

Natural variation in GmRVE4d facilitates soybean latitudinal adaptation by regulating GmPRR5a-dependent flowering and maturity

Corresponding Author: Dr Baiquan Sun

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Review to Nature Communications 709534: 'Natural variation in GmRVE4d facilitates soybean latitudinal adaptation by regulating GmPRR5a-dependent flowering and maturity' by B. Sun et al.

This work makes an interesting contribution to the regulation of flowering time and maturity in soybean through the circadian clock. In the model Arabidopsis many mutants of the clock show flowering defects that also seems to be pervasive to other species like soybean. In this case, a REVILLE homolog (GmRVE4d) seems to present natural variants that affect flowering time and maturity. By cloning, making mutants and altering its expression, the authors have identified a homolog of PRR5 (another member of the clock) whose mutation seems to recapitulate some of the effects observed by the authors in the natural variants.

Although the paper makes some interesting observations about the effects of flowering time and maturity (two independent, but certainly related developmental stages) by a REVILLE homolog in soybean, it lacks the in-depth analysis to produce some of the conclusions raised. I will divide my comments into major and minor points.

Major comments.

1. Figures, sup figures, figure legends, and sup data sets should be explained in more detail and clearly identified. Otherwise, it is very hard to understand the statements of the authors.
2. It is not clear what does the different haplotypes have in flowering and maturation states of soybean. Are they all equally late? Or early? Apparently, the original mutations caused a delay in flowering time and maturation, but the CRISPR mutation causes an acceleration of both phenotypes. What do authors explain this to? The phenotypes associated to GmRVE4d ox and mutation are clear, but are those of the natural variants the same? Have authors checked PRR5 and FT expression in the natural variants?
3. Assumption in line 267 that GmRVE4d is independent of photoperiod is not clear. It just shows that the overexpression has an effect in both photoperiods, but they still can sense the photoperiod. Similarly, the assumption in line 268 could be produced by the 35S promoter being very effective, more so than the mutation. This also applies to lines 279-280, not clear assumption on the role on vegetative phase can be deduced from the effect of a 35S-promoter based phenotype (it is not regulated by the clock, the tissue or the photoperiod).
4. The version used for overexpression is the allele H3, why do authors did not express the native form without modifications?
5. The analysis of the binding site of GmRVE4d clearly indicates that it is an evening element, only it is normally refereed as AAAATATCT. Do authors found this in the MEME analysis and why do not name it as an EE?
6. Do authors used an antibody against GFP to do the ChIP-seq? Please indicate so.
7. Authors assume then that H3 natural version has a summative effect by repressing more GmPRR5 than H1, and that would explain the delay in the natural variants? It does not happen to be seen in the natural environment (Fig 2d). If GmPRR5 is a repressor as in Arabidopsis (as most PRRs including TOC1) how do authors explain this effect?
8. Further, the mutation of GmPRR5 delays flowering in soybean by decreasing GmFT homologs. This effect in Arabidopsis, a LD plant, is caused by the repression of CYCLING DOF FACTOR (CDF) genes, which are inhibitors of the flowering

activator CONSTANS (CO) and this causes downregulation of FT, so it is a rather indirect effect. I suggest checking CO and CDF homologs. In conclusion, GmPRR5 could be a repressor of a repressor.

9. A further explanation could be caused by the effect of PRR5 stabilizing CONSTANS in the evening and the fact that its interaction with PRR5 and HY5 affects the clock in a photoperiod-dependent manner. This way the clock has feedback from the photoperiod to the clock to set it at the correct seasonal time. I suggest the authors include this in the possible model.

Minor comments.

1. While the term photothermal refers to the combined regulatory effect of temperature and light on gene expression or any other regulatory layer, it cannot be widespread used in other sentences i.e. A photothermal environment does not mean any specific situation since all plant living situations are photothermal. Also, the photothermal adaptation is an elusive concept. In general terms, there is an abuse of the concept in the abstract. Please correct or be more specific. In most of the times you refer to light and temperature (not a close combination of both).
2. Please review use of gene and number throughout the text. i.e. Herein, we performed genome-wide association study (GWAS)...
3. In the abstract, what do authors mean by '...and its favorable allele...' favorable to what?
4. Please use the past tense when describing results, as it makes the distinction between results and deductions much simpler. i.e. Soybean (*Glycine max* (L.) Merr.) was originally domesticated in the temperate... (line 61). 'but these were not explored further here.' (line 177).
5. When making a deduction or discussion, try to evaluate the comment i.e. rather than 'Thus, GmRVE4d is a novel negative regulator of...' 'Thus, we propose that GmRVE4d is a novel negative regulator of...'.
6. Has GmPRR5 been demonstrated to be the ortholog of PRR5 in *Arabidopsis*? (line 45) I would suggest referring to it as a homolog, as it is difficult to ascertain the same function for both genes in the system.
7. Please refer to abbreviations once, the first time used in the text, only i.e. GWAS.
8. Line 127. Please define 'soybean growth period structure' as it is not done anywhere else in the text. Does it mean the morphology of the plant in any given developmental stage?
9. Please define in the introduction the flowering and maturity phenotype of GmRVE4d mutants and variants. Without this data, it is difficult to follow the discussion.
10. Please define the concept 'but also may be involved in the formation of environmental adaptation' line 133. Do authors mean it may be involved in the process of artificial adaptation of the species to the environment?
11. Line 142, what do the authors mean by 'but no significant differences between years at the same location'? Does it mean that maturity was more dependent on the site of experiment rather than on the year of replication?
12. Please use different abbreviations for Long Days (LD) and Linkage disequilibrium (LD) line 164.
13. The deduction of DeepGOPlus cannot ascertain that a gene is involved in photoperiod response, only suggests that (line 187).
14. the Y axis in Figure 1d is almost impossible to see. Please modify or repeat in sup data.
15. I could not find sup data set 5 (line 188), please indicate clearly in the files.
16. Although in 2 days it is difficult to ascertain, I would say that the thioredoxin candidate: Glyma.19G177600, also seemed to show a circadian pattern. Please discuss.
17. Please explain sup Fig. 10 in more detail in the figure legend.
18. Lines 197-199. Authors can deduce that the protein is expressed in the nucleus and that it is regulated by the clock, but it is too early to say (without the following homology studies) that it is a TF and that it regulates the clock. One could say that it could be a good candidate to do so.
19. In sup fig 11, what does the X axis represent, amino acids or bp? If it represents amino acids, then it should say amino or carboxy ends, not 5' and 3' ends.
20. Line 210. Could authors explain the source of GmRVE4dH1 and GmRVE4dH3 alleles in more detail?
21. Please change figure 2 legend to: a Gene structure and coding region variants of GmRVE4d in the different haplotypes. Figure 2b is important- It should be explained in more detail. There are certain legends that are not explained in the text. i.e. What is W82?
22. In figure 2d we see the maturity of Haplotypes 1 and 3, but cannot see the flowering time and the comparison to the WT. Are they significantly delayed?
23. In sup fig 12, what do the pink and magenta colors indicate? Please define the source of homologues.
24. Do the GmRVE4d ox lines have a GFP fused at the amino part? Please describe this in the text or in sup figs 15-16.
25. Please refer to JACK as the WT plants used throughout the text (if this is the case).
26. In line 290-297 authors refer to GmPRR5 as the ortholog of PRR5 without giving any reference. Do they mean it is a homolog?
27. Figures 4g and 4i are changed in the figure legend.
28. In figure 5g please use NSD, not SANYA conditions.
29. Please check out English in the discussion section, particularly reflecting quality contributions to the sentences. i.e. may or might, possibly, etc

Reviewer #2

(Remarks to the Author)

In this manuscript the authors identify new QTLs involved in regulating flowering time in cultivated soybeans. Within this QTL they demonstrate that there is variation in the a gene homologous to the *Arabidopsis* clock components RVE4 and

RVE8, that they name RVE4d. They observe different haplotypes of RVE4d in their population. Moreover, they also observe variation in flowering time associated with two different alleles of RVE4d in their population. Loss of function mutants and overexpressors of RVE4d display changes of flowering time. They also identify PRR5a as a target of RVE4d regulation, which, when mutated also displays differences in flowering time.

This is a generally clear and well written manuscript. I have a couple of comments to ensure the information is accurate and useful to the readers.

In the abstract the authors mention that GmRVE4d is a circadian clock gene, however they do not show any circadian phenotypes in the mutants or overexpressors. They also call the expression of RVE4d as circadian controlled (also line 191-192), but the only data shown is under light/dark cycles, so there is no current evidence of circadian control. The expression is definitely diel regulated, as is the expression of PRR5a, but there is not evidence in this manuscript their expression is clock controlled. For this manuscript it is not important whether or not RVE4d and PRR5 are clock genes or clock controlled, so that text can simply be modified (the summary fig. 6 also needs to be modified to indicate diel regulation not free running oscillations). The notation of RVE4d and PRR5a indicates that there are multiple members of each orthogroup, it would be useful to show a phylogenetic tree that supports the naming used. A reference to another paper describing these genes would be sufficient if that is the case. Such phylogeny can be used to support the hypothesis that RVE4d and PRR5a might be putative clock genes.

Figure 2 shows 7 haplotypes for GmRVE4d, however, differences in flowering time are only shown for two of them (H1 and H3). What is the maturity for the other haplotypes? This will be important to fully describe the variation observed. In the abstract the term 'photothermal adaptation' is used, it is not explained what they mean by photothermal and how it is related to this study. Explain how the temperature is involved would be useful in the introduction as they describe the current understanding of flowering time regulation in soybean.

On line 69 in the introduction there is a brief description of flowering time regulation. It is unclear if this section is specific to soybean. The description is very generic and vague. Please explain what is known about how flowering time is regulated in soybean. Just a list of Arabidopsis light signaling genes and the GI-CO-FT pathway is not very helpful, in particular because the work presented does not seem to imply any direct light regulation. It

On line 103 the authors describe the previous identification of putative clock related genes. It would be useful to specify the genes with an actual clock function and which ones were discovered as influencers of flowering time. It is important to separate both functions because orthologs of Arabidopsis clock genes can have non-clock functions in other species and is important to be specific about this information.

Line 158: the authors list tests to analyzed genetic relationships among the accessions studied, please link each to a figure panel or data table.

Supp. Fig. 16 shows a western blot of RVE4d-GFP, a negative control is absolutely necessary if the authors want to state that the top band is RVE4d-GFP. A size marker is also necessary.

For all the figures showing maturity values please explain what VE-R7 means, the explanations are different (sometimes absent) in the different figures, including supplemental figures. In particular most supplemental figures lack any explanations about the data shown. Figure legends should provide sufficient information to understand the figures without having to read the main text.

Supp. Fig. 3 and Supp. Fig. 4 Have the same title although it seems that represent different type of data. Please provide a clear explanation of how each of the graphs were generated.

All figure legends should include a brief explanation of how the experiment/analyses used to generate the data shown.

Line 128, for clarity please add a "," after GmPRR5a

Line 228, it would be useful to briefly specify what kind of mutants are the Gmrve4d mutants analyzed.

Please use non-capitalized 'p' for p-values (in italics)

Line 145: I am not sure what the authors mean by 'advancement of maturity groups', what is 'advancement'?

Fig. 6 model implied that all temperature and light signals regulating flowering time are transmitted via GmRVE4d. It would be more useful to integrate these proteins, if possible, with other known components, or simply make a model using the new data: RVE4d-PRR5-FT2a/FT5a.

Reviewer #3

(Remarks to the Author)

I read with great interest the report by Sun et al., that presents novel regulators of flowering and maturation in soybean, proposing new candidates for plant adaptation and potential breeding targets. The combination of genomic analyses, phenotyping, and ample molecular assays is a notable strength and has the potential to provide meaningful insights into the genetic architecture underlying these traits.

However, several aspects of the genomic analyses and their interpretation require clarification and, in some cases, more

rigorous treatment. Addressing the following points will greatly improve the robustness and impact of the manuscript.

Major Comments

1. Population structure and definition of clusters.

The study includes 257 accessions; however, the basis for defining three genetic clusters is not sufficiently supported. The phylogenetic tree suggests additional structure, and the PCA presented in Supplementary Fig. 3 does not clearly resolve three groups. Moreover, the proportion of variance explained by each principal component is not reported.

A more robust characterization of population structure is needed. The authors should justify the number of clusters using appropriate statistical approaches and provide clearer visualization and quantitative support. The lack of clarity in this part, has an impact in the GWAS section.

2. GWAS methodology and model specification

The GWAS was conducted using ~3 million SNPs, but it is unclear whether LD pruning was applied to reduce redundancy. This raises concerns about both statistical power and false-positive associations.

In addition, the mixed linear model appears insufficiently controlled for population structure and kinship – poor adjustment depicted in the QQ-plots (GAPIT itself allows to check for critical K value).

It would also strengthen the analysis to:

- Report the proportion of phenotypic variance explained by associated loci
- Provide minor allele frequencies for significant SNPs

Further, key details regarding phenotypic data are missing: number of biological replicates (plants per accession per trial), statistical treatment of phenotypes prior to GWAS.

3. Evidence for selection at the GmRVE4d locus

L130: “nucleotide polymorphisms in the promoter region of GmRVE4d have undergone strong artificial selection during domestication”. The inference is based on F_{st} and nucleotide diversity (π) calculated over the promoter region only, while the apparent selective sweep extends beyond this interval – it needs to be detailed that the calculations are done per site. To substantiate this claim, the authors should use other suitable selection tests. One simple option is to use Tajima’s D across the entire LD block; other options could be π , or EHH.

If selection is indeed acting on the promoter, functional evidence should be provided, such as differences in GmRVE4d expression among wild, landrace, and cultivated accessions.

4. Haplotype structure and domestication history

The manuscript would benefit from a clearer analysis of haplotype evolution. Specifically:

Are haplotypes H1 and/or H3 present in wild accessions?

Can the ancestral haplotype be inferred?

These analyses would provide important context for interpreting domestication-related selection.

5. LD and diversity patterns

The visualization of diversity and LD patterns requires improvement:

In Fig. 2F, π values in wild accessions dominate the scale, obscuring variation in landraces and cultivars. Alternative visualization strategies should be considered.

A comparison of LD structure between wild and domesticated groups would help support the proposed selection scenario.

6. RNA-seq and differential expression analyses

The manuscript lacks essential methodological details for RNA-seq and differential gene expression analyses (L589).

Information on experimental design, replication, data processing, normalization, and statistical testing must be included to ensure reproducibility.

7. Consideration of additional candidate genes

Given its annotation as a transcription factor involved in flowering and reproductive system development, the potential role of Glyma.19G178200 could be also addressed, at least theoretically. Does it exhibit signatures of selection comparable to GmRVE4d? are there different haplotypes in different geographic groups? Is it homologous to any other relevant flowering regulator?

Minor Comments

- The genomic analyses are generally described at a relatively superficial level. A more detailed and integrative discussion would improve accessibility and interpretability.
- PCA plots should include the percentage of variance explained by each component.
- Figure annotations and scaling should be improved for clarity.
- Suppl. Fig 4 is incorrectly

Reviewer #4

(Remarks to the Author)

In the manuscript under review, the authors reported a novel gene GmRVE4d in soybean which regulate flowering and maturity. Genome-wide association study (GWAS) and a deep learning tool that predicts protein function from sequence information were applied to identify GmRVE4d as a candidate gene. The function of the gene was experimentally validated by comparing the maturity of mutant and over-expression lines with the corresponding wild type cultivar. The authors also

found that the gene directly binds to the promoter of GmPRR5a, an ortholog of a circadian clock gene from Arabidopsis, and inhibits its expression.

Overall, I found that most parts of the manuscript are nicely written, the experimental results reported in this study are convincing and very insightful. However, I have some major concerns about the statistical analysis. In addition, some data showed in certain figures seemed not consistent with each other. The following are detailed explanations:

1) Statistical analysis: how was the candidate gene GmRVE4d identified?

First of all, it was described in Methods that the threshold of GWAS was $1/n$, where n is the total number of SNPs (line 509). This means, if we consider Bonferroni correction for multiple testing (which is a very common method applied in many GWAS studies), the family-wise error rate would be one. In other words, the results are almost sure to contain at least one false-positive. Of course, different thresholds (more stringent or more liberal ones) can be applied. But the threshold chosen by the authors seems to be too liberal.

Secondly, it is not clear whether the locus on Chr. 19 is the only significant signal across all three environments. If not (very likely there is a peak on Chr. 6 across environments, from Fig. 1c), why did the authors only focus on this one?

Thirdly, it is not clear how the QTL region was determined. It was only stated “an 84.1-kb interval (43.672 – 43.756 Mb)” in the text (line 180). The procedure from marker-trait associations to QTL region or interval should be clearly specified.

I do not doubt the experimental results showed later in the manuscript about the candidate gene GmRVE4d. But we know that usually it is a big challenge to search the “right” candidate genes from GWAS results. It seems that the authors did it in a very successful way that they landed at only one candidate gene. Thus, I would strongly suggest that more details are provided.

2) Inconsistencies between some phenotypic data.

Looking at Fig. 1a and Fig. 2d, I am a bit confused as the phenotypic range of the accessions in BJ2020 is clearly different. In Fig. 1a, all accessions were shown together and the upper-bound should be less than 125. However, in Fig. 2d, accessions were shown in two subgroups, one carrying Haplotype 1 and the other carrying haplotype 3. Many accessions in the second group have a value much larger than 125 (even exceeds 150). Why did this happen?

The performance of “Jack” as shown in Supplementary Fig. 14 and 17 is also not consistent. In both figures, days from VE to R7 stage were shown for “Jack” under controlled long-day as well as short-day conditions (In one figure, it was compared with the Gmrve4d mutants and in the other it was contrasted with the over-expression lines). But the performance of this cultivar is quite different in these two figures. For example, in Supplementary Fig. 14, the maturity of “Jack” plants was between 80 and 85 days, whereas in Supplementary Fig. 17 it was between 66 and 70 days. Were these plants really grown in the same controlled conditions?

I also have the following minor comments/suggestions which may help the authors to improve the quality of the paper:

Lines 140-142: Since there are only three environments, I would suggest to make the statement less general, like “The results revealed significant differences in maturity phenotypes between the two locations, but no significant differences between the two years at the location Beijing”.

Line 168: The critical value 0.4 written here in the text is not consistent with what was shown in Supplementary Fig. 5 (indicated by the dashed line). Please check. In addition, I would also suggest to write directly “... declines to half of its maximum value” as described in Methods (line 501).

Lines 191-192: Looking at Fig. 1e and 1f, I would not be so convinced about the circadian rhythmic pattern. The peak expression occurred both during the day and the night. Considering that the figure only showed results for two consecutive days (48 hours), the result is not completely convincing.

Lines 220-223: The number of cultivated soybeans (214) is much larger than that of wild soybeans (10), according to the description in Methods (lines 444-446). Are there possibly biases in the comparison of the molecular diversities of the two subpopulations due to the sample sizes (because the frequencies of the sequences are relevant in the calculation of Π value)?

Line 228: Which haplotype of GmRVE4d did the cultivar “Jack” carry, H1 or H3? Please specify here.

Lines 273-274: The occurrence of GmPRR5a in the interpretation here is very confusing as it is only revealed in the next subsection that GmRVE4d inhibits the transcription of GmPRR5a.

Lines 333-336: There were only results under natural short-day conditions, but no experimental results under natural long-day conditions?

Line 505: Please also cite the original paper of GAPIT version 3 (Wang and Zhang 2021): <https://doi.org/10.1016/j.gpb.2021.08.005>.

Line 513: It seems that the reference cited here (Yang et al. 2023) is only a study applying the tool DeepGOPlus. The original study developed the tool is Kulmanov and Hoehndorf (2020) and it should be cited: <https://doi.org/10.1093/bioinformatics/btz595>.

Fig. 1d: It is impossible to see the labels under the x-axis.

Supplementary Fig. 3: Please report the proportion of variance explained by the three PCs.

Supplementary Fig. 4: The caption of the figure is the same as Supplementary Fig. 3.

Please check throughout the manuscript whether the *Italic* font has been applied to all gene names.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

Comments

The authors have addressed most of my initial comments but there are a few revisions needed:

A. All methods description should enable fully reproducibility of the experiments. For example, please indicate the extraction buffer used and the exact Western blot protocol (for example: amount of protein loaded, gel type/concentration, membrane used, blocking agent used, exact antibody etc.). The authors should also indicate how many times this experiment was conducted. All protocols should enable exact reproduction of the experiment or provide a reference to a reproducible procedure. Please make sure to do that for all the methods provided (eg, RNA extraction, DAP profiling, EMSA...).

B. The phylogenetic trees are insufficient. All gene IDs need to be provided in the figure or legend. For GmRVE4d, the tree should demonstrate that the gene they studied is closely related to Arabidopsis RVE4, that means it is necessary to add all the other RVE genes (and CCA1/LHY) as well. Please add all the current gene names (if available) as well. The tree should include all the 'RVE' like genes identified in soybean, Arabidopsis and the other species shown in the current tree. Please address the same issues on the figure showing the GmPRR5 phylogenetic tree. The legends for these figures should indicate the methods used for alignment and estimating the phylogenetic tree, as well as what sequences were used (CDS, peptide sequence...). The trees should also include the bootstrap confidence levels.

Supplementary Fig. 21. Instead of indicating the protein shown, please indicate the antibody used for each panel.

In Line 77 the authors write: "E1 acts as a soybean-specific flowering repressor by suppressing the florigen genes GmFT2a and GmFT5a, whereas E2/GmGla, E3/GmPhyA3, E4/GmPhyA2, and J/GmELF3 connect photoperiod perception, circadian regulation, and downstream flowering responses. These regulators converge on the conserved GI-CO-FT module, in which GmFT2a and GmFT5a function as major florigens that promote floral transition under inductive conditions."

The sentence "These regulators converge..." does not make sense, since 'these regulators' include the FT and GI genes in soybean but not CO. It might be more useful to the reader to actually just explain what they mean by 'converge'.

Reviewer #3

(Remarks to the Author)

I appreciate the effort made by the authors to address the concerns of all the four reviewers. From my perspective, however, there are many open questions.

The population structure analysis is generally well presented; however, the rationale for selecting $K = 3$ as the primary representation of genetic structure is not entirely convincing. The cross-validation error does not appear to identify a clear optimum at $K = 3$ and instead suggests additional hierarchical structure within the population. It would therefore be informative to include ADMIXTURE plots for a range of K values (e.g., $K = 2-10$) in the Supplementary Information. Presenting multiple K values is common practice in population genomic studies and would allow readers to better assess the complexity of the inferred population structure and the robustness of the chosen clustering. In addition, it is unclear whether highly related individuals were identified and removed prior to the population genomic and GWAS analyses. Sequencing panels often contain duplicated or closely related accessions that can bias estimates of population structure, linkage disequilibrium, and association statistics if not properly accounted for. The authors should clarify whether relatedness filtering was performed and, if so, describe the criteria used to exclude closely related samples.

Inspection of the Q-Q plots suggests that the GWAS models are not consistently well calibrated across the three experimental conditions. Rather than following the null expectation and deviating only in the extreme tail, the observed p-values show an early and systematic departure from the expected distribution, indicating residual inflation or other sources

of confounding that are not fully accounted for by the current model. This suggests that additional efforts to control population structure or relatedness, or the evaluation of alternative GWAS models (e.g., FarmCPU or BLINK), may be needed.

Among the three analyses, the GWAS performed in Beijing (both years) appears to be better calibrated, although some inflation is still evident. In contrast, the third experimental condition exhibits a pronounced early deviation from the null expectation, raising concerns that the association statistics may be substantially influenced by confounding rather than true genotype–phenotype associations. Consequently, I would interpret the GWAS results from this dataset with considerable caution, as the quality control metrics suggest that the model does not adequately account for the underlying structure of the data.

The presentation of the GWAS summary statistics could also be improved. Supplementary Data 3 does not clearly indicate the experimental trial(s) in which each association reached significance, making it difficult to relate the summary statistics to the corresponding Manhattan plots. Explicitly annotating the trial(s) for each significant association would greatly improve the readability of the results.

In addition, the Manhattan plots suggest that chromosome 6 harbors a recurrent association signal across all three trials, similar to the signal reported on chromosome 19. This observation is not fully reflected in the text, which emphasizes chromosome 19 as the only locus consistently detected across environments. Given the apparent extent of linkage disequilibrium in this population, including the large LD block surrounding the chromosome 19 locus (almost 1Mb if not longer), it would be useful to characterize the LD structure around both regions. This would help determine whether chromosome 19 truly represents the only robust cross-environment signal or whether chromosome 6 also contains a stable association. Moreover, the association signal on chromosome 19 appears considerably weaker in the third trial, and therefore the statement that this locus is consistently detected across all three environments should be qualified accordingly.

Several conclusions rely on comparisons among the HR, NR and SR populations. However, the ADMIXTURE analysis does not appear to support these groups as genetically discrete units. Instead, individuals sampled within each geographic region show varying proportions of ancestry from multiple inferred populations. In other words, because the ADMIXTURE analysis indicates that geographic origin is an imperfect proxy for genetic ancestry, differences in haplotype frequencies among regions may reflect variation in ancestry composition rather than geographically restricted evolutionary processes. Clarification of how ancestry was accounted for in these analyses would strengthen the interpretation of the geographic distribution of candidate haplotypes.

The selection analyses are also somewhat limited in scope. While I appreciate that a comprehensive population genomic analysis is not the primary objective of this study, the current evidence falls short of supporting the strong conclusions drawn regarding domestication and selection, particularly for a journal such as Nature Communications. The authors present selection statistics only for the focal LD block, without providing the surrounding genomic context or genome-wide background distribution. As a result, it is difficult to assess whether this region represents an exceptional signal of selection or simply reflects the generally high levels of linkage disequilibrium observed in this population.

Indeed, the extensive LD in the study population makes this contextual information even more important. Without examining the selection signatures in the broader genomic neighborhood, it remains unclear whether the proposed candidate gene is the most likely target of selection or whether the observed signal extends across a larger haplotype containing multiple plausible candidates. Presenting the local genomic landscape, together with the genome-wide distribution of the selection statistics, would substantially strengthen the evidence supporting the identification of this candidate and the conclusions regarding domestication.

Reviewer #4

(Remarks to the Author)

The authors have carefully addressed my comments and suggestions in their revised manuscript. I have no further points to add.

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

I have read the adjustments and edits to the manuscript and I would say I am satisfied with the way the authors incorporated my suggestions and addressed the issues that concerned me in the earlier version.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Responses to Reviewers

Major comments from Reviewer #1:

1. Figures, sup figures, figure legends, and sup data sets should be explained in more detail and clearly identified. Otherwise, it is very hard to understand the statements of the authors.

Response:

We thank the reviewer for this important comment. We agree that the previous version provided insufficient guidance for interpreting the figures and supplementary materials. In the revised manuscript, we have (i) rewritten all figure legends and supplementary figure legends to include clear, panel-by-panel descriptions; (ii) expanded the explanations of all supplementary datasets; and (iii) improved in-text cross-referencing to specific figures and supplementary items. These revisions should substantially improve the accessibility and flow of the results.

Major comments from Reviewer #1:

2. It is not clear what does the different haplotypes have in flowering and maturation states of soybean. Are they all equally late? Or early? Apparently, the original mutations caused a delay in flowering time and maturation, but the CRISPR mutation causes an acceleration of both phenotypes. What do authors explain this to? The phenotypes associated to GmRVE4d ox and mutation are clear, but are those of the natural variants the same? Have authors checked PPR5 and FT expression in the natural variants?

Response:

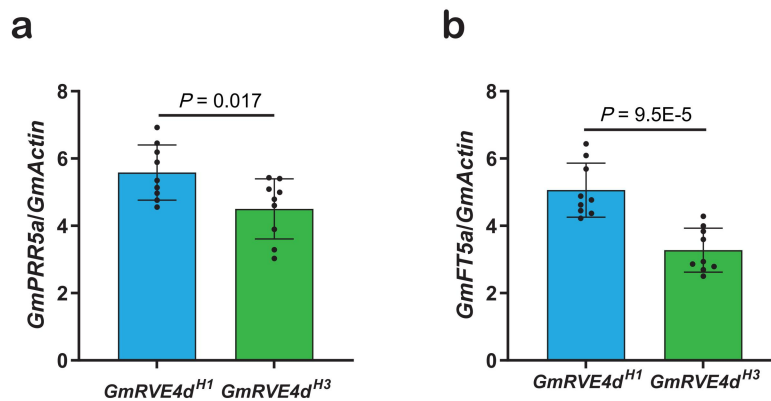
We thank the reviewer for this insightful comment. We agree that the previous version failed to clearly distinguish the phenotypic effects of natural haplotypes from those of engineered loss-of-function alleles. We have addressed this in three parts.

First, the natural haplotypes are not phenotypically equivalent. We have now added flowering-time data for the major haplotypes. As shown in Fig. 2c, accessions carrying *GmRVE4d^{H3}* flower significantly later than those carrying *GmRVE4d^{H1}* in all three tested environments; the maturity data in Fig. 2d show a consistent trend. These results indicate that *GmRVE4d^{H3}* is associated with later flowering and later maturity relative to *GmRVE4d^{H1}*.

Second, the apparent discrepancy between natural variants and CRISPR/Cas9 mutants reflects differences in allelic strength. The CRISPR/Cas9 lines are complete loss-of-function mutants that show early flowering and early maturity, consistent with *GmRVE4d* acting as a negative regulator of these traits. By contrast, *GmRVE4d^{H3}* is not a loss-of-function allele. Our data demonstrate that *GmRVE4d^{H3}* binds the *GmPRR5a* promoter with greater affinity and represses *GmPRR5a*

transcription more strongly than *GmRVE4d^{H1}*. Consequently, this gain-of-repression natural allele is associated with delayed flowering and maturity.

Third, to directly address the reviewer's question about gene expression in natural variants, we examined *GmPRR5a* and *GmFT5a* transcript levels in representative accessions carrying *GmRVE4d^{H1}* or *GmRVE4d^{H3}* (Response Fig. 1). Accessions carrying *GmRVE4d^{H3}* showed significantly lower *GmPRR5a* and *GmFT5a* transcript accumulation than *GmRVE4d^{H1}* accessions under the same sampling regime, consistent with the proposed regulatory model.



Response Fig. 1. qRT-PCR validation of *GmPRR5a* and *GmFT5a* expression in soybean accessions carrying distinct *GmRVE4d* haplotypes. Relative transcript levels of *GmPRR5a* (a) and *GmFT5a* (b) in representative accessions harboring *GmRVE4d^{H1}* (Clay, Heihe 43, and Mengdou 30) or *GmRVE4d^{H3}* (Zhonghuang 301, Zhongdou 19, and Dowling). All samples were collected under the same sampling regime. Gene expression was quantified by qRT-PCR and normalized to *GmActin*. Three independent biological replicates were performed per accession. Significance was assessed by a two-tailed Student's *t*-test.

Major comments from Reviewer #1:

3. Assumption in line 267 that *GmRVE4d* is independent of photoperiod is not clear. It just shows that the overexpression has an effect in both photoperiods, but they still can sense the photoperiod. Similarly, the assumption in line 268 could be produced by the 35S promoter being very effective, more so than the mutation. This also applies to lines 279-280, not clear assumption on the role on vegetative phase can be deduced from the effect of a 35S-promoter based phenotype (it is not regulated by the clock, the tissue or the photoperiod).

Response:

We thank the reviewer for this important critique. We agree that observing delayed flowering in 35S::*GmRVE4d* lines under both short-day and long-day conditions does not demonstrate photoperiod independence; it indicates only that elevated *GmRVE4d* expression delayed flowering under both photoperiod regimes. We have revised the text accordingly and removed all statements implying photoperiod independence.

We also agree that the 35S promoter bypasses native circadian, tissue-specific, and photoperiodic regulation, and that the stronger phenotype of overexpression lines cannot be interpreted as evidence for a more prominent physiological role relative to the loss-of-function lines. We have revised the relevant text to avoid this overinterpretation and have restricted our conclusions to the directly measured developmental traits, VE-R1 and VE-R7.

Major comments from Reviewer #1:

4. The version used for overexpression is the allele H3, why do authors did not express the native form without modifications?

Response:

We thank the reviewer for this helpful comment and apologize for the lack of clarity in the original manuscript. The designation '*GmRVE4d^{H3}*' refers to a naturally occurring haplotype of *GmRVE4d*, not an artificially modified version of the gene. In our population analysis, *GmRVE4d^{H1}* corresponds to the reference allele represented by Williams 82 and is also the haplotype present in the recipient cultivar Jack.

We therefore overexpressed the naturally occurring *GmRVE4d^{H3}* coding sequence in the Jack background, which already carries the endogenous *GmRVE4d^{H1}* allele, to test whether the late-flowering- and late-maturity-associated *GmRVE4d^{H3}* allele is sufficient to alter these traits when ectopically expressed. The *GmRVE4d^{H3}* coding sequence used in the overexpression construct was not artificially modified. We have clarified this point in the revised Methods and Results sections.

Major comments from Reviewer #1:

5. The analysis of the binding site of GnRVE4d clearly indicates that it is an evening element, only it is normally refereed as AAAATATCT. Do authors found this in the MEME analysis and why do not name it as an EE?

Response:

We thank the reviewer for this helpful observation. We agree that the enriched motif corresponds to the Evening Element. In our MEME analysis, the motif was primarily detected in the reverse-complement orientation as AGATATTTT, which corresponds to the canonical EE sequence AAAATATCT. We have revised the manuscript to refer to this motif as the Evening Element throughout the text and figure legends, and we now report both orientations explicitly to avoid ambiguity.

Major comments from Reviewer #1:

6. Do authors used an antibody against GFP to do the ChIP-seq? Please indicate so.

Response:

We thank the reviewer for this clarification request. We did not perform ChIP-seq in this study. We performed ChIP-qPCR using chromatin prepared from 35Spro:*GmRVE4d*-GFP plants, with immunoprecipitation carried out using an anti-GFP antibody. We have revised the Methods section to state this explicitly and to prevent any confusion with ChIP-seq.

Major comments from Reviewer #1:

7. Authors assume then that H3 natural version has a summative effect by repressing more GmPRR5 than H1, and that would explain the delay in the natural variants? It does not happen to be seen in the natural environment (Fig 2d). If GmPRR5 is a repressor as in Arabidopsis (as most PRRs including TOC1) how do authors explain this effect?

Response:

We thank the reviewer for this important question. Our results suggest that *GmPRR5a* functions as a positive regulator of flowering under the tested conditions in soybean, which differs from the repressive roles reported for several *Arabidopsis* PRR family members in the photoperiodic flowering pathway.

Specifically, CRISPR/Cas9-mediated knockout of *GmPRR5a* delayed both flowering and maturity and reduced *GmFT2a* and *GmFT5a* transcript accumulation under natural short-day conditions in Sanya, China, in 2025. These results indicate that *GmPRR5a* positively contributes to floral induction under this condition. Accordingly, stronger repression of *GmPRR5a* by *GmRVE4d^{H3}* would be expected to reduce *GmPRR5a* activity and thereby delay flowering and maturity—an interpretation consistent with the phenotypic comparison of natural haplotypes, in which *GmRVE4d^{H3}* accessions exhibited later flowering and maturity than *GmRVE4d^{H1}* accessions. We have clarified this mechanistic interpretation in the revised Results and Discussion sections.

Major comments from Reviewer #1:

8. Further, the mutation of GmPRR5 delays flowering in soybean by decreasing GmFT homologs. This effect in Arabidopsis, a LD plant, is caused by the repression of CYCLING DOF FACTOR (CDF) genes, which are inhibitors of the flowering activator CONSTANS (CO) and this causes downregulation of FT, so it is a rather indirect effect. I suggest checking CO and CDF homologs. In conclusion, GmPRR5 could be a repressor of a repressor.

Response:

We thank the reviewer for this helpful suggestion. We examined the expression of soybean CDF homologs (*GmCDF1* and *GmCDF2*) and CO-like homologs (*GmCOL1a* and *GmCOL1b*) in wild-type Jack and two independent *Gmprp5a* mutant lines. Relative to Jack, *GmCDF1* and *GmCDF2* were significantly downregulated in the *Gmprp5a* mutants, whereas *GmCOL1a* and

GmCOL1b were upregulated. In the same mutant backgrounds, *GmFT2a* and *GmFT5a* were significantly downregulated.

These results indicate that *GmPRR5a* influences the expression of CDF and CO-like homologs in soybean. However, because these data are based on transcript abundance, they do not distinguish direct from indirect regulatory effects, nor do they demonstrate that the soybean pathway functions identically to the *Arabidopsis* PRR–CDF–CO–FT module. We have therefore revised the manuscript to present the *GmPRR5a*–GmCDF–GmCOL–GmFT pathway as a putative regulatory module that requires further functional validation. The new data have been included as Supplementary Fig. 29.

Major comments from Reviewer #1:

9. A further explanation could be caused by the effect of PRR5 stabilizing CONSTANS in the evening and the fact that its interaction with PRR5 and HY5 affects the clock in a photoperiod-dependent manner. This way the clock has feedback from the photoperiod to the clock to set it at the correct seasonal time. I suggest the authors include this in the possible model.

Response:

We thank the reviewer for this insightful suggestion. We agree that *GmPRR5a* may influence flowering through CO-like modules and that this could involve regulation of CO-like protein stability or *HY5*-related pathways, as reported in other systems. We have revised Fig. 6 to incorporate a putative GmCDF–GmCOL module downstream of *GmPRR5a*. In this working model, the direct molecular evidence supports repression of *GmPRR5a* transcription by *GmRVE4d* protein, while the downstream relationships among *GmPRR5a*, *GmCDFs*, *GmCOLs*, and *GmFT2a/GmFT5a* are based on expression patterns and prior knowledge of photoperiodic flowering pathways. These unresolved mechanistic links are depicted as dashed lines in the revised model to reflect their putative nature.

Minor comments from Reviewer #1:

1. While the term photothermal refers to the combined regulatory effect of temperature and light on gene expression or any other regulatory layer, it cannot be widespread used in other sentences i.e. A photothermal environment does not mean any specific situation since all plant living situations are photothermal. Also, the photothermal adaptation is an elusive concept. In general terms, there is an abuse of the concept in the abstract. Please correct or be more specific. In most of the times you refer to light and temperature (not a close combination of both).

Response:

We thank the reviewer for pointing out the overuse and ambiguity of the term 'photothermal' in the Abstract. We agree that, in a broad sense, most plant growth conditions involve both light and

temperature, making 'photothermal environment' imprecise without an explicit definition. Our intent was to describe the regulatory roles of photoperiod and ambient temperature on flowering time and maturity, rather than to use 'photothermal' as a general environmental descriptor. We have therefore revised the Abstract and relevant main-text sentences by removing vague phrases such as 'photothermal signals' and 'photothermal adaptation,' replacing them with the specific terms 'photoperiod' and 'temperature' where appropriate.

Minor comments from Reviewer #1:

2. Please review use of gene and number throughout the text. i.e. Herein, we performed genome-wide association study (GWAS)...

Response:

Thank you for the suggestion. We have thoroughly proofread the manuscript and corrected the formatting and usage of gene symbols, gene names, and numerals throughout the main text, figure legends, and supplementary information. We have also ensured that all abbreviations are defined at their first mention and that number–unit presentation is consistent throughout.

Minor comments from Reviewer #1:

3. In the abstract, what do authors mean by ‘..and its favorable allele...’ favorable to what?

Response:

Thank you for this observation. We agree that the term 'favorable allele' was ambiguous, as it did not specify the trait to which the allele confers an advantage. We have revised the Abstract to clarify the phenotypic effect and avoid subjective language. The sentence now reads: 'Genetic analysis revealed that *GmRVE4d* underwent domestication-related artificial selection, and a late-maturing haplotype, *GmRVE4d*^{H3}, was characterized.'

Minor comments from Reviewer #1:

4. Please use the past tense when describing results, as it makes the distinction between results and deductions much simpler. i.e. Soybean (*Glycine max* (L.) Merr.) was originally domesticated in the temperate... (line 61). ‘but these were not explored further here.’ (line 177).

Response:

Thank you for this suggestion. We agree that consistent use of the past tense when reporting experimental results helps readers distinguish data from interpretation. We have revised the manuscript to use the past tense for all descriptions of our own results and observations, while retaining the present tense for established facts and general background statements. We also

corrected the specific examples highlighted by the reviewer and applied consistent tense usage throughout the Results section.

Minor comments from Reviewer #1:

5. When making a deduction or discussion, try to evaluate the comment i.e. rather than ‘Thus, GmRVE4d is a novel negative regulator of...’ ‘Thus, we propose that GmRVE4d is a novel negative regulator of...’.

Response:

Thank you for this important comment. We agree that the original wording was too assertive for a conclusion based on the available experimental evidence. We have revised the sentence to: 'Taken together, our findings suggest that GmRVE4d is a novel negative regulator of soybean flowering and maturity that functions by directly repressing *GmPRR5a* expression, and is a promising molecular target for expanding cultivation latitudes through molecular breeding.'

Minor comments from Reviewer #1:

6. Has GmPRR5 been demonstrated to be the ortholog of PRR5 in Arabidopsis? (line 45) I would suggest referring to it as a homolog, as it is difficult to ascertain the same function for both genes in the system.

Response:

Thank you for this comment. We agree that designating *GmPRR5a* as an 'ortholog' is too strong without comprehensive phylogenetic, syntenic, and cross-species functional evidence. We have revised the text throughout to refer to *GmPRR5a* as a soybean homolog of *Arabidopsis PRR5*, and have removed any language implying functional equivalence between the two genes.

Minor comments from Reviewer #1:

7. Please refer to abbreviations once, the first time used in the text, only i.e. GWAS.

Response:

Thank you for the suggestion. We have revised the manuscript to ensure that all abbreviations, including genome-wide association study (GWAS), are defined only at their first occurrence and are not re-defined subsequently.

Minor comments from Reviewer #1:

8. Line 127. Please define ‘soybean growth period structure’ as it is not done anywhere else in the text. Does it mean the morphology of the plant in any given developmental stage?

Response:

Thank you for the comment. We agree that 'growth period structure' was insufficiently defined. In this study, the term refers to the ratio of reproductive to vegetative growth duration, not to plant morphology. We have added a clear definition at its first occurrence and revised the sentence to improve clarity.

Minor comments from Reviewer #1:

9. Please define in the introduction the flowering and maturity phenotype of *GmRVE4d* mutants and variants. Without this data, it is difficult to follow the discussion.

Response:

We thank the reviewer for this important point. We agree that the flowering and maturity phenotypes of *GmRVE4d* mutants, overexpression lines, and natural haplotypes should be introduced clearly before the mechanistic interpretation is presented. We have therefore added a concise summary at the beginning of the relevant Results subsection describing the main phenotypic trends observed in *Gmrve4d* mutants, 35Spro:*GmRVE4d*-GFP overexpression lines, and natural *GmRVE4d* haplotypes under the tested photoperiod conditions. This revision should make the subsequent Discussion easier to follow.

Minor comments from Reviewer #1:

10. Please define the concept 'but also may be involved in the formation of environmental adaptation' line 133. Do authors mean it may be involved in the process of artificial adaptation of the species to the environment?.

Response:

We thank the reviewer for this comment. We agree that the phrase 'may be involved in the formation of environmental adaptation' was ambiguous and potentially misleading. Our intended meaning was that *GmRVE4d* may have contributed to soybean adaptation to diverse environmental conditions during domestication. However, because this statement remains difficult to convey concisely without risking misinterpretation, we have deleted it from the revised manuscript.

Minor comments from Reviewer #1:

11. Line 142, what do the authors mean by 'but no significant differences between years at the same location'? Does it mean that maturity was more dependent on the site of experiment rather than on the year of replication?

Response:

Thank you for this helpful comment. The reviewer's interpretation is correct. Our intent was to convey that differences in photoperiod among geographically distinct locations had a greater effect on maturity than year-to-year variation at the same location. We have revised the text to state explicitly that maturity showed significant differences among locations but not between years at the same location, indicating that site of evaluation was a more important source of variation than year of evaluation. We have also avoided overinterpreting this observation as direct evidence of environmental adaptation.

Minor comments from Reviewer #1:

12. Please use different abbreviations for Long Days (LD) and Linkage disequilibrium (LD) line 164.

Response:

Thank you for this comment. To eliminate ambiguity, we have revised the manuscript to spell out 'linkage disequilibrium' in full throughout the text. The abbreviation 'LD' is now reserved exclusively for 'long days.'

Minor comments from Reviewer #1:

13. The deduction of DeepGOPlus cannot ascertain that a gene is involved in photoperiod response, only suggests that (line 187).

Response:

Thank you for the insightful comment. We agree that DeepGOPlus provides computational Gene Ontology predictions and cannot conclusively demonstrate involvement in photoperiod response. We have revised the text to avoid causal or definitive wording and now describe the result as a predicted functional association.

Minor comments from Reviewer #1:

14. the Y axis in Figure 1d is almost impossible to see. Please modify or repeat in sup data.

Response:

Thank you for the observation. We agree that the y-axis in the original Fig. 1d was difficult to read. We have reformatted Fig. 1d by adjusting the layout, figure dimensions, and axis presentation so that the y-axis labels are now clearly legible.

Minor comments from Reviewer #1:

15. I could not find sup data set 5 (line 188), please indicate clearly in the files.

Response:

Thank you for noting this. We have re-uploaded Supplementary Dataset 5 and ensured that it is clearly listed and cross-referenced in the supplementary file list.

Minor comments from Reviewer #1:

16. Although in 2 days it is difficult to ascertain, I would say that the thioredoxin candidate: Glyma.19G177600, also seemed to show a circadian pattern. Please discuss.

Response:

Thank you for this perceptive observation. We agree that *Glyma.19G177600* displayed time-dependent expression changes over the two-day sampling period. However, DeepGOPlus-based functional prediction did not assign this gene to biological processes directly associated with the target trait, and no haplotype-phenotype association comparable to that of *GmRVE4d* was detected for this gene. Therefore, although *Glyma.19G177600* exhibited rhythmic expression, the cumulative evidence did not support it as the primary causal candidate within this genomic interval. We have incorporated this point into the revised Results section and supplementary materials.

Minor comments from Reviewer #1:

17. Please explain sup Fig. 10 in more detail in the figure legend.

Response:

Thank you for the suggestion. We agree that the previous legend for Supplementary Fig. 10 was too brief. We have revised it to provide a comprehensive description of the data shown, the meaning of all labels, and the interpretation of each panel.

Minor comments from Reviewer #1:

18. Lines 197-199. Authors can deduce that the protein is expressed in the nucleus and that it is regulated by the clock, but it is too early to say (without the following homology studies) that it is a TF and that it regulates the clock. One could say that it could be a good candidate to do so.

Response:

Thank you for this important comment. We agree that nuclear localization alone does not establish transcription factor function or direct circadian clock regulation. We have revised the text to read: 'The *GmRVE4d*-GFP fusion protein was predominantly localized to the nucleus (Supplementary Fig. 17), consistent with its predicted function as a transcription factor that regulates downstream gene expression.' This revision avoids claiming circadian clock regulation on the basis of subcellular localization data alone.

Minor comments from Reviewer #1:

19. In sup fig 11, what does the X axis represent, amino acids or bp? If it represents amino acids, then it should say amino or carboxy ends, not 5' and 3' ends.

Response:

Thank you for identifying this labeling error. The x-axis in Supplementary Fig. 11 represents amino acid residue positions, not nucleotide base pairs. The use of 5' and 3' end labels was therefore inappropriate. We have revised the figure by replacing '5' and '3' with 'N terminus' and 'C terminus,' respectively, and have updated the figure legend accordingly.

Minor comments from Reviewer #1:

20. Line 210. Could authors explain the source of GmRVE4dH1 and GmRVE4dH3 alleles in more detail?

Response:

Thank you for this helpful comment. These two alleles represent naturally occurring coding-sequence haplotypes identified from resequencing data of our soybean natural population. Haplotypes were defined by extracting sequence variants within the *GmRVE4d* coding region and combining the linked polymorphic sites.

Relative to *GmRVE4d*^{H1}, *GmRVE4d*^{H3} carries four nucleotide substitutions: c.230G>A, c.771G>A, c.780A>G, and c.793C>A, corresponding to the amino acid changes p.Ser77Asn, p.Glu257Glu, p.Ser260Ser, and p.His265Asn, respectively. Among these, two substitutions (Ser77Asn and His265Asn) are nonsynonymous, and two are synonymous. This information has been added to the revised Results section and Methods.

Minor comments from Reviewer #1:

21. Please change figure 2 legend to: a Gene structure and coding region variants of GmRVE4d in the different haplotypes. Figure 2b is important- It should be explained in more detail. There are certain legends that are not explained in the text. i.e. What is W82?

Response:

Thank you for the valuable suggestion. We have revised the legend of Fig. 2a as recommended: 'Gene structure and coding region variants of *GmRVE4d* in the different haplotypes.' We agree that Fig. 2b required a more detailed explanation and have expanded its legend accordingly. The following definitions have been added: 'W82 denotes Williams 82, the soybean reference cultivar; CDS, coding sequence; nt, nucleotide; AA, amino acid; #, nucleotide or amino acid position.' All abbreviations and labels in the figure are now clearly defined.

Minor comments from Reviewer #1:

22. In figure 2d we see the maturity of Haplotypes 1 and 3, but cannot see the flowering time and the comparison to the WT. Are they significantly delayed?

Response:

Thank you for this comment. *GmRVE4d^{H1}* and *GmRVE4d^{H3}* in Fig. 2d represent naturally occurring allelic variants from our soybean population rather than engineered lines; therefore, an isogenic wild-type control is not available for direct comparison. To address the reviewer's concern about flowering time, we have added flowering-time data for the same accessions analyzed in Fig. 2d (now shown in Fig. 2c). The flowering-time pattern was consistent with the maturity trend: accessions carrying *GmRVE4d^{H3}* flowered significantly later than those carrying *GmRVE4d^{H1}* across all tested environments, and the difference was statistically significant.

Minor comments from Reviewer #1:

23. In sup fig 12, what do the pink and magenta colors indicate? Please define the source of homologues.

Response:

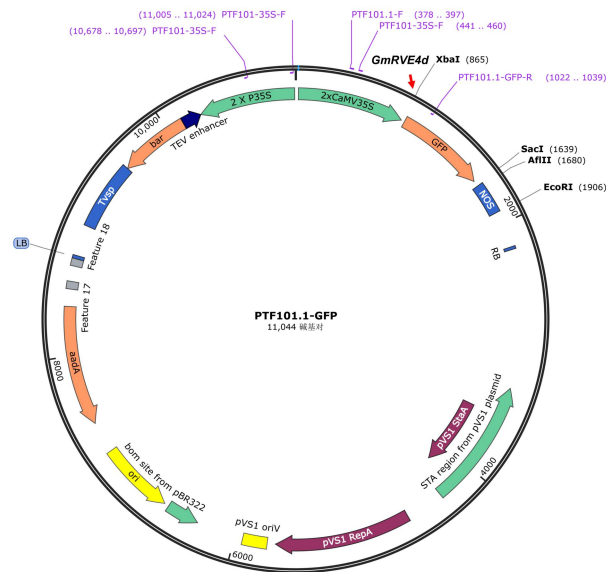
Thank you for the question. In Supplementary Fig. 12, black, pink, and blue indicate different degrees of amino acid conservation in the multiple sequence alignment. Magenta asterisks mark the positions of amino acid polymorphisms between *GmRVE4d^{H1}* and *GmRVE4d^{H3}*. Regarding the source of homologs, homologous protein sequences were retrieved from publicly available genome annotations for representative plant species—including Arabidopsis, rice, wheat, and maize—using *GmRVE4d* as the query in BLASTp-based homology searches. We have clarified all of these points in the revised figure legend.

Minor comments from Reviewer #1:

24. Do the GmRVE4d ox lines have a GFP fused at the amino part? Please describe this in the text or in sup figs 15-16.

Response:

We apologize for the ambiguity caused by our original construct notation, PTF101.1-GFP-*GmRVE4d*, which could be interpreted as an N-terminal GFP fusion. In fact, GFP is fused to the C-terminus of *GmRVE4d*: the *GmRVE4d* coding sequence was inserted in-frame upstream of GFP to produce a C-terminal *GmRVE4d*-GFP fusion protein. We have corrected the construct name throughout the revised manuscript and Supplementary Information to PTF101.1-*GmRVE4d*-GFP and have clarified this point in Response Fig. 2, which provides a schematic diagram of the overexpression vector.



Response Fig. 2 Schematic representation of the PTF101.1-*GmRVE4d*-GFP overexpression vector. The coding sequence of *GmRVE4d* was inserted in-frame upstream of GFP to generate a C-terminal GmRVE4d-GFP fusion. The red arrow marks the insertion site of *GmRVE4d*.

Minor comments from Reviewer #1:

25. Please refer to JACK as the WT plants used throughout the text (if this is the case).

Response:

Thank you for the suggestion. We confirm that Jack is the wild-type soybean cultivar used throughout this study. We have revised the manuscript to introduce 'Jack wild type (WT)' at its first comparison with mutant or overexpression lines, and to use 'Jack' consistently thereafter.

Minor comments from Reviewer #1:

26. In line 290-297 authors refer to *GmPRR5* as the ortholog of *PRR5* without giving any reference. Do they mean it is a homolog?

Response:

Thank you for pointing this out. We agree that, without explicit phylogenetic and orthology inference, *GmPRR5a* should be described as a homolog—not an ortholog—of *Arabidopsis PRR5*. We have revised the text accordingly throughout the manuscript. We have also added a phylogenetic analysis of the *PRR5* gene family as Supplementary Fig. 26 to support the homology designation.

Minor comments from Reviewer #1:

27. Figures 4g and 4i are changed in the figure legend.

Response:

Thank you for the careful review. We have corrected the figure legends for Figs. 4g and 4i accordingly.

Minor comments from Reviewer #1:

28. In figure 5g please use NSD, not SANYA conditions.

Response:

Thank you for the suggestion. We have corrected Fig. 5d and Fig. 5g to indicate 'NSD conditions,' replacing 'SY2025 conditions,' and have revised the corresponding figure legends accordingly.

Minor comments from Reviewer #1:

29. Please check out English in the discussion section, particularly reflecting quality contributions to the sentences. i.e. may or might, possibly, etc

Response:

Thank you for the suggestion. We have carefully revised the Discussion to improve clarity, readability, and appropriate epistemic calibration. We have replaced overly assertive language with appropriately hedged terms—such as 'may,' 'might,' 'possibly,' and 'we propose'—throughout, and have avoided overstatement of results or interpretations.

Comments from Reviewer #2:

1. In the abstract the authors mention that GmRVE4d is a circadian clock gene, however they do not show any circadian phenotypes in the mutants or overexpressors. They also call the expression of RVE4d as circadian controlled (also line 191-192), but the only data shown is under light/dark cycles, so there is no current evidence of circadian control. The expression is definitely diel regulated, as is the expression of PRR5a, but there is no evidence in this manuscript that their expression is clock controlled. For this manuscript it is not important whether or not RVE4d and PRR5 are clock genes or clock controlled, so that text can simply be modified (the summary fig. 6 also needs to be modified to indicate diel regulation not free running oscillations). The notation of RVE4d and PRR5a indicates that there are multiple members of each orthogroup, it would be useful to show a phylogenetic tree that supports the naming used. A reference to another paper describing these genes would be sufficient if that is the case. Such phylogeny can be used to support the hypothesis that RVE4d and PRR5a might be putative clock genes.

Response:

We thank the reviewer for this important comment. We fully agree that our expression data were generated under light/dark cycles and therefore demonstrate diel regulation, not endogenous

circadian control under free-running conditions. We have revised the manuscript accordingly in three respects.

First, we revised the Abstract and all relevant main-text sections, including the statements around lines 191–192, removing the terms 'circadian clock gene,' 'clock-controlled,' and 'circadian-controlled.' These have been replaced with accurate descriptors such as 'regulated under the light–dark cycle' or 'showing time-dependent expression' where appropriate.

Second, we revised Fig. 6 to remove any implication of free-running circadian oscillations. The revised model now indicates that *GmRVE4d* and *GmPRR5a* display diel expression patterns under the tested light/dark conditions, without claiming circadian clock control. We fully agree that whether these genes are bona fide clock components is not central to the conclusions of this study.

Third, to support the nomenclature of *GmRVE4d* and *GmPRR5a* and to clarify their relationships with related paralogs, we added phylogenetic analyses as new supplementary figures. These analyses included homologs from *Arabidopsis thaliana* and representative crops, including *Zea mays*, *Oryza sativa* and *Triticum aestivum*. Specifically, Supplementary Fig. 15 shows an RVE4-family phylogeny supporting the designation of *GmRVE4d*, and Supplementary Fig. 26 shows a PRR5-family phylogeny supporting the designation of *GmPRR5a*. In both trees, soybean contains multiple closely related paralogs, and *GmRVE4d* and *GmPRR5a* group with the corresponding *AtRVE4* and *AtPRR5* clades, respectively, supporting our naming scheme and their annotation as putative clock-related homologs.

Comments from Reviewer #2:

2. Figure 2 shows 7 haplotypes for *GmRVE4d*, however, differences in flowering time are only shown for two of them (H1 and H3). What is the maturity for the other haplotypes? This will be important to fully describe the variation observed.

Response:

Thank you for this important suggestion. In our panel, all haplotypes other than *GmRVE4d*^{H1} and *GmRVE4d*^{H3} were rare, each represented by fewer than five accessions. Given these very small sample sizes, statistical comparisons of maturity or flowering time among the rare haplotypes would lack statistical power and could be misleading. We therefore restricted phenotypic comparisons to the two major haplotypes, which had adequate sample sizes. We have clarified this reasoning explicitly in the revised Results section and now state that rare haplotypes were excluded from statistical comparisons due to insufficient sample size.

Comments from Reviewer #2:

3. In the abstract the term 'photothermal adaptation' is used, it is not explained what they mean by photothermal and how it is related to this study. Explain how the temperature is involved would be

useful in the introduction as they describe the current understanding of flowering time regulation in soybean.

Response:

Thank you for the helpful comment. We agree that 'photothermal adaptation' was not sufficiently defined and could imply a direct temperature-response function for *GmRVE4d*, which was not tested in the present study. We have removed this term from the Abstract and revised the related text accordingly. In the Introduction, we have added a clarification that soybean flowering and maturity are influenced by both photoperiod and temperature: photoperiod is a primary environmental signal governing key developmental transitions such as floral induction. By contrast, temperature largely modulates developmental rate, affecting how rapidly plants progress through vegetative and reproductive phases under a given photoperiod.

Comments from Reviewer #2:

4. On line 69 in the introduction there is a brief description of flowering time regulation. It is unclear if this section is specific to soybean. The description is very generic and vague. Please explain what is known about how flowering time is regulated in soybean. Just a list of *Arabidopsis* light signaling genes and the GI-CO-FT pathway is not very helpful, in particular because the work presented does not seem to imply any direct light regulation.

Response:

We thank the reviewer for this helpful comment. We agree that the previous section was too generic and was largely based on the general photoperiod flowering pathway described in *Arabidopsis*. We have rewritten this section to focus specifically on soybean. The revised Introduction now describes the roles of the major soybean flowering and maturity loci, including *E1-E4*, the long-juvenile gene *J*, and the downstream florigen genes *GmFT2a* and *GmFT5a*, explaining how these genes contribute to photoperiod sensitivity, flowering time, maturity, and latitudinal adaptation. We have removed the generic list of *Arabidopsis* light-signaling components to better align the Introduction with the scope and findings of our study.

Comments from Reviewer #2:

5. On line 103 the authors describe the previous identification of putative clock related genes. It would be useful to specify the genes with an actual clock function and which ones were discovered as influencers of flowering time. It is important to separate both functions because orthologs of *Arabidopsis* clock genes can have non-clock functions in other species and is important to be specific about this information.

Response:

We thank the reviewer for this important and constructive comment. We agree that the original wording was imprecise and failed to distinguish bona fide clock components from flowering-time regulators that happen to be homologs of clock genes. We have revised this paragraph in the Introduction to make this distinction explicit. The revised text separately discusses genes with evidence for clock or clock-associated rhythmic functions—such as *J*, *GmLHYs*, *LUX1*, *LUX2*, *E2*, and *GmTOC1b*—from genes that are homologous to clock components but have been characterized primarily as regulators of flowering time, maturity, or photoperiod adaptation in soybean, including *Tof11/GmPRR3a*, *Tof12/GmPRR3b/GmPRR37*, *GmEID1*, and *Dgt1/GmRVE1*. This revision avoids implying that all clock-gene homologs function as core clock components in soybean.

Comments from Reviewer #2:

6. Line 158: the authors list tests to analyzed genetic relationships among the accessions studied, please link each to a figure panel or data table.

Response:

We thank the reviewer for this helpful suggestion. We have revised the sentence at line 158 to explicitly link each analytical approach to its corresponding figure: principal component analysis is now cited with Supplementary Fig. 3, population genetic structure analysis with Supplementary Fig. 4, and phylogenetic analysis with Fig. 1b.

Comments from Reviewer #2:

7. Supp. Fig. 16 shows a western blot of RVE4d-GFP, a negative control is absolutely necessary if the authors want to state that the top band is RVE4d-GFP. A size marker is also necessary.

Response:

We thank the reviewer for this important comment. We agree that the identity of the upper band could not be unambiguously assigned to GmRVE4d-GFP without appropriate controls and a molecular-weight reference. We have repeated the experiment and updated Supplementary Fig. 21 to include a negative-control lane processed in parallel and clearly visible molecular-weight markers on the same blot. The figure legend has been revised to indicate the expected molecular weight of the *GmRVE4d*-GFP fusion protein and to identify the band corresponding to this protein. These additions support the assignment of the upper band to GmRVE4d-GFP.

Comments from Reviewer #2:

8. For all the figures showing maturity values please explain what VE-R7 means, the explanations are different (sometimes absent) in the different figures, including supplemental figures. In particular most supplemental figures lack any explanations about the data shown.

Figure legends should provide sufficient information to understand the figures without having to read the main text.

Response:

We thank the reviewer for pointing this out. We agree that the definition of VE–R7 was not presented consistently and that several supplementary figure legends were insufficient for independent interpretation. We have standardized all figure legends reporting maturity-related traits: VE–R7 is now consistently defined as the number of days from emergence (VE) to physiological maturity (R7). We have also expanded all relevant supplementary figure legends to ensure that each figure can be understood without reference to the main text.

Comments from Reviewer #2:

9. Supp. Fig. 3 and Supp. Fig. 4 Have the same title although it seems that represent different type of data. Please provide a clear explanation of how each of the graphs were generated.

Response:

We thank the reviewer for this comment. We confirm that Supplementary Figs. 3 and 4 were inadvertently assigned the same title in the original submission. We have corrected the title of each figure to accurately reflect its distinct content and have substantially revised both legends to explain clearly how each analysis was performed and how the data were generated.

Comments from Reviewer #2:

10. All figure legends should include a brief explanation of how the experiment/analyses used to generate the data shown.

Response:

We thank the reviewer for this suggestion and agree entirely. In the original submission, several figure legends—particularly those for supplementary figures—lacked sufficient methodological context. We have revised all main and supplementary figure legends to include a brief description of the experimental or analytical approach, the data shown, and their interpretation, ensuring that each figure can be understood independently of the main text.

Comments from Reviewer #2:

11. Line 128, for clarity please add a "," after *GmPRR5a*.

Response:

Thank you for pointing this out. We have added a comma after *GmPRR5a* at the indicated position.

Comments from Reviewer #2:

12. Line 228, it would be useful to briefly specify what kind of mutants are the *Gmrve4d* mutants analyzed.

Response:

We have revised the text at this position to specify the nature of the *Gmrve4d* mutants. These are CRISPR/Cas9-generated homozygous deletion mutants in the Jack background: *Gmrve4d-1* carries a 1-bp deletion, *Gmrve4d-2* carries a 5-bp deletion, and *Gmrve4d-3* carries a 33-bp deletion at the respective target sites. Detailed mutation information is provided in Supplementary Fig. 18.

Comments from Reviewer #2:

13. Please use non-capitalized 'p' for p-values (in italics)

Response:

Done. We have revised the entire manuscript to use a lowercase italic *p* for all *p*-values, including the main text, figure legends, and supplementary information.

Comments from Reviewer #2:

14. Line 145: I am not sure what the authors mean by 'advancement of maturity groups', what is 'advancement'?

Response:

We thank the reviewer for flagging this ambiguity. We agree that 'advancement of maturity groups' was unclear. Our intended meaning was that accessions belonging to higher maturity groups showed increasingly delayed maturity. We have revised the sentence to: 'Moreover, maturity was significantly delayed with increasing MG ($p < 0.05$), indicating substantial natural variation in maturity-related traits among soybean accessions from different MGs.'

Comments from Reviewer #2:

15. Fig. 6 model implied that all temperature and light signals regulating flowering time are transmitted via *GmRVE4d*. It would be more useful to integrate these proteins, if possible, with other known components, or simply make a model using the new data: *RVE4d*-*PRR5*-*FT2a*/*FT5a*.

Response:

We thank the reviewer for this important point. We agree that the previous version of Fig. 6 could be misinterpreted as implying that all photoperiod and temperature signals converge exclusively through *GmRVE4d*, which is not supported by our data. We have revised the model accordingly.

In the revised Fig. 6, we removed the broad photoperiod and temperature input elements to avoid overinterpretation. The model now emphasizes the experimentally supported *GmRVE4d-GmPRR5a* regulatory module, clearly distinguishing the directly demonstrated molecular interaction from putative downstream expression-associated relationships. We have retained the putative GmCDF–GmCOL module to place *GmPRR5a* within a broader photoperiod-related flowering regulatory context; however, because the downstream relationships among *GmPRR5a*, *GmCDFs*, *GmCOLs*, *GmFT2a*, and *GmFT5a* were not directly tested in this study, they are shown as dashed lines representing putative associations rather than confirmed direct interactions.

Major comments from Reviewer #3:

1. Population structure and definition of clusters.

The study includes 257 accessions; however, the basis for defining three genetic clusters is not sufficiently supported. The phylogenetic tree suggests additional structure, and the PCA presented in Supplementary Fig. 3 does not clearly resolve three groups. Moreover, the proportion of variance explained by each principal component is not reported.

A more robust characterization of population structure is needed. The authors should justify the number of clusters using appropriate statistical approaches and provide clearer visualization and quantitative support. The lack of clarity in this part, has an impact in the GWAS section.

Response:

We thank the reviewer for this important and constructive suggestion. We have strengthened the population structure analysis and provided clearer justification for three major genetic clusters.

First, we revised Supplementary Fig. 3 to include the percentage of variance explained by each principal component on the corresponding axes: PC1, PC2, and PC3 explain 10.24%, 6.78%, and 4.35% of the total genetic variance, respectively.

Second, we performed a model-based population structure analysis using a linkage disequilibrium-pruned set of approximately independent SNPs, evaluating $K = 2$ to $K = 15$. The cross-validation error was minimized at $K = 3$, statistically supporting the presence of three major genetic groups. The ADMIXTURE results and cross-validation error plot are now provided in Supplementary Fig. 4.

At $K = 3$, the three clusters broadly correspond to major soybean eco-regions in China: Cluster 1 (112 accessions) mainly represented the Northern Spring Planting Region; Cluster 2 (72 accessions) mainly represented the Huang-Huai-Hai Summer Planting Region; and Cluster 3 (73 accessions) mainly represented the Southern Multiple Cropping Region. Accession numbers have been added to the revised Results and Methods sections.

Finally, we have clarified in the GWAS section that population stratification was controlled by incorporating population structure covariates together with a genomic relationship matrix in the mixed linear model framework.

Major comments from Reviewer #3:

2. GWAS methodology and model specification

The GWAS was conducted using ~3 million SNPs, but it is unclear whether LD pruning was applied to reduce redundancy. This raises concerns about both statistical power and false-positive associations.

In addition, the mixed linear model appears insufficiently controlled for population structure and kinship – poor adjustment depicted in the QQ-plots (GAPIT itself allows to check for critical K value).

It would also strengthen the analysis to:

- Report the proportion of phenotypic variance explained by associated loci
- Provide minor allele frequencies for significant SNPs

Further, key details regarding phenotypic data are missing: number of biological replicates (plants per accession per trial), statistical treatment of phenotypes prior to GWAS.

Response:

We thank the reviewer for this comprehensive comment. We have revised the Methods section to address all of these points.

Regarding linkage disequilibrium pruning: we did not apply linkage disequilibrium pruning to the full SNP set used for GWAS, as pruning may remove informative markers and reduce mapping resolution. linkage disequilibrium pruning (PLINK --indep-pairwise 50 10 0.2) was applied only for population structure analyses, including PCA, ADMIXTURE, and phylogenetic analysis.

Regarding model specification: GWAS was performed in GAPIT v3 using a mixed linear model with the first three principal components as fixed-effect covariates and the genomic relationship matrix as a random effect. We re-examined the quantile–quantile plots to assess the adequacy of model correction and report the updated results.

Regarding variance explained and minor allele frequency: we have added the minor allele frequency and the proportion of phenotypic variance explained for each significant lead SNP to Supplementary Data 3.

Regarding phenotypic data: we have added the following clarification to the Methods: 'For the phenotypic data used in GWAS, two biological replicates were evaluated per accession per trial, with five plants per replicate. Individual plant values were averaged to obtain accession-level means, which were used for association analysis.'

Major comments from Reviewer #3:

3. Evidence for selection at the GmRVE4d locus

L130: “nucleotide polymorphisms in the promoter region of GmRVE4d have undergone strong artificial selection during domestication”. The inference is based on F_{ST} and nucleotide diversity (π) calculated over the promoter region only, while the apparent selective sweep extends beyond this interval – it needs to be detailed that the calculations are done per site.

To substantiate this claim, the authors should use other suitable selection tests. One simple option is to use Tajima’s D across the entire LD block; other options could be π , or EHH.

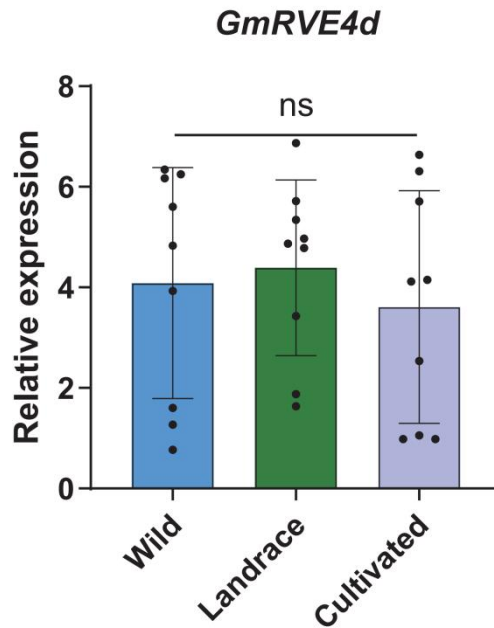
If selection is indeed acting on the promoter, functional evidence should be provided, such as differences in GmRVE4d expression among wild, landrace, and cultivated accessions.

Response:

We thank the reviewer for this insightful suggestion. We have substantially revised this analysis.

First, we redefined the analysis window by characterizing the full linkage disequilibrium block surrounding *Glyma.19G178000*, which spans approximately 200 kb. We performed selection scans across this entire block using Tajima’s D , F_{ST} , and nucleotide diversity π , while retaining per-site F_{ST} and π analyses for the 2.7-kb promoter region to describe local patterns. These analyses revealed elevated genetic differentiation between wild and domesticated soybeans and reduced nucleotide diversity in landraces and cultivated soybeans relative to wild soybean, supporting domestication-associated selection at the *Glyma.19G178000* genomic region. We have revised the text to avoid implying that selection acted exclusively on the promoter.

Second, following the reviewer’s suggestion, we examined *GmRVE4d* transcript levels by qRT-PCR in trifoliolate leaves from representative wild, landrace, and cultivated accessions under short-day conditions at 13 days after emergence. No significant difference in expression was detected among the three groups (Response Fig. 3). We therefore do not infer that the promoter polymorphisms alter basal *GmRVE4d* expression under this condition, and have revised the manuscript accordingly to present the promoter variants as components of a domestication-selected haplotype rather than as expression-altering mutations.



Response Fig. 3 *GmRVE4d* expression in wild, landrace, and cultivated soybean under short-day conditions. Relative expression of *GmRVE4d* in trifoliolate leaves of wild soybean, landrace, and cultivated soybean grown under short-day conditions. Three independent accessions were selected for each group: wild soybean, HeiheSoja, BayanSoja and GuangxiSoja; landrace, Changjihuangdou, Miqianhuangdou and Fengchengniupidou; cultivated soybean, Lonoke, Rhodes and Dillon. Gene expression was quantified by qRT-PCR and normalized to the internal reference gene *GmActin*. Each accession was analyzed with three independent biological replicates. Pairwise statistical comparisons between groups were performed using two-tailed Student's *t*-tests; “ns” denotes no significant difference, $p \geq 0.05$.

Major comments from Reviewer #3:

4. Haplotype structure and domestication history

The manuscript would benefit from a clearer analysis of haplotype evolution. Specifically:

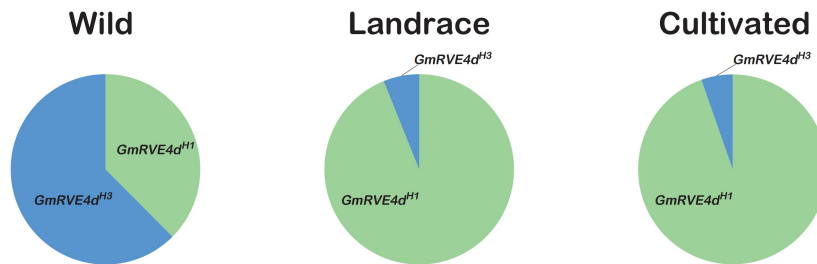
- (1) Are haplotypes H1 and/or H3 present in wild accessions?
- (2) Can the ancestral haplotype be inferred?

These analyses would provide important context for interpreting domestication-related selection.

Response:

We thank the reviewer for this helpful suggestion. We clarified the haplotype distribution and domestication context using the SoyOmics database (<https://ngdc.cncb.ac.cn/soyomics/index>). Both haplotypes are present in wild soybean: *GmRVE4d*^{H1} was detected in 66 wild accessions and *GmRVE4d*^{H3} in 110 wild accessions. In landraces, *GmRVE4d*^{H1} and *GmRVE4d*^{H3} were detected in 1,936 and 126 accessions, respectively; in cultivated accessions, in 3,272 and 186 accessions, respectively (Response Fig. 4).

Because both haplotypes occur in wild soybean, the ancestral state cannot be unambiguously inferred from presence–absence patterns alone. However, the higher relative frequency of *GmRVE4d^{H3}* in wild soybean and the marked enrichment of *GmRVE4d^{H1}* from wild soybean through landraces to cultivated accessions are consistent with domestication favoring the *GmRVE4d^{H1}* haplotype. We have revised the manuscript accordingly and avoided overstatement regarding ancestral-state inference.



Response Fig. 4 Distribution of *GmRVE4d* haplotypes in wild, landrace and cultivated soybean. Pie charts showing the relative frequencies of the two major *GmRVE4d* haplotypes, *GmRVE4d^{H1}* and *GmRVE4d^{H3}*, in wild, landrace and cultivated populations. In the wild soybean population, the proportion of *GmRVE4d^{H3}* is higher than that of *GmRVE4d^{H1}*; in both the landrace and cultivated soybean populations, the proportion of *GmRVE4d^{H1}* is significantly higher than that of *GmRVE4d^{H3}*, indicating that *GmRVE4d^{H1}* was strongly enriched in domesticated and improved populations.

Major comments from Reviewer #3:

5. LD and diversity patterns

The visualization of diversity and LD patterns requires improvement:

In Fig. 2F, π values in wild accessions dominate the scale, obscuring variation in landraces and cultivars. Alternative visualization strategies should be considered.

A comparison of LD structure between wild and domesticated groups would help support the proposed selection scenario..

Response:

We thank the reviewer for these helpful suggestions. We have made two improvements.

First, we replaced the original nucleotide diversity plot in Fig. 2f with an improved visualization in which the higher π values in wild soybean no longer dominate the scale, allowing clear comparison of π patterns among wild soybean, landraces, and cultivated soybeans. The updated result is shown in Fig. 2g.

Second, we added linkage disequilibrium decay analyses for wild soybean, landraces, cultivated soybeans, and the total panel. As expected under domestication, wild soybean showed the fastest linkage disequilibrium decay, while landraces and cultivated soybeans showed progressively

slower decay, consistent with increased linkage disequilibrium in domesticated populations. These results are provided in Supplementary Fig. 5.

Major comments from Reviewer #3:

6. RNA-seq and differential expression analyses

The manuscript lacks essential methodological details for RNA-seq and differential gene expression analyses (L589). Information on experimental design, replication, data processing, normalization, and statistical testing must be included to ensure reproducibility.

Response:

We thank the reviewer for this important comment. We have substantially revised the RNA-seq and differential expression analysis section in the Methods to include all details required for reproducibility. Specifically, we now describe the experimental design, biological replication, read quality control and trimming, reference genome alignment, gene-level read counting, normalization strategy, statistical testing procedure, multiple-testing correction, and the criteria used to define differentially expressed genes. These revisions make the RNA-seq analyses fully transparent and reproducible.

Major comments from Reviewer #3:

7. Consideration of additional candidate genes

Given its annotation as a transcription factor involved in flowering and reproductive system development, the potential role of *Glyma.19G178200* could be also addressed, at least theoretically. Does it exhibit signatures of selection comparable to *GmRVE4d*? are there different haplotypes in different geographic groups? Is it homologous to any other relevant flowering regulator?

Response:

We thank the reviewer for this helpful suggestion. We have now explicitly considered *Glyma.19G178200* in the revised manuscript.

Glyma.19G178200 showed time-dependent expression under both short-day and long-day conditions and was predicted by DeepGOPlus to be associated with photoperiod response and flowering regulation, similar to *Glyma.19G178000*. We therefore included it in the candidate-gene prioritization process.

However, unlike *Glyma.19G178000*, we did not detect nonsynonymous coding variants in *Glyma.19G178200* among the surveyed accessions. Consequently, this gene does not define protein-altering haplotypes comparable to those of *Glyma.19G178000*, and no haplotype–phenotype association or geographic distribution pattern analogous to that of

GmRVE4d was observed. These results support *Glyma.19G178000/GmRVE4d* as the stronger candidate gene at this locus. The expression profile, functional prediction, and coding-sequence variation of *Glyma.19G178200* have been incorporated into the revised Results section and supplementary figures (Supplementary Figs. 8, 9, and Supplementary Data 5).

Minor comments from Reviewer #3:

1. The genomic analyses are generally described at a relatively superficial level. A more detailed and integrative discussion would improve accessibility and interpretability.

Response:

We thank the reviewer for this constructive suggestion. We have revised the genomic analysis sections to provide more detailed and integrated descriptions of resequencing and variant calling, population structure and linkage disequilibrium analyses, and their relationship to the multi-environment GWAS and linkage disequilibrium-based candidate interval refinement on chromosome 19. We have also integrated expression profiling, functional prediction, haplotype–phenotype and haplotype–geography analyses, and population-genetic selection scans into a coherent narrative supporting *Glyma.19G178000/GmRVE4d* as the leading candidate gene. Revised text, additional supplementary figures, and supplementary data have been included accordingly.

Minor comments from Reviewer #3:

2. PCA plots should include the percentage of variance explained by each component.

Response:

We agree and have revised the PCA plot to include the proportion of variance explained by each component on the corresponding axes. PC1, PC2, and PC3 now explain 10.24%, 6.78%, and 4.35% of the total genetic variance, respectively, as indicated on the updated axis labels. The figure legend has been updated accordingly.

Minor comments from Reviewer #3:

3. Figure annotations and scaling should be improved for clarity.

Response:

We thank the reviewer for this suggestion. We have revised figure annotations, axis labels, legends, and scaling throughout the manuscript to improve clarity and readability. In particular, we optimized the visualization of nucleotide diversity and linkage disequilibrium-related results to ensure that differences among wild soybean, landraces, and cultivated soybeans are clearly distinguishable.

Minor comments from Reviewer #3:

4. Suppl. Fig 4 is incorrectly.

Response:

We thank the reviewer for noting this error. Supplementary Fig. 4 was inadvertently duplicated from Supplementary Fig. 3 during figure assembly. We have replaced it with the correct figure and updated the corresponding legend in the revised Supplementary Information.

Major comments from Reviewer #4:

1. Statistical analysis: how was the candidate gene GmRVE4d identified?

First of all, it was described in Methods that the threshold of GWAS was $1/n$, where n is the total number of SNPs (line 509). This means, if we consider Bonferroni correction for multiple testing (which is a very common method applied in many GWAS studies), the family-wise error rate would be one. In other words, the results are almost sure to contain at least one false-positive. Of course, different thresholds (more stringent or more liberal ones) can be applied. But the threshold chosen by the authors seems to be too liberal.

Secondly, it is not clear whether the locus on Chr. 19 is the only significant signal across all three environments. If not (very likely there is a peak on Chr. 6 across environments, from Fig. 1c), why did the authors only focus on this one?

Thirdly, it is not clear how the QTL region was determined. It was only stated “an 84.1-kb interval (43.672 – 43.756 Mb)” in the text (line 180). The procedure from marker-trait associations to QTL region or interval should be clearly specified.

I do not doubt the experimental results showed later in the manuscript about the candidate gene GmRVE4d. But we know that usually it is a big challenge to search the “right” candidate genes from GWAS results. It seems that the authors did it in a very successful way that they landed at only one candidate gene. Thus, I would strongly suggest that more details are provided.

Response:

We thank the reviewer for this important and constructive comment. We have substantially revised the Results and Methods sections to provide a more transparent and reproducible description of the GWAS threshold, candidate locus selection, and interval definition.

First, we have clarified that the threshold of $p < 1/n$ (corresponding to $-\log_{10}(p) \geq 6$ in our dataset) was used as a suggestive threshold for initial marker–trait association screening, not as a strict Bonferroni-corrected genome-wide significance threshold. We acknowledge that this threshold is relatively liberal if interpreted in isolation. Therefore, we explicitly state in the revised manuscript that candidate loci were prioritized not solely on the basis of exceeding this threshold, but on

reproducibility across environments, local linkage disequilibrium support, and subsequent biological evidence.

Second, we have clarified that the chromosome 19 locus was not the only association signal detected. Additional peaks were observed on other chromosomes, including chromosome 6; however, these were not consistently detected at comparable genomic positions across all three environments. By contrast, the chromosome 19 locus showed a reproducible association peak at a consistent position in all three environments and was therefore selected for further candidate-gene analysis.

Third, we have provided a clear explanation of how the candidate interval was determined. Based on the estimated genome-wide linkage disequilibrium decay distance of approximately 50 kb, we examined the local linkage disequilibrium block surrounding the lead SNPs on chromosome 19 (Chr19:43,721,425; Chr19:43,687,363; Chr19:43,684,936). These SNPs co-localized within a single linkage disequilibrium block spanning an 84.1-kb interval (43.672–43.756 Mb) containing nine annotated genes. Among these, *Glyma.19G178000/GmRVE4d* was prioritized on the basis of functional annotation, time-dependent expression, haplotype–phenotype association, geographic haplotype distribution, and selection signatures.

Major comments from Reviewer #4:

2. Inconsistencies between some phenotypic data.

Looking at Fig. 1a and Fig. 2d, I am a bit confused as the phenotypic range of the accessions in BJ2020 is clearly different. In Fig. 1a, all accessions were shown together and the upper-bound should be less than 125. However, in Fig. 2d, accessions were shown in two subgroups, one carrying Haplotype 1 and the other carrying haplotype 3. Many accessions in the second group have a value much larger than 125 (even exceeds 150). Why did this happen?

The performance of “Jack” as shown in Supplementary Fig. 14 and 17 is also not consistent. In both figures, days from VE to R7 stage were shown for “Jack” under controlled long-day as well as short-day conditions (In one figure, it was compared with the *Gmrve4d* mutants and in the other it was contrasted with the over-expression lines). But the performance of this cultivar is quite different in these two figures. For example, in Supplementary Fig. 14, the maturity of “Jack” plants was between 80 and 85 days, whereas in Supplementary Fig. 17 it was between 66 and 70 days. Were these plants really grown in the same controlled conditions?

Response:

We thank the reviewer for carefully identifying these inconsistencies. We have investigated and resolved both issues.

Regarding the discrepancy between Fig. 1a and Fig. 2d: this was caused by inconsistent treatment of late-maturing accessions in the BJ2020 dataset. Several accessions did not reach R7 before the first frost. In Fig. 1a, these accessions were excluded and only directly observed maturity records were used. However, in the original haplotype analysis shown in Fig. 2d, we inadvertently included estimated maturity dates for these late-maturing accessions, causing some values to exceed the range shown in Fig. 1a. We have now removed all estimated maturity values and applied directly observed records consistently throughout. Fig. 2d and the associated statistical analysis have been updated accordingly. Importantly, the haplotype-associated phenotypic trend is unchanged after this correction.

Regarding the difference in Jack's performance between Supplementary Figs. 14 and 17: the mutant and overexpression materials were evaluated in two independent controlled-environment experiments conducted at different times, as these materials were generated sequentially. Although the same nominal photoperiod settings were applied, batch effects between independent growth-chamber experiments can influence absolute VE–R7 values. Critically, in each experiment, the transgenic lines and the Jack control were grown simultaneously in the same chamber and batch; all conclusions are therefore based on within-experiment comparisons between each transgenic line and its contemporaneous Jack control, not on cross-experiment comparisons of absolute values. We have clarified this point in the revised Methods and figure legends.

Minor comments from Reviewer #4:

1. Lines 140-142: Since there are only three environments, I would suggest to make the statement less general, like “The results revealed significant differences in maturity phenotypes between the two locations, but no significant differences between the two years at the location Beijing”.

Response:

We agree. The revised text now reads: 'The results revealed significant differences in maturity phenotypes between the two locations, but no significant differences between the two years at the Beijing location.'

Minor comments from Reviewer #4:

2. Line 168: The critical value 0.4 written here in the text is not consistent with what was shown in Supplementary Fig. 5 (indicated by the dashed line). Please check. In addition, I would also suggest to write directly “... declines to half of its maximum value” as described in Methods (line 501).

Response:

Thank you for identifying this inconsistency. We have re-examined and re-plotted the linkage disequilibrium decay figure. The linkage disequilibrium decay distance is now consistently

defined as the physical distance at which linkage disequilibrium declines to half of its maximum value, and this definition is used uniformly in the text, Methods, and Supplementary Fig. 5.

Minor comments from Reviewer #4:

3. Lines 191-192: Looking at Fig. 1e and 1f, I would not be so convinced about the circadian rhythmic pattern. The peak expression occurred both during the day and the night. Considering that the figure only showed results for two consecutive days (48 hours), the result is not completely convincing.

Response:

We thank the reviewer for this important comment. We agree that 48-hour time-course data are insufficient to confirm a circadian rhythmic pattern, particularly when peak expression is observed at both daytime and nighttime time points. We have replaced 'circadian rhythmic pattern' with 'time-dependent expression changes' throughout the manuscript. The revised text now states that *GmRVE4d* and *GmPRR5a* exhibited temporal variation in expression over the 48-hour sampling period, without claiming confirmed circadian regulation.

Minor comments from Reviewer #4:

4. Lines 220-223: The number of cultivated soybeans (214) is much larger than that of wild soybeans (10), according to the description in Methods (lines 444-446). Are there possibly biases in the comparison of the molecular diversities of the two subpopulations due to the sample sizes (because the frequencies of the sequences are relevant in the calculation of π value)?

Response:

We thank the reviewer for raising this important methodological concern. We agree that the unequal sample sizes between wild and cultivated soybean accessions may affect the precision of nucleotide diversity (π) estimates, particularly given the very small wild soybean sample. We have therefore avoided interpreting the absolute difference in π as independent evidence for selection. In the revised Discussion, we explicitly acknowledge this limitation and clarify that the inference of domestication-associated selection at the *GmRVE4d* region is based on multiple concordant lines of evidence—elevated F_{ST} , Tajima's D , local haplotype structure, haplotype–phenotype associations, and geographic haplotype differentiation—rather than on π alone.

Minor comments from Reviewer #4:

5. Line 228: Which haplotype of *GmRVE4d* did the cultivar “Jack” carry, H1 or H3? Please specify here.

Response:

Thank you for this helpful comment. The cultivar Jack, used for all functional validation experiments in this study, carries the *GmRVE4d^{HI}* haplotype. We have now added this information explicitly to the revised manuscript.

Minor comments from Reviewer #4:

6. Lines 273-274: The occurrence of GmPRR5a in the interpretation here is very confusing as it is only revealed in the next subsection that GmRVE4d inhibits the transcription of GmPRR5a.

Response:

Thank you for this observation. We agree that introducing *GmPRR5a* at this point was premature and disrupted the logical flow of the text. We have deleted the sentence referring to *GmPRR5a* at lines 273–274 and revised the surrounding text to maintain coherence.

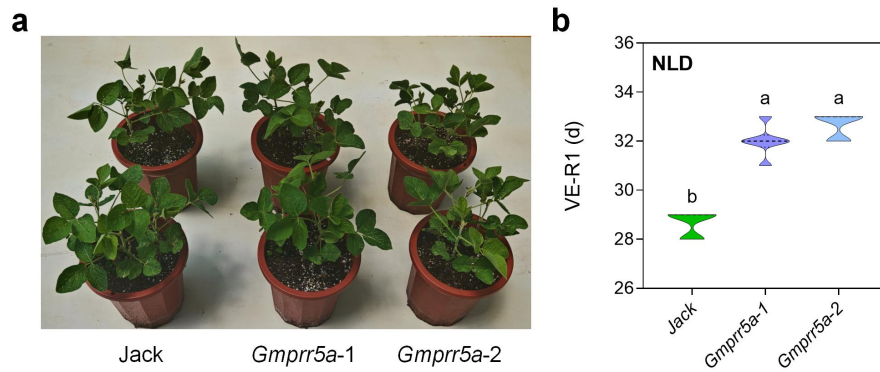
Minor comments from Reviewer #4:

7. Lines 333-336: There were only results under natural short-day conditions, but no experimental results under natural long-day conditions?

Response:

We thank the reviewer for this important suggestion. We agree that experiments under natural long-day conditions would further strengthen the environmental interpretation of our results. In our location, natural long-day conditions require spring sowing, and the spring field season commenced after the initial manuscript submission.

We have completed the flowering time phenotyping of a spring-sown pot experiment under natural long-day (NLD) conditions in Beijing (Response Fig. 5). Seeds of Jack and two independent *Gmprr5a* mutant lines (*Gmprr5a-1* and *Gmprr5a-2*) were sown on 1 May, and seedlings emerged on 5 May. Under NLD conditions, both mutant lines flowered significantly 3–4 days later than Jack (one-way ANOVA followed by Tukey's test, $p < 0.05$). Because the plants require more than 100 days from emergence to physiological maturity at this location, complete maturity data could not be obtained within the current revision timeline and will be reported in future work.



Response Fig. 5 Flowering time phenotypes of *Gmprr5a* mutants under natural long-day conditions in Beijing. **a** Representative images of Jack and two independent *Gmprr5a* mutant lines, *Gmprr5a-1* and *Gmprr5a-2*, grown in pots under natural long-day (NLD) conditions in Beijing, China. Seeds were sown on 1 May, and seedlings emerged on 5 May. For each genotype, two pots were planted, and seedlings were thinned to three plants per pot after emergence. **b** Flowering time, measured as days from emergence (VE) to beginning bloom (R1), under NLD conditions. $n = 6$ plants. Different lowercase letters indicate significant differences among genotypes, as determined by one-way ANOVA followed by Tukey's test, $p < 0.05$.

Minor comments from Reviewer #4:

8. Line 505: Please also cite the original paper of GAPIT version 3 (Wang and Zhang 2021): <https://doi.org/10.1016/j.gpb.2021.08.005>.

Response:

Thank you for the suggestion. We have added the original reference for GAPIT version 3 (Wang and Zhang, 2021) to the Methods section and included it in the reference list.

Minor comments from Reviewer #4:

9. Line 513: It seems that the reference cited here (Yang et al. 2023) is only a study applying the tool DeepGOPlus. The original study developed the tool is Kulmanov and Hoehndorf (2020) and it should be cited: <https://doi.org/10.1093/bioinformatics/btz595>.

Response:

Thank you for this correction. We have added the primary reference (Kulmanov and Hoehndorf, 2020) to the Methods section and included it in the reference list. The previously cited application study (Yang et al., 2023) has been retained alongside it.

Minor comments from Reviewer #4:

10. Fig. 1d: It is impossible to see the labels under the x-axis.

Response:

Thank you for pointing this out. We have revised Fig. 1d to increase the x-axis label font size and optimized the figure layout so that all x-axis labels are now clearly legible.

Minor comments from Reviewer #4:

11. Supplementary Fig. 3: Please report the proportion of variance explained by the three PCs.

Response:

We have revised Supplementary Fig. 3 to include the proportion of variance explained by each principal component on the axis labels: PC1 = 10.24%, PC2 = 6.78%, and PC3 = 4.35%. The figure legend has been updated accordingly.

Minor comments from Reviewer #4:

12. Supplementary Fig. 4: The caption of the figure is the same as Supplementary Fig. 3.

Response:

We thank the reviewer for noting this error. Supplementary Fig. 4 was inadvertently duplicated from Supplementary Fig. 3 during figure assembly. We have replaced it with the correct figure and updated the legend in the revised Supplementary Information. This correction does not affect any analyses, results, or conclusions.

Minor comments from Reviewer #4:

13. Please check throughout the manuscript whether the Italic font has been applied to all gene names.

Response:

Thank you for the suggestion. We have carefully reviewed the entire manuscript and ensured that all gene names—including *GmRVE4d*, *GmPRR5a*, and all other gene symbols—are formatted in italics consistently throughout the main text, figure legends, and supplementary information.

Responses to Reviewers

Comments from Reviewer #2:

1. All methods description should enable fully reproducibility of the experiments. For example, please indicate the extraction buffer used and the exact Western blot protocol (for example: amount of protein loaded, gel type/concentration, membrane used, blocking agent used, exact antibody etc.). The authors should also indicate how many times this experiment was conducted. All protocols should enable exact reproduction of the experiment or provide a reference to a reproducible procedure. Please make sure to do that for all the methods provided (eg, RNA extraction, DAP profiling, EMSA...).

Response:

We thank the reviewer for this constructive comment. We have thoroughly revised the Methods section to improve reproducibility. Specifically, we added detailed information for the Western blot assay, including the extraction buffer composition, amount of tissue used, protein loading amount, denaturation conditions, gel concentration, membrane type, blocking agent, antibody information with catalog numbers and dilutions, washing conditions, ECL substrate, imaging system, molecular weight marker, and the number of independent biological replicates.

We also revised the descriptions of the other experimental and computational procedures, including DNA extraction and sequencing, RNA extraction, qRT-PCR, RNA-seq, DAP-seq, EMSA, ChIP-qPCR, yeast one-hybrid assays, subcellular localization, dual-luciferase reporter assays, and statistical analyses. We added sample amounts, replicate information, reagent and kit information, software versions, reaction conditions, and references to published protocols or manufacturer's instructions where appropriate. These revisions ensure that the experiments and analyses can be reproduced more precisely.

Comments from Reviewer #2:

2. The phylogenetic trees are insufficient. All gene IDs need to be provided in the figure or legend. For GmRVE4d, the tree should demonstrate that the gene they studied is closely related to Arabidopsis RVE4, that means it is necessary to add all the other RVE genes (and CCA1/LHY) as well. Please add all the current gene names (if available) as well. The tree should include all the 'RVE' like genes identified in soybean, Arabidopsis and the other species shown in the current tree. Please address the same issues on the figure showing the GmPRR5 phylogenetic tree. The legends for these figures should indicate the methods used for alignment and estimating the phylogenetic tree, as well as what sequences were used (CDS, peptide sequence...). The trees should also include the bootstrap confidence levels.

Response:

Thank you for this helpful suggestion. We have reconstructed the GmRVE4d and GmPRR5a phylogenetic trees and revised the corresponding figure legends.

For the GmRVE4d tree, we included all identifiable RVE-like proteins from soybean (*Glycine max*), Arabidopsis (*Arabidopsis thaliana*), wheat (*Triticum aestivum*), maize (*Zea mays*), and rice (*Oryza sativa*), together with Arabidopsis CCA1 and LHY. For the GmPRR5a tree, all identifiable PRR-like proteins from these five species were included. Homologues were identified from complete proteomes through orthogroup inference using OrthoFinder v3 with default parameters. GmRVE4d^{H1} and GmRVE4d^{H3} cluster with Arabidopsis RVE4, whereas GmPRR5a clusters with Arabidopsis PRR5.

For both analyses, full-length protein sequences were aligned using MAFFT, and maximum-likelihood trees were constructed using IQ-TREE 2 with the parameters "-m MFP -bb 1000 -alrt 1000." Branch support was evaluated using 1,000 ultrafast bootstrap and 1,000 SH-aLRT replicates. All database protein IDs are shown in the trees, and previously published or officially assigned gene names are provided where available. Proteins without published or officially assigned gene names are labeled using their database protein IDs only. The target proteins are highlighted in light blue, and branch colors represent different species. These methodological details and the meaning of the node symbols have been added to the revised figure legends.

Comments from Reviewer #2:

3. Supplementary Fig. 21. Instead of indicating the protein shown, please indicate the antibody used for each panel.

Response:

Thank you for this helpful suggestion. We have revised the relevant figure (formerly Supplementary Fig. 21, now Supplementary Fig. 22 in the revised manuscript) to label each panel according to the antibody used rather than the protein detected. Specifically, the upper panel is now labeled "anti-GFP," and the lower panel is labeled "anti-β-actin." The corresponding figure legend has also been revised accordingly.

Comments from Reviewer #2:

4. In Line 77 the authors write: "E1 acts as a soybean-specific flowering repressor by suppressing the florigen genes GmFT2a and GmFT5a, whereas E2/GmGIa, E3/GmPhyA3, E4/GmPhyA2, and J/GmELF3 connect photoperiod perception, circadian regulation, and downstream flowering responses. These regulators converge on the conserved GI-CO-FT module, in which GmFT2a and GmFT5a function as major florigens that promote floral transition under inductive conditions."

The sentence "These regulators converge..." does not make sense, since 'these regulators' include the FT and GI genes in soybean but not CO. It might be more useful to the reader to actually just explain what they mean by 'converge'.

Response:

Thank you for pointing this out. We agree that the original wording was ambiguous, especially because the term "these regulators" could be interpreted as including both upstream photoperiod/circadian regulators and the downstream florigen genes themselves. Our intention was not to imply that all of these genes are components of the *GI-CO-FT* module, but rather to indicate that photoperiod- and circadian-related regulators ultimately affect flowering by modulating the expression of the florigen genes *GmFT2a* and *GmFT5a*. We have therefore removed the vague term "converge" and revised the text as follows:

"*E1* acts as a soybean-specific flowering repressor by suppressing the florigen genes *GmFT2a* and *GmFT5a*, whereas *E2/GmGIa*, *E3/GmPhyA3*, *E4/GmPhyA2*, and *J/GmELF3* link photoperiod perception and circadian regulation with downstream flowering responses. These pathways ultimately regulate the florigen genes *GmFT2a* and *GmFT5a*, which function as key downstream outputs of the photoperiodic flowering network and promote floral transition under inductive conditions."

Major comments from Reviewer #3:

1. The population structure analysis is generally well presented; however, the rationale for selecting $K = 3$ as the primary representation of genetic structure is not entirely convincing. The cross-validation error does not appear to identify a clear optimum at $K = 3$ and instead suggests additional hierarchical structure within the population. It would therefore be informative to include ADMIXTURE plots for a range of K values (e.g., $K = 2-10$) in the Supplementary Information. Presenting multiple K values is common practice in population genomic studies and would allow readers to better assess the complexity of the inferred population structure and the robustness of the chosen clustering.

Response:

We thank the reviewer for this valuable suggestion. In response, we have expanded Supplementary Fig. 4 to include ADMIXTURE results across a broader range of K values. The revised figure presents the cross-validation error and the corresponding ADMIXTURE bar plots for $K = 2-11$. The cross-validation error progressively decreased from $K = 2$ to $K = 10$, reached its minimum at $K = 10$, and then increased slightly at $K = 11$. In addition, the ADMIXTURE bar plots show progressively finer genetic subdivisions and admixture patterns as K increases, indicating hierarchical complexity in the population structure.

We agree that $K = 3$ should not be interpreted as the statistically optimal or uniquely supported estimate of the number of ancestral populations. We have therefore revised the manuscript to remove wording that characterized $K = 3$ as the optimal solution. $K = 3$ is retained in the main figure as a simplified visualization of the broad-scale population structure rather than as a formal estimate of the true number of ancestral populations. The complete results for $K = 2$ –11 are now presented in Supplementary Fig. 4 to illustrate the hierarchical and admixed nature of the population structure. Together, these revisions enable readers to assess the complexity and robustness of the inferred population structure without overinterpreting any single K value.

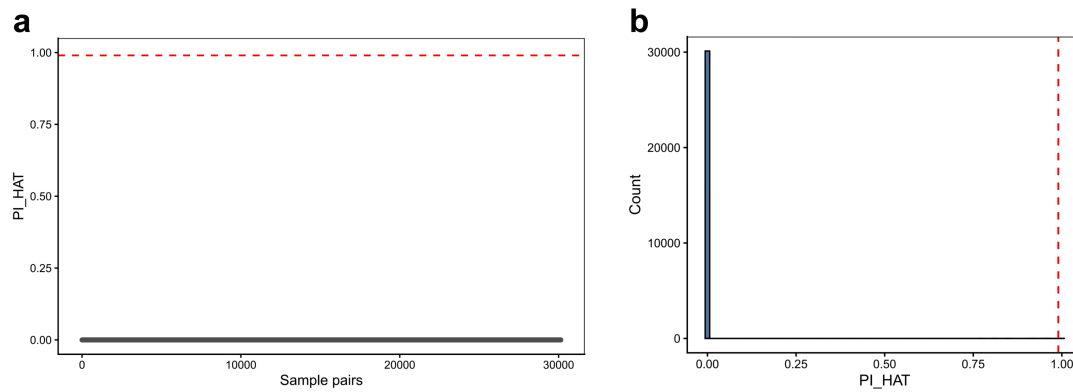
Major comments from Reviewer #3:

2. In addition, it is unclear whether highly related individuals were identified and removed prior to the population genomic and GWAS analyses. Sequencing panels often contain duplicated or closely related accessions that can bias estimates of population structure, linkage disequilibrium, and association statistics if not properly accounted for. The authors should clarify whether relatedness filtering was performed and, if so, describe the criteria used to exclude closely related samples.

Response:

We thank the reviewer for this helpful comment. We agree that duplicated or closely related accessions may bias estimates of population structure, linkage disequilibrium, and GWAS statistics. Therefore, before conducting the population genomic and GWAS analyses, we performed a pairwise identity-by-descent (IBD) analysis among all 257 accessions to identify potential duplicates or closely related individuals. Pairwise PI_HAT values were calculated using LD-pruned SNPs. The PI_HAT values were concentrated near zero, well below the PI_HAT = 0.99 threshold typically used to flag duplicate or genetically identical samples, indicating an overall low level of genetic relatedness among the accessions (Response Fig. 1).

In addition, we manually checked the accession identifiers, source information, and sequencing records and found no evidence of sample duplication. Taken together, these results indicate that the dataset contains no duplicated, nearly identical, or highly related accessions that would warrant exclusion from the subsequent analyses. Therefore, no accession was removed on the basis of duplication or close relatedness.



Response Fig. 1. Pairwise identity-by-descent assessment of relatedness among 257 soybean accessions. Pairwise PI_HAT values were calculated for all sample pairs using LD-pruned SNPs. The red dashed line indicates the reference value expected for duplicated or genetically identical samples (PI_HAT = 0.99).

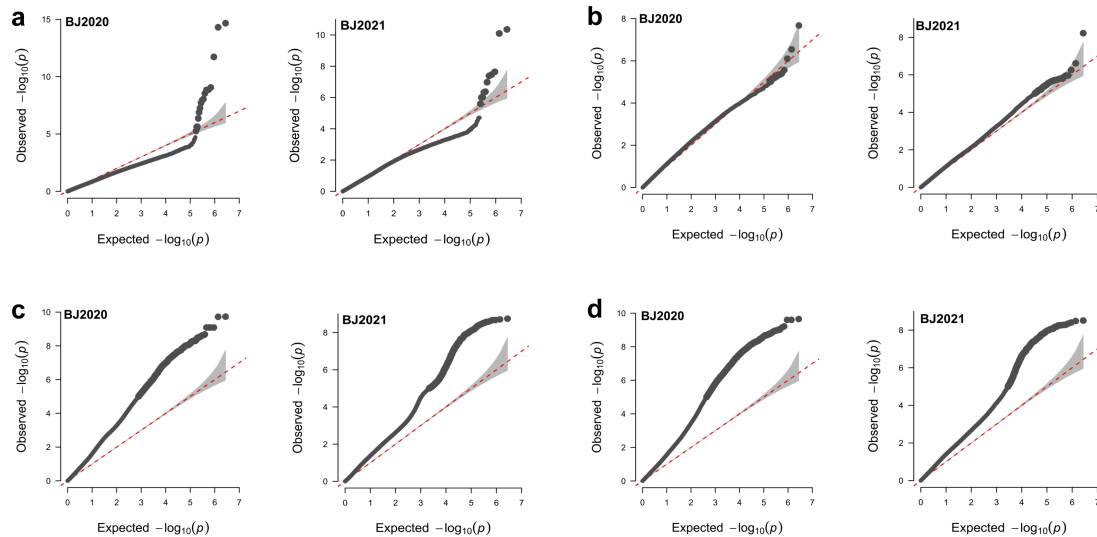
Major comments from Reviewer #3:

3. Inspection of the Q-Q plots suggests that the GWAS models are not consistently well calibrated across the three experimental conditions. Rather than following the null expectation and deviating only in the extreme tail, the observed p-values show an early and systematic departure from the expected distribution, indicating residual inflation or other sources of confounding that are not fully accounted for by the current model. This suggests that additional efforts to control population structure or relatedness, or the evaluation of alternative GWAS models (e.g., FarmCPU or BLINK), may be needed.

Response:

We thank the reviewer for raising this important point. To evaluate whether alternative GWAS models could improve model calibration, we have reanalyzed the GWAS data using four alternative models: FarmCPU, BLINK, SUPER, and GLM. For each model, we generated Q-Q plots to assess the agreement between the observed and expected distributions of association *p*-values (Response Fig. 2). However, none of the four alternative models fully eliminated the systematic deviation from the null expectation in BJ2020 or BJ2021, and none showed consistently improved calibration compared with the MLM model.

We agree that these results suggest that some residual inflation or confounding may remain and cannot be completely addressed simply by changing the GWAS model. Nevertheless, because MLM explicitly accounts for both population structure and kinship and performed no worse than the alternative models evaluated here, we have retained MLM for the GWAS analyses presented in the main text.



Response Fig. 2 Comparison of Q-Q plots obtained using different GWAS models for BJ2020 and BJ2021. Quantile–quantile plots of GWAS results generated using four models for experiments conducted in Beijing, China, in 2020 (BJ2020) and 2021 (BJ2021): **a** FarmCPU, **b** BLINK, **c** SUPER, and **d** GLM. The observed $-\log_{10}(p)$ values are plotted against the expected $-\log_{10}(p)$ values. The dashed diagonal line represents the expected distribution under the null hypothesis of no association.

Major comments from Reviewer #3:

4. Among the three analyses, the GWAS performed in Beijing (both years) appears to be better calibrated, although some inflation is still evident. In contrast, the third experimental condition exhibits a pronounced early deviation from the null expectation, raising concerns that the association statistics may be substantially influenced by confounding rather than true genotype–phenotype associations. Consequently, I would interpret the GWAS results from this dataset with considerable caution, as the quality control metrics suggest that the model does not adequately account for the underlying structure of the data.

Response:

We thank the reviewer for raising this important concern. We agree that the Q-Q plot for the XX2021 dataset showed a pronounced early deviation from the null expectation, indicating that the association statistics from this environment may have been affected by residual confounding effects or environmental noise.

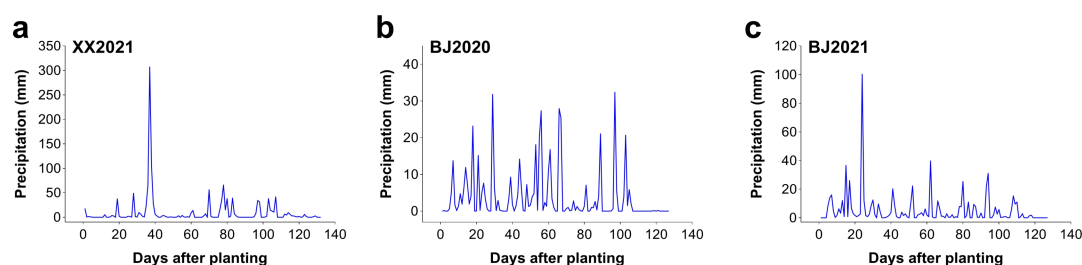
To further investigate potential environment-specific factors contributing to the abnormal distribution of the association statistics, we have re-examined the field records and meteorological data for all three experimental environments. Daily precipitation data were downloaded from the Climate Data Online platform of the NOAA National Centers for Environmental Information (NOAA NCEI CDO; <https://www.ncei.noaa.gov/cdo-web/>). The downloaded data covered the

period from sowing to the first frost for each environment, thereby spanning the entire soybean growth period.

This analysis showed that the XX2021 environment experienced an extreme rainfall event during the full-flowering stage, with daily precipitation exceeding 300 mm (Response Fig. 3a). In contrast, no comparable extreme rainfall event was observed during the corresponding growth period in BJ2020 or BJ2021 (Response Fig. 3b, c).

Excessive rainfall or waterlogging during flowering can substantially affect soybean growth, reproductive development, and maturity-related phenotypes. Therefore, the maturity phenotype measured in XX2021 may have been strongly affected by severe environmental stress and associated field-level spatial heterogeneity, rather than solely reflecting genetic differences among accessions. Although this observation does not by itself establish that the Q-Q plot deviation was caused by the extreme rainfall event, it provides evidence that the XX2021 field trial was exposed to an unusually strong environmental disturbance that could have introduced additional phenotypic noise and confounding.

We therefore agree with the reviewer that the GWAS results from XX2021 should be interpreted with considerable caution. To avoid overinterpretation, we have removed the XX2021 GWAS results from the main GWAS analyses and no longer use this dataset to support the core association signals. The revised manuscript now focuses on the better-calibrated BJ2020 and BJ2021 environments.



Response Fig. 3 Precipitation patterns during the soybean growth periods in the three experimental environments. Daily precipitation was recorded from sowing to the first frost in each of the following experimental environments: **a** Xinxiang, Henan Province, China, in 2021 (XX2021); **b** Beijing, China, in 2020 (BJ2020); and **c** Beijing, China, in 2021 (BJ2021). Precipitation data were obtained from the Climate Data Online platform of the NOAA National Centers for Environmental Information (NOAA NCEI CDO; <https://www.ncei.noaa.gov/cdo-web/>) and span the entire soybean growth period for each environment.

Major comments from Reviewer #3:

5. The presentation of the GWAS summary statistics could also be improved. Supplementary Data 3 does not clearly indicate the experimental trial(s) in which each association reached significance, making it difficult to relate the summary statistics to the corresponding Manhattan

plots. Explicitly annotating the trial(s) for each significant association would greatly improve the readability of the results.

Response:

Thank you for this helpful suggestion. We have revised Supplementary Data 3 by adding an "Experimental trial" column that explicitly identifies the trial in which each SNP–trait association reached significance. These identifiers are consistent with those used in the corresponding Manhattan plots, allowing readers to readily link each significant association to the relevant experimental trial and GWAS result. The corresponding changes have been incorporated into the revised Supplementary Data 3.

Major comments from Reviewer #3:

6. In addition, the Manhattan plots suggest that chromosome 6 harbors a recurrent association signal across all three trials, similar to the signal reported on chromosome 19. This observation is not fully reflected in the text, which emphasizes chromosome 19 as the only locus consistently detected across environments. Given the apparent extent of linkage disequilibrium in this population, including the large LD block surrounding the chromosome 19 locus (almost 1Mb if not longer), it would be useful to characterize the LD structure around both regions. This would help determine whether chromosome 19 truly represents the only robust cross-environment signal or whether chromosome 6 also contains a stable association. Moreover, the association signal on chromosome 19 appears considerably weaker in the third trial, and therefore the statement that this locus is consistently detected across all three environments should be qualified accordingly.

Response:

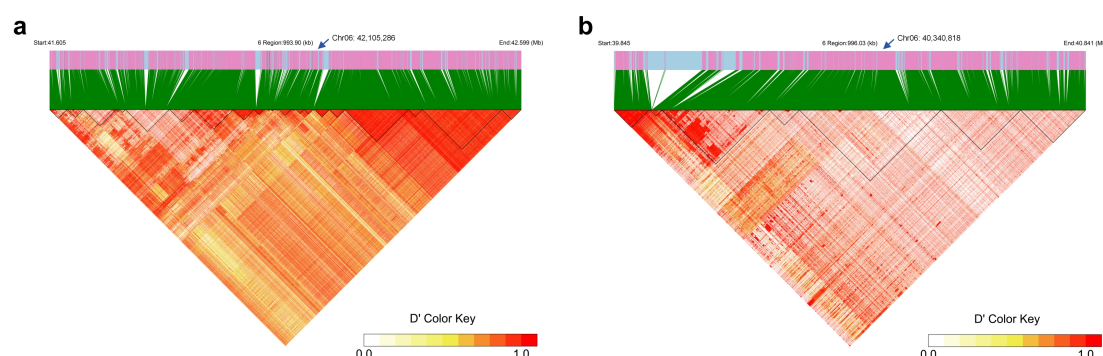
Thank you for this insightful comment. We agree that the recurrent association signals on chromosome 6 warrant further examination and that our original description of the chromosome 19 locus as being consistently detected across all three environments was too strong.

To characterize the local LD structure surrounding the chromosome 6 associations, we have performed LD analyses of the regions extending approximately 500 kb upstream and downstream of the lead SNP identified in each relevant trial. The lead SNP detected in BJ2020 was located at Chr06:42,105,286, whereas the lead SNP detected in BJ2021 was located at Chr06:40,340,818. These two SNPs are approximately 1.76 Mb apart, which is substantially greater than the LD decay distance of our population. These results suggest that the two signals likely represent distinct association regions rather than the same locus detected across trials.

Moreover, the region surrounding Chr06:42,105,286 exhibits relatively extensive LD, whereas the region surrounding Chr06:40,340,818 displays a more fragmented and generally weaker LD structure (Response Fig. 4). Thus, although chromosome 6 harbored association signals in two

trials, the lead SNPs were located in distinct genomic regions and exhibited different local LD patterns. Therefore, the available evidence does not support interpreting these signals as representing the same stable, colocalized locus across environments. In a follow-up study, we will further investigate and functionally characterize the significant association loci identified on chromosome 6.

We have also removed the statement that the chromosome 19 locus was consistently detected across all three environments. The revised manuscript now states that this locus was detected in two environments, thereby providing a more accurate and appropriately qualified interpretation of the results.



Response Fig. 4 Local linkage disequilibrium structures surrounding the significant association signals on chromosome 6. LD was analyzed within regions extending approximately 500 kb upstream and downstream of the lead SNPs detected in the two retained trials. **a** LD structure surrounding Chr06:42,105,286, identified in Beijing, China, in 2020 (BJ2020), covering the 41.605–42.599 Mb region. **b** LD structure surrounding Chr06:40,340,818, identified in Beijing, China, in 2021 (BJ2021), covering the 39.845–40.841 Mb region. LD is represented by D' , with values ranging from 0 to 1. Blue arrows indicate the positions of the lead SNPs.

Major comments from Reviewer #3:

7. Several conclusions rely on comparisons among the HR, NR and SR populations. However, the ADMIXTURE analysis does not appear to support these groups as genetically discrete units. Instead, individuals sampled within each geographic region show varying proportions of ancestry from multiple inferred populations. In other words, because the ADMIXTURE analysis indicates that geographic origin is an imperfect proxy for genetic ancestry, differences in haplotype frequencies among regions may reflect variation in ancestry composition rather than geographically restricted evolutionary processes. Clarification of how ancestry was accounted for in these analyses would strengthen the interpretation of the geographic distribution of candidate haplotypes.

Response:

Thank you for this important comment. We agree that the ADMIXTURE results do not support HR, NR, and SR as genetically discrete populations. They represent three major soybean-producing agroecological regions of China, and the accessions were assigned to these groups according to their geographic origins and the established agroecological regionalization of soybean production in China. Thus, the geographic grouping and the ADMIXTURE analysis describe different aspects of the germplasm—agroecological origin and genetic ancestry, respectively.

The comparison among these regional groups has a clear agroecological rationale. Soybean is a highly photoperiod-sensitive short-day crop, and different genotypes have different critical photoperiod requirements; because photoperiod conditions vary with latitude, cultivars adapted to one region may flower or mature too early or too late in another, constraining cross-regional cultivation. NR, HR, and SR span different latitudinal and agroecological zones, so comparing *GmRVE4d* haplotype distributions across them provides biologically relevant information: because *GmRVE4d* is associated with soybean maturity, differences in haplotype distribution across these regions may reflect differences in the suitability of accessions for regions with different photoperiod conditions. However, these results indicate only an association between haplotype distribution and agroecological region, and do not allow this association to be specifically attributed to regional adaptation.

We acknowledge, however, that geographic origin is not equivalent to genetic ancestry. The ancestry proportions inferred by ADMIXTURE were not explicitly included as covariates in the regional haplotype-frequency comparisons, because these comparisons were intended as descriptive analyses of haplotype distributions across agroecological regions rather than as formal tests of geographically restricted selection. Therefore, the present regional comparison cannot fully separate the potential effects of latitude-dependent agroecological adaptation from those of ancestry composition, population structure, demographic history, or breeding history. The observed differences in *GmRVE4d* haplotype frequencies among NR, HR, and SR may therefore reflect the combined influence of these factors. Accordingly, we interpret these results primarily as geographic differences in the distributions of accessions carrying different *GmRVE4d* haplotypes, rather than as evidence of geographically restricted evolutionary processes.

We appreciate the reviewer's comment, which allowed us to clarify the purpose, biological rationale, and interpretive scope of the regional comparison.

Major comments from Reviewer #3:

8. The selection analyses are also somewhat limited in scope. While I appreciate that a comprehensive population genomic analysis is not the primary objective of this study, the current evidence falls short of supporting the strong conclusions drawn regarding domestication and selection, particularly for a journal such as Nature Communications. The authors present selection

statistics only for the focal LD block, without providing the surrounding genomic context or genome-wide background distribution. As a result, it is difficult to assess whether this region represents an exceptional signal of selection or simply reflects the generally high levels of linkage disequilibrium observed in this population.

Indeed, the extensive LD in the study population makes this contextual information even more important. Without examining the selection signatures in the broader genomic neighborhood, it remains unclear whether the proposed candidate gene is the most likely target of selection or whether the observed signal extends across a larger haplotype containing multiple plausible candidates. Presenting the local genomic landscape, together with the genome-wide distribution of the selection statistics, would substantially strengthen the evidence supporting the identification of this candidate and the conclusions regarding domestication.

Response:

We thank the reviewer for this valuable suggestion. We have now added the genome-wide distribution of XP-CLR scores and a local XP-CLR profile covering the 2-Mb region centered on the candidate locus to provide both genome-wide and local genomic context for the comparison between wild soybeans and landraces (Supplementary Fig. 11). The candidate locus on chromosome 19 shows an XP-CLR score above the genome-wide top 5% threshold, whereas most surrounding windows show lower scores. Thus, the selection signal at this locus is elevated relative to both the genome-wide background and most of the surrounding genomic region, rather than merely reflecting the generally high level of LD in the wild soybean–landrace comparison.