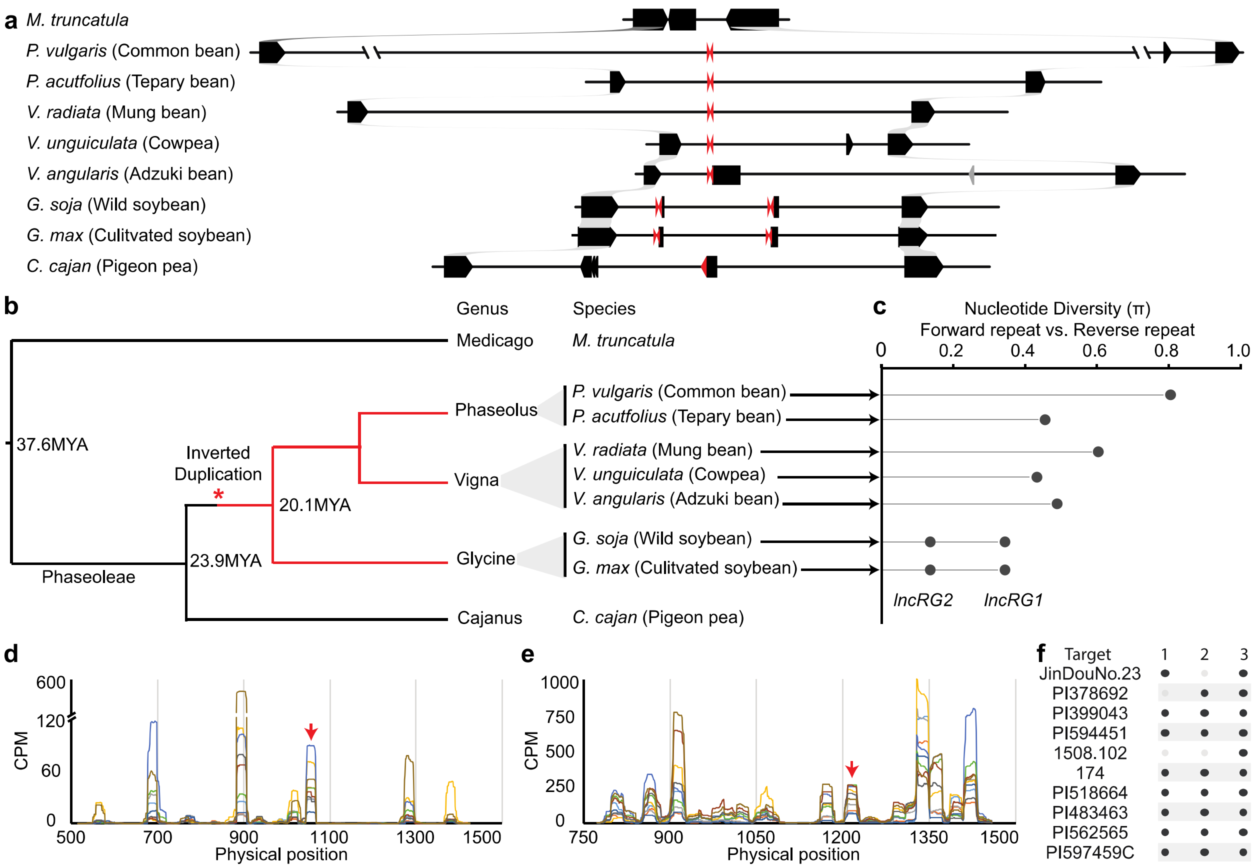
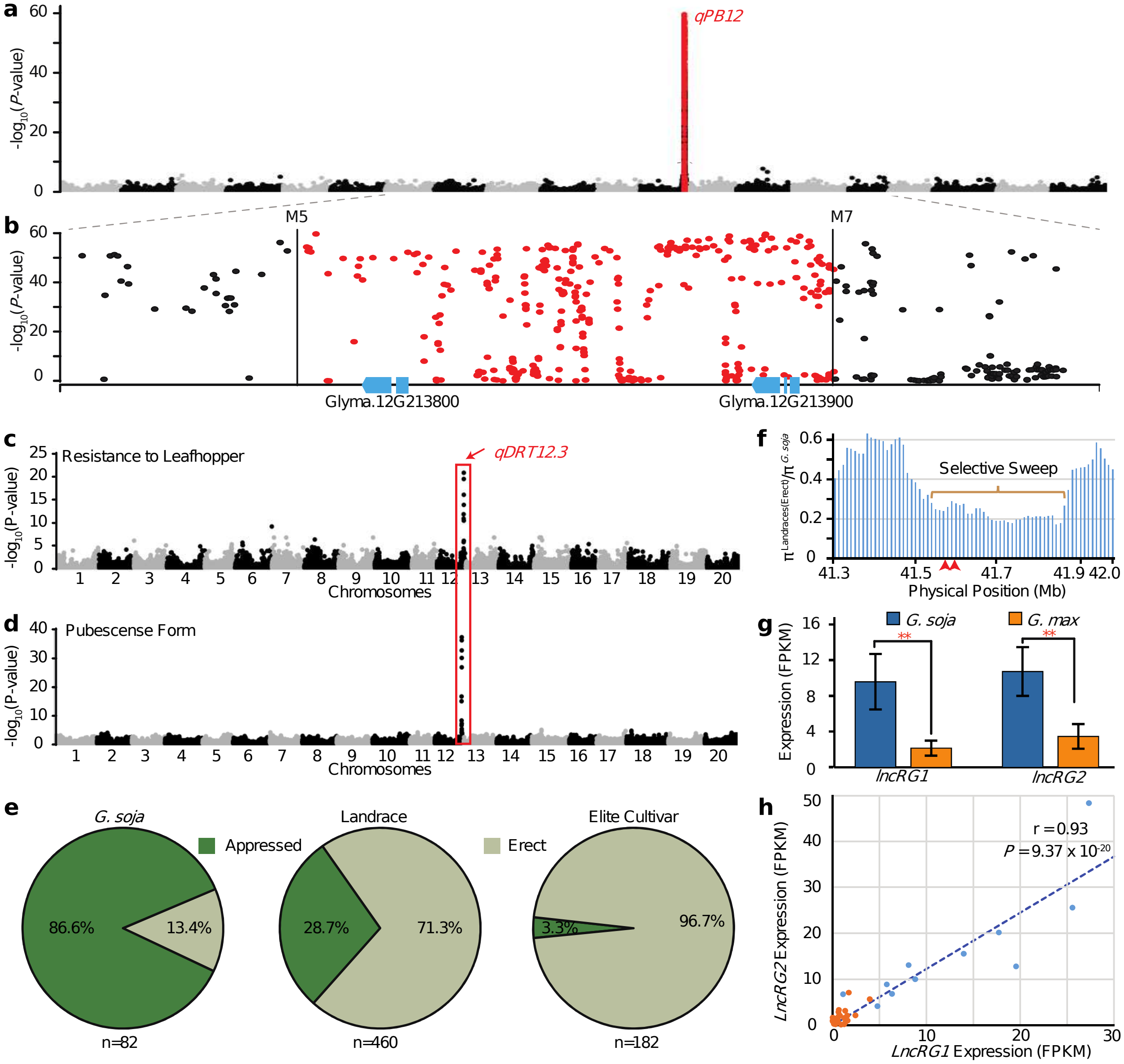
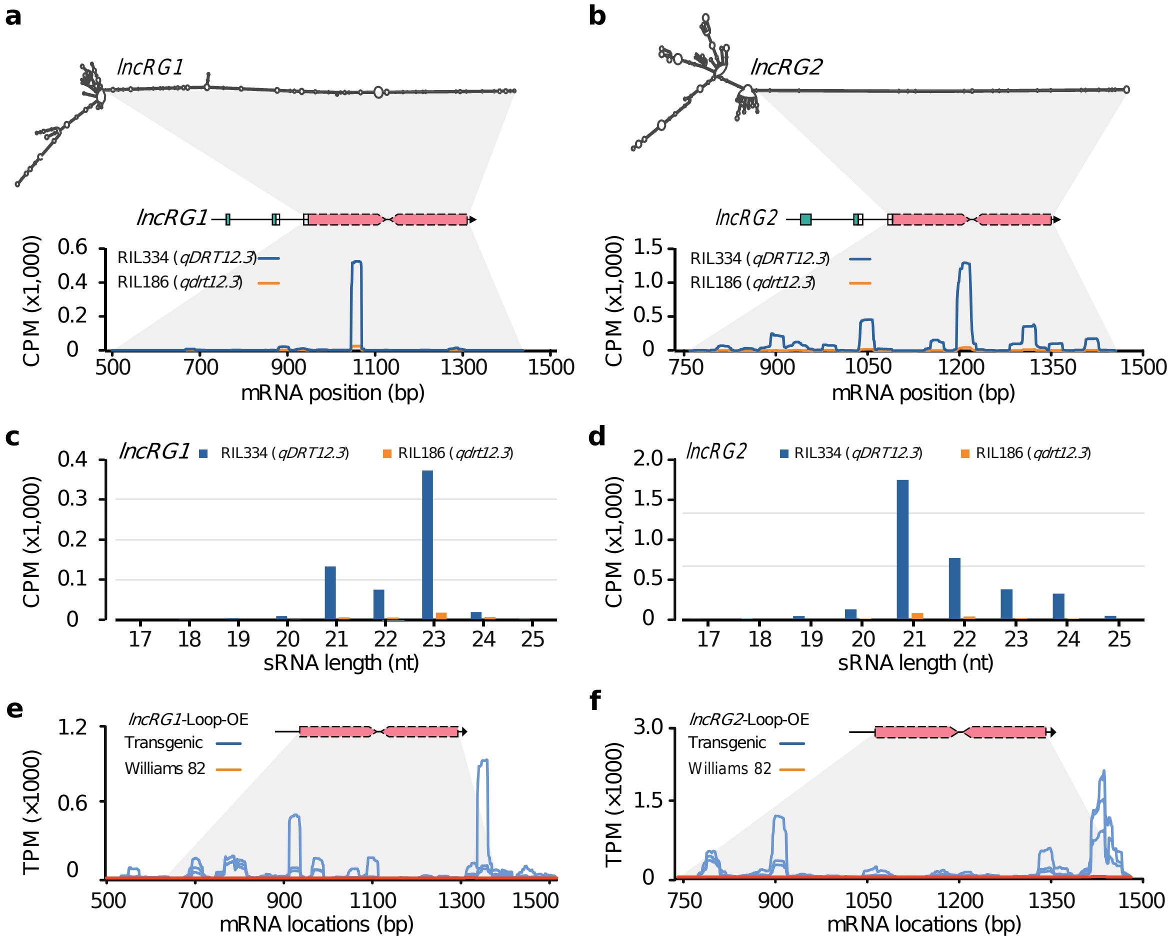


Fig. 5: The birth and evolutionary consequences of the lncRGs in legumes. a, Collinearity analysis of nine legume species at the region harboring the orthologs of *lncRG1* and *lncRG2*. Black boxes present genes and grey shades connect the ortholog genes between species. Red triangles represent the IRs. **b**, Phylogenetic relationships of the nine legume species as determined in previous studies[10,11]. Red lines highlight the genera that carry the IRs and the asterisk indicates the deduced timepoint when the original IRs occurred. **c**, Nucleotide diversity between the forward and reverse repeats in each species. **d**-**e**, The distribution patterns of the sRNAs

produced by *lncRG1* (d) and *lncRG2* (e) in ten diverse soybean accessions as indicated by different colors. Arrows points 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 sRNAs produced by *lncRG1* and *lncRG2* in each of the ten soybean accessions. Black dots indicate predicted targets, while grey dots indicate they are not predicted to be targets.

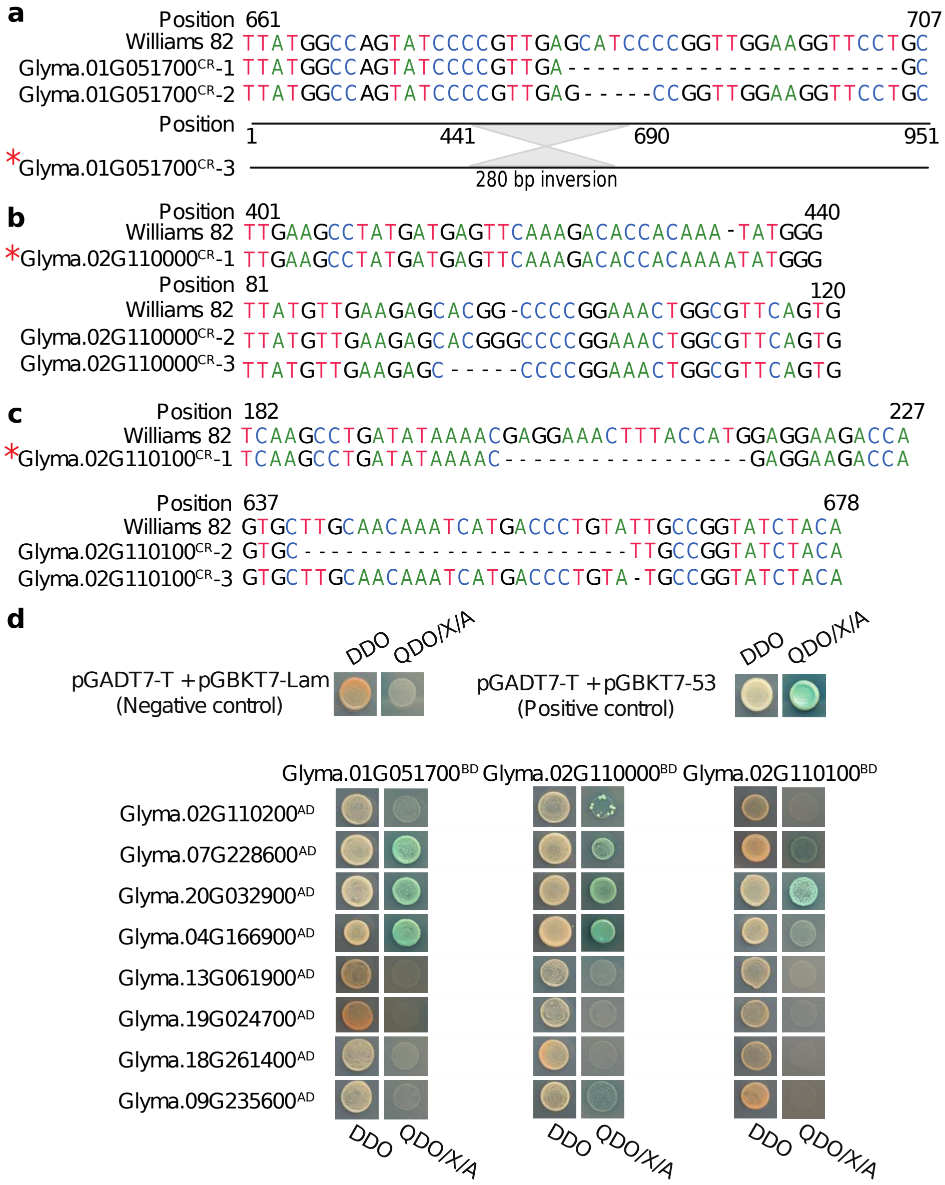


Extended Data Fig. 1 Association studies, selection analyses and expression analyses. a-b Association between genetic variations and expression levels of *lncRG1* and *lncRG2* within the final mapping region. Manhattan plot displays the result of genome-wide association study (GWAS) on pubescence form. The y-axes are the negative log10 of the *P*-values and the red color highlight markers within the final mapping region. **c**-**d**, Manhattan plots displaying the results of genome-wide association studies (GWAS) on leafhopper resistance (c) and pubescence form (d). The *y*-axes represent the negative log10 of the *P*-values from GWAS. The *x*-axes represent the twenty soybean chromosomes. The rectangle highlights the *qDRT12.3* locus on chromosome 12. The genotypic and phenotypic data of the soybean accessions (n=784) used for the GWAS are from the USDA soybean germplasm collection. **e**, Frequencies of erect and appressed pubescence form in *G. soja*, landrace and elite cultivar sub-populations. n indicates the number of soybean accessions in each sub-population. **f**, Selective sweep surrounding the fine-mapped *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). Each vertical bar represents a 100-kb window (with 10-kb sliding step). 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**, Comparison of expression levels of *lncRG1* and *lncRG2* between *G. soja* (n=9) and *G. max* (n=36) accessions. The *y*-axis represents the expression level as measured from RNA-seq data and the unit is fragments per kilobase of transcript per million mapped reads (FPKM). The red asterisks indicate the significant level at $P < 0.01$ (Student's *t*-test) and data are represented as mean ± SEM. **h**, Co-expression between *lncRG1* and *lncRG2*. The *x*-axis and y-axis represent the expression levels of *lncRG1* and *lncRG2*, respectively, as measured from RNA-seq data of 45 highly diverse soybean accessions. 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). Unit is fragments per kilobase of transcript per million mapped reads (FPKM). Dashed line is the trend line. The Pearson correlation value and the corresponding *P*-value were labeled.

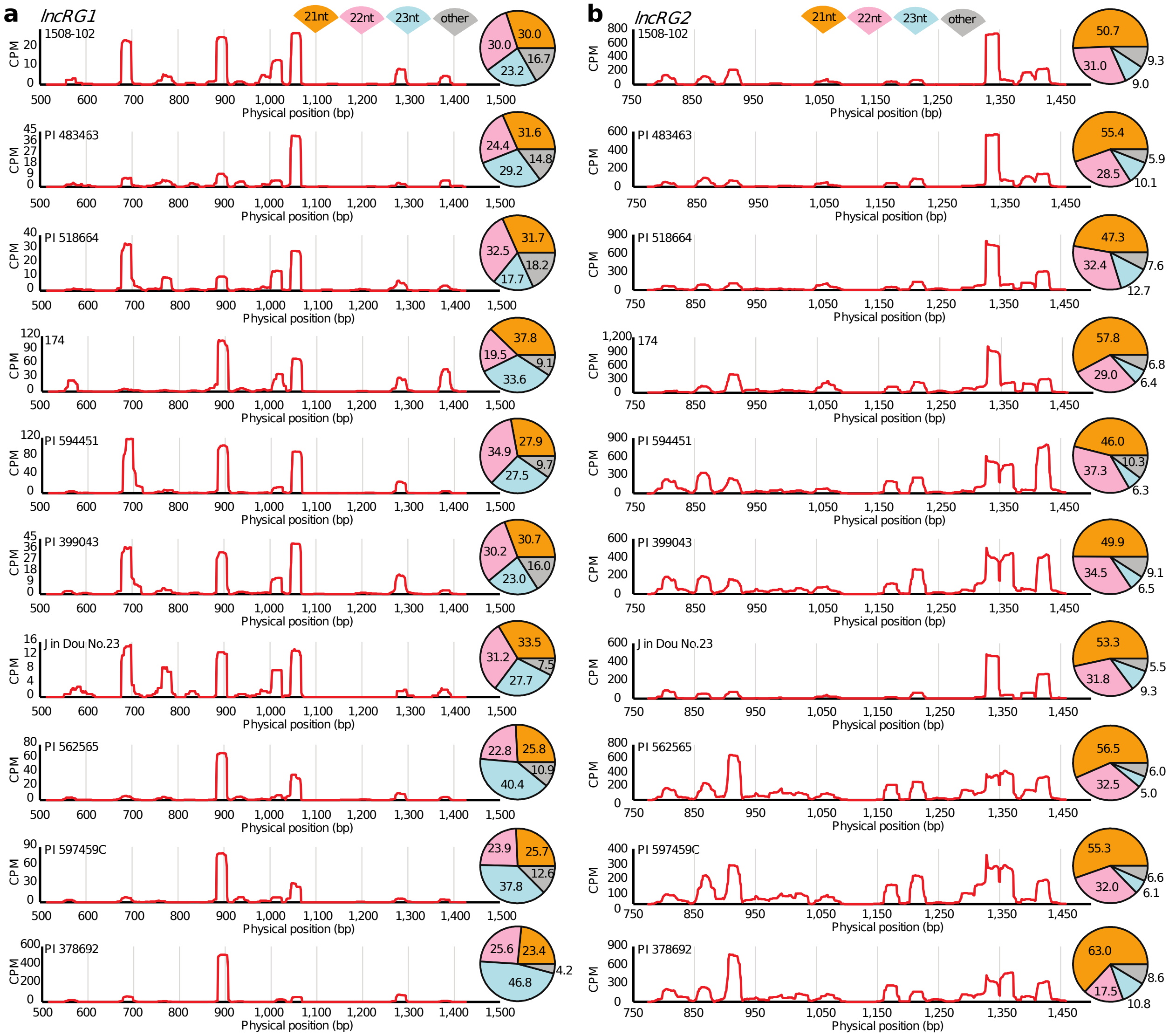


Extended Data Fig. 2 Abundance and distribution of sRNAs produced by *lncRG1* and *lncRG2* in a pair of RILs and the transgenic lines. a, Abundance and distribution of sRNAs produced by *lncRG1* in RIL186 (*qdrt12.3*) and RIL334 (*qDRT12.3*). The *x*-axis shows the position on the *lncRG1* transcript, 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).


a
Position 661 707
Williams 82 TTATGGCCAGTATCCCCGTTGAGCATCCCCGGTTGGAAGGTTCCTGC
Glyma.01G051700CR-1 TTATGGCCAGTATCCCCGTTGA-----------------------GC
Glyma.01G051700CR-2 TTATGGCCAGTATCCCCGTTGAG-----CCGGTTGGAAGGTTCCTGC
Position 1 441 690 951
*Glyma.01G051700CR-3
280 bp inversion
b
Position 401 440
Williams 82 TTGAAGCCTATGATGAGTTCAAAGACACCACAAA-TATGGG
*Glyma.02G110000CR-1 TTGAAGCCTATGATGAGTTCAAAGACACCACAAAATATGGG
Position 81 120
Williams 82 TTATGTTGAAGAGCACGG-CCCCGGAAACTGGCGTTCAGTG
Glyma.02G110000CR-2 TTATGTTGAAGAGCACGGGCCCCGGAAACTGGCGTTCAGTG
Glyma.02G110000CR-3 TTATGTTGAAGAGC-----CCCCGGAAACTGGCGTTCAGTG
c
Position 182 227
Williams 82 TCAAGCCTGATATAAAACGAGGAAACTTTACCATGGAGGAAGACCA
*Glyma.02G110100CR-1 TCAAGCCTGATATAAAAC-----------------GAGGAAGACCA
Position 637 678
Williams 82 GTGCTTGCAACAAATCATGACCCTGTATTGCCGGTATCTACA
Glyma.02G110100CR-2 GTGC-----------------------TTGCCGGTATCTACA
Glyma.02G110100CR-3 GTGCTTGCAACAAATCATGACCCTGTA-TGCCGGTATCTACA
d
DDO QDO/X/A
pGADT7-T +pGBKT7-Lam (Negative control)
DDO QDO/X/A
pGADT7-T +pGBKT7-53 (Positive control)
Glyma.01G051700BD Glyma.02G110000BD Glyma.02G110100BD
Glyma.02G110200AD
Glyma.07G228600AD
Glyma.20G032900AD
Glyma.04G166900AD
Glyma.13G061900AD
Glyma.19G024700AD
Glyma.18G261400AD
Glyma.09G235600AD
DDO QDO/X/A
DDO QDO/X/A
DDO QDO/X/A

Extended Data Fig. 3 Mutations created by CRISPR-Cas9 and protein-protein interaction as detected by Y2H. 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**, 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.



Extended Data Fig. 4 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 are shown in percentage (%) in individual pies.

Supplemental information

Supplementary Table 1. sRNAs produced by *lncRG1* and *lncRG2* with CPM>10.

Supplementary Table 2. List of genes targeted by 27 sRNAs (CPM>100) produced by *lncRG1* and *lincRG2*.

Supplementary Table 3. Expression levels (FPKM) of the 163 target genes in shoots, stems, and leaves of Williams 82 and PI 479752.

Supplementary Table 4. List of top 20 sRNAs produced by *lncRG1* and *lncRG2* in 10 diverse soybean accessions with *G. soja* haplotype.

Supplementary Table 5. List of genes targeted by sRNAs (top 20) produced by *lncRG1* and *lincRG2* in 10 soybean accessions.

Supplementary Table 6. List of primers used in this study.

Supplementary Movie 1. Erected pubescence confers resistance to leafhopper.

Supplementary Movie 2. Appressed pubescence is susceptible to leafhopper.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- MovieS1.mov
- MovieS2.mov