

## Peer Review Information

---

**Journal:** Nature Genetics

**Manuscript Title:** Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance in soybean

**Corresponding author name(s):** Professor Jianxin Ma, Professor Xianzhong Feng

### Reviewer Comments & Decisions:

|                                          |
|------------------------------------------|
| <b>Decision Letter, initial version:</b> |
|------------------------------------------|

10th Aug 2023

Dear Professor Ma,

Your Article, "Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team with a view to identifying key priorities that should be addressed in revision. In this case, we think all three referees have provided constructive reviews aimed at strengthening the analyses. We particularly ask that you improve the mechanistic aspects by additional experiments, provide more detailed description of data, methodology and statistical analysis, and include more molecular data to strengthen your conclusions. Please address all referee comments as thoroughly as possible with appropriate revisions.

We therefore invite you to revise your manuscript taking into account all reviewer and editor comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

\*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each

referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

\*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available [here](#).

Refer also to any guidelines provided in this letter.

\*3) Include a revised version of any required Reporting Summary:

<https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our [guidelines on digital image standards](#).

Please use the link below to submit your revised manuscript and related files:

[redacted]

**Note:** This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within 4 to 6 months. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit [www.springernature.com/orcid](http://www.springernature.com/orcid).

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,  
Wei

Wei Li, PhD  
Senior Editor  
Nature Genetics  
New York, NY 10004, USA

www.nature.com/ng

## Reviewers' Comments:

### Reviewer #1:

#### Remarks to the Author:

This is a very interesting paper that I enjoyed reading. The authors identified two inverted repeated elements associated with domestication-related traits in soybean. Basically, the authors showed that these two IRs, named lncRG1 and 2 produce abundant small RNAs in wild soybean, silencing three MYB transcription factors (TFs). Conversely, in domesticated soybean, the expression of both IRs is reduced, releasing the repression of the TFs and leading to characteristic phenotypic traits. The genetics and the experiments with transgenic wild soybean are well-conducted and convincing. Overall, I'm positive about this article. Still, there are a few mechanistic aspects that would benefit from some additional analysis.

1. Lines 119-120: it is interesting that lncRG3 and lncRG4 are not associated with any trait. Is there any explanation? are these two copies permanently silenced or activated in wild and domesticated soybean accession?

2. One missing piece of this story is why lncRG1 and lncRG2 are less expressed in domesticated soybeans. Considering that these IRs are likely to trigger TGS, see my following comments, I suggest analyzing the DNA methylation status of both lncRNAs in wild and domesticated plants. Perhaps a difference in the RdDM machinery between accession impact differentially the epigenetic landscape of these loci, turning these IRs into trans Obligatory epialleles, which could be an amazing discovery.

3. L129. The sRNA size profile over the two IRs is curious. given the large complementarity of the IR I would expect their transcript to be mainly processed by DCL3/DCL3-like enzymes to produce 24 nt sRNAs as previously shown (doi: /10.1038/nplants.2016.163; /10.1111/tpj.12545; /10.1186/s12864-022-08364-4; and /10.1073/pnas.1903131116). However, you detect none. This is strange. As a matter of fact, the most abundant specie at lncRG1 are 23 nt, also present in lncRG1, is 23 nt sRNAs, which according to Craig Pikaard papers are a subproduct of 24 nt biogenesis. Thus, the presence of 23 nt suggests there should be 24s, although it is true IR may not need RDR2, which activity is responsible for the 23 nt species.

Did you intentionally exclude 24 nt sRNA from the analysis, or is there no 24 nt mapping there? Could this be an analysis artifact, perhaps by filtering out multimappers? Actually, the material and methods explanation of sRNA library preparation and analysis is very poor. Is the typical double peak 21/24 sRNA size profile visible in the entire dataset?

4. A recent paper showed that polymorphisms in IRs are associated with phenotypic traits (doi.org/10.1016/j.celrep.2023.112029). Although these are not polymorphic, it is possible they act as epialleles, something interesting to discuss. The same paper shows that IRs affect the expression of neighbor genes, this is something it was shown for many IRs (doi.org/10.1093/molbev/msr282; /10.1016/j.celrep.2023.112029; /10.1093/plcell/koac220; /10.1111/nph.18068; /10.1016/j.molp.2020.02.009; 10.1073/pnas.1903131116).

This makes me wonder: a. Have you checked the expression of the lncRG1 and 2' adjacent genes that are conserved according to Figure 5A?

b. Is the genomic vicinity of lncRG1 and 2 conserved at lncRG3 and loci? Can this explain the lack of phenotype of L3 and 4 if L1 and 2 regulate in cis neighbor genes? In other words, could genes next to lncRG1 and 2, but absent in the region encoding lncRG3 and 4, help explain the phenotype?

In your paper, the two lncRNA seem to work as miRNAs mostly at PTGS level. However, long IRs were previously associated with TGS (see references mentioned before). It would be interesting to see the methylation patterns over the lncRNAs loci and targets. I'm sure there should be publicly available BS-seq datasets to consult.

The M&M is too poor, it needs to be expanded to include more details. A clear example is the mRNA, sRNA, and degradome session which gives almost no information about libraries preparation and analysis.

There is a very limited introduction/discussion of the current literature about gene regulation mediated by IRs.

Reviewer #2:

Remarks to the Author:

The study demonstrated that the recent birth of two long non-coding RNAs from MYB genes, their evolutionary pathway and their functionality in regulating several traits associated with domestication. The genes regulate the traits through small RNA pathway that suppress the transcript accumulation of the MYB genes by RNA degradation. Authors provide solid evidence supporting the conclusion. In addition, the linkage mapping and GWAS pinpoint the multiple QTLs for the traits into a 29 kb region where the two genes are located. The 29 kb region was also located in a sweeping region. The accumulation of the two lncRNAs, and their major small RNAs and their targeted MYB RNAs are different in soja and max accessions, two parental lines of the RIL population. The authors concluded that the QTLs 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. The study has a good originality and significance. The manuscript is well written. Overall, the data supports the conclusion. The author needs to provide more detailed and comprehensive description of data, methodology and statistic analysis. Introduction and Discussion could be improved to articulate the originality and significance related to the previous publication.

Suggested Overall Improvement.

1. Authors should provide more detailed and comprehensive description of the data and methods used and the results observed in the study. It is important at the situation that the causal allele was not identified. Many results are based on association, gene expression (affected by environments and genetic backgrounds) and different populations.

a. Please provide the data resource, all phenotypic data, genotyping data, accession IDs and transcriptome data used in the study in supplementary files or make them available in the public database. A significant amount of data used in the study is from the public. It should be encouraged by our research community. However, please try to clearly state the data used in the study and if the data was validated.

b. Please present results from the linkage mapping and GWAS for all domestication-related traits that the two genes regulate, and their phenotypes in the over-expression lines. Not all of the stated traits were presented in each analysis (Linkage mapping, GWAS using resequencing data, and transgenic analysis), for example, leafhopper resistance in transgenic plants. If there is no consistence in mapping and transgenic study, please remove the traits or state the results.

c. Please provide more details for the material and methods and figure legend such as specific population used in each analysis. Please describe how to determine the phenotypes for the traits. Since many results were from different populations and based on association. For examples, Extended Data Fig 1 c, d, e, f, g, and h seem that different populations were used.

2.Considering that soja and max have sequence difference of whole genomes, two lncRNAs, Myb genes. RNA-seq from soja and max should be aligned to their own genomes for determining gene expression, not only on Wm82. The primers of the QRT-PCRs should be validated to amplify the targeted RNAs at the same efficiency to determine their expression in soja and max.

3.Optional: Determine the accumulation of the lncRNAs and their targeted Myb transcripts in the collection of accessions used in Extended Figure 1g, determine their accumulation correlation of lncRNAs vs targeted RNAs in the accessions such as PCC, and conduct TWAS for the traits.

#### Suggested Specific Improvement

4.Figure 1 a, b. c, d: please mark the specific lines instead of just soja and max.

5.It is not clear for Extended Data Figure 1a and 1b: It is likely a mistake.

Figure legend states that "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" However, in the Result (Line 71-74) "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 resequencing data from 74 G. soja and 594 G. max accessions5 (Extended Data Fig. 1a and 1b)."

6.Extended Figure 1f: the bars should overlap based on figure legend. Why are they distinct bars without overlapping.

7.Mark the tissues and accessions used for RNA seq and accessions in Ex Figure 1g and 1h. Are they the same tissue?

8.P79-80: "According to the Wm82 reference genome, this region harbors only two genes, Glyma.12G213800 and Glyma.12G213900, both lncRNAs".

a.Please aligned all RNA reads of RNA-seq from Wm82 and PI479752 back to genome sequences of Wm82 and PI 479752 respectively to exclude the possibility that there are other genes in the region.

9.Figure 2a and 2j, please provide evidence that qRT-PCR amplified the genes in soja and max at the same efficiency. Including the sequence alignments of the primers against their targeted gene sequences of PI479752 and Wm82.

10. Please examine if number of RI gene (2 vs 1 genes) varies in soja accessions using available re-sequencing data. It will help to determine if duplication of the lncRNA genes occurred after soybean speciation and before domestication.
11. Please describe if other phenotypic changes in transgenic plants in addition to the traits that the authors described. If yes, please describe them.
12. Please merge figure 4 d, f, h, and also merge Figure 4 e, g, i to better compare their effects.
13. Please describe or articulate why those traits have been selected over soybean domestication.

Reviewer #3:

Remarks to the Author:

Remarks to the Author

The authors cloned multiple quantitative trait loci (QTL) underlying different domestication-related traits (DRTs) by the recombinant inbred population from a cross between cultivated soybean and wild soybean. Intriguingly, they nicely showed that three QTLs for DRTs and the resistance to leafhoppers were also delimited in the same 29kb physical region using a genome-wide association study (GWAS). This region includes two tandemly duplicated lncRNA genes, lncRG1, and lncRG2, which produce abundant sRNAs primarily targeting three closely related MYB genes, thereby regulating the DRTs, such as pubescence form, plant height, and leaf size.

On the whole, I felt the experiments are technically solid, and the manuscript is generally well written with good data presentation, especially for genetic evidence. It will be perfect if more molecular data could be included.

1. The lncRNAs and sRNAs may have distinct mechanisms in regulating targeting gene expression. Although sRNAs were generated in the lncRNA region, the authors didn't clearly demonstrate which one is more important in manipulating the detected domestication-related traits.
2. Did authors check the MYB genes whether they have transcript activity? What kind of genes were regulated by these three MYB genes? Are there any genes associated with domestication-related traits? This may help the author figure out how to explain the DRTs and the leafhopper's resistance.
3. In Figures 2b and 2c, what kind of sequence (amino acid or nucleotides) were used for analysis? Please explain and describe in detail.
4. For the sRNA sequencing data, the sRNAs showed significantly different, if possible, please confirm sRNA expression by northern blot and testify sRNAs (miRNA or siRNA) production in this region by in vitro or in vivo assay.
5. In Figures 3b, c and 3e, f, the phenotypic images of transgenic plants should be included in extended data. Also, Figure 4d-4i.

6. As mentioned above, whether MYB genes are transcript factors or not? If yes, Figures 4J and 4K are not correct, please replace them with other assays but not Y2H. Also including Extended Data Fig. 3d, it doesn't make sense that two transcript factors are used in the Y2H assay.

7. Line 172-173 there is no data is support to this conclusion, please add it.

**Author Rebuttal to Initial comments**

**RESPONSES TO REVIEWERS' COMMENTS**

**Response Table 1.** Summary of the changes made in/though figures and tables in the revision.

| Categories            | Original submission      | Revised version                                                            | Notes                                     |
|-----------------------|--------------------------|----------------------------------------------------------------------------|-------------------------------------------|
| Main Figures          | Main Figs. 1-5           | Main Figs. 1-5                                                             | Same                                      |
| Extended Data Figures | Extended Data Fig. 1a-1h | Extended Data Fig. 1a-1h<br>Extended Data Fig. 1i                          | Same<br>Newly added                       |
|                       | Extended Data Fig. 2a-2f | Extended Data Fig. 2a-2f<br>Extended Data Fig. 2g-2j                       | Same<br>Newly added                       |
|                       | Extended Data Fig. 3a-3c | Extended Data Fig. 3a-3c<br>Extended Data Fig. 3d                          | Same<br>Newly added                       |
|                       | Extended Data Fig. 3d    | Extended Data Fig. 3e<br>Extended Data Fig. 3f-3k                          | Re-numbered<br>Newly added                |
|                       | Extended Data Fig. 4     | Extended Data Fig. 4a-e                                                    | Newly added                               |
|                       |                          | Extended Data Fig. 5<br>Extended Data Fig. 6a-c                            | Re-numbered<br>Newly added                |
| Supplementary Tables  |                          | Supplementary Table 1                                                      | Newly added                               |
|                       |                          | Supplementary Table 2                                                      | Newly added                               |
|                       |                          | Supplementary Table 3                                                      | Newly added                               |
|                       |                          | Supplementary Table 4                                                      | Newly added                               |
|                       |                          | Supplementary Table 5                                                      | Newly added                               |
|                       |                          | Supplementary Table 6                                                      | Newly added                               |
|                       |                          | Supplementary Table 7                                                      | Newly added                               |
|                       | Supplementary Table 1    | Supplementary Table 8                                                      | Re-numbered                               |
|                       | Supplementary Table 2    | Supplementary Table 9                                                      | Re-numbered                               |
|                       | Supplementary Table 3    | Supplementary Table 10<br>Supplementary Table 11<br>Supplementary Table 12 | Re-numbered<br>Newly added<br>Newly added |
|                       | Supplementary Table 4    | Supplementary Table 13                                                     | Re-numbered                               |
|                       | Supplementary Table 5    | Supplementary Table 14                                                     | Re-numbered                               |

|                      |                                                |                                                |                            |
|----------------------|------------------------------------------------|------------------------------------------------|----------------------------|
|                      | Supplementary Table 6                          | Supplementary Table 15                         | Re-numbered                |
| Supplementary Movies | Supplementary Movie 1<br>Supplementary Movie 2 | Supplementary Movie 1<br>Supplementary Movie 2 | Re-numbered<br>Re-numbered |

**Reviewer #1 (Remarks to the Author):**

This is a very interesting paper that I enjoyed reading. The authors identified two inverted repeated elements associated with domestication-related traits in soybean. Basically, the authors showed that these two IRs, named *lncRG1* and 2 produce abundant small RNAs in wild soybean, silencing three MYB transcription factors (TFs). Conversely, in domesticated soybean, the expression of both IRs is reduced, releasing the repression of the TFs and leading to characteristic phenotypic traits. The genetics and the experiments with transgenic wild soybean are well-conducted and convincing. Overall, I'm positive about this article. Still, there are a few mechanistic aspects that would benefit from some additional analysis.

1. Lines 119-120: it is interesting that *lncRG3* and *lncRG4* are not associated with any trait. Is there any explanation? are these two copies permanently silenced or activated in wild and domesticated soybean accession?

**RESPONSE #1.1:** To address the reviewer's question, we have investigated the expression levels of *lncRG3* and *lncRG4* and their capabilities to produce sRNAs. *lncRG4* exhibited an exceptionally low expression level in both Williams 82 (Wm82) and PI 479752 (**Supplementary Table 6** and **Extended Data Fig. 4b**), and no sRNAs were found to be associated with *lncRG4* in either Wm82 or PI 479752 (**Extended Data Fig. 4c**), suggesting that *lncRG4* has been silenced in both lines. *lncRG3* exhibited a relatively higher level of expression, but its expression was significantly higher in Wm82 than in PI 479752 (**Supplementary Table 6** and **Extended Data Fig. 4b**). This contrasts with the expression patterns of *lncRG1* and *lncRG2* detected between the two parents. Additionally, the sRNAs mapped to *lncRG3* were very low in Wm82 and almost none in PI 479752 (**Extended Data Fig. 4d**). Specially, the highest sRNA peak mapped to *lncRG3* is less than 1/10 of the peak height seen on *lncRG1* and even less than 1/50 of the peak height on *lncRG2* (**Fig. 2h, 2i** and **Extended Data Fig. 4d**). Furthermore, we have measured the level of nucleotide diversities within the IRs of the *lncRGs* using the cultivated and wild soybean pan-genome data (Liu et al., 2020) and found that the IRs of *lncRG3* and *lncRG4* have evolved at significantly higher rates than those of *lncRG2* and *lncRG1*, respectively (**Extended Data Fig. 4e**). These data potentially explain why *lncRG3* and *lncRG4* are not associated with these domestication traits. Please also see descriptions in lines 129-131; 252-263.

2. One missing piece of this story is why *lncRG1* and *lncRG2* are less expressed in domesticated soybeans. Considering that these IRs are likely to trigger TGS, see my following comments, I suggest analyzing the DNA methylation status of both *lncRNAs* in wild and domesticated plants. Perhaps a difference in the RdDM machinery between accession impact differentially the

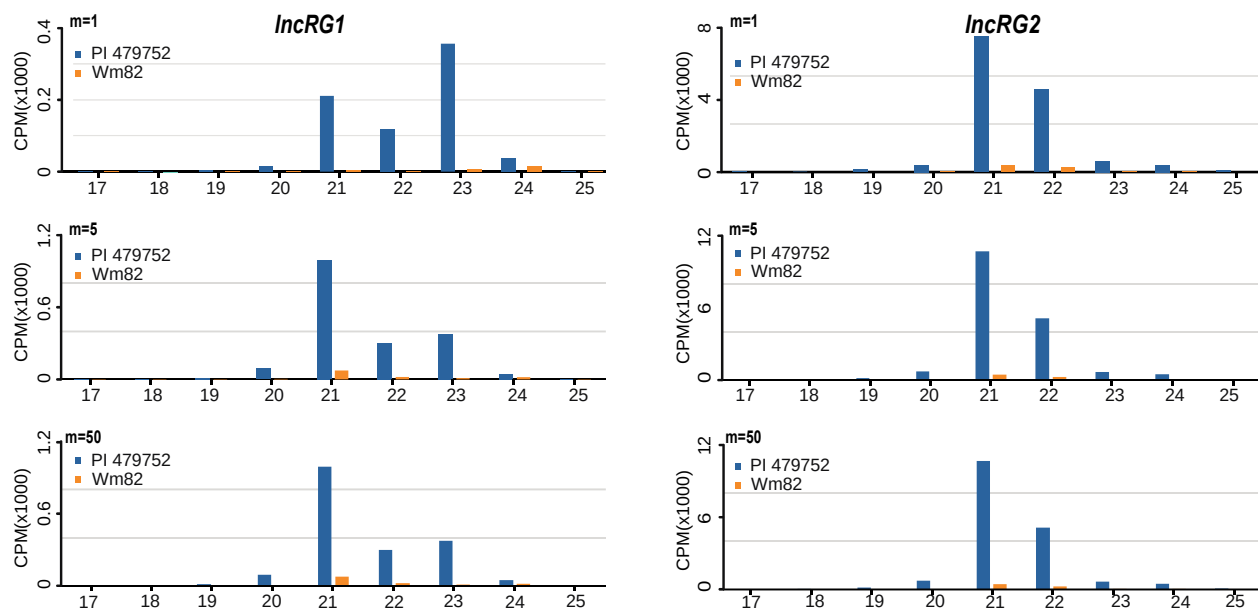
epigenetic landscape of these loci, turning these IRs into trans Obligatory epialleles, which could be an amazing discovery.

**RESPONSE #1.2:** To address this comment, we have analyzed the whole genome bisulfite sequencing (WGBS) data from a panel of 45 diverse *G. soja* and *G. max* accessions from a previous study (Shen et al., 2018), with a specific focus on the fine-mapped *lncRG1* and *lncRG2* regions. The new analysis reveals that the CpG methylation levels in the promoter regions of both *lncRG1* and *lncRG2* in the *G. max* haplotypes are significantly higher than those in the *G. soja* haplotypes (**Extended Data Fig. 6a**). In addition, we observed significant negative correlations between the CpG methylation levels in the promoter regions of *lncRG1* and *lncRG2* and the expression levels of the two genes (**Extended Data Fig. 6b and 6c**). These observations support the reviewer's speculation that the elevated levels of CpG methylation in the promoter regions of *lncRG1* and *lncRG2* could be responsible for the reduced expression levels of the two genes in cultivated soybeans, and thus for the underlying changes of the DRTs. We appreciate the comment and have now integrated these data into the manuscript to provide more insights. However, we acknowledge that it remains to be determined whether this epigenetic difference is fully or partially responsible for the phenotypic differences modulated by the two genes. Please also see descriptions in lines 281-293; 371-373.

3. L129. The sRNA size profile over the two IRs is curious. given the large complementarity of the IR I would expect their transcript to be mainly processed by DCL3/DCL3-like enzymes to produce 24 nt sRNAs as previously shown (doi:/10.1038/nplants.2016.163; /10.1111/tpj.12545; /10.1186/s12864-022-08364-4; and /10.1073/pnas.1903131116). However, you detect none. This is strange. As a matter of fact, the most abundant specie at lncRG1 are 23 nt, also present in lncRG1, is 23 nt sRNAs, which according to Craig Pikaard papers are a subproduct of 24 nt biogenesis. Thus, the presence of 23 nt suggests there should be 24s, although it is true IR may not need RDR2, which activity is responsible for the 23 nt species. Did you intentionally exclude 24 nt sRNA from the analysis, or is there no 24 nt mapping there? Could this be an analysis artifact, perhaps by filtering out multimappers? Actually, the material and methods explanation of sRNA library preparation and analysis is very poor. Is the typical double peak 21/24 sRNA size profile visible in the entire dataset?

**RESPONSE #1.3:** We apologize for the lack of clarity in our original submission regarding the methods used in our analysis and have now included detailed descriptions in the Methods of our revised submission to help clarify our observations. However, we would like to emphasize that, as shown in **Fig. 2h** and **2i**, we did not exclude 24-nt sRNAs in our analysis and did detect 24-nt sRNAs from *lncRG1* and *lncRG2*, albeit in small proportions (5% and 3%, respectively). Yes, we only counted reads uniquely mapped to these two genes of respective parental lines. To ensure that 24-nt sRNAs were not unintentionally filtered out in our original analysis, we have reanalyzed the sRNA data by allowing multiple mapping sites using the option '-m' to 5 or 50. The abundances of 21-nt to 23-nt sRNAs exhibited slight variations, but the proportions of 24-nt

sRNAs remained consistent and small, as shown in the **Response Fig. 1** (next page), which was drawn in the same way as described in **Fig. 2h** and **2i**. Therefore, the low proportions of 24-nt sRNAs from *lncRG1* and *lncRG2* were not due to analysis artifacts but rather reflected the facts. We would also like to note that, among the sRNAs produced by *lncGR2*, which was expressed at a higher level (**Fig 2a, 2f, 2g, 2h, and 2i**) and which is evolutionarily more conserved than the other copy of the lncRGs in the mapped QTL region (**Fig. 5c**), 21-nt and 22-nt sRNAs were seen as the predominant forms, in all soybean accessions examined, with 5-15% sRNAs of 23-nt only (**Extended Data Fig. 5**). Regarding the comment from the reviewer about the IRs associated with transposon elements (TE), as the IRs in *lncRG1* and *lncRG2* were derived from coding sequences of ancient MYB genes, it is very likely that *lncRG1* and *lncRG2* represent the evolutionary intermediates of miRNA processors, similar to the IR-containing *SULF* gene in snapdragon, which also produced few 24-nt sRNAs (<https://doi.org/10.1126/science.aao3526>). Therefore, we speculate that the biogenesis of sRNAs from IRs in *lncRG1* and *lncRG2* may not require DCL3/DCL3-like enzymes that are used for biogenesis of 24-nt sRNAs from transposon-containing or transposon-derived IRs. Please also see descriptions in lines 144-146; 311-332



**Response Fig. 1.** Abundance of sRNAs in different sizes produced by *lncRG1* and *lncRG2*.

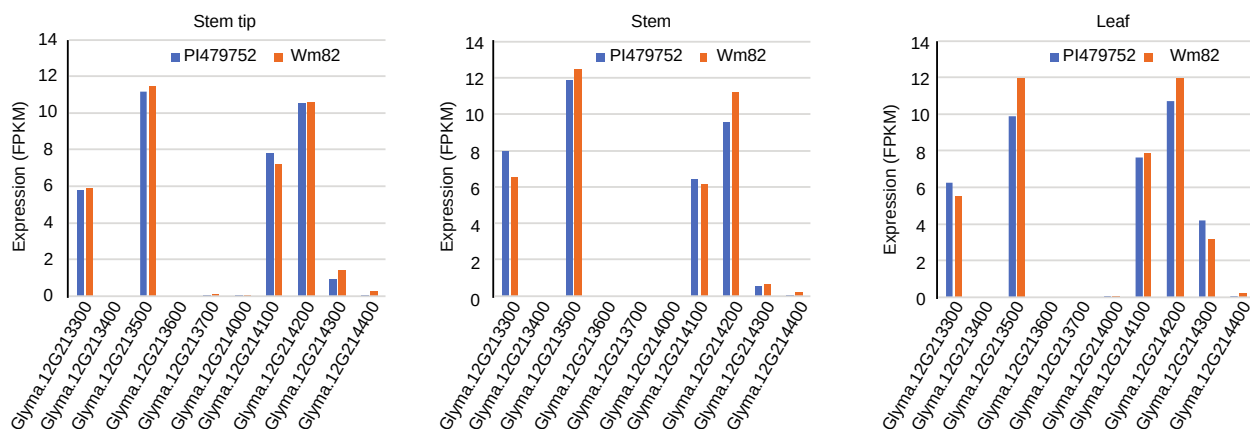
4. A recent paper showed that polymorphisms in IRs are associated with phenotypic traits ([doi.org/10.1016/j.celrep.2023.112029](https://doi.org/10.1016/j.celrep.2023.112029)). Although these are not polymorphic, it is possible they act as epialleles, something interesting to discuss. The same paper shows that IRs affect the expression of neighbor genes, this is something it was shown for many IRs

([doi.org/10.1093/molbev/msr282](https://doi.org/10.1093/molbev/msr282); /10.1016/j.celrep.2023.112029; /10.1093/plcell/koac220; /10.1111/nph.18068; /10.1016/j.molp.2020.02.009; 10.1073/pnas.1903131116).

This makes me wonder:

a. Have you checked the expression of the *lncRG1* and 2' adjacent genes that are conserved according to Figure 5A?

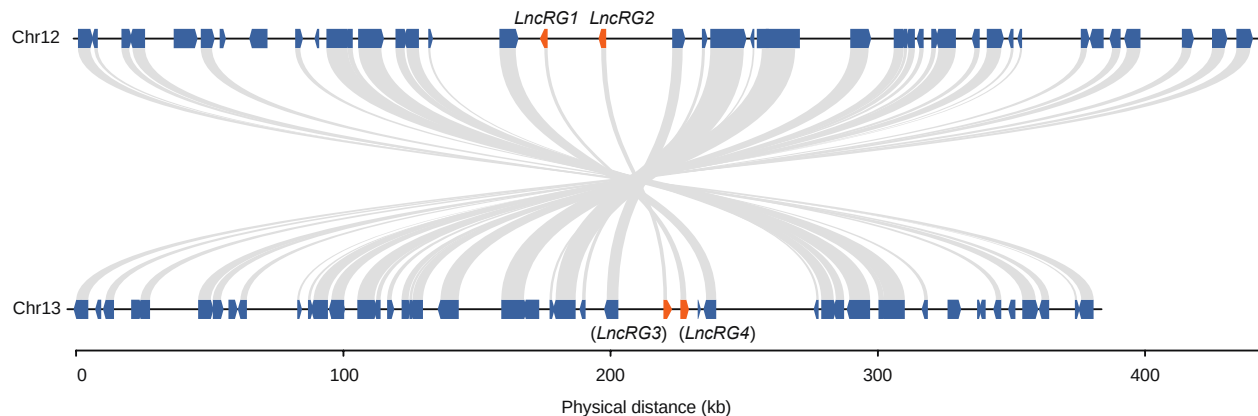
**RESPONSE #1.4a:** We have examined the expression levels of the genes adjacent to *lncRG1* and *lncRG2*, which include five genes on each side, in the stem tip, stem, and leaf tissues of both PI 479752 and Wm82, using our RNA-seq data. Our analysis reveals that none of these adjacent genes displayed significant differences between PI 479752 and Wm82 in any of these tissues, as illustrated in the **Response Fig. 2** (next page). Therefore, it appears that the adjacent genes are not affected by the IRs in *lncRG1* and *lncRG2*. As we explained in the **RESPONSE #1.3**, the IRs in *lncRG1* and *lncRG2* were derived from the coding regions of ancient MYB genes, potentially functioning through mechanisms distinct from those of transposon element (TE)-derived IRs in the papers referenced by the reviewer. We appreciate the reviewer's comment and sharing of relevant papers, which helped us to enhance our Discussion by including this point. Please also see descriptions in lines 311-332.



**Response Fig. 2.** Expression of genes flanking the *lncRG1* and *lncRG2* region in the two parents based on RNA-seq data. FPKM, fragments per kilobase of transcript per million mapped reads.

b. Is the genomic vicinity of *lncRG1* and 2 conserved at *lncRG3* and loci? Can this explain the lack of phenotype of L3 and 4 if L1 and 2 regulate in cis neighbor genes? In other words, could genes next to *lncRG1* and 2, but absent in the region encoding *lncRG3* and 4, help explain the phenotype?

**RESPONSE #1.4b:** We have compared the *lncRG1-lncRG2* region with the *lncRG3-lncRG4* region. As shown in **Response Fig. 3** below, these duplicated regions derived from the WGD event retain nearly perfect genic collinearity, so the lack of phenotypes of *lncRG3* and *lncRG4* was unlikely the cause of cis elements associated with their neighbor genes. Additionally, as described in **RESPONSE #1.4a**, the genes flanking *lncRG1* and *lncRG2* were unaffected by the IRs. Furthermore, as demonstrated in **RESPONSE #1.1**, the lack of a discernible phenotype in *lncRG3* and *lncRG4* appear to be either silencing or more rapid divergence between their IRs. Lastly, our genetic mapping work defined causal mutation(s) to the 29-kb fine-mapped *lncRG1* and *lncRG2* region, regardless of whether genetic or epigenetic variation within the region was causal for the reduced expression of *LncRG1* and *LncRG2*. Please also see descriptions in lines 252-263.



**Response Fig. 3.** Genic collinearity between the *lncRG1* and *lncRG2* region and the *lncRG3* and *lncRG4* region. Duplicated genes derived from the WGD event are linked with the grey curve lines.

In your paper, the two lncRNA seem to work as miRNAs mostly at PTGS level. However, long IRs were previously associated with TGS (see references mentioned before). It would be interesting to see the methylation patterns over the lncRNAs loci and targets. I'm sure there should be publicly available BS-seq datasets to consult.

**RESPONSE #1.5:** Thanks again for the suggestion. We think we have addressed this question by adding the suggested analysis as described in **RESPONSE #1.2**, and by discussing potential mechanisms for biogenesis of sRNAs from TE and IRs associated with genic sequences as described in **RESPONSE #1.3**.

The M&M is too poor, it needs to be expanded to include more details. A clear example is the mRNA, sRNA, and degradome session which gives almost no information about libraries preparation and analysis.

**RESPONSE #1.6:** Thank you for pointing this out. We have now included more detailed descriptions for the methods used in this study (Please see highlighted changes through the whole Methods section).

There is a very limited introduction/discussion of the current literature about gene regulation mediated by IRs.

**RESPONSE #1.7:** As responded above, particularly in **RESPONSE #1.3**, we thought that the biogenesis of the 24nt-lacking sRNAs from the IRs of *lncRG1* and *lncRG2* may not involve DCL3/DCL3-like enzymes that are generally needed for biogenesis of 24-nt sRNAs from transposon-containing or transposon-derived IRs, so we didn't include the literatures about gene regulation mediated by IRs associated with transposons. We have now included such literatures to the Discussion to strengthen the manuscript. Please see added descriptions in lines 311-332

#### **Reviewer #2 (Remarks to the Author):**

The study demonstrated that the recent birth of two long non-coding RNAs from MYB genes, their evolutionary pathway and their functionality in regulating several traits associated with domestication. The genes regulate the traits through small RNA pathway that suppress the transcript accumulation of the MYB genes by RNA degradation. Authors provide solid evidence supporting the conclusion. In addition, the linkage mapping and GWAS pinpoint the multiple QTLs for the traits into a 29 kb region where the two genes are located. The 29 kb region was also located in a sweeping region. The accumulation of the two lncRNAs, and their major small RNAs and their targeted MYB RNAs are different in soja and max accessions, two parental lines of the RIL population. The authors concluded that the QTLs 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. The study has a good originality and significance. The manuscript is well written. Overall, the data supports the conclusion. The author needs to provide more detailed and comprehensive description of data, methodology and statistical analysis. Introduction and Discussion could be improved to articulate the originality and significance related to the previous publication.

Suggested Overall Improvement.

1. Authors should provide more detailed and comprehensive description of the data and methods used and the results observed in the study. It is important at the situation that the causal allele was

not identified. Many results are based on association, gene expression (affected by environments and genetic backgrounds) and different populations.

**RESPONSE #2.1.** Thanks for the valuable suggestions. We have now expanded the description of the data and methods and results, with some details, as described in more details in **RESPONSES #2.1a, b, c and 2.3.**

a. Please provide the data resource, all phenotypic data, genotyping data, accession IDs and transcriptome data used in the study in supplementary files or make them available in the public database. A significant amount of data used in the study is from the public. It should be encouraged by our research community. However, please try to clearly state the data used in the study and if the data was validated.

**RESPONSE #2.1a:** We have now articulated or added descriptions of the data sources used in each analysis in newly added **Supplementary Tables 1-7, 11, and 12**, in addition to citation of specific references.

b. Please present results from the linkage mapping and GWAS for all domestication-related traits that the two genes regulate, and their phenotypes in the over-expression lines. Not all of the stated traits were presented in each analysis (Linkage mapping, GWAS using resequencing data, and transgenic analysis), for example, leafhopper resistance in transgenic plants. If there is no consistence in mapping and transgenic study, please remove the traits or state the results.

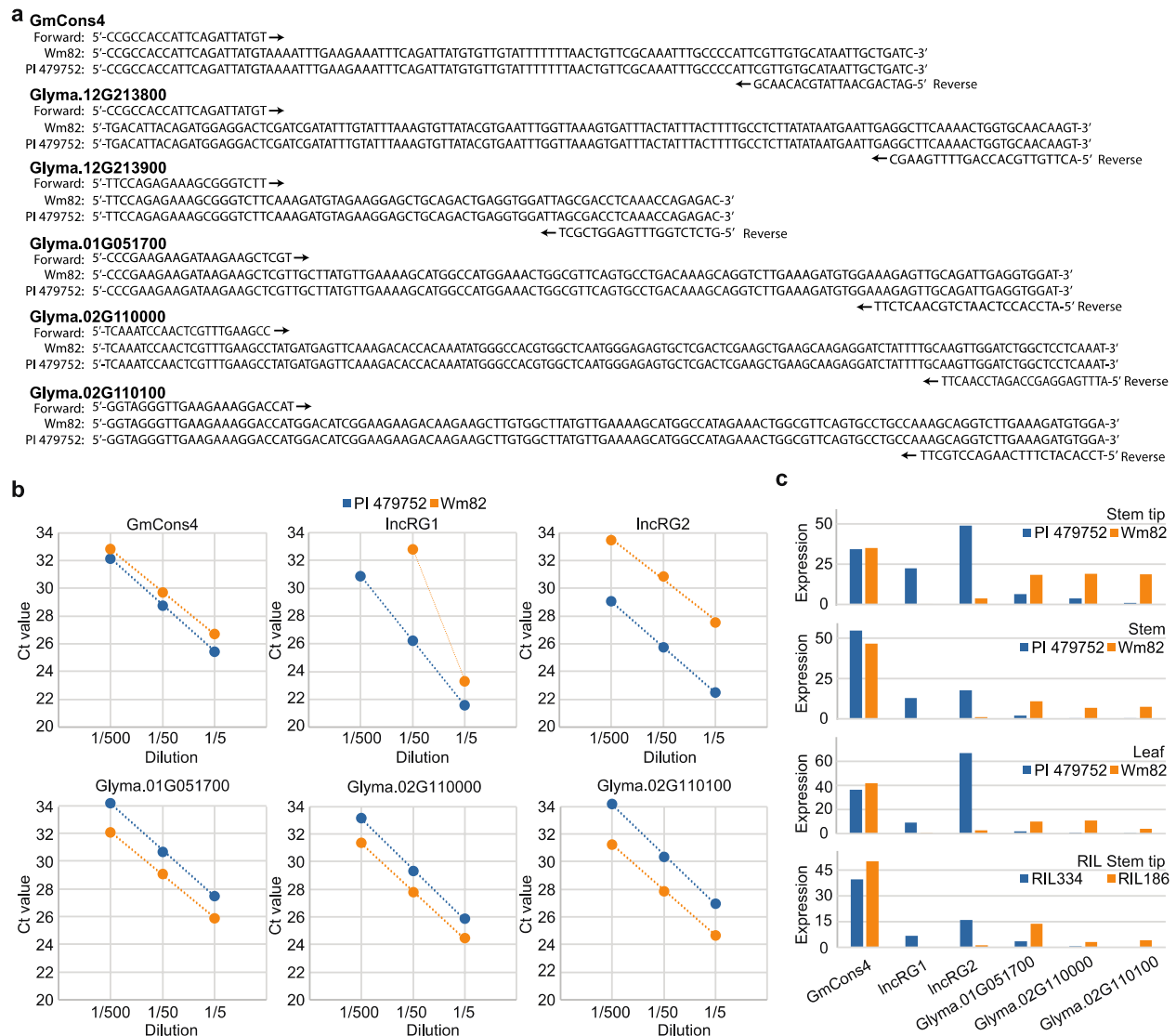
**RESPONSE #2.1b:** As previously described and cited, the initial mapping of these QTLs was published in **Reference #4**. Therefore, this manuscript didn't include linkage mapping but instead focused on fine mapping with key recombinants identified using two insertion/deletion (InDel) markers, M1 and M10, that initially defined the boundaries of the qDRT12.3 region. We have now added a supplemental table (**Supplementary Table 1**) to show the genotypes and phenotypes, which pinpointed the three major QTLs for the three traits to the 29-kb region, as shown in **Fig. 1f-1h**. We have also articulated that only these three major QTLs were defined into this 29-kb region by fine mapping. GWAS didn't have the power to define any of the QTLs to this 29-kb region, so it was not used for analysis of these already fine-mapped QTLs. However, GWAS was performed for the pubescence form and leafhopper resistance traits to demonstrate that there was only a single, overlapped, major QTL for each of these two traits (**Extended Data Fig. 1c and 1d**). Therefore, it was reasonable to deduce that the previously reported "leafhopper resistance attributed by erect pubescence" (**Reference #6**) was controlled by *IncRG1* and *IncRG2*. This deduction was further validated by phenotyping the double mutants of two major MYB target genes created by gene editing, as exemplified in **Movie S2**. We acknowledge that the purposes of these different types of analyses for specific traits were not clearly articulated in the original submission. Therefore, we have improved the descriptions of these analyses for clarity in the revised manuscript. Restructured descriptions in lines 86-95 exemplify such an improvement.

c. Please provide more details for the material and methods and figure legend such as specific population used in each analysis. Please describe how to determine the phenotypes for the traits. Since many results were from different populations and based on association. For examples, Extended Data Fig 1 c, d, e, f, g, and h seem that different populations were used.

**RESPONSE #2.1c:** We have expanded the descriptions of the materials and methods, as well as the phenotypic data used in each analysis in the Methods and Results sections or figure legends for clarity.

2. Considering that soja and max have sequence difference of whole genomes, two lncRNAs, Myb genes. RNA-seq from soja and max should be aligned to their own genomes for determining gene expression, not only on Wm82. The primers of the QRT-PCRs should be validated to amplify the targeted RNAs at the same efficiency to determine their expression in soja and max.

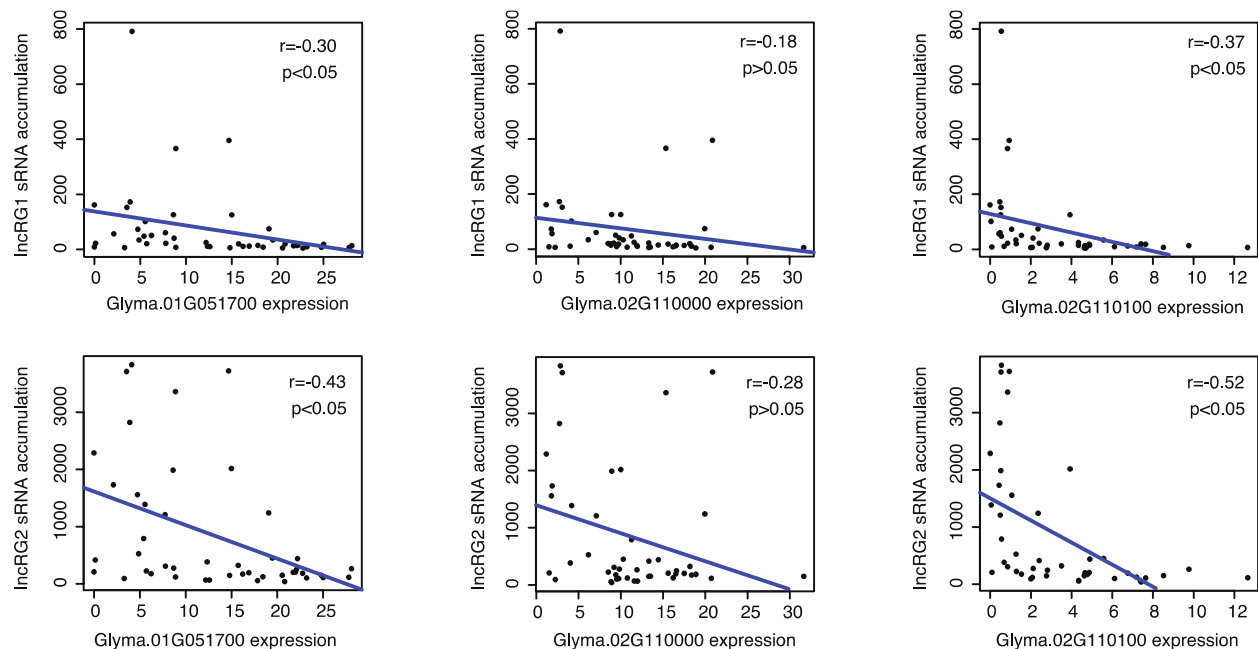
**RESPONSE #2.2:** Thanks for the comment. To enhance the mapping quality of the non-reference accession, we created a “SNP-corrected” reference for PI 479752 using genome re-sequencing data. We then aligned all corresponding sequencing to this “corrected” genome to determine gene expression levels. To ensure the desirable efficiency of our qRT-PCR primers, we first confirmed the absence of polymorphisms in the primers between Wm82 and PI 479752, as shown in **Response Fig. 4a** (next page). We then evaluated their amplification efficiency by generating three cDNA dilution series (1/5, 1/50, 1/500). As illustrated in **Response Fig. 4b**, all primers consistently demonstrated efficiency in amplifying their respective genes in both Wm82 and PI 479752, with the exception of the primers for evaluating *lncRG1* expression in Wm82. We attribute this anomaly to its extremely low level of expression in Wm82 (FPKM=0.2), which may have caused amplification failure with further dilution. To further address this issue, we have examined the expression levels of *lncRG1*, *lncRG2*, and the three target MYB genes based on our RNA-seq data in stem tips, stems, and leaves of both Wm82 and PI 479752. Additionally, we have assessed their expression in stem tips of a pair of RILs from the mapping population, RIL186 (*qrt12.3*) and RIL334 (*qDRT12.3*), as shown in **Response Fig. 4c**. The expression patterns of these genes revealed by qRT-PCR and RNA-seq have been found to be consistent. We have now included the gene expression data from RNA-seq in the newly added **Supplementary Table 6** to support our observations on gene expression.



**Response Fig. 4.** Gene and primer sequences (a) and expression analyses by qRT-PCR (b) and RNA-seq (c). Fragments per kilobase of transcript per million mapped reads (FPKM) was used to normalize the expression levels (*GmCons4* is a house-keeping gene routinely used for qRT-PCR in soybean research).

3.Optional: Determine the accumulation of the lncRNAs and their targeted Myb transcripts in the collection of accessions used in Extended Figure 1g, determine their accumulation correlation of lncRNAs vs targeted RNAs in the accessions such as PCC, and conduct TWAS for the traits.

**RESPONSE #2.3:** Thanks for the suggestions. We have determined the accumulation of *lncRG1*- and *lncRG2*-produced sRNAs and their targeted Myb transcripts in the accessions used in **Extended Fig. 1g**, and then computed the Pearson correlation coefficient between the sRNAs and the three target genes. As expected, and as shown in **Response Fig. 5** (next page), we observed significant negative correlations between accumulation of sRNAs and the expression levels of their respective target genes Glyma01G051700 and Glyma02G110100. Such a correlation between accumulation of sRNAs and the expression level of their target gene Glyma.02G110000 was not statistically significant. This could be partly due to the varied distribution patterns of sRNAs derived from *lncRG1* and *lncRG2* within this population, as shown in **Extended Data Fig. 5**, and partly due to the relatively small size (45 accessions) of the population. Due to the small size of the population, we didn't attempt to contact the research lab in China for phenotyping their accessions, which are not available in the USDA soybean germplasm collection. As a result, we did not perform Transcriptome-Wide Association Studies (TWAS) for the traits. We fully agree that such an analysis would help to further understand the functionality of the two *lncRGs*.



**Response Fig. 5.** Pearson correlations between *lncRG1*- and *lncRG2*-produced sRNAs and their respective MYB targets, Glyma01G051700, Glyma02G110100, and Glyma.02G110000.

Suggested Specific Improvement

4. Figure 1 a, b, c, d: please mark the specific lines instead of just soja and max.

**RESPONSE #2.4:** Thanks. We have marked the specific lines (Wm82 and PI 479752) in Figure 1a-1d as suggested.

5. It is not clear for Extended Data Figure 1a and 1b: It is likely a mistake. Figure legend states that “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 log<sub>10</sub> of the P-values and the red color highlight markers within the final mapping region” However, in the Result (Line 71-74) “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 (Extended Data Fig. 1a and 1b).”

**RESPONSE #2.5:** We would like to express our sincere apologies for the mistake made in the legend of **Extended Data Fig. 1a** and **1b**. These two panels depict the association between the genetic variations and the pubescence phenotypes. The mistake has been corrected.

6. Extended Figure 1f: the bars should overlap based on figure legend. Why are they distinct bars without overlapping.

**RESPONSE #2.6:** In **Extended Fig. 1f**, we calculated the ratios of nucleotide diversity for each sliding window, with the bars representing the values at the midpoint of each window. Thus, individual bars didn't exhibit overlaps along the investigated region. We have now expanded the figure legend for clarity.

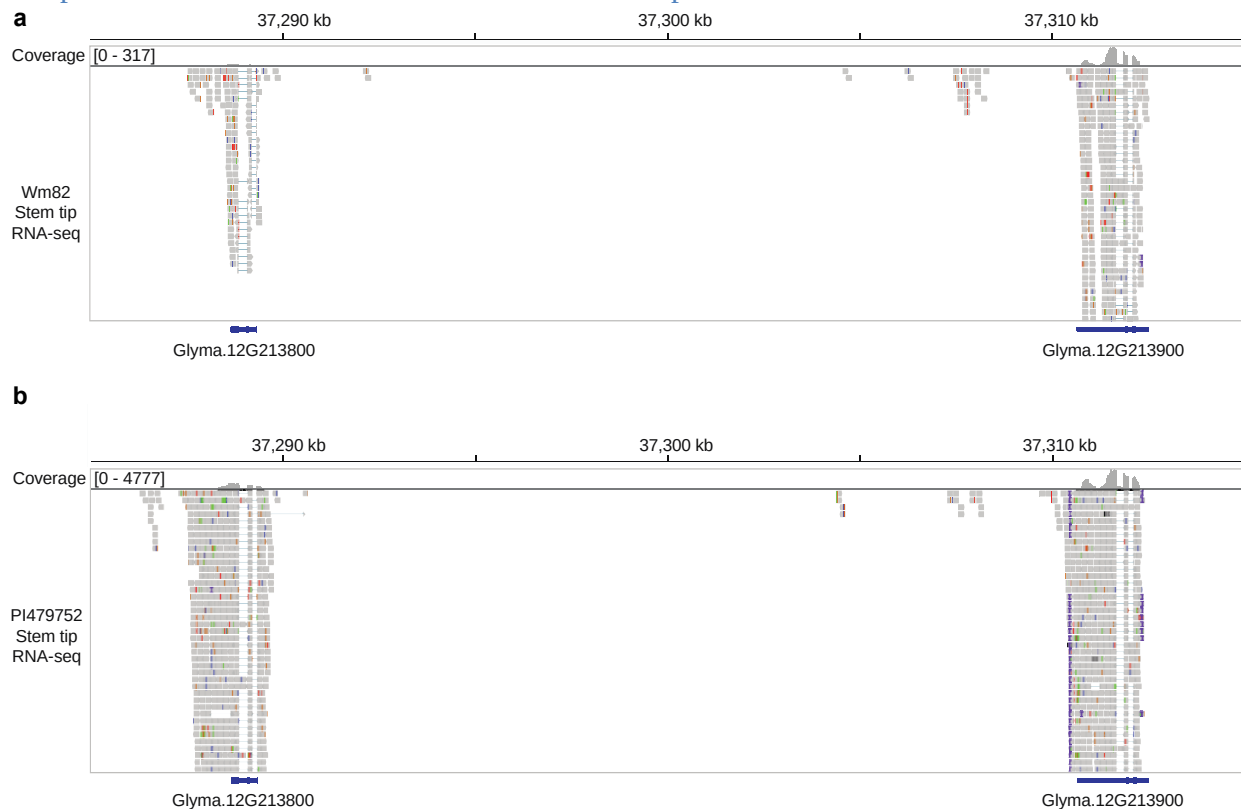
7. Mark the tissues and accessions used for RNA seq and accessions in Ex Figure 1g and 1h. Are they the same tissue?

**RESPONSE #2.7:** All the RNA-seq data used in **Extended Figure 1g** and **1h** are from the same tissue (stem tips) at the same developmental stage (first trifoliate expanded). We have now mentioned this in the legend and provided the list of accessions used for these two panels in the newly added supplementary table, **Supplementary Table 7**.

8. P79-80: “According to the Wm82 reference genome, this region harbors only two genes, Glyma.12G213800 and Glyma.12G213900, both lncRNAs”.

a. Please align all RNA reads of RNA-seq from Wm82 and PI479752 back to genome sequences of Wm82 and PI 479752 respectively to exclude the possibility that there are other genes in the region.

**RESPONSE #2.8:** As suggested, we have aligned the RNA-seq data from Wm82 and PI 479752 back to their respective genomes. As shown in **Response Fig. 6** (next page), the vast majority of reads from either Wm82 or PI 479752 were aligned to *lncRG1* and *lncRG2* in respective genomes. As expected, the relative abundance of the aligned reads in Wm82 was substantially lower than in PI 479752, which is consistent to the results shown in **Extended Data Fig. 1g**. Small, similar proportions of reads were aligned to small pieces of intergenic region between *lncRG1* and *lncRG2* in Wm82 and PI 479752, but these pieces were not annotated as genes and didn't match any proteins based on BLASTX search against the protein database in NCBI. Therefore, there is no evidence to support the existence of additional genes in the region. It is unlikely that such small pieces of sequences with few RNA-seq reads contribute to the phenotypic differences between the two parental lines. Please see modified/added descriptions in lines 245-247



**Response Fig. 6.** Alignments of RNA-seq reads from stem tips of Wm82 (a) and PI 479752 (b) to the fine-mapped QTL regions of respective genomes.

9. Figure 2a and 2j, please provide evidence that qRT-PCR amplified the genes in soja and max at the same efficiency. Including the sequence alignments of the primers against their targeted gene sequences of PI 479752 and Wm82.

**RESPONSE #2.9:** We have addressed this question in **RESPONSE #2.2**.

10. Please examine if number of RI gene (2 vs 1 genes) varies in soja accessions using available re-sequencing data. It will help to determine if duplication of the lncRNA genes occurred after soybean speciation and before domestication.

**RESPONSE #2.10:** We have carefully examined the corresponding genomic regions of three wild soybean accessions, nine landraces and 14 cultivars used for construction of the soybean pan-genome (Liu et al. 2020) and found that the genes in the corresponding regions are completely shared among these accessions without addition or deletion of any genes, as illustrated in the newly added **Extended Data Fig. 4a**. This piece of data, together with the results shown in **Fig. 5a and 5b**, suggest that duplication of the lncRNA genes occurred after soybean speciation and before domestication. We have integrated this point in the Results. Please see lines 126-130; 245-247; 252-263.

11. Please describe if other phenotypic changes in transgenic plants in addition to the traits that the authors described. If yes, please describe them.

**RESPONSE #2.11:** No noticeable other phenotypic changes were observed in the transgenic plants.

12. Please merge figure 4 d, f, h, and also merge Figure 4 e, g, i to better compare their effects.

**RESPONSE #2.12:** Considering that the phenotypes of the single, double, and triple mutants were assessed in different years, and each genotype was compared to its respective control line grown in the same environment, we prefer to keep them as separate panels. However, if the reviewer still wants us to merge them, we will be happy to do that in the next round.

13. Please describe or articulate why those traits have been selected over soybean domestication.

**RESPONSE #2.13:** We intended not to speculate too much, but it appears that erected pubescence for leafhopper resistance and larger leaflets were the major targets for selection over domestication. We have added this point in the Discussion. Please see lines 359-362.

**Reviewer #3 (Remarks to the Author):**

The authors cloned multiple quantitative trait loci (QTL) underlying different domestication-related traits (DRTs) by the recombinant inbred population from a cross between cultivated soybean and wild soybean. Intriguingly, they nicely showed that three QTLs for DRTs and the resistance to leafhoppers were also delimited in the same 29kb physical region using a genome-wide association study (GWAS). This region includes two tandemly duplicated lncRNA genes, *lncRG1*, and *lncRG2*, which produce abundant sRNAs primarily targeting three closely related MYB genes, thereby regulating the DRTs, such as pubescence form, plant height, and leaf size. On the whole, I felt the experiments are technically solid, and the manuscript is generally well written with good data presentation, especially for genetic evidence. It will be perfect if more molecular data could be included.

1. The lncRNAs and sRNAs may have distinct mechanisms in regulating targeting gene expression. Although sRNAs were generated in the lncRNA region, the authors didn't clearly demonstrate which one is more important in manipulating the detected domestication-related traits.

**RESPONSE #3.1:** Thanks for the comment. *lncRG1* and *lncRG2* are structurally different from classic lncRNA genes, which generally do not possess inverted repeats (IRs). As we mentioned in the **RESPONSE #1.3** to the first reviewer, “the IRs in *lncRG1* and *lncRG2* were derived from coding sequences of ancient MYB genes and it is very likely that *lncRG1* and *lncRG2* represent the evolutionary intermediates of miRNA processors, similar to the IR-containing *SULF* gene in snapdragon (<https://doi.org/10.1126/science.aao3526>).” In both our study on *lncRG1* and *lncRG2* and the *SULF* study, IR-derived sRNAs were validated to determine the traits. Specifically, when a key sRNA targeting the MYB gene was overexpressed in Wm82, the wild soybean (PI 479752)-type of pubescence form was observed (**Fig 3d**). Similarly, when a MYB target was knocked out in Wm82, the PI 479752-type of pubescence form was also observed (**Fig. 4a**). More notably, the double mutants and triple mutant m1,2,3 in the Wm82 background exhibited nearly identical pubescence form as PI 479752, suggesting the sRNA-mediated posttranscriptional regulation of their MYB targets is the key determinant for the domestication-related traits. We have added this point in the Discussion section. Please see lines 329-332.

2. Did authors check the MYB genes whether they have transcript activity? What kind of genes were regulated by these three MYB genes? Are there any genes associated with domestication-related traits? This may help the author figure out how to explain the DRTs and the leafhopper's resistance.

**RESPONSE #3.2:** We have now demonstrated that the three MYB genes function as transcription factors that bind to the promoter regions of many genes through Chromatin Immunoprecipitation Sequencing (ChIP-seq) analysis. We generated data for this analysis right after submitting the manuscript and have now included it in the revised version (see newly added **Extended Data Fig. 4f-k**). Briefly, we fused the coding sequences of Glyma.01G051700 and Glyma.02G110000, two of the three MYB targets with major effects on the domestication traits, with the FLAG epitope

sequence, and then generated Wm82 transgenic lines that overexpressed these fused proteins under the control of the 35S promoter, separately. We then conducted ChIP-seq using stem tips collected from the transgenic plants that overexpressed Glyma.01G051700-FLAG and Glyma.02G110000-FLAG, respectively. As any two of the three MYB proteins directly interact, we thought that ChIP-seq for the third protein was unnecessary. In total, 8,167 genes were detected in both ChIP-seq assays. We further conducted functional annotation of these 8,167 genes putatively targeted by the MYB proteins using gene ontology (GO) analysis. Our analysis revealed significant enrichment of genes associated with photosystem/photosynthesis and the auxin-activated signaling pathway for plant growth (**Extended Data Fig. 4k**). This finding may explain some of the domestication traits observed in cultivated soybeans such as larger leaves. To date, only a few genes underlying soybean domestication traits have been reported, and they were not regulated by these MYB proteins according to the ChIP-seq data. We'd also like to note that the leafhopper resistance in cultivated soybeans is attributed to erect pubescence, which serves as a physical mechanism that makes insect climbing more difficult. We have elaborated this mechanism in **RESPONSE #2.1b**. Please also see modified/added descriptions in lines 208-211; 221-234; 359-362.

3. In Figures 2b and 2c, what kind of sequence (amino acid or nucleotides) were used for analysis? Please explain and describe in detail.

**RESPONSE #3.3:** Shared nucleotide sequences corresponding to transcripts were used for the phylogenetic tree construction in **Fig. 2b** and the full gene sequences were used for alignment and comparison in **Fig. 2c**. We have added this information in the figure legend and the Method section as well.

4. For the sRNA sequencing data, the sRNAs showed significantly different, if possible, please confirm sRNA expression by northern blot and testify sRNAs (miRNA or miRNA) production in this region by in vitro or in vivo assay.

**RESPONSE #3.4:** Thanks for the comment. We agree that northern blot is an alternative approach for measuring and comparing sRNA abundance. However, we feel this approach may not be necessary for two facts: Firstly, our results show that small RNAs along the inverted repeats of the *lncRG1* and *lncRG2* transcripts vary in abundance among varieties, as demonstrated in **Fig. 5d** and **5e**. Therefore, we believe that sRNA-seq is an efficient approach for measuring relative sRNA abundances, especially when extreme differences are observed between compared lines, as shown in **Fig. 2f** and **2g**. Secondly, we found significant differences in the abundance of *lncRG1*- and *lncRG2*-produced sRNAs between three additional *G. soja* accessions, and 23 *G. max* accessions used for constructing the soybean pan-genome, as reported by Liu et al. in 2020. These repeatable observations based on sRNA-seq from multiple lines in different experiments suggest that the differences in sRNA abundance between the two parental lines PI 479752 (*G. soja*) and Wm82

(*G. max*) that we detected are reliable. We hope that our explanations are satisfactory to the reviewer.

5. In Figures 3b, c and 3e, f, the phenotypic images of transgenic plants should be included in extended data. Also, Figure 4d-4i.

**RESPONSE #3.5:** Thanks for the suggestion. We have now incorporated phenotypic images of the transgenic plants that overexpressed the entire stem loops of *lncRG1* and *lncRG2* in the newly added **Extended Data Fig. 2g** and **2h**. However, we decide not to include phenotypic images for Fig. 3e, 3f, and 4d-4i. Although significant differences were statistically detected in the compared lines, the phenotypic changes observed in individual transgenic plants, which either overexpressed a single sRNA or contained a mutation in a single target gene, were not visually striking when compared to respective control lines. Picking up some plants from a specific line showing a bigger difference than other plants compared with the control lines would not be appropriate. Therefore, we prefer to show the distribution pattern of these quantitative traits from all examined plants of each line as illustrated in the original figure panels.

6. As mentioned above, whether MYB genes are transcript factors or not? If yes, Figures 4J and 4K are not correct, please replace them with other assays but not Y2H. Also including Extended Data Fig. 3d, it doesn't make sense that two transcript factors are used in the Y2H assay.

**RESPONSE #3.6:** As described in **RESPONSE #3.2**, the MYB genes function as transcription factors. Prior to assessing potential interactions among these MYB genes using the Y2H system, we conducted a pre-test to determine if any of these MYB genes activated the reporter gene independently. The results of this pre-test are shown in the newly added **Extended Data Fig. 3d**, which demonstrates that none of the MYB genes activated the reporter gene independently. Therefore, the Y2H system is suitable for detecting interactions among these MYB transcription factors. We apologize for not including these pre-test results in our original submission and have now added them. It is worth noting that the Y2H system has been widely used to detect protein-protein interactions involving transcription factors under similar conditions, as exemplified in various studies. Here, we'd cite two of recent publications (<https://doi.org/10.1038/s41588-022-01240-7>; <https://doi.org/10.1038/s41467-022-34153-4>).

7. Line 172-173 there is no data is support to this conclusion, please add it.

**RESPONSE #3.7:** We have now included phenotypic images in newly added **Extended Data Fig. 2i** and **2j** to support this conclusion.

Decision Letter, first revision:

21st Nov 2023

Dear Professor Ma,

Your Article, "Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance" has now been seen by 3 referees. You will see from their comments below that while they find your work of interest, some important points are raised. We are interested in the possibility of publishing your study in Nature Genetics, but would like to consider your response to these concerns in the form of a revised manuscript before we make a final decision on publication.

We therefore invite you to revise your manuscript taking into account all reviewer comments. Please highlight all changes in the manuscript text file. At this stage we will need you to upload a copy of the manuscript in MS Word .docx or similar editable format.

We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

When revising your manuscript:

\*1) Include a "Response to referees" document detailing, point-by-point, how you addressed each referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be sent back to the referees along with the revised manuscript.

\*2) If you have not done so already please begin to revise your manuscript so that it conforms to our Article format instructions, available [here](#).

Refer also to any guidelines provided in this letter.

\*3) Include a revised version of any required Reporting Summary:

<https://www.nature.com/documents/nr-reporting-summary.pdf>

It will be available to referees (and, potentially, statisticians) to aid in their evaluation if the manuscript goes back for peer review.

A revised checklist is essential for re-review of the paper.

Please be aware of our [guidelines on digital image standards](#).

Please use the link below to submit your revised manuscript and related files:

[redacted]

**Note:** This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised manuscript within four to eight weeks. If you cannot send it within this time, please let us know.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Nature Genetics is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as 'corresponding author' on published papers create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System (MTS), prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the MTS by clicking on 'Modify my Springer Nature account'. For more information please visit [www.springernature.com/orcid](http://www.springernature.com/orcid).

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Sincerely,  
Wei

Wei Li, PhD  
Senior Editor  
Nature Genetics  
New York, NY 10004, USA  
[www.nature.com/ng](http://www.nature.com/ng)

#### Reviewers' Comments:

##### Reviewer #1:

##### Remarks to the Author:

Thank you for answering my questions, and congratulations on the paper. Overall, all my concerns were addressed. I only would like to include two minor changes.

-Please state in extended figure 6a whether the observed methylation is in the CG, CHG, or CHH context. I suggest either color-coding each context in the same plots or using individual plots for each context.

-The discussion between lines 308-321 leaves some room for improvement. Here is a suggestion to replace the same paragraph:

IRs integration in the genome can result from processes like DNA replication repair or transposable elements (TEs) [15]. IRs derived from TEs are typically processed by enzymes such as DCL3/DCL3-like, resulting in the production of 24-nt small RNAs [16,17]. These small RNAs often have a noticeable impact on the expression of nearby genes and phenotypic traits [17,18] (beside 17 and 18, please add references: 10.1016/j.celrep.2023.112029; 10.1111/nph.18068; 10.1093/plcell/koac220; 10.1016/j.molp.2020.02.009). In contrast, the IRs found in lncRG1 and lncRG2 predominantly give rise to 21-23-nt small RNAs (Fig. 2h and 2i). It's interesting to note that lncRG1 generates 23-nt

sRNAs, while lncRG2 primarily yields 21-22-nt sRNAs (Fig. 2h and 2i). In concordance, lncRG1, which produces 23-nt sRNAs, displays signs of gene body methylation (see Ext. Data Fig. 6a), as reported for other IRs. This suggests that in this instance, 23-nt sRNAs 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 IRs (10.1016/j.celrep.2023.112029), we didn't observe differential gene expression in the vicinity of lncRG1 and lncRG2 when comparing Wm82 and PI according to the RNA-seq data (Figure X?). This further supports the idea that lncRG1 and lncRG2 operate through a distinct mechanism. In this context, the 21-nt and 22-nt sRNAs generated from the conserved lncRG2, which account for over 80-89% of the total sRNAs produced in these loci in *G. soja* accessions, are likely the major contributors to gene regulation. Consequently, the mechanism responsible for processing IRs in lncRG1 and lncRG2 to generate sRNAs is probably distinct from the process seen in TE-derived IRs and more akin to the miRNA pathway, as previously reported for evolutionarily young miRNAs.

Reviewer #2:

Remarks to the Author:

The authors made careful and sincere efforts to address my concerns. My concerns were addressed. Again, the manuscript is clearly written and well demonstrates that the recently arisen "young" lncRG is involved in determining a set of traits. It is a wonderful study.

Just a suggestion: Considering that the study failed to identify the causal mutation underlying the QTLs and failed to get mutually supporting results for all traits in those analysis. I suggest that the authors tone down the conclusion about pleiotropic effects of the two genes on those traits and selection of the traits during domestication, such as using "likely", "hypothesize" or "supporting hypothesis". Those conclusions are mainly from quantitative and association studies, which just show a probability and evidence supporting an evolutionary hypothesis. Although it is very likely that the conclusion is correct, it is hard to completely approve those conclusions. For example, there are a large number of genes showing differential expression and methylation between soja and max. The study cannot exclude the possibility that the causal allele in an unknown trans-acting regulatory component is present in the QTL region, which regulate the two genes and many other genes. However, I will leave authors to decide how to address the issue. It should not affect quality of the manuscript.

In addition, there is a couple of publications showing pleiotropic regulation of domestication associated traits in soybean:

Goettel, POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean | Nature Communications.

Reviewer #3:

Remarks to the Author:

This is a revised version of the manuscript "Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance" by Wang et al.. The current version is substantially improved regarding data presentation and quality. I have no further concerns and agree with the publication of the manuscript.

## RESPONSES TO REVIEWERS' COMMENTS

## Reviewer #1 (Remarks to the Author)

Thank you for answering my questions, and congratulations on the paper. Overall, all my concerns were addressed. I only would like to include two minor changes.

-Please state in extended figure 6a whether the observed methylation is in the CG, CHG, or CHH context. I suggest either color-coding each context in the same plots or using individual plots for each context.

**RESPONSE#1.1:** As shown on Page 49, we have now displayed the CG, CHG and CHH methylation separately in individual plots in Extended Fig. 6a.

-The discussion between lines 308-321 leaves some room for improvement. Here is a suggestion to replace the same paragraph:

IRs integration in the genome can result from processes like DNA replication repair or transposable elements (TEs) [15]. IRs derived from TEs are typically processed by enzymes such as DCL3/DCL3-like, resulting in the production of 24-nt small RNAs [16,17]. These small RNAs often have a noticeable impact on the expression of nearby genes and phenotypic traits [17,18] (beside 17 and 18, please add references: 10.1016/j.celrep.2023.112029; 10.1111/nph.18068; 10.1093/plcell/koac220; 10.1016/j.molp.2020.02.009). In contrast, the IRs found in lncRG1 and lncRG2 predominantly give rise to 21-23-nt small RNAs (Fig. 2h and 2i). It's interesting to note that lncRG1 generates 23-nt sRNAs, while lncRG2 primarily yields 21-22-nt sRNAs (Fig. 2h and 2i). In concordance, lncRG1, which produces 23-nt sRNAs, displays signs of gene body methylation (see Ext. Data Fig. 6a), as reported for other IRs. This suggests that in this instance, 23-nt sRNAs 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 IRs (10.1016/j.celrep.2023.112029), we didn't observe differential gene expression in the vicinity of lncRG1 and lncRG2 when comparing Wm82 and PI according to the RNA-seq data (Figure X?). This further supports the idea that lncRG1 and lncRG2 operate through a distinct mechanism. In this context, the 21-nt and 22-nt sRNAs generated from the conserved lncRG2, which account for over 80-89% of the total sRNAs produced in these loci in *G. soja* accessions, are likely the major contributors to gene regulation. Consequently, the mechanism responsible for processing IRs in lncRG1 and lncRG2 to generate sRNAs is probably distinct from the process seen in TE-derived IRs and more akin to the miRNA pathway, as previously reported for evolutionarily young miRNAs.

**RESPONSE#1.2:** As shown on pages 14 and 15, we have now modified the discussion, and added all the references as suggested.

#### **Reviewer #2 (Remarks to the Author)**

The authors made careful and sincere efforts to address my concerns. My concerns were addressed. Again, the manuscript is clearly written and well demonstrates that the recently arisen “young” LncRG is involved in determining a set of traits. It is a wonderful study.

Just a suggestion: Considering that the study failed to identify the causal mutation underlying the QTLs and failed to get mutually supporting results for all traits in those analysis. I suggest that the authors tone down the conclusion about pleiotropic effects of the two genes on those traits and selection of the traits during domestication, such as using “likely”, “hypothesize” or “supporting hypothesis”. Those conclusions are mainly from quantitative and association studies, which just show a probability and evidence supporting an evolutionary hypothesis. Although it is very likely that the conclusion is correct, it is hard to completely approve those conclusions. For example, there are a large number of genes showing differential expression and methylation between soja and max. The study cannot exclude the possibility that the causal allele in an unknown trans-acting regulatory component is present in the QTL region, which regulate the two genes and many other genes. However, I will leave authors to decide how to address the issue. It should not affect quality of the manuscript.

In addition, there is a couple of publications showing pleiotropic regulation of domestication associated traits in soybean:

Goettel, POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean | Nature Communications.

**RESPONSE #2.1:** We have now added “likely” and changed a few “most likely” to “very likely” on Pages 5 and 6. In addition, we further articulated the three traits validated by transformation and gene-editing on Page 9.

We feel that the pleiotropic effects of POWR1 on seed compositions and seed weight belong to a category of pleiotropy that is quite different from what was described in our study, as those seed traits are closely related to each other, for instance, a single gene increases seed protein content almost always decreases seed oil content and seed weight. Therefore, we didn’t cite this paper, and felt further discussion on the phenomena of pleiotropy may be neither necessary nor helpful.

#### **Reviewer #3 (Remarks to the Author)**

This is a revised version of the manuscript "Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance" by Wang et al.. The current version is substantially improved regarding data presentation and quality. I have no further concerns and agree with the publication of the manuscript.

**Decision Letter, second revision:**

12th Feb 2024

Dear Dr. Ma,

Thank you for submitting your revised manuscript "Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance" (NG-A62922R1). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Genetics, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements soon. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Thank you again for your interest in Nature Genetics Please do not hesitate to contact me if you have any questions.

Sincerely,

Wei Li, PhD  
Senior Editor  
Nature Genetics  
New York, NY 10004, USA  
[www.nature.com/ng](http://www.nature.com/ng)

**Author Rebuttal, Second revision:**

Reviewer #1 (Remarks to the Author):

All my concerns have been addressed. I'm happy with this new version. This is a very interesting paper, congratulation to the authors.

Reviewer #2 (Remarks to the Author):

My concern was addressed.

Yong-qiang Charles An

## Responses to Reviewers' Comments

Reviewer #1 (Remarks to the Author):

All my concerns have been addressed. I'm happy with this new version. This is a very interesting paper, congratulation to the authors.

Reviewer #2 (Remarks to the Author):

My concern was addressed.

**Response:** We cordially appreciated all the comments provided by the reviewers, which have helped us substantially improve the quality of our work.

### Final Decision Letter:

28th Mar 2024

Dear Dr. Ma,

I am delighted to say that your manuscript "Long non-coding RNAs underlie multiple domestication traits and leafhopper resistance in soybean" has been accepted for publication in an upcoming issue of Nature Genetics.

Over the next few weeks, your paper will be copyedited to ensure that it conforms to Nature Genetics style. Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

After the grant of rights is completed, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at [rjsproduction@springernature.com](mailto:rjsproduction@springernature.com) immediately.

You will not receive your proofs until the publishing agreement has been received through our system.

Due to the importance of these deadlines, we ask that you please let us know now whether you will be difficult to contact over the next month. If this is the case, we ask you provide us with the contact information (email, phone and fax) of someone who will be able to check the proofs on your behalf, and who will be available to address any last-minute problems.

Your paper will be published online after we receive your corrections and will appear in print in the next available issue. You can find out your date of online publication by contacting the Nature Press Office ([press@nature.com](mailto:press@nature.com)) after sending your e-proof corrections.

You may wish to make your media relations office aware of your accepted publication, in case they consider it appropriate to organize some internal or external publicity. Once your paper has been scheduled you will receive an email confirming the publication details. This is normally 3-4 working days in advance of publication. If you need additional notice of the date and time of publication, please let the production team know when you receive the proof of your article to ensure there is sufficient time to coordinate. Further information on our embargo policies can be found here: <https://www.nature.com/authors/policies/embargo.html>

Before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may very well include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Genetics. Our Press Office may contact you closer to the time of publication, but if you or your Press Office have any enquiries in the meantime, please contact [press@nature.com](mailto:press@nature.com).

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Please note that *Nature Genetics* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. [Find out more about Transformative Journals](#)

**Authors may need to take specific actions to achieve [compliance](#) with funder and institutional open access mandates.** If your research is supported by a funder that requires immediate open access (e.g. according to [Plan S principles](#)) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including [a href="https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish"](https://www.nature.com/nature-portfolio/editorial-policies/self-archiving-and-license-to-publish). Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact [ASJournals@springernature.com](mailto:ASJournals@springernature.com)

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to

read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

An online order form for reprints of your paper is available at <https://www.nature.com/reprints/author-reprints.html>. Please let your coauthors and your institutions' public affairs office know that they are also welcome to order reprints by this method.

If you have not already done so, we invite you to upload the step-by-step protocols used in this manuscript to the Protocols Exchange, part of our on-line web resource, [natureprotocols.com](http://natureprotocols.com). If you complete the upload by the time you receive your manuscript proofs, we can insert links in your article that lead directly to the protocol details. Your protocol will be made freely available upon publication of your paper. By participating in [natureprotocols.com](http://natureprotocols.com), you are enabling researchers to more readily reproduce or adapt the methodology you use. [Natureprotocols.com](http://natureprotocols.com) is fully searchable, providing your protocols and paper with increased utility and visibility. Please submit your protocol to <https://protocolexchange.researchsquare.com/>. After entering your [nature.com](http://nature.com) username and password you will need to enter your manuscript number (NG-A62922R2). Further information can be found at <https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards#protocols>

Sincerely,  
Chiara

Chiara Anania, PhD  
Associate Editor  
Nature Genetics  
<https://orcid.org/0000-0003-1549-4157>