

Deep-resequencing of 984 soybean accessions unveils the critical role of structural variations in agronomic traits

Corresponding Author: Professor Huixia Shou

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)

This study generated a high-confidence genomic structural variation (SV) map from 984 diverse soybean accessions via whole-genome re-sequencing, revealed SVs as a core driver of soybean domestication and improvement, and identified multiple functional SVs underlying key agronomic traits including seed coat color, flower color, and seed size, providing a valuable resource for soybean molecular breeding. However, the manuscript still needs some improvements before publication.

1. What criteria were used to define high-confidence variants in this study?
2. What is the quality of the 56 previously published genomes used in this study? Since the release of the first soybean reference genome in 2010, numerous genomes have been published. However, some of these assemblies were generated relatively early and are constrained by the sequencing platforms and assembly strategies available at the time, resulting in variable assembly quality. Were these genomes systematically evaluated prior to use, and do they meet current standards for high-quality genome assemblies?
3. What is the relationship between Figure 1C and gene families? Figure 1 appears to focus on variation analysis rather than gene family characterization.
4. What is the biological and analytical significance of constructing the pangenome in this study? The authors report the identification of an additional 58,760 PAVs based on the pangenome. Are all variants identified from the pangenome absent from the resequencing-based variant set? If not, how many variants overlap between the two approaches? Among the overlapping variants, are they completely identical, or do some show partial discrepancies? If discrepancies exist, would experimental validation (e.g., PCR) be necessary to determine which call is more accurate? Furthermore, do the variants identified exclusively from the pangenome but absent in resequencing data segregate within the population, and were they incorporated into downstream analyses?
5. Within the same main figure, the font sizes of subpanels are inconsistent (e.g., Figures 2A and 2B). Moreover, as Figures 2A and 2B represent the same type of analysis, their axes and scaling should be consistent in order to clearly demonstrate that SVs capture greater genetic variation. Why does SV-based PCA exhibit larger genetic divergence compared with SNP-based PCA?
6. Figure 2 mainly illustrates previously well-established patterns and lacks sufficient novelty as a main figure. In addition, this part is not mentioned in either the abstract or the highlights, which implies that the author also considers this part to be unimportant. It is therefore suggested that Figure 2 be moved to the supplementary materials.
7. In the population-wide SV identification, SVs were defined using a 50 bp threshold, and SURVIVOR merge parameters were set as "1000 2 1 1 0 50". However, a 100 bp threshold was applied during validation. Please clarify the rationale for using inconsistent thresholds. In addition, only 24 published genomes and 24 re-sequenced accessions were used for accuracy evaluation. Is this sample size sufficient? What criteria were used to select these materials, were they chosen randomly or based on representativeness?
8. Why was the transcriptional impact of TE1 and TE2 insertions investigated only for CHS1-3? Do these insertions have regulatory effects on other CHS genes?
9. The gene regulating flower color is a previously characterized gene. What is the distribution of flower color phenotypes among wild, landrace, and cultivated accessions? Has this variant already been used as a molecular marker in breeding programs?
10. In Case 3, the authors state that a gene affected by the variant was identified through fine-mapping. However, based on the description, the variant appears to directly disrupt the gene, and no fine-mapping procedure is described in the Methods

section.

11. In Case 3, functional validation of the candidate gene relies on only a single EMS mutant line, and the mutation affects a splice site. This evidence is insufficient, as EMS mutants typically harbor numerous background mutations. Therefore, it cannot be conclusively demonstrated that the observed phenotype is solely caused by mutation of this gene.

12. The authors should improve the accuracy and quality of figure presentation. For example, an error such as "Plant weight 3-1" appears in Figure 6A.

13. The logical flow in the first half of the manuscript, particularly in Figure 1 is extremely confusing. The author should reorganize it.

14. The line 144-149, the author reported 800 (81% out of 984 accessions) could capture 92% of SVs. While, 77.5% of SVs represented rare alleles, "implying that the majority of SVs were specific to individual soybean accessions". Those two data or conclusion are conflict with each other.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

This manuscript presents a comprehensive structural variation analysis of 984 soybean accessions with integrated genomic and phenotypic data. Through genome-wide association studies on 27 agronomic traits, the authors uncovered potential roles of multiple SVs in regulating agronomic traits. However, the novelty of the study is limited, as several reported examples have been previously established. Additionally, the reliance on short-read sequencing for SV detection raises concerns about accuracy. Below are my specific comments.

1 Many of the cases presented (e.g., the locus in Figure 4, the CHS mechanism in L298, and the flower color gene in Figure 5D) have been reported in prior studies. The authors should more clearly articulate the novel insights provided by their SV-based approach compared to existing SNP-based studies.

2 The use of short-read sequencing for SV identification is a significant limitation. The authors should discuss the potential for false positives/negatives and consider validating key SVs with long-read sequencing or PCR.

3 The associated locus for seed quality traits has been previously identified via SNP-GWAS. To demonstrate added value, the authors should analyze whether the SV signal resides in a linkage disequilibrium (LD) block distinct from previously reported SNP signals.

4 The siRNA-mediated silencing of CHS genes is a well-established mechanism (Jia et al., 2020). The authors should provide a deeper discussion reconciling their proposed TE-mediated methylation model with this existing pathway.

5 L301, To clarify the role of the identified TE haplotype in domestication, the authors should quantify and compare its frequency in landraces versus modern cultivars.

6 A 259 bp intronic deletion was identified alongside TE insertions. As intronic deletions can affect gene expression, the authors should explicitly justify why the TE insertion, rather than this deletion, is considered the causal variant.

7 Fig 5D, The genetic basis of soybean flower color has been previously documented (Yan et al., 2016) and should be cited. The proposed mechanism requires clarification: Does the TE insertion alter GmW1 expression levels, or does it disrupt the gene's function (e.g., through protein truncation)?

8 Fig 5E, Instead of relying solely on overexpression lines, comparing the enzymatic activity of the two natural haplotypes would strengthen the functional conclusion.

9 Fig 6, The first two examples primarily validate the SV-GWAS framework. The authors should emphasize more explicitly how this approach facilitates the discovery of loci underlying quantitative traits.

10 More detailed background on the Gmmate1 EMS mutant is needed. Does the point mutation result in a complete loss of function, or does the protein retain partial activity? The use of an EMS mutant (point mutation) to represent a large deletion identified in the GWAS is not ideal. Generating CRISPR/Cas9 lines that precisely mimic the identified deletion would provide more rigorous functional validation.

Minor issues:

Figure 1B: Inconsistent formatting: "G. Soja" vs. "G. soja".

Figure 2: Typo: "Landrance" should be "Landrace".

Figure 2C-F: Inconsistent unit formatting: "Kb" vs. "kb".

Figure 4C: Typo: "Upsteram" should be "Upstream".

Figure 6B: Typo: "Delection" should be "Deletion".

Figure 6B: Please verify the appropriateness of the unit "g/DW".

L177: Citation formatting error.

L266: Grammar: "these gene" should be "these genes".

L289–291: The claim that SVs "outperformed SNPs" appears overstated and should be toned down.

L304: Grammar: "associations signals" should be "association signals".

L486: Please clarify the sample composition upfront. Explicitly state the relationship between the 940 newly sequenced accessions and the 44 public ones, and specify the sources and sequencing methods of the public data.

L495: Typo: "Shengzheng" should be "Shenzhen".

L498: Word choice: "databases" is likely incorrect; use "datasets" or "accessions".

L510: Inconsistency: The text mentions "Twenty-four" traits, but the abstract/title implies 27.

L526–528: This sentence should be rephrased for clarity.

L555: Typo: "detecte" should be "detect".

L560: The phrasing here is unclear and should be revised.

L655: "NOT peach" appears to be a note/error; please correct.
L677: Terminology: Consider using "quantitative traits" or "continuous traits" consistently.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

In the manuscript entitled "Deep-resequencing of 984 soybean accessions unveils the critical role of structural variations in agronomic traits", the authors analyze 984 soybean accessions—largely of Chinese origin—and present genome-wide variant discovery, population analyses, GWAS, and several case studies purportedly involving structural variants (SVs). While the study appears ambitious in scope and the figures are visually appealing, the manuscript lacks essential methodological detail, exhibits inconsistencies between claims and methods (especially regarding "graph-based pangenome" analyses), and includes insufficiently supported functional validations. These deficiencies impede reproducibility and interpretation. Before the work can be evaluated on scientific merit, the authors must address the issues listed below.

Major comments:

1. Provenance and accessibility of germplasm (traceability and reproducibility).

The manuscript does not clearly state from which institutions the soybean accessions were obtained or under what material transfer agreements. Many accessions are labeled with ZDD/WDD identifiers or through SoyOD—locally used codes that are not internationally recognized gene bank accessions—which limits reproducibility. For a global journal, authors should (i) provide a mapping table from their internal IDs to international germplasm accessions where applicable, (ii) use official cultivar/line names in the original language/romanization of the source country, and (iii) describe how external researchers can re-access the same lines (contact points, repositories, or terms of use). Without this, independent replication is not feasible.

2. Claimed "graph-based pangenome" is not evidenced by Methods.

The Abstract states "Using a graph-based genome constructed from 56 assembled accessions...", and Section headings also invoke "graph-based pangenome construction." However, the Methods do not provide a reproducible workflow for graph construction, read mapping to the graph, or graph-based genotyping. Instead, the described pipeline relies on short-read alignment to the linear ZH13 reference and merging SV calls from multiple callers. If a graph-based analysis was genuinely performed, the authors must specify:

- Graph construction: software, versions, input assemblies, alignment strategy, file formats (e.g., GFA/GBZ), node/edge/path definitions, and any pruning/subsampling strategy.
- Graph mapping: mapper/tool, key parameters, and quality filters.
- Graph-based genotyping and population summarization: how SV/PAV genotypes were derived from graph paths and aggregated to allele frequencies/LD/GWAS inputs.
- As written, the manuscript cannot be evaluated on this central claim.

3. SURVIVOR merging thresholds need justification.

The Methods mention SURVIVOR merging with parameters such as a 1,000-bp distance and ≥ 2 tools. This risks over-merging distinct SVs and blurring breakpoints, which may be problematic for GWAS and downstream causal interpretation. The authors should justify these thresholds with type-specific and size-specific validations (e.g., DEL/INS/INV/TRA; <1 kb, 1–10 kb, >10 kb) and consider region-aware tuning (e.g., high-repeat versus unique regions). Robustness analyses or positive controls would strengthen confidence in the merged callset.

4. Domestication origin—avoid categorical statements.

Lines 63–64 appear to make definitive claims about the geographic origin of soybean domestication. Given ongoing debate in the literature, the wording should be tempered and supported by balanced citation. Categorical phrasing is not appropriate.

5. Section title inconsistency vs. Methods content.

The subtitle "Deep-resequencing of 984 germplasms and graph-based pan-genome construction" (from Line 119) is not supported by the Methods, which lack graph details. Please either provide complete graph methods (see Comment 2) or revise the text to accurately describe a linear-reference-based SV analysis.

6. Use of 56 assemblies for graph—complexity not addressed.

The manuscript claims to have used 56 assemblies to build a graph for SV discovery (Lines 131–133), but again no methods are given. Even if such a graph exists, incorporating 56 assemblies would substantially increase graph complexity and generally complicate genotyping. How did the authors mitigate this? No description is provided, so the approach cannot be assessed.

7. Population composition limits cross-country comparisons.

Lines 211–214: The purported comparisons among countries/regions are not convincing because the sampling is heavily skewed toward Chinese accessions. With such a biased panel, country-level inferences are not robust; the text should be

revised to avoid over-interpretation.

8. Phenotypes across different seasons and BLUP integration are unclear.

Phenotyping spans winter (2021) and summer (2022, 2023), implying substantial environmental differences (day length, temperature). The manuscript seems to use BLUPs, but does not specify the mixed model, fixed/random effects, or how year/season and G×E were modeled. I recommend:

- Provide the exact model specification (effects, variance components, software).
- Report GWAS per environment/year and highlight reproducible peaks across environments; BLUPs can then be presented as a complement, not a substitute.

9. SV-GWAS vs. SNP-GWAS inconsistency with stated claims.

In Supplementary Figures 16–42, SV-GWAS peaks appear largely co-located with SNP-GWAS, and SNP-GWAS often shows clearer peaks. This seems inconsistent with the repeated claim that “InDels, PAVs, and SVs outperformed SNPs in pinpointing causal genes” (Lines 290–291). The authors should include direct side-by-side comparisons (same locus, same trait) in the main text and reconcile these results with their claim.

10. Case 1 (Fig. 4): scale of reported events vs. expectations from a graph approach.

If a true graph-based SV framework was employed across wild vs. cultivated soybeans, one would expect recovery of known large-scale structural polymorphisms in addition to small insertions/deletions. However, the events highlighted are on the order of a few hundred bp to ~4 kb, which is underwhelming relative to the promise of a pangenome graph and raises doubt that graph-based discovery underlies the result. The exact analytical route from raw data to these events remains ambiguous.

11. Case 2 (Fig. 5, GmW1): overexpression experiment lacks essential information.

The overexpression (OE) validation does not report (i) the number of independent transgenic events, (ii) qRT-PCR confirmation of overexpression levels in those lines, nor (iii) any correlation between expression dosage and phenotype across independent events. These are standard requirements to support causality.

12. Case 3 (Fig. 6, GmMATE1): EMS line acquisition/validation unclear; GWAS signal weak.

PCR validation is appreciated, but the GWAS peak is not compelling. Crucially, the Methods do not describe how the EMS mutant was identified, confirmed (Sanger/WGS), background genotype, generation, or backcrossing. Given the heavy background mutational load in EMS lines and the environmental sensitivity of seed traits, single-generation evidence is insufficient. At minimum, one expects M2 genotyping and M3 bulk phenotyping, and genetic segregation tests linking the candidate mutation to the trait.

13. Typographical error.

Line 655: “peach evolution” appears to be a typo and should read “soybean” (please check surrounding headings for consistency).

14. Global assessment of Methods and presentation.

Overall, the Methods are too vague to enable reproduction or rigorous evaluation. Given the repeated use of “graph-based,” the paper should either (i) provide a complete, reproducible graph workflow and re-analyze results accordingly, or (ii) remove graph-based claims and frame the study explicitly as a linear-reference-based SV analysis with SURVIVOR-merged callsets. Even in the latter case, the functional validations must be strengthened as outlined above. In its current form, the manuscript does not meet Nature Communications standards.

(Remarks on code availability)

In the “Remarks on code availability,” the authors provide a figshare link. However, when accessed, the link results in a “403 Forbidden” error, meaning the resource is not publicly accessible.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

My comments have been adequately addressed and discussed, and the manuscript has been greatly improved.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

All concerns have been addressed. No more comments.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

The authors have substantially improved the manuscript, particularly by providing a clearer and more reproducible description of the graph-based pangenome methodology. The addition of detailed workflows for graph construction (vg, MUMmer, and SyRI), mapping, and genotyping is appreciated and resolves previous concerns regarding reproducibility. However, despite these improvements, several core conceptual and interpretational issues remain unresolved, limiting the scientific clarity and overall impact of the study.

The remaining major concerns and required revisions are outlined below.

1. Conceptual inconsistency: two SV frameworks without clear integration

The manuscript currently combines two distinct variant discovery frameworks: (i) resequencing-based SVs identified using short reads and multiple callers, and (ii) assembly-based PAVs derived from graph/pangenome analyses. However, these two variant systems remain only loosely connected and are not conceptually integrated. Specifically, only a small subset (~4,523 variants) overlaps between the two datasets, whereas the vast majority (~54,237 variants) are unique to one approach. The manuscript does not sufficiently clarify whether these differences reflect complementary biological signals or methodological discrepancies (e.g., false positives or false negatives). At present, the manuscript implicitly assumes that non-overlapping PAVs represent variants inaccessible to short-read approaches, but this conclusion is not adequately demonstrated.

The authors should explicitly address the discrepancy between the two SV discovery systems. This should include validation or comparison of overlapping versus non-overlapping variants, analyses of size distributions and genomic contexts for each variant set, and evidence demonstrating that PAV-specific variants are truly undetectable using short-read data. Alternatively, the authors may consider refocusing the manuscript around a single, coherent SV framework or providing a systematic integration showing how the two approaches complement one another at shared loci, as the current presentation does not clearly achieve either.

2. Ambiguity in the use of the term “SV” throughout the manuscript

A major issue affecting readability and interpretation is the inconsistent use of the term “SV.” Because the study involves fundamentally different classes of variants—including short-read SVs, assembly-derived PAVs, and graph-based genotypes—the blanket use of the term “SV” creates substantial ambiguity. In several sections, particularly at transitions between paragraphs, it is unclear whether “SV” refers to resequencing-based calls or pangenome-derived variants. This ambiguity makes it difficult for readers to evaluate the biological conclusions and undermines the interpretability of the entire study. The authors should introduce clear and consistent terminology (e.g., “reseq-SV,” “PAV,” and “graph-SV”), explicitly indicate which variant set is used in each analysis and figure, and ensure that conclusions are not conflated across fundamentally different variant classes.

3. The added value of the graph-based pangenome remains unconvincing

Although the methodological descriptions have improved, the manuscript still does not convincingly demonstrate the unique biological insights enabled by the graph-based approach. Specifically, no quantitative comparison between linear-reference and graph-based SV detection is provided. Furthermore, the representative case studies (CHS, GmW1, and GmMATE1) all involve relatively small variants (hundreds of base pairs to a few kilobases) that are readily detectable using standard short-read pipelines. No example is presented in which the graph-based approach reveals structural features that cannot be recovered through conventional linear mapping.

The authors should provide at least one of the following: (i) a direct comparison of SV detection performance between linear and graph-based approaches, (ii) a demonstration of a biologically relevant SV uniquely recoverable using the graph-based framework, or (iii) characterization of larger-scale structural variation, such as inversions, CNVs, or repeat-associated structures. Without such evidence, the graph-based component appears descriptive and not clearly functionally justified.

4. Interpretation of the CHS locus remains incomplete and insufficiently reconciled with existing models

The CHS locus is presented as a central example in the manuscript, yet the current analysis focuses primarily on relatively small deletions (259 bp, 772 bp, and 4 kb), while previously reported large-scale structural haplotypes at the I locus are neither reconstructed nor systematically compared. Given that CHS regulation is known to involve complex structural configurations and RNAi-mediated silencing, it remains unclear whether the identified TE-associated variants represent primary causal structures or secondary modifiers within larger haplotypic backgrounds.

The authors should explicitly place their findings within the context of previously reported CHS/I locus structural models, clarify whether the identified SVs define novel haplotypes or act within existing structural backgrounds, and provide a clear structural model integrating large-scale genomic configuration, TE insertions, and methylation/siRNA-associated regulatory effects.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)

The authors have carefully addressed my previous concerns. In particular, the clarification of terminology has improved the readability and interpretability of the manuscript. The authors have also provided additional analyses and examples to support the complementary roles of resequencing-based SVs and pangenome-derived PAVs, which strengthens the overall narrative.

I believe the manuscript has been sufficiently improved and now meets the standard for publication.

(Remarks on code availability)

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/>

Response to reviewers

The following two sections are the modifications and some explanations for the manuscript.

Responses to Reviewer #1:

Reviewer #1 (Remarks to the Author):

This study generated a high-confidence genomic structural variation (SV) map from 984 diverse soybean accessions via whole-genome re-sequencing, revealed SVs as a core driver of soybean domestication and improvement, and identified multiple functional SVs underlying key agronomic traits including seed coat color, flower color, and seed size, providing a valuable resource for soybean molecular breeding. However, the manuscript still needs some improvements before publication.

1. What criteria were used to define high-confidence variants in this study?

Response: We employed stringent criteria for each variant. Supplementary Fig. 2 illustrates the workflow of the variant detection pipeline and the description of the workflow is added in Results (lines 121-129) and in Methods (lines 604-621).

1. For conventional SNPs and InDels: Reads were aligned to the reference genome using BWA, followed by stringent filtering using SAMtools (depth (QD) < 2.0, mapping quality (MQ) > 40.0). Detailed methods are described in the “Read mapping and small variants calling” section. Lines 604-621
2. For SVs: We employed four tools: GRIDSS (v2.13.2), Manta (v1.6.0), Lumpy (v0.2.13), and Delly (v1.1.6), which are the most frequently used and have relatively high detection accuracy based on read-pair and split-read methods (Tattini et al., 2015). SVs detected by at least two of these tools were retained as candidates.
3. For PAVs: We utilized a graph-based pan-genome approach and obtained bi-allelic variant information using the classic MUMmer alignment method.

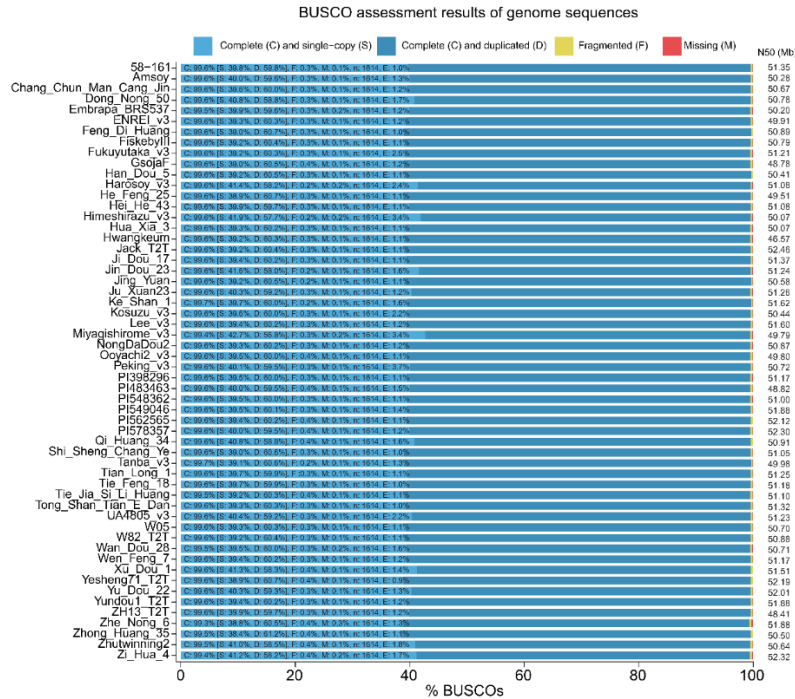
To further ensure reliability, we performed PCR amplification for validation

(Supplementary Fig. 46, 49C; Fig. 7E). Finally, for the variants identified above, we applied Minor Allele Frequency (MAF) and missing rate filters for subsequent GWAS analysis. Detailed procedures are available in the Methods section.

2. What is the quality of the 56 previously published genomes used in this study? Since the release of the first soybean reference genome in 2010, numerous genomes have been published. However, some of these assemblies were generated relatively early and are constrained by the sequencing platforms and assembly strategies available at the time, resulting in variable assembly quality. Were these genomes systematically evaluated prior to use, and do they meet current standards for high-quality genome assemblies?

Response: We thank the reviewer for raising this important issue. We carefully considered genome quality during the selection process. Initially, we collected 62 publicly available genomes and filtered for annotation completeness and assembly continuity.

For reference genomes initially assembled using earlier sequencing technologies (e.g., Williams 82 Glyma1.0), we prioritized the most updated versions (e.g., the T2T version, 10.1016/j.molp.2023.08.012). The genomes included in our analysis were all published within approximately the past six years and represent contemporary standards of genome quality. Systematic evaluation using BUSCO assessment demonstrated that all genomes had a scaffold N50 values exceeded 46.57 Mb with an average single-copy Embryophyta gene completeness of 99.58% in genome sequences. This collection includes five T2T-level assemblies, thus meeting current standards for high-quality genome assemblies. Complete assembly statistics and quality assessments are provided in newly added Supplementary Table 7 and Supplementary Fig 7 (Lines 188-195).



Supplementary Fig. 7. BUSCO evaluation of 56 soybean genome and assembly-stats calculating N50. C: Complete BUSCOs; S: Complete and single-copy BUSCOs; D: Complete and duplicated BUSCOs; F: Fragmented BUSCOs; M: Missing BUSCOs; n: Total BUSCO groups searched. Average N50 = 50.81 Mb. Average single-copy Embryophyta genes = 99.58% (in genome sequences) and 96.67% (in protein sequences).

3. What is the relationship between Figure 1C and gene families? Figure 1 appears to focus on variation analysis rather than gene family characterization.

Response: We sincerely appreciate the reviewer's insightful observation regarding the logical disconnect in the original Figure 1C. To resolve this issue and enhance the manuscript's coherence, we have comprehensively reorganized Figures 1–3 (see Lines 119–260 for details). The revised structure is as follows:

(1) New Figure 1 now strictly focuses on structural variation (SV) analysis, presenting the SVs directly identified from the population-level data, and the systematic classification of SV types.

(2) New Figure 2 presents gene family and pan-genome analyses. It includes pan-genome composition based on the 56 genomes, gene family metrics to evaluate the completeness of the pan-genome, and identification of PAVs.

(3) New Figure 3 highlights functional implications of PAVs, specifically

highlighting impacts of PAVs located to promoter and exon regions on gene function.

By decoupling SV analysis (Figure 1) from gene family characterization (Figure 2), we ensure a clearer narrative flow and eliminate prior ambiguities. The original Figure 1C (pertaining to gene families) has been removed, as its content is now fully integrated into the restructured Figure 2. All revised figures and legends have been rigorously updated for accuracy and alignment with the text.

4. What is the biological and analytical significance of constructing the pangenome in this study? The authors report the identification of an additional 58,760 PAVs based on the pangenome. Are all variants identified from the pangenome absent from the resequencing-based variant set? If not, how many variants overlap between the two approaches? Among the overlapping variants, are they completely identical, or do some show partial discrepancies? If discrepancies exist, would experimental validation (e.g., PCR) be necessary to determine which call is more accurate? Furthermore, do the variants identified exclusively from the pangenome but absent in resequencing data segregate within the population, and were they incorporated into downstream analyses?

Response: The primary objective of constructing the pangenome in this study was to obtain comprehensive SV information from high-quality genome assemblies to complement the SVs detected from short-read resequencing data. While 4,523 variants were shared between pangenome and resequencing-based SV calls, 54,237 non-overlapping variants primarily, reflecting the pangenome's ability to resolve complex or repetitive SVs inaccessible to short-read sequencing. We have incorporated this clarification into the revised version in lines 212-216.

To address the reviewer's concern regarding discrepancies among the overlapping variants, we performed validation using 24 accessions with both assembled genomes and resequencing data. After merging SVs from both datasets using SURVIVOR, we PCR-verified a randomized subset of 200 loci. The analysis revealed 87% genotype concordance between dataset pairs, while pangenome-specific PAVs showed a 92% validation rate (Supplementary Table 19) (lines 473-476).

The variants identified exclusively from the pangenome (PAVs) do segregate within the population and were incorporated into the subsequent PAV-GWAS analysis to identify potential causal variants underlying agronomic traits.

5. Within the same main figure, the font sizes of subpanels are inconsistent (e.g., Figures 2A and 2B). Moreover, as Figures 2A and 2B represent the same type of analysis, their axes and scaling should be consistent in order to clearly demonstrate that SVs capture greater genetic variation. Why does SV-based PCA exhibit larger genetic divergence compared with SNP-based PCA?

Response: As suggested, we have standardized the formatting of all figures to maintain uniform font sizes, axis labels, and scaling across subpanels in the revised manuscript. This ensures clarity and direct visual comparison of the genetic variation captured by SNPs and SVs.

The greater genetic divergence observed in SV-based PCA reflects the biological expectation that SVs often have broader functional impacts than SNPs. While SNPs are abundant and often evolutionarily neutral, SVs frequently span larger genomic segments, which are strongly linked to domestication and improvement traits. Due to their pronounced phenotypic effects, SVs undergo stronger selective pressures during breeding, resulting in sharper genetic differentiation between subpopulations (e.g., wild versus cultivated) in PCA. These findings underscore the utility of SVs as powerful markers for detecting artificial selection signatures.

6. Figure 2 mainly illustrates previously well-established patterns and lacks sufficient novelty as a main figure. In addition, this part is not mentioned in either the abstract or the highlights, which implies that the author also considers this part to be unimportant. It is therefore suggested that Figure 2 be moved to the supplementary materials.

Response: In response to the reviewer's comment, we have removed Figure 2 and renumbered the remaining figures as Figures 1-3. Additionally, we have optimized the layout to enhance clarity and reader engagement.

7. In the population-wide SV identification, SVs were defined using a 50 bp threshold, and SURVIVOR merge parameters were set as “1000 2 1 1 0 50”. However, a 100 bp threshold was applied during validation. Please clarify the rationale for using inconsistent thresholds. In addition, only 24 published genomes and 24 re-sequenced accessions were used for accuracy evaluation. Is this sample size sufficient? What criteria were used to select these materials, were they chosen randomly or based on representativeness?

Response: We appreciate the reviewer’s thoughtful questions. Below, we address each point in detail.

- 1) Regarding consistency of SV detection threshold:

The 100 bp parameter used in validation (-l 100 in delta-filter) refers to the minimum alignment length, not the SV size threshold. This parameter is used to filter out short, potentially spurious alignments while retaining SVs ≥ 50 bp—for example, a 60 bp deletion flanked by alignment blocks >100 bp would still be detected. Thus, our SV definition (≥ 50 bp) remained consistent throughout identification and validation. We have updated the Methods to explicitly clarify this distinction (lines 633-640).

- 2) Regarding the sample size and selection for validation:

Notably, among the total 984 accessions in our study, only 24 overlaps with previously published genomes; therefore, these were specifically utilized as a gold-standard benchmark. Despite this limited number of available references, the validation results from these 24 genomes are highly representative of the accuracy for the majority of the variants in our dataset. Furthermore, to provide additional orthogonal confirmation, we also performed experimental PCR validation on a subset of the variants.

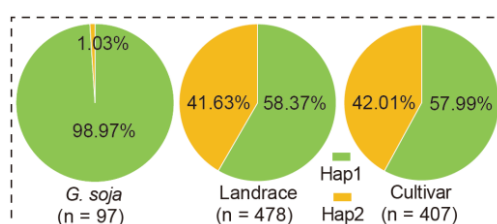
8. Why was the transcriptional impact of TE1 and TE2 insertions investigated only for CHS1-3? Do these insertions have regulatory effects on other CHS genes?

Response: We thank the reviewer for this question. While we investigated the impact of TE insertions across all *CHS* genes (Fig. 5I, Supplementary Fig. 48), significant methylation changes induced by TE insertions were specifically observed at *CHS1*,

CHS2, and *CHS3*. Other *CHS* family members did not exhibit statistically meaningful methylation differences linked to TE1/TE2 insertions. Therefore, we focused on *CHS1-3* in Figure 6 to highlight the TE-driven regulatory mechanism underlying seed coat color variation.

9. The gene regulating flower color is a previously characterized gene. What is the distribution of flower color phenotypes among wild, landrace, and cultivated accessions? Has this variant already been used as a molecular marker in breeding programs?

Response: The gene underlying flower color variation was indeed previously reported (Zabala and Vodkin, 2007). In the study, the causal variants for flower color, including multiple insertions in 3rd exon of *SoyZH13_13G057501*, were detected across SNP-, InDel-, and SV-GWAS analyses, explaining both flower color and stem color pleiotropy ($p < 1 \times 10^{-8}$). We performed haplotype analysis based on the variation in the third exon, classifying the accessions into Hap1 (purple flower) and Hap2 (White flower) haplotypes (Supplementary Fig. 51). The distributions of purple flower color among wild soybeans, landraces, and cultivated accessions were 99%, 58.4% and 58%, respectively (Supplementary Figure 51). We have incorporated this clarification into the revised version in lines 418-424. While flower/stem color has long been used as a visual marker, our work defines the multiple insertions could be used as a molecular marker.



Supplementary Fig. 51. Frequency distribution of the two haplotypes of *SoyZH13_13G057501* in *G. soja*, landraces and cultivars.

10. In Case 3, the authors state that a gene affected by the variant was identified through fine-mapping. However, based on the description, the variant appears to directly disrupt the gene, and no fine-mapping procedure is described in the Methods section.

Response: We appreciate the reviewer's careful reading and comment. The original phrasing inaccurately implied the use of fine-mapping, which was not performed in this study. As correctly noted, the causal variant was first identified through SV-GWAS, followed by confirmation via comparative genomics and PCR validation. Its functional impact was further supported by EMS mutant phenotyping. To ensure accuracy, we have revised the text to clarify that the variant was identified through GWAS-based discovery followed by experimental validation rather than fine-mapping (Lines 418-424).

11. In Case 3, functional validation of the candidate gene relies on only a single EMS mutant line, and the mutation affects a splice site. This evidence is insufficient, as EMS mutants typically harbor numerous background mutations. Therefore, it cannot be conclusively demonstrated that the observed phenotype is solely caused by mutation of this gene.

Response: While we acknowledge that background mutations may complicate phenotypic interpretation, our conclusions are supported by multiple lines of evidence beyond the EMS mutant. The candidate gene (*GmMATE1*) was initially identified through SV-GWAS, and haplotype analysis in the natural population strongly linking this locus to seed weight variation (Figure 7D). We validated the existence and effect of this variant via PCR and sequencing in 24 randomly selected accessions, confirming the correlation between the genotype and phenotype (Figure 7D-E).

While we acknowledge that CRISPR-Cas9 knockout mutants of *GmMATE1* would provide definitive functional validation, soybean transformation remains a time-intensive process. The *GmMATE1* knock-out mutants are currently in development but require an additional generation for phenotypic characterization. As preliminary supporting evidence, we used a homozygous EMS mutant with a characterized mutation that disrupts a critical protein domain. Importantly, the phenotypic consistency observed between this EMS mutant and *GmMATE1 Hap2* accessions (Supplementary Figures 54-57) strongly supports the gene's role in seed weight regulation. These methodological considerations have been explicitly addressed in the

discussion of the revised manuscript to ensure proper interpretation of our current findings while awaiting transgenic confirmation (Lines 431-434; 441-445).

12. The authors should improve the accuracy and quality of figure presentation. For example, an error such as “Plant weight 3-1” appears in Figure 6A.

Response: Following reviewer suggestions, we have streamlined Figure 7A by removing the 'Plant weight 3-1' annotation to focus specifically on seed weight-associated loci. The revised figure now exclusively presents genomic regions with direct seed weight associations identified through our SV-GWAS. Additionally, we have conducted a full audit of all figures to ensure consistency in labeling, scaling, and terminology.

13. The logical flow in the first half of the manuscript, particularly in Figure 1 is extremely confusing. The author should reorganize it.

Response: We thank the reviewer for the constructive criticism regarding the logical flow. The first half of the manuscript is organized into two complementary parts focused on structural variation (SV) analysis.

Section 1 presents SV detection based on resequencing data from 984 accessions, followed by population genetic analysis and functional characterization of SVs. The corresponding results are presented in Figure 1, including: 1) The geographic origin of the accessions used for resequencing, along with their phylogenetic relationships, to contextualize diversity (Figure 1A-B). 2) SV-Driven Population Differentiation. The SV-PCA clearly separates wild and cultivated soybean populations (Figure 1C), while SV-LD analysis reveals accelerated linkage disequilibrium (LD) decay from wild to cultivated soybeans (Figure 1D) and region-specific LD patterns (Figure 1E). 3) SV size distribution and annotations (Figure 1F-G) demonstrate how SVs likely influence phenotypic evolution by disrupting functional genomic regions.

Section 2 is pan-genome construction and PAV identification. Building on the population-level SV data, the second section transitions to a species-wide perspective by integrating *de novo* assemblies from 56 high-quality genomes to construct a graph-

based pan-genome. Compared to the individual-level SV exploration in Section 1, this approach systematically catalogs presence-absence variation (PAV) across the species (Figure 2D) and reveals evolutionary dynamics (Figure 2A-C) and structural features of wild soybeans (Figure 2F).

14. The line 144-149, the author reported 800 (81% out of 984 accessions) could capture 92% of SVs. While, 77.5% of SVs represented rare alleles, “implying that the majority of SVs were specific to individual soybean accessions”. Those two data or conclusion are conflict with each other.

Response: We thank the reviewer for this insightful comment. We apologize if the connection between these two statements was not clear.

There is actually a logical consistency between the capture efficiency and the allele frequency distribution:

Capture Efficiency (92%): This indicates that our sample size (800 accessions) was sufficient to saturate the discovery of SVs in this population.

Allele Frequency (77.5% rare): This describes the distribution of the discovered SVs—specifically, that they are predominantly low-frequency variants.

These two observations are complementary rather than conflicting. The fact that 77.5% of SVs are rare alleles explains why a relatively large sample size was required to capture them. Our analysis demonstrates that with 800 accessions, we successfully captured the majority (92%) of these variants, including the rare ones. We have revised the text (lines 165-170) to better articulate this relationship and clarify that the high capture rate was achieved precisely because our sample size was large enough to detect these prevalent rare variants.

Responses to Reviewer #2:

Reviewer #2 (Remarks to the Author):

1. Many of the cases presented (e.g., the locus in Figure 4, the CHS mechanism in L298, and the flower color gene in Figure 5D) have been reported in prior studies.

The authors should more clearly articulate the novel insights provided by their SV-based approach compared to existing SNP-based studies.

Response: We thank the reviewer for this insightful comment. While acknowledging that discussed loci have been documented previously, we wish to emphasize that our structural variation-centric methodology has yielded significant novel findings that transcend conventional SNP-based analyses. We highlighted these distinctions in the revised manuscript (Lines 361-375)

1. Regarding the *CHS* locus (Seed coat color):

Previous studies primarily attributed seed coat color variation to the inversion at the *I* locus, which leads to *CHS* silencing (Supplementary Fig. 45). However, we found inconsistency between the inversion and seed coat color, suggesting additional mechanism may exist. Our SV-GWAS approach identified a distinct significant locus on Chromosome 8. Through PCR validation and further analysis, we discovered that a TE insertion in this region affects the methylation levels of the nearby *CHS* gene, thereby regulating its expression. This suggests a novel regulatory mechanism beyond the well-known *I* locus inversion.

2. Regarding the flower color gene (*W1*):

While the *W1* gene is known to regulate flower color, our study systematically clarified the specific structural basis of this trait. Although both SNP-, InDel- SV - GWAS identified the locus, our analysis precisely pinpointed a multiple insertion (total of 53-bp) in the third exon as the causal variant. Furthermore, we demonstrated that this variant is a pleiotropic gene simultaneously regulating flower color and hypocotyl color. This finding provides a molecular explanation for the traditional breeding practice of using stem and flower color traits to assess hybrid purity and success rates.

2. The use of short-read sequencing for SV identification is a significant limitation.

The authors should discuss the potential for false positives/negatives and consider validating key SVs with long-read sequencing or PCR.

Response: While short-read sequencing remains widely used for SV detection, we acknowledge that short-read sequencing has inherent limitations in resolving complex

structural variations (e.g., large-scale translocations, highly repetitive regions). To maximize accuracy within our target size range (50 bp–10 kb), we implemented a strict quality control process (Supplementary Fig. 2). Four complementary tools, including Lumpy, Manta, Delly, and GRIDSS were used for breakpoint detection. High-confidence SVs were retained using SURVIVOR filtering criteria of at least detected with 2 supporting tools and size between 50 bp and 10 Mb. MUMmer and Syri were employed to independently validate the accuracy of the detected SVs with 24 assembled genomes, followed by manual validation of 200 randomly selected SVs using IGV (Lines 467-476).

3. The associated locus for seed quality traits has been previously identified via SNP-GWAS. To demonstrate added value, the authors should analyze whether the SV signal resides in a linkage disequilibrium (LD) block distinct from previously reported SNP signals.

Response: We appreciate the reviewer's suggestion. Our LD analysis revealed both connections and key differences between our SV signals and previously reported SNP associations. While some identified SVs were in weak to moderate LD with nearby SNPs, several critical SVs exhibited statistically independent effects. Most importantly, the SV-based approach allowed us to pinpoint causal SVs with clear functional impacts. While SNPs are fundamental to genetic studies, they do not capture the full spectrum of causal variation; SVs represent another critical class of variants driving phenotypic diversity. Consequently, these newly identified SVs successfully bridge the gap between genotype and phenotype. We have expanded our discussion of these comparative analyses in the revised manuscript (Lines 324-333).

4. The siRNA-mediated silencing of CHS genes is a well-established mechanism (Jia et al., 2020). The authors should provide a deeper discussion reconciling their proposed TE-mediated methylation model with this existing pathway.

Response: We appreciate the reviewer's thoughtful reference to the well-established siRNA-mediated silencing pathway for CHS genes (Jia et al., 2020). In the revised

manuscript, we have explicitly reconciled our TE-mediated methylation model with this existing pathway. We propose that these two mechanisms are not mutually exclusive but rather complementary. The identified TE insertion may serve as genomic of siRNAs, potentially through double-stranded RNA structures or inverted repeats, which are processed into siRNAs via the RNA-directed DNA methylation machinery to reinforce CHS repression. By linking this structural variant to siRNA production, our findings provide a genetic basis for the epigenetic silencing previously observed, offering a more unified mechanistic framework for *CHS* regulation. We have revised the Discussion section to elaborate on this relationship in lines 511-533.

5. L301, To clarify the role of the identified TE haplotype in domestication, the authors should quantify and compare its frequency in landraces versus modern cultivars.

Response: We thank the reviewer for this constructive suggestion. To better elucidate the role of the identified TE haplotype in soybean domestication, we have performed frequency comparisons across three quantified and compared its frequency across wild soybeans, landraces, and modern cultivars. The results reveal a clear frequency differentiation among these groups, providing new insights into the domestication history of this locus. These data have been added to the revised manuscript (see Figure 5E).

6. A 259 bp intronic deletion was identified alongside TE insertions. As intronic deletions can affect gene expression, the authors should explicitly justify why the TE insertion, rather than this deletion, is considered the causal variant.

Response: We thank the reviewer for this insightful comment. While both the 259 bp intronic deletion and the TE insertion co-occur at the *I* locus, the TE insertion is the more likely causal variant. The TE insertion showed near-perfect genotype-phenotype concordance across accessions whereas the 259 bp deletion exhibits incomplete linkage (Supplementary Fig. 45). Furthermore, the TE is disruptively positioned, overlapping a known regulatory element/splice site. In contrast, the 259 bp deletion falls within the intron region. We have revised the manuscript to explicitly discuss the relationship

between these variants and justify our focus on the TE insertion. Lines 361-375

7. Fig 5D, The genetic basis of soybean flower color has been previously documented (Yan et al., 2016) and should be cited. The proposed mechanism requires clarification: Does the TE insertion alter GmW1 expression levels, or does it disrupt the gene's function (e.g., through protein truncation)?

Response: As suggested, we have added the citation for Yan et al. (2016) as reference [48] in the revised manuscript. To clarify the proposed mechanism, this variant is not caused by a TE insertion but rather the insertion of multiple fragments into the third exon, which leads to a premature stop codon and protein truncation. We also observed differential expression between the two haplotypes. Therefore, this locus affects both gene structure (via protein truncation) and expression levels.

8. Fig 5E, Instead of relying solely on overexpression lines, comparing the enzymatic activity of the two natural haplotypes would strengthen the functional conclusion.

Response: As stated in the response to Q7, in *Hap 2*, the variant not only introduces a premature termination in the F3'5'H protein-coding sequence but also reduces expression, likely due to its inability to produce a functional protein. Therefore, we conclude that *Hap 2* lack functional F3'5'H protein and catalytic activity. This conclusion is further supported by anthocyanin extraction data. In the revised manuscript, we included the anthocyanin levels of both *OE* and *WT* (Fig. 6H). We therefore speculate that the variation in flower color across different haplotypes aligns with this molecular mechanism. We have already planted a set of diverse germplasm accessions, which will be used for further phenotypic and genetic validation in our future work.

9. Fig 6, The first two examples primarily validate the SV-GWAS framework. The authors should emphasize more explicitly how this approach facilitates the discovery of loci underlying quantitative traits.

Response: As suggested, in the revised manuscript, we expanded the discussion to explicitly highlight how SV-GWAS overcomes limitations of SNP-based GWAS and

accelerates causal gene discovery for quantitative traits (lines 511-519) as followed.

“Beyond validating the framework, our SV-GWAS approach demonstrates significant advantages in uncovering novel QTLs. Unlike SNPs, SVs often exhibit distinct LD patterns and can act as causal variants by directly altering gene dosage or regulatory structure (Fig. 1C, Supplementary Fig. 5). For instance, SVs in *GmMATE1* and other candidate genes provided enhanced resolution in pinpointing causal variants compared to flanking SNPs, thereby offering a direct molecular mechanism explaining continuous variation in seed size and weight (Fig. 7). This highlights the potential of SV-GWAS to uncover hidden genetic architecture underlying complex quantitative traits.”

10. 10 More detailed background on the *Gmmate1* EMS mutant is needed. Does the point mutation result in a complete loss of function, or does the protein retain partial activity? The use of an EMS mutant (point mutation) to represent a large deletion identified in the GWAS is not ideal. Generating CRISPR/Cas9 lines that precisely mimic the identified deletion would provide more rigorous functional validation.

Response: The EMS-induced *Gmmate1* mutant carries a mutation in a splice site, and caused a wrong splicing of its fifth exons (Supplementary Fig. 54). This suggests that the *Gmmate1* mutant is likely loss-of-function effect rather than the retention of partial activity. The phenotypic convergence between the EMS mutant and natural deletion haplotype (Fig. 7D vs 7F) provides mutually reinforcing evidence for *GmMATE1*'s role in seed size regulation. Although CRISPR knockout would provide more definitive evidence, soybean transformation remains time-intensive. While CRISPR knockout lines are currently under development, the *Gmmate1* EMS mutant provides independent genetic validation by functionally mimicking the natural deletion phenotype despite arising from a distinct molecular lesion. Lines 431-434.

Minor issues:

Figure 1B: Inconsistent formatting: "G. Soja" vs. "G. soja".

Figure 2: Typo: "Landrance" should be "Landrace".

Figure 2C-F: Inconsistent unit formatting: "Kb" vs. "kb".

Figure 4C: Typo: "Upsteram" should be "Upstream".

Figure 6B: Typo: "Delection" should be "Deletion".

Figure 6B: Please verify the appropriateness of the unit "g/DW".

L177: Citation formatting error.

L266: Grammar: "these gene" should be "these genes".

L289–291: The claim that SVs "outperformed SNPs" appears overstated and should be toned down.

L304: Grammar: "associations signals" should be "association signals".

L486: Please clarify the sample composition upfront. Explicitly state the relationship between the 940 newly sequenced accessions and the 44 public ones, and specify the sources and sequencing methods of the public data.

L495: Typo: "Shengzheng" should be "Shenzhen".

L498: Word choice: "databases" is likely incorrect; use "datasets" or "accessions".

L510: Inconsistency: The text mentions "Twenty-four" traits, but the abstract/title implies 27.

L526-528: This sentence should be rephrased for clarity.

L555: Typo: "detecte" should be "detect".

L560: The phrasing here is unclear and should be revised.

L655: "NOT peach" appears to be a note/error; please correct.

L677: Terminology: Consider using "quantitative traits" or "continuous traits" consistently.

Response: We gratefully acknowledge the reviewer's careful identification of formatting and grammatical errors. A comprehensive proofreading has been performed to address these issues.

Responses to Reviewer #3:

Reviewer #3 (Remarks to the Author):

In the manuscript entitled "Deep-resequencing of 984 soybean accessions unveils the

critical role of structural variations in agronomic traits", the authors analyze 984 soybean accessions —largely of Chinese origin— and present genome-wide variant discovery, population analyses, GWAS, and several case studies purportedly involving structural variants (SVs). While the study appears ambitious in scope and the figures are visually appealing, the manuscript lacks essential methodological detail, exhibits inconsistencies between claims and methods (especially regarding “graph-based pangenome” analyses), and includes insufficiently supported functional validations. These deficiencies impede reproducibility and interpretation. Before the work can be evaluated on scientific merit, the authors must address the issues listed below.

1. Provenance and accessibility of germplasm (traceability and reproducibility).

The manuscript does not clearly state from which institutions the soybean accessions were obtained or under what material transfer agreements. Many accessions are labeled with ZDD/WDD identifiers or through SoyOD—locally used codes that are not internationally recognized gene bank accessions—which limits reproducibility. For a global journal, authors should (i) provide a mapping table from their internal IDs to international germplasm accessions where applicable, (ii) use official cultivar/line names in the original language/romanization of the source country, and (iii) describe how external researchers can re-access the same lines (contact points, repositories, or terms of use). Without this, independent replication is not feasible.

Response: We thank the reviewer for emphasizing the importance of germplasm provenance and accessibility. In response, we have clarified the origin and traceability of all soybean accessions used in this study.

The ZDD and WDD codes are China’s nationally standardized accession numbers for soybean germplasm. According to China's crop germplasm data standards, ZDD (Zhongguo Da-Dou) and WDD (Waiguo Da-Dou) denotes as Chinese and foreign-origin soybean accessions maintained within the China’s Germplasm System, respectively. These identifiers are fully searchable through the Chinese Crop Germplasm Resources Information System (CGRIS) To facilitate global access, we have supplemented our dataset with corresponding international accession numbers

(e.g., PI numbers from the USDA Soybean Germplasm Collection) for accessions where such information is available. These PI numbers can be searched through the GRIN-Global database (<https://npgsweb.ars-grin.gov/gringlobal/search>). All germplasm metadata—including original identifiers, ZDD/WDD codes, and PI numbers—are now fully integrated into our Soybean Omics Database (SoyOD) through its dedicated Germplasm Search module.

To ensure reproducibility, we provide a cross-referenced supplementary table (Supplementary Table 1) mapping internal IDs to ZDD/WDD and PI identifiers in the revised manuscript. We have updated the manuscript to include: A mapping table linking our internal IDs to ZDD/WDD and PI numbers (where available), provided as a supplementary table; Clear descriptions of the germplasm sources and the procedures for requesting materials from the corresponding gene banks (Supplementary Table 1).

2. Claimed “graph-based pangenome” is not evidenced by Methods.

The Abstract states “Using a graph-based genome constructed from 56 assembled accessions...”, and Section headings also invoke “graph-based pangenome construction.” However, the Methods do not provide a reproducible workflow for graph construction, read mapping to the graph, or graph-based genotyping. Instead, the described pipeline relies on short-read alignment to the linear ZH13 reference and merging SV calls from multiple callers. If a graph-based analysis was genuinely performed, the authors must specify:

- Graph construction: software, versions, input assemblies, alignment strategy, file formats (e.g., GFA/GBZ), node/edge/path definitions, and any pruning/subsampling strategy.
- Graph mapping: mapper/tool, key parameters, and quality filters.
- Graph-based genotyping and population summarization: how SV/PAV genotypes were derived from graph paths and aggregated to allele frequencies/LD/GWAS inputs.
- As written, the manuscript cannot be evaluated on this central claim.

Response: We thank the reviewer for pointing out the lack of methodological detail regarding our graph-based pangenome analysis. We have now revised the Methods

section to provide a reproducible workflow as followed:

- 1) Graph Construction: We constructed a reference-guided graph pangenome using the *vg* toolkit (v1.46.0). Assemblies of the 56 accessions were first aligned to the reference genome using MUMmer (v4.0). PAVs were then identified using SYRI (v1.0). The resulted variant calls (VCF format) were integrated with the reference genome to build a variation graph using *vg construct*. The final graph was indexed into an *xg* format (~1.8 GB), which encodes nodes, edges, and paths for both reference and alternative alleles.
- 2) Graph Mapping: The short reads of the 984 resequenced accessions were mapped to the constructed graph using *vg map*. We applied standard quality filters (e.g., $\text{MAPQ} \geq 20$) to retain high-confidence alignments. The mapping results were stored in GAM format for subsequent analysis.
- 3) Graph-based Genotyping and Population Summarization: We performed genotyping on the graph using *vg call* (or *vg genotype*) to determine the allelic status of SVs/PAVs for each accession. The genotypes were extracted and transformed into a standard matrix format, which served as the input for population structure analysis, LD calculation, and GWAS.

These methodological details have been thoroughly documented in the revised Methods section to ensure full reproducibility of our graph-based pangenome analysis (lines 658-695).

3. SURVIVOR merging thresholds need justification.

The Methods mention SURVIVOR merging with parameters such as a 1,000-bp distance and ≥ 2 tools. This risks over-merging distinct SVs and blurring breakpoints, which may be problematic for GWAS and downstream causal interpretation. The authors should justify these thresholds with type-specific and size-specific validations (e.g., DEL/INS/INV/TRA; <1 kb, 1–10 kb, >10 kb) and consider region-aware tuning (e.g., high-repeat versus unique regions). Robustness analyses or positive controls would strengthen confidence in the merged callset.

Response: We appreciate the reviewer's thoughtful comment regarding the

SURVIVOR merging parameters. We understand the concern that the 1,000-bp distance threshold might potentially lead to over-merging of distinct SVs. However, this parameter was carefully selected based on careful technical consideration of short-read SV detection and validation data. This merging approach has been widely adopted and validated in multiple large-scale genomic studies by Liu et al., (2025 Nature Commun), Abel et al., (2024 Nature Commun); Guo et al., (2020 Genome Biol) and Ji et al., (2024 Genome Biol).

The selection of the 1,000-bp distance threshold accounts for inherent uncertainty in short-read mapping, particularly in repetitive regions or for larger SVs. A strict distance threshold (e.g., 100 bp) could fail to merge biologically identical events due to minor breakpoint shifts between different detection tools. The 1,000 bp threshold, commonly adopted in large-scale SV studies (e.g., utilizing SURVIVOR), balances sensitivity and precision in capturing the true biological signal.

Our implementation includes stringent merging criteria to minimize false-positive combinations. Specifically, the SURVIVOR command `merge 1000 2 1 1 0 50` enforces type-specific merging: Only SVs of the same type (e.g., DEL with DEL) are merged, ensuring that only SVs of the same category are combined. Additionally, we require detections by at least two independent callers, reducing the risk of merging random artifacts. These safeguards enhance the reliability of our merged SV calls.

4. Domestication origin—avoid categorical statements.

Lines 63–64 appear to make definitive claims about the geographic origin of soybean domestication. Given ongoing debate in the literature, the wording should be tempered and supported by balanced citation. Categorical phrasing is not appropriate.

Response: We agree that the original wording may have overstated the certainty, given ongoing scholarly discussions in this complex topic. Accordingly, we revised the statement as “Cultivated soybean is thought to have be domesticated from its wild progenitor (*Glycine soja* Sieb. and Zucc.) in East Asia approximately 5,000 years ago, before its subsequent spread and extensive global cultivation” (Lines 62–65).

5. Section title inconsistency vs. Methods content.

The subtitle “Deep-resequencing of 984 germplasms and graph-based pan-genome construction” (from Line 119) is not supported by the Methods, which lack graph details. Please either provide complete graph methods (see Comment 2) or revise the text to accurately describe a linear-reference-based SV analysis.

Response: We appreciate the reviewer’s careful assessment of the section title’s alignment with methods. We would like to clarify that our study actually integrates two complementary analytical approaches: a standard linear-reference-based SV analysis (for the 984 accessions) and a graph-based pan-genome analysis (constructed from the 56 assembled genomes).

We acknowledge that the Methods section in the previous version did not adequately describe the graph-based component. As suggested, we have now expanded the Methods to include full technical details of our graph-based workflow, including pan-genome graph construction methodology, graph-based read mapping procedures, and variant genotyping using the vg toolkit to the revised Methods section (lines 658-676).

6. Use of 56 assemblies for graph—complexity not addressed.

The manuscript claims to have used 56 assemblies to build a graph for SV discovery (Lines 131–133), but again no methods are given. Even if such a graph exists, incorporating 56 assemblies would substantially increase graph complexity and generally complicate genotyping. How did the authors mitigate this? No description is provided, so the approach cannot be assessed.

Response: We thank the reviewer for raising this critical issue regarding graph complexity. We agree that incorporating 56 de novo assemblies poses a risk of creating an overly complex graph, which could hamper read mapping and genotyping (Supplementary Fig. 4, 7). To mitigate this, we adopted the following strategies:

Reference-guided graph construction: Using the vg toolkit, we first constructed a reference-based graph. We then assessed the completeness of this pan-genome based on the graph structure, before selectively integrating the PAVs identified from the 56

assemblies (Lines 187–260).

Targeted genotyping optimization: During genotyping using `vg call`, we specifically focused on PAV variants. This avoided the immense computational burden of traversing highly complex sub-graphs.

Together, these measures allowed us to leverage the genetic diversity of the 56 assemblies while maintaining a streamlined graph structure suitable for efficient and accurate genotyping of the 984 resequenced accessions. We have expanded the Methods section with these details (Lines 658-676).

7. Population composition limits cross-country comparisons.

Lines 211–214: The purported comparisons among countries/regions are not convincing because the sampling is heavily skewed toward Chinese accessions. With such a biased panel, country-level inferences are not robust; the text should be revised to avoid over-interpretation.

Response: We sincerely thank the reviewer for pointing out the sampling bias in our population composition. We fully agree that the heavy skew toward Chinese accessions makes direct cross-country comparisons unbalanced and prone to over-interpretation.

To mitigate this issue, our analysis employed a classification strategy based on ecological regions rather than political boundaries. Specifically, to prevent the over-representation of certain regions from skewing the results, we divided the Chinese accessions into six distinct sub-populations (CHN1–CHN6) based on their ecological origins (e.g., Northeast, Huang-Huai-Hai, and Southern regions). Additionally, we merged accessions from Korea, North Korea, and Japan into a single group (JAP&KOR) due to their similar latitudes and ecological backgrounds, and excluded countries with extremely limited sample sizes (e.g., Russia, Thailand).

The final grouping used for analysis is as follows: CHN1 (228), CHN2 (284), CHN3 (159), CHN4 (77), CHN5 (34), CHN6 (23), NA (North America, 103), and JAP&KOR (59).

We have revised the manuscript to clarify that our comparisons focus on these ecologically defined groups, which provide a more robust scientific interpretation than

simple country-level comparisons. This ecological classification strategy is well-supported by previous soybean pan-genome and population genetic studies (Liu et al., 2020; He et al., 2024 and so on).

Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-Genome of Wild and Cultivated Soybeans. *Cell*. 2020;182(1):162-176.

He J, Fu L, Hao X, Wu Y, Wang M, Zhang Q, et al. Identification of QTL-allele systems of seed size and oil content for simultaneous genomic improvement in Northeast China soybeans. *Front Plant Sci*. 2024;15:1483995.

8. Phenotypes across different seasons and BLUP integration are unclear.

Phenotyping spans winter (2021) and summer (2022, 2023), implying substantial environmental differences (day length, temperature). The manuscript seems to use BLUPs, but does not specify the mixed model, fixed/random effects, or how year/season and G×E were modeled. I recommend:

- Provide the exact model specification (effects, variance components, software).
- Report GWAS per environment/year and highlight reproducible peaks across environments; BLUPs can then be presented as a complement, not a substitute.

Response: We thank the reviewer for the insightful comments regarding our phenotypic analysis. We acknowledge that the mixed model specifications were not clearly detailed in the previous version. We have now revised the Methods section to include the exact model specifications. Specifically, quality traits were scored based on direct phenotypic observations. For quantitative traits, we utilized BLUPs for a subset of traits, while simple arithmetic means were calculated for the remaining traits.

1. Model Specification for BLUP Calculation: Phenotypic data across different seasons were integrated using a linear mixed model implemented in the R package phenotype. Consistent with the previous method Fang et al., Piepho et al., and Buckler et al.

2. Model Specification for GWAS:

Genome-wide association studies were performed using EMMAX (Efficient Mixed-Model Association eXpedited). The model is defined as: $y = X\beta + Zu + e$. Where y is the vector of BLUP or mean values; X is the design matrix for fixed effects

(including the SNP marker and population structure covariates); β is the vector of fixed effects; u is the vector of random polygenic effects with variance structure $u \sim N(0, K\sigma_g^2)$, where K is the kinship matrix; and e is the residual error.

3. Response to “Report GWAS per environment/year”:

During this comparative analysis, we observed that the two years of data from the Changxing environment acted as highly complementary datasets, showing consistent genetic architectures. In contrast, the Hainan environment represented a distinctly different ecological background. Therefore, to provide a clear, concise, and representative overview in the main text, our revised GWAS figures now display three key panels: (1) the Hainan environment, (2) the Changxing environment (integrating the complementary data from both years), and (3) the multi-environment BLUP results. By comparing these single-environment GWAS results with the BLUP-based GWAS, we demonstrated that the core significant peaks identified by the BLUP approach are highly reproducible and stable across these distinct ecological contexts (Supplementary Fig. 27-29, Supplementary Fig. 52). The method description can be found from lines 768-794.

Reference

- 1) Fang C, Ma Y, Wu S, Liu Z, Wang Z, Yang R, et al., Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 2017;18(1):161.
- 2) Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al., The genetic architecture of maize flowering time. *Science.* 2009;325(5941):714-718.
- 3) Piepho, H.P., Möhring, J., Melchinger, A.E. et al. BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica.* 2008;161, 209–228.

9. SV-GWAS vs. SNP-GWAS inconsistency with stated claims.

In Supplementary Figures 16–42, SV-GWAS peaks appear largely co-located with SNP-GWAS, and SNP-GWAS often shows clearer peaks. This seems inconsistent with the repeated claim that “InDels, PAVs, and SVs outperformed SNPs in pinpointing

causal genes” (Lines 290–291). The authors should include direct side-by-side comparisons (same locus, same trait) in the main text and reconcile these results with their claim.

Response: We agree with the reviewer that a large proportion of SV-GWAS signals co-localize with SNP-GWAS signals. As stated in our revised text, approximately 50% of the signals overlap. This is biologically reasonable because many SNPs are in high linkage disequilibrium (LD) of SVs. However, there is still a 50% overlap missing (Supplementary Table 14).

However, the key advantage of SV-GWAS does not lie in these overlapping signals, but rather in its ability to resolve the precise causal mechanism that SNP-GWAS cannot. For instance, at the W1 locus on chromosome 13, significant association signals were detected across SNP-, InDel-, SV-, and PAV-GWAS analyses. While the top SNPs in this region do not cause any obvious alterations to the gene, we discovered that multiple insertions leading to the premature termination of the third exon are the actual causal variant underlying the flower color variation. This perfectly illustrates that SNPs merely flag the associated region, whereas SVs pinpoint the exact structural change causing the phenotype. Figure 6B

To directly address your concern and reconcile this with our claim, we have now added a side-by-side comparison for specific loci, including the W1 locus mentioned above. This direct comparison clearly demonstrates that while SNP-GWAS may show sharper peaks in some overlapping regions, SV-GWAS uniquely identifies the causal structural variants that directly alter gene structure. Furthermore, as highlighted in the revised Discussion, we identified 550 trait-related SVs that were completely missed by SNP-GWAS. Lines 324-333.

10. Case 1 (Fig. 4): scale of reported events vs. expectations from a graph approach. If a true graph-based SV framework was employed across wild vs. cultivated soybeans, one would expect recovery of known large-scale structural polymorphisms in addition to small insertions/deletions. However, the events highlighted are on the order of a few hundred bp to ~4 kb, which is underwhelming relative to the promise of a pangenome

graph and raises doubt that graph-based discovery underlies the result. The exact analytical route from raw data to these events remains ambiguous.

Response: We thank the reviewer for this insightful comment regarding the scale of structural variations detected. We would like to clarify that while the specific case study highlighted in Fig. 5 involves SVs in the range of a few hundred base pairs to ~4 kb, our pan-genome analysis detected a broad spectrum of structural variations, including larger events (>10 kb) present in the dataset (as shown in the genome-wide distribution in Supplementary Fig. 45).

The decision to highlight these specific events in Fig. 4 was driven by their biological significance rather than their physical size. For soybean, many functionally relevant variations, particularly those associated with gene presence/absence or transposon insertions, fall within the 1 – 5 kb range.

We have clarified the workflow from raw data to SV discovery in the revised Methods to demonstrate how the graph structure facilitated the detection of these functional variants. Lines 658-691

11. Case 2 (Fig. 5, GmW1): overexpression experiment lacks essential information.

The overexpression (OE) validation does not report (i) the number of independent transgenic events, (ii) qRT-PCR confirmation of overexpression levels in those lines, nor (iii) any correlation between expression dosage and phenotype across independent events. These are standard requirements to support causality.

Response: We thank the reviewer for pointing out the missing details regarding the overexpression experiment. In the revised manuscript, we have provided the complete data to support the causality of *GmW1*:

Number of Independent Events: We generated two independent transgenic overexpression lines (OE-1, and OE-2).

Expression Confirmation: We performed qRT-PCR to confirm the expression levels in all two lines. The results show significantly elevated transcript levels compared to the wild type.

Dosage-Phenotype Correlation: We observed a consistent phenotypic trend across

the three lines. Furthermore, the degree of phenotypic change correlated positively with the expression levels of *GmWI* in these lines, reinforcing the conclusion that *GmWI* regulates the trait (Figure 5).

12. Case 3 (Fig. 6, *GmMATE1*): EMS line acquisition/validation unclear; GWAS signal weak.

PCR validation is appreciated, but the GWAS peak is not compelling. Crucially, the Methods do not describe how the EMS mutant was identified, confirmed (Sanger/WGS), background genotype, generation, or backcrossing. Given the heavy background mutational load in EMS lines and the environmental sensitivity of seed traits, single-generation evidence is insufficient. At minimum, one expects M2 genotyping and M3 bulk phenotyping, and genetic segregation tests linking the candidate mutation to the trait.

Response: We thank the reviewer for this professional suggestion regarding the functional validation of *GmMATE1*. We have revised this section to provide more robust evidence addressing the concerns about the EMS mutant and environmental sensitivity.

1. Multi-environment phenotypic validation:

To address the concern regarding the environmental sensitivity of seed traits, we phenotyped the EMS mutant line in two distinct environments: a field station in Hainan (natural field conditions) and a greenhouse in Zhejiang (controlled environment). The mutant consistently exhibited significantly altered seed weight compared to the wild type in both environments (Supplementary Fig. 56, Figure 7F). This stability across diverse conditions strongly suggests that the phenotypic variation is genetically determined rather than being an artifact of environmental noise.

2. Validation of the mutation:

Regarding the nature of the EMS mutant, we confirmed the point mutation via Sanger sequencing. This mutation results in an amino acid substitution that disrupts the conserved domain of the *GmMATE1* protein (Supplementary Fig. 54), suggesting a likely loss-of-function effect.

3. Limitations and Future Work:

We fully agree that CRISPR/Cas9 lines and detailed segregation tests would be the gold standard. While *GmMATE1* transgenic lines are currently under construction, the genetic transformation cycle in soybean is lengthy. Therefore, we utilized the available EMS mutant to provide preliminary genetic evidence. The consistent phenotype observed across two locations, combined with our population genetic evidence, supports the conclusion that *GmMATE1* plays a critical role in regulating seed weight. Lines 418-445.

13. Typographical error.

Line 655: “peach evolution” appears to be a typo and should read “soybean” (please check surrounding headings for consistency).

Response: We gratefully acknowledge the reviewer's careful identification of formatting and grammatical errors. A comprehensive proofreading has been performed to address these issues.

14. Global assessment of Methods and presentation.

Overall, the Methods are too vague to enable reproduction or rigorous evaluation. Given the repeated use of “graph-based,” the paper should either (i) provide a complete, reproducible graph workflow and re-analyze results accordingly, or (ii) remove graph-based claims and frame the study explicitly as a linear-reference-based SV analysis with SURVIVOR-merged callsets. Even in the latter case, the functional validations must be strengthened as outlined above. In its current form, the manuscript does not meet Nature Communications standards.

Response: We thank the reviewer for the comprehensive assessment and the constructive criticism regarding the reproducibility and rigor of our study. We acknowledge that the Methods section in the previous version was insufficient. In response to this critical feedback, we have chosen to implement the first option proposed by the reviewer: providing a complete, reproducible graph-based workflow. We have made substantial revisions to address the specific deficiencies:

Complete Graph Workflow: We have added a detailed description of the graph construction (using vg), mapping, and genotyping steps, including software versions, parameters, and file formats.

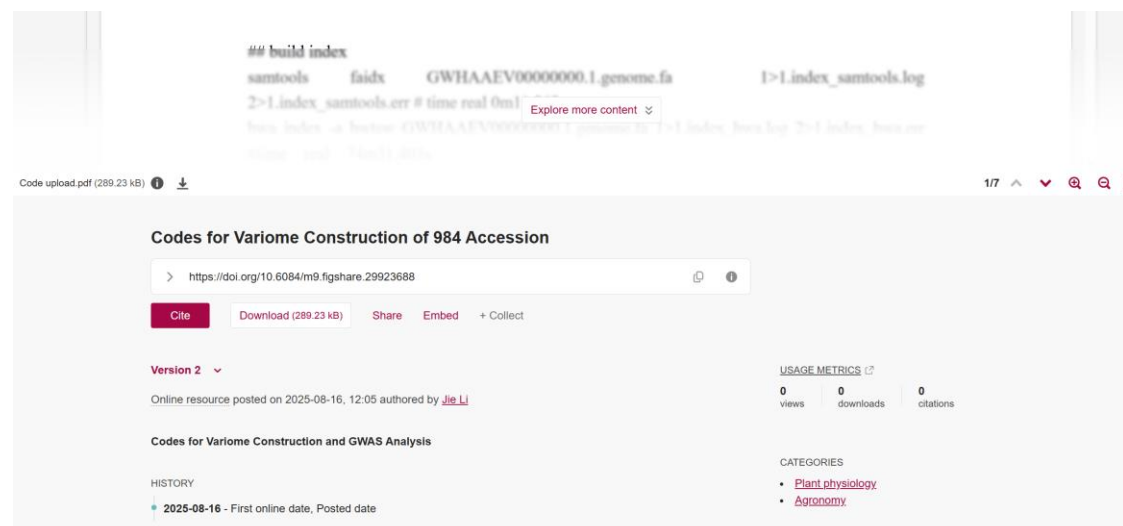
Justification of Parameters: We have provided validation and reasoning for the SURVIVOR merging thresholds to ensure the robustness of our SV callset.

Strengthened Functional Validation: We have refined our interpretations and provided additional evidence to support the functional analysis, ensuring the conclusions are robust. We believe these revisions have significantly strengthened the manuscript, making the methodology transparent and reproducible. Lines 658-691.

Reviewer #3 (Remarks on code availability):

In the “Remarks on code availability,” the authors provide a figshare link. However, when accessed, the link results in a “403 Forbidden” error, meaning the resource is not publicly accessible.

Response: We thank the reviewer for checking our code and data availability. We apologize for the earlier technical issue with the figshare repository access settings due to an access permission issue on the repository. We have now updated the access restriction settings, and the link is publicly accessible. The current link is: <https://doi.org/10.6084/m9.figshare.29923688>. We have verified that the resource can be accessed without login, and the updated link has also been provided in the revised manuscript.



Response to reviewers

Reviewer #1 & #2: We are grateful for your approval of our revisions.

Responses to Reviewer #3:

The authors have substantially improved the manuscript, particularly by providing a clearer and more reproducible description of the graph-based pangenome methodology. The addition of detailed workflows for graph construction (vg, MUMmer, and SyRI), mapping, and genotyping is appreciated and resolves previous concerns regarding reproducibility. However, despite these improvements, several core conceptual and interpretational issues remain unresolved, limiting the scientific clarity and overall impact of the study. The remaining major concerns and required revisions are outlined below.

Q1. Conceptual inconsistency: two SV frameworks without clear integration

The manuscript currently combines two distinct variant discovery frameworks: (i) resequencing-based SVs identified using short reads and multiple callers, and (ii) assembly-based PAVs derived from graph/pangenome analyses. However, these two variant systems remain only loosely connected and are not conceptually integrated. Specifically, only a small subset (~4,523 variants) overlaps between the two datasets, whereas the vast majority (~54,237 variants) are unique to one approach. The manuscript does not sufficiently clarify whether these differences reflect complementary biological signals or methodological discrepancies (e.g., false positives or false negatives). At present, the manuscript implicitly assumes that non-overlapping PAVs represent variants inaccessible to short-read approaches, but this conclusion is not adequately demonstrated.

The authors should explicitly address the discrepancy between the two SV discovery systems. This should include validation or comparison of overlapping versus non-overlapping variants, analyses of size distributions and genomic contexts for each variant set, and evidence demonstrating that PAV-specific variants are truly undetectable using short-read data. Alternatively, the authors may consider refocusing the manuscript around a single, coherent SV framework or providing a systematic integration showing how the two approaches complement one another at shared loci, as the current presentation does not clearly achieve either.

Response: We are grateful for the positive feedback from the reviewers. As suggested, we now explicitly demonstrate their complementarity through systematic integration and validation, rather than assuming that non-overlapping PAVs are inherently undetectable in short-read data.

Size distribution and technological biases dictate detection capability: As shown in Figure 1F, the discrepancy is largely driven by fundamental differences in the detection principles of the two approaches. Resequencing-based SV detection relies on short-read local alignment, which inherently struggles to resolve large insertions because the short reads cannot span the entire variant or its breakpoints. In contrast, assembly-based PAV detection (using MUMmer and Syri) relies on long-

read/assembly-based global collinearity, which captures large structural variations by identifying macroscopic shifts and rearrangements between whole-genome alignments. Consequently, with our 150 bp paired-end short-read data, detecting insertions is highly restricted; even after merging calls from four different SV callers using SURVIVOR, only 57 reliable insertions were identified, whereas assembly-based methods detected 34528 insertions. For deletions, the vast majority (32.3%) detected by short-read data are small events (<100 bp) (Figure 1F). In stark contrast, the 79.8% of deletions identified by PAVs were larger than 100 bp (Supplementary Figure S9). The different length distributions of assembly- and resequencing-based SVs demonstrate the complementarity of the two methods. We now frame this technological distinction in Results (Lines 219-222) and Discussion (514-522).

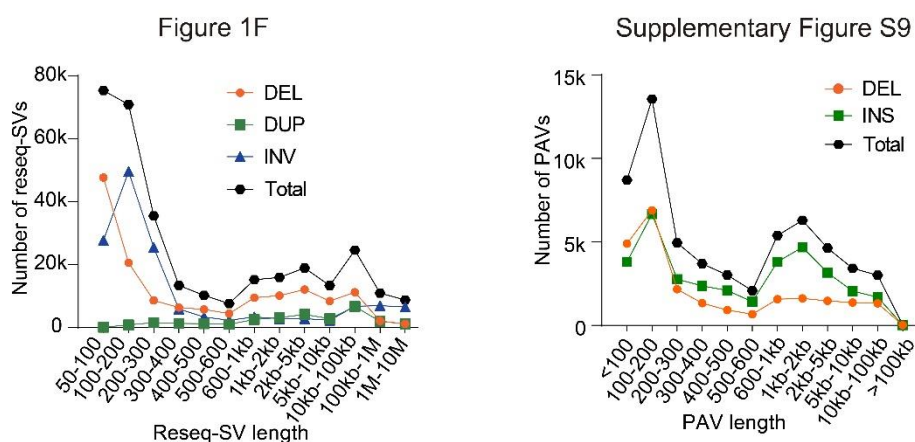


Figure 1F and Supplementary Figure S9. Size distribution of reseq-SVs and PAVs by types.

Q2. Ambiguity in the use of the term “SV” throughout the manuscript

A major issue affecting readability and interpretation is the inconsistent use of the term “SV.” Because the study involves fundamentally different classes of variants—including short-read SVs, assembly-derived PAVs, and graph-based genotypes—the blanket use of the term “SV” creates substantial ambiguity. In several sections, particularly at transitions between paragraphs, it is unclear whether “SV” refers to resequencing-based calls or pangenome-derived variants. This ambiguity makes it difficult for readers to evaluate the biological conclusions and undermines the interpretability of the entire study.

The authors should introduce clear and consistent terminology (e.g., “reseq-SV,” “PAV,” and “graph-SV”), explicitly indicate which variant set is used in each analysis and figure, and ensure that conclusions are not conflated across fundamentally different variant classes.

Response: We sincerely thank the reviewer for highlighting this critical issue. We have revised the terminology throughout the manuscript, and the updated terms are indicated as follows:

- 1) Reseq-SV: Resequencing-based structural variants.

2) PAV: Presence/absence variations based on the pan-genome.

Specifically, we have comprehensively revised the entire manuscript to ensure that all instances referring to resequencing-based structural variants are uniformly changed to “Reseq-SV”, and all instances referring to pan-genome-derived presence/absence variations are uniformly changed to “PAV”.

To further clarify our analytical framework, we have explicitly restricted the general term “SV” throughout the main analyses to refer exclusively to reseq-SVs derived from short-read data. For pangenome-derived variations, we exclusively retained and used the term “PAV,” deliberately avoiding a generic “SV” label to prevent any conceptual conflation.

Q3. The added value of the graph-based pangenome remains unconvincing

Although the methodological descriptions have improved, the manuscript still does not convincingly demonstrate the unique biological insights enabled by the graph-based approach. Specifically, no quantitative comparison between linear-reference and graph-based SV detection is provided. Furthermore, the representative case studies (CHS, GmW1, and GmMATE1) all involve relatively small variants (hundreds of base pairs to a few kilobases) that are readily detectable using standard short-read pipelines. No example is presented in which the graph-based approach reveals structural features that cannot be recovered through conventional linear mapping.

The authors should provide at least one of the following: (i) a direct comparison of SV detection performance between linear and graph-based approaches, (ii) a demonstration of a biologically relevant SV uniquely recoverable using the graph-based framework, or (iii) characterization of larger-scale structural variation, such as inversions, CNVs, or repeat-associated structures. Without such evidence, the graph-based component appears descriptive and not clearly functionally justified.

Response: We sincerely thank the reviewer for this rigorous critique. We agree that our initial manuscript failed to demonstrate the unique biological insights provided by the graph-based pan-genome. We provided both quantitative evidence and functionally validated case studies enabled by graph-based pan-genome.

To demonstrate the functional necessity of the graph-based approach, we provided compelling experimental and genetic evidence for PAVs affecting gene promoter regions largely unresolved in our short-read reseq-SV dataset. We identified PAVs causing the loss of promoter regions in *SoyZH13_15G048502* and *SoyZH13_12G099200*. Through dual-luciferase assays, we demonstrated that these promoter deletions significantly reduce the LUC/REN ratio, demonstrating that these structural variations directly impair gene expression (Figure 3F, G), establishing a direct causal relationship between these structural variations and gene expression modulation.

Using PAV-GWAS approach, we identified significant loci for leaf shape and 100-seed weight that were undetected by SNP-based or short-read SV-based GWAS. For leaf shape, a 245 bp promoter deletion perfectly co-localizes with known leaf morphology QTLs (leaflet shape 7-2, leaflet length 10-3, and leaflet width 7-1). The associated gene shows leaf-specific expression pattern, strongly supporting its role as a

key regulator of leaf morphology (Figure 3I). For 100-seed weight, a 142-bp deletion in the 3'UTR region was identified in a gene highly expressed during early seed development. The expression pattern precisely matches the developmental timing of known seed weight QTLs (24-3/29-5/36-5/50-13), suggesting this variant plays a crucial role in determining seed weight (Figure 3H). Through this comprehensive integration of short-read data and pan-genome data, we have not only enhanced the completeness of variant detection but also uncovered fundamental biological insights that remain inaccessible using traditional approaches (Lines 270-280).

Q4. Interpretation of the CHS locus remains incomplete and insufficiently reconciled with existing models

The CHS locus is presented as a central example in the manuscript, yet the current analysis focuses primarily on relatively small deletions (259 bp, 772 bp, and 4 kb), while previously reported large-scale structural haplotypes at the I locus are neither reconstructed nor systematically compared. Given that CHS regulation is known to involve complex structural configurations and RNAi-mediated silencing, it remains unclear whether the identified TE-associated variants represent primary causal structures or secondary modifiers within larger haplotypic backgrounds.

The authors should explicitly place their findings within the context of previously reported CHS/I locus structural models, clarify whether the identified SVs define novel haplotypes or act within existing structural backgrounds, and provide a clear structural model integrating large-scale genomic configuration, TE insertions, and methylation/siRNA-associated regulatory effects.

Response: We appreciate the reviewer's insight critique regarding the CHS/I locus complexity. Following the reviewer's comment, we have re-evaluated this locus by integrating our pan-genome with the classic CHS structural models (Supplementary Figure 47). INV is not the only factor affecting seed coat color in observed in previous report³³.

In our analysis, all accessions with INV events and TE insertions (1st and 2nd columns from left side in Supplementary Figure 47) showed yellow colors. Notably, 77.8% (14/18) of the accessions containing TE1/TE2 insertions but lacking INV events also showed a yellow seed coat, suggesting that TE insertions can affect seed coat color independently of INV events. Furthermore, around half (4/7) of the accessions containing a single TE insertion but no INV events were yellow, suggesting that the number of TEs may affect methylation levels, thereby influencing CHS gene expression and subsequent seed coat color transition. Accessions lacking both INV and TE insertions displayed a black seed coat.

Considering that TE-methylation acts at the transcriptional level through DNA methylation, whereas the INV-siRNA axis acts at the post-transcriptional level through mRNA degradation, our analysis suggests that TE insertions and INVs within the CHS cluster region could synergistically contribute to the seed coat color transition from black to yellow. Therefore, we conclude that seed coat color is a comprehensive trait influenced by multiple factors. (Lines 554-570).

Reviewer #3 (Remarks to the Author):

The authors have carefully addressed my previous concerns. In particular, the clarification of terminology has improved the readability and interpretability of the manuscript. The authors have also provided additional analyses and examples to support the complementary roles of resequencing-based SVs and pangenome-derived PAVs, which strengthens the overall narrative.

I believe the manuscript has been sufficiently improved and now meets the standard for publication.

Response: We thank Reviewer #3 for their constructive feedback and support for publication.