nature portfolio

# Peer Review File

# Genetic architecture of sugarcane traits in a polyploid genomics framework

Corresponding Author: Professor Xingtan Zhang

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Wang et al. primarily describe the assembly of a complex sugarcane genome (POJ2878) and the population resequencing of 981 accessions. They also propose three effective tools for polyploid genome analysis. The development of a novel assembly strategy for such a complex genome is both exciting and significant, and the proposed tools indeed offer promising approaches to tackle challenges in complex polyploid genomes. However, there are several major concerns that should be addressed:
Major Concerns:
1. Overinterpretation of gene function: The sections on domestication and GWAS are highly condensed and overly ambitious. The authors frequently draw conclusions about gene functions that are not substantiated by either the cited literature or their own experimental data. For example, functions are attributed to the TB1 and SRC2 genes in sugarcane, yet no experimental evidence is presented, and the supporting literature refers to studies in maize and Arabidopsis. The manuscript also overlooks the significant challenges involved in inferring orthology at both the sequence and functional levels—particularly in a complex polyploid like sugarcane. This issue is exemplified by the statement in the Discussion: “These findings redefine sugarcane domestication as not only a story of metabolic optimization but also one of cellular architecture evolution.” This is a substantial overstatement of the data and detracts from the otherwise solid contributions of the paper. Such speculative claims, based on weak or indirect evidence, contribute to an impression of carelessness that is at odds with the impressive technical achievement of the genome assembly. In conclusion, the manuscript would benefit from the removal or substantial revision of these sections to avoid undermining the credibility of the work.
2. Assembly Validation: Could the authors provide additional validation beyond sequencing data? At a minimum, validation of several large structural variants would be helpful. Additionally, can any small HE (homoeologous exchange) regions be experimentally validated, for example by FISH?
3. Population Analysis: The manuscript discusses three groups—S. hybrid, S. spontaneum (S.s), and S. officinarum (S.o). Based on the mapping method described (line 1290), all accessions were aligned to the POJ2878 reference genome. Theoretically, there should be no shared SNPs between S.s and S.o. This casts doubt on the interpretation of some of the population structure in Figure 2. A more detailed explanation and methodological clarification are needed.
4. Domestication Analysis: According to Healey et al. (https://www.nature.com/articles/s41586-024-07231-4#Sec1), domestication should be discussed in the context of S. officinarum and S. robustum, rather than modern cultivars. The breeding process involving current cultivars should be classified as crop improvement, similar to how rye was used to enhance disease resistance in wheat. Thus, the interpretation and analysis of domestication here appear to be inaccurate.
5. Genome Assembly Focus: The assembly is arguably the paper’s highlight. While tools like Allele-Express and KMERIA are valuable for polyploid genome analysis, the manuscript lacks a clear connection between these tools and the assembly itself. There is insufficient explanation of how results from these tools feed back into genome construction or interpretation.
Minor Comments and Suggestions:
6. Please provide a coverage distribution plot for the whole-genome assembly. Additionally, indicate the proportion of IBD (identity-by-descent) regions across the haploid genome. How many regions are shared among haplotypes? What is the maximum and average number of haplotypes sharing a region?
7. Provide an example demonstrating how methylation data helped resolve complex regions during genome assembly.
8. How are Pore-C connections allocated among shared haplotypes?

9. For Iso-Seq data, include a saturation curve showing the relationship between sequencing depth and transcript discovery. This is necessary to confirm whether the data are sufficient for reliable quantification.
10. Please describe in detail the method used for identifying favorable haplotypes.
11. Figure 3d contains confusing labels. The relationship between genome haplotypes and labels such as so_hap and ss_hap needs to be clarified. Additionally, clarify what determines haplotype count—are they based on genome haplotypes, isoforms, or copy number variation?
12. Of the three companion papers mentioned for the major methods, only the paper on C-Phasing is cited. The other two are not provided and should be included.
13. The Allele-Express GitHub repository lacks English documentation. Please add an English-language explanation of the tool.
14. A workflow diagram for the Allele-Express method should be provided. Also, explain how homologous gene groups are defined and what filtering parameters are used (e.g., mapping quality, edit distance). How does Allele-Express associate transcripts with complex genomic regions? How are alleles assigned to homologous gene groups and genomic loci?
15. Can Allele-Express and Iso-Seq quantification results be validated—either experimentally or using RNA-seq data? If so, please provide supporting evidence. Is this method reliable enough to replace traditional RNA-seq for stable quantification?
16. In Figure 2 (panels a–d), the same colors are used for the three species, but the color assignments differ in panels e–f. Please use consistent color schemes across all subfigures.
17. What is the reference genome used in Figure 4a?
18. In line 667, label (e) appears to be incorrect.
19. Lines 64-6: To provide a balanced discussion, it should be acknowledged that the easy availability of sugar has contributed to a global health crisis. The phrase “reshaped human migration patterns” is also too euphemistic a reference to the transatlantic slave trade.
20. The terms “noble” and “nobilization” should be explicitly defined upon first use. These are technical terms specific to sugarcane breeding and may not be familiar even to geneticists working with other plant species.

(Remarks on code availability)
The code is accessible and properly documented.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
This manuscript describes an impressive advancement in sugar cane genomics, really the last, biggest, and hardest agriculturally significant plant genome out there. Overall I was impressed with both how well the new genome assembly worked and how clearly the paper was written. I think this study would be of wide interest and the genome assembly is highly impactful with a lot of potential for reuse. I have a number of minor concerns and suggestions, primarily focused on the downstream analyses presented using this genome assembly and one modern concern regarding the novelty of the k-mer based association method.

Moderate:
1. Lines 437-446 (and 497): Kmer based GWAS is a technique that has been used in a number of previous contexts. Is KMERIA significantly different from the method of He et al 2024 which was also using kmers to link traits to copy number variation via GWAS? (He, C., Washburn, J. D., Schleif, N., Hao, Y., Kaeppler, H., Kaeppler, S. M., ... & Liu, S. (2024). Trait association and prediction through integrative k-mer analysis. The Plant Journal, 120(2), 833-850
Note: I don’t question the biological results obtained via the author's method, I just don’t believe this can correctly be framed as a novel solution, let alone a paradigm change

Minor:
2. Line 212: “exceeding” rather than “exeeds”? Something is not quite right in this sentence.
3. 215-218: these dates would place the transposon bursts prior the hybridization events that produced POJ2878 which occurred less than 200 years ago. Does that suggest that the true progenitor lineages of used in producing POJ2878 were significantly different from those generated in this study?
4. 281-283: If it is necessary to use long read iso-seq technologies to measure gene expression, can short reads be confidently used for SNP/indel discovery?
5. 304-306: 30 Mb out of a 10.37 Gb genome is actually not very much. I don’t think this really does underscore the widespread use of POJ2878 in breeding.
6. 323-326: Typically optimal alleles for fitness/yield will not be those with the most extreme expression levels. See Kremling et al 2018. (Kremling, K. A., Chen, S. Y., Su, M. H., Lepak, N. K., Romay, M. C., Swarts, K. L., ... & Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature, 555(7697), 520-523.)
7. 425-436: If there isn’t a significant correlation across the entire dataset it is fair to try a few other ways of slicing and dicing the data, but it does introduce a multiple testing problem. P values of only 0.04 and 0.01 aren’t high enough to be convincing in the context of likely running a number of different tests with different cutoffs and approaches.
8. For an international journal, describing populations of sugar cane as Chinese/non-Chinese will aid intelligibility rather

than national/international which are terms whose meaning will vary depending on the country of the reader.
9. Typo in selective on line 388.

Referee #4

(Remarks to the Author)
Wang et al. present an impressive study of sugarcane genomics, offering valuable insights into its complex polyploid genome through haplotype-resolved chromosomal-level assemblies. The approaches used are novel and powerful. The results shown are solid and engaging. This work represents a state-of-the-art in resolving highly complex polyploid genomic studies using a combination of long-read and short-read methods. To further push the boundaries, this study would benefit from stronger integration across novel methods, particularly in leveraging the phased genome to investigate subgenome and homoeolog-specific dynamics, which represents one of its most novel contributions. Additionally, greater emphasis on biological interpretation and synthesis, especially in the context of evolutionary and functional genomics, would enhance its overall impact. Nonetheless, this is a substantial and exciting contribution to the field that, with some refinement, could serve as a foundational resource for sugarcane biology and comparative genomics. Detailed comments are shown below.

The first paragraph of the introduction is beautifully written. The second paragraph presents some technical challenges, but mostly focuses on sequencing technologies. They should have mentioned the progress of using long reads in sugarcane genome assemblies, such as Healay et al. 2024, Nature. The third paragraph is largely a duplication of the abstract. Overall, the introduction brought about the importance of sugarcane research and describes challenges in sugarcane genomics, but missed the opportunities to introduce the current progress and challenges in sugarcane biology and evolutionary studies, which are the major selling points of this study.

This study proposed an approach to identify the subgenome origin of hybrid sugarcane using T2T monoploid genome assemblies and resequenced hybrid sugarcane genotypes. How does the current method and finding represent an improvement over an existing study from Zhang et al. 2018, Nature Genetics?

ePore-C seems an improved version as described by the companion paper, but only Pore-C was mentioned in the manuscript. It is not very clear which one was used.

C-Phasing was developed and introduced as a new method in this manuscript, yet the authors also submit a companion manuscript describing the C-Phasing algorithm and benchmarks. This is double-dipping. They should either report the C-Phasing method as a stand-alone work and USE/cite it in this study, or describe the development of C-Phasing in this study without a companion paper.

Similarly, based on the kmeria github and L1380-84, there is a manuscript in progress for this method as well. So, claiming the development of KMERIA as the novelty in this study seems improper. I want to point out that this work still represents great progress in polyploid genomics and population and quantitative genetics without describing C-Phasing, KMERIA, and Allele-Express as part of the work.

L154, intra-chromosomal interaction is almost always the strongest. The authors may refer to intra-subgenome interactions that look stronger than inter-subgenome interactions.

L201-204, is 25% enriched or depleted when compared to the genome background? If not comparing to the genome background or random expectation, you may say: “X% of the Ss-So recombined chromosomes were telocentric and acrocentric.”

L229-231, The difference in allele counts between So- Ss- is interesting. In Fig. 1b, the hybrid sugarcane genome has about two copies of the Ss subgenome and ~11 copies of the So subgenome. It seems that homoelogous genes are mostly retained in the Ss subgenome but not in the So subgenome. What may be the underlying mechanism causing this? Is it pseudogenization or biased fractionation? If so, which mechanisms prevail (frame shifts, premature stops, TE disruption, etc.)?

Table S4 shows over 99% short read mapping rate, which contradicts the results in L238 – 247, claiming mapping issues of short reads. Maybe Table S4 did not show the unique mapping rate, if so, please also show the unique mapping rate.

You have phased chromosomes and the allele-specific highly expressed genes (ASHEG) gives you power to quantify homo(eo)log/subgenome expression. What is the distribution of ASHEGs across homo(eo)logous chromosomes/subgenome? Are there any biased activations of chromosomes for expression (chromosomes with enriched ASHEG)?

How are branches colored in Figure 2d? There are “78 S. officinarum (SO), 290 S. spontaneum (SS), and 613 hybrid sugarcane cultivars”. I thought these were red, green, and purple, respectively, but those counts do not appear to match branch counts. For example, samples located on the right of the cladogram with a high proportion of So-genomic components were colored purple on the branch.

Fig3a legend references red and blue, but the figure appears red and green. Also, I find this figure confusing. How are edge colors determined? How are nodes spatially oriented? Does the distance from POJ2878 tell us anything? L304-306 reported

that 312/573 cultivars shared over 30 Mb of IBD sequences with POJ2878, which seems to be a very small number given the over 10 Gb genome size, the relatively short breeding history, and the lack of sexual recombination in sugarcane production. How is this not due to random noise?

Fig3c, is there a way to indicate which subgenome these are? Consider highlighting SUS2 so that it's more easily referenced in e-f.

L282, are SNPs called from unique read alignments?

L376-381. Fig 4b is XPCLR not Fst.

Fig4a has peaks and labels, but many of the most distinct peaks are not labeled? Are they not containing genes?

Fig 4efgh, lack of data to further support selection over random drifting

Fig 5efg are stretched and have a distorted ratio.

How to use kmeria-identified genomic regions to find genes? How was fine mapping done? Line 1380-1384 are not detailed.

L462-467, OE varies too much (7.2% vs 28.92%). Why?

The approach of the current manuscript has been focusing on finding genomic regions with varying methods (IBD/PAV/introgression/expression/selection), then performing GO/KEGG analysis to identify patterns. The integration between MAS-ISO-seq, GWAS, selection, and phased genome is weak. How biological questions can be answered by combining these data? The paper so far presents results from only one or two methods without organically combining them to strengthen the conclusion.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have thoroughly revised their manuscript and provided a detailed response letter. Overall, we are satisfied with most of the points, but two issues remain.
(1) The oligo-FISH analysis does not appear to have worked as intended, although the authors do not state this explicitly. Several panels show no signal or only a single colour, and the red probe frequently failed to yield a signal. Where signals are present, arrows should be added to indicate them clearly.
We appreciate the authors' willingness to present negative results; however, Extended Data Fig. 1 may confuse readers who do not also consult the Peer Review file. The presentation should therefore be improved, possibly by omitting negative results. Since Nature uses an open peer review policy, these results are already appropriately documented. We would also be curious to know why so many probes failed to produce a signal.
(2) The section of the population analysis referring to "a large number of evolutionarily conserved loci" remains unclear. Specifically, we ask the authors to provide additional information:
How many of these conserved loci show overlap between S. s. and S. o. in the POJ2878 genome, and where are these overlapping regions located?
Why these signals cannot be explained by ambiguous read mapping rather than true evolutionary conservation. Additionally, we wonder whether this could be related to aneuploidy—i.e., S. s. and S. o. reads both mapping to the same regions of the POJ2878 genome. Theoretically, the proportion of such shared regions should not be very high, and it is difficult to see how this would represent genome-wide differences.
Without these clarifications, it remains challenging to fully understand the biological significance of the population analysis in this context.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #3

(Remarks to the Author)
I appreciate the substantial amount of work the authors have put into revising and updating the manuscript. They have fully addressed the concerns I raised in the previous round of review and I have no additional points to raise.

Referee #4

(Remarks to the Author)
Wang et al. present a revised manuscript with greatly improved quality and clarity. The reviewer still has several concerns about the manuscript's conceptual framing, data interpretation, use of selection methods, and gene mapping. The selection analyses are improved, but may still suffer from the overly permissive cutoffs, which likely produced extensive false positives, undermining conclusions about selective sweeps. Concerns are also raised about the insufficient detail in fine mapping, inadequate discussions of biological insights, and several minor issues. Detailed comments are shown below.

The introduction is still mostly focusing on sugarcane genomics, not sugarcane biology and evolutionary genetics. The reasoning behind GWAS studies of parenchyma cell size and sucrose storage capacity is unclear. What is the significance of focusing on these traits? Are these poorly studied traits? Are there any related genes reported? For the population genetics analyses, are there any genes reported previously as undergone domestication? Without these information, it is difficult to comprehend the significance of this study and the necessity of conducting the reported analyses.

For subgenome allelic counts, $1.96 / 5 \approx 5.03 / 12$, so allelic counts appear primarily driven by gene #. On the other hand, the Ss subgenome has an average of ~2 copies, while the So subgenome has an average of ~11 copies. The number of alleles per gene copy is $1.96/2 = 0.98$ for Ss and $5.03/11 = 0.46$ for So, suggesting higher per-copy allelic diversity in Ss, not So (Line 264 – 266). This observation contradicts the higher average sequence identity in So (~95.51%) than in Ss (~95.00%) (Lines 266 – 273).

Table 1 and Suppl. Table 1 shows the N50 of ONT UL reads as 155.5 kb, but Suppl Figure 1b does not seem to have such a read distribution. Also, the longest reads from ONT sequencing are usually at the megabase level, even for regular ONT sequencing, but in the Suppl. Table 1, the longest ONT UL read shows 812 kb. Please double-check your data and statistics.

Figure 3a is still confusing and not straightforward, and the main text citing this result (L367 - 372) carries more information than the figure itself. The links between cultivars are almost all covered by nodes. The log2 transformation inappropriately enlarges smaller nodes.

The selection analyses are not performed reasonably. For the genome assembly size of 10.37Gb, they identified 11.43 million SNPs, corresponding to an average of 907 bp per SNP. They used 10kb windows with 2kb sliding steps to calculate XP-CLR and Fst, effectively using 11 SNPs per window and 2 SNPs per step. Such a shallow number of data points in each calculation will result in extremely high false positives, despite these methods may be developed to account for random drifting. Furthermore, the use of 5% and 10% cutoff values to call selective sweeps are very loose criteria, which basically calls the whole genome undergone selective sweeps, as shown in their Figure 4a-d, where the majority of chromosomes are beyond the critical value(s).

The revision provides more detail on how to use KMERIA identify candidate genomic intervals for gene discovery. They mentioned using LD to adjust the genomic interval. However, details regarding fine mapping are still missing. How were the reported genes fine-mapped from the at-least 100kb region? Without further details, these results may not be reproducible.

The discussion still focuses heavily on the new methods and reads as though these are products of this paper.

Minor comments:
Figures 1b,d,e are essentially the same. Some of them can become supplementary to reduce the size of this figure. Currently, the fonts are very small and impossible to read without zooming in.

There is a typo on line 48 – "profling with by our newly"

Figure 4j, species names are not labeled.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have satisfactorily addressed our outstanding concerns. Although we doubt that the approach followed by the authors is generally applicable to other polyploid systems, they provide sufficient empirical evidence that the detected SNP loci can be used to study population structure in Saccharum. We therefore recommend acceptance.

Referee #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Referee #4

(Remarks to the Author)
We appreciate the authors continued improvements. Their manuscript represents a significant contribution to sugarcane genomics, and they have thoroughly addressed our comments.

Minor:
ASHEG was not defined (L228).
For non-sugarcane researchers, the manuscript should state more clearly that Brix is a measure of sugar content.

Referee #5

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Point-by-point responses to Referees' comments

Dear reviewers,

We appreciate the thoughtful and constructive feedback from the referees and the editor. We have now substantially improved our manuscript and our key changes are:

1) a thorough refinement of the manuscript's language to temper claims and ensure they are fully supported by the presented data;

2) the inclusion of new experimental validations for key genes, including transgenic functional validation in both rice and sugarcane, to substantiate their roles in the biological processes discussed;

3) the experimental validation of large structural variants using haplotype-specific oligo-FISH probes, providing strong support for the accuracy of our genome assembly;

4) the incorporation of new data from a recent study on wild *Saccharum* species to enrich the context for understanding the domestication of *S. officinarum*;

5) the addition of a detailed explanation and a new workflow diagram to clarify our integrated analytical framework and illustrate the synergy between the computational tools employed;

6) the inclusion of a detailed algorithm description and validation for Allele-Express in Supplementary Note 2;

7) a careful reframing of the manuscript's scope and novelty, removing claims of new contributions for the C-Phasing and KMERIA tools and instead presenting them as established methods with proper citations to their respective companion papers.

Please see our point-by-point responses to Referees' comments:

Referee #1 (Remarks to the Author):

Wang et al. primarily describe the assembly of a complex sugarcane genome (POJ2878) and the population resequencing of 981 accessions. They also propose three effective tools for polyploid genome analysis. The development of a novel assembly strategy for such a complex genome is both exciting and significant, and the proposed tools indeed offer promising approaches to tackle challenges in complex polyploid genomes. However, there are several major concerns that should be addressed:

Major Concerns:

1. Overinterpretation of gene function: The sections on domestication and GWAS are highly condensed and overly ambitious. The authors frequently draw conclusions about gene functions that are not substantiated by either the cited literature or their own experimental data. For example, functions are attributed to the TB1 and SRC2 genes in sugarcane, yet no experimental evidence is presented, and the supporting literature refers to studies in maize and Arabidopsis. The manuscript also overlooks the significant challenges involved in inferring orthology at both the sequence and functional levels—particularly in a complex polyploid like sugarcane. This issue is exemplified by the statement in the Discussion: “These findings redefine sugarcane domestication as not only a story of metabolic optimization but also one of cellular architecture evolution.” This is a substantial overstatement of the data and detracts from the otherwise solid contributions of the paper. Such speculative claims, based on weak or indirect evidence, contribute to an impression of carelessness that is at odds with the impressive technical achievement of the genome assembly. In conclusion, the manuscript would benefit from the removal or substantial revision of these sections to avoid undermining the credibility of the work.

Response: We sincerely thank the reviewer for their insightful and constructive feedback. We agree that our original manuscript contained overstatements regarding gene function and the broader implications of our findings. In the revised version, we have carefully addressed these issues by substantially revising the relevant sections, incorporating new experimental data, and integrating additional literature to provide stronger support for our interpretations.

***Comprehensive Revision of Language and Claims.*** We have systematically revised the manuscript, particularly the domestication and GWAS sections, to temper our claims and more accurately reflect the supporting evidence.

(1) Softened Language for Inferred Gene Functions: For genes whose functions were previously inferred by homology (e.g., *SRC2*, *CCD8*, *UBP14*, *Mob1A*), we have replaced definitive language ("functions as," "regulates") with more cautious and appropriate phrasing ("is predicted to," "may play a role," "suggests a function in"). We now explicitly state that these functions are

predictions based on homology.

For example, we have modified the description of *SRC2* function as follows: “The *SRC2* gene (**Figure 4h**) encodes a type II membrane protein in *Arabidopsis* involved in vacuolar protein trafficking, suggesting the homology in sugarcane may related to vacuole development”.

(2) Moderation of High-Impact Statements: We have removed or substantially revised statements that previously overstated the impact of our findings. For instance, the "redefine sugarcane domestication" statement and "transformative approach" have been removed. Other strong claims have been moderated (e.g., "new paradigm" or “paradigm shift” have been moderated to more appropriate term "integrated genomic framework").

(3) Acknowledgment of Orthology Challenges: We have added a sentence to the Discussion that explicitly addresses the challenges of inferring orthology and functional conservation in a complex polyploid like sugarcane, which reads: “Nevertheless, given the complexity of polyploid genomes and potential functional divergence, it should be noted that experimental validation in sugarcane is required to confirm this predicted function. ” (**Lines 646-648, Page 24**)

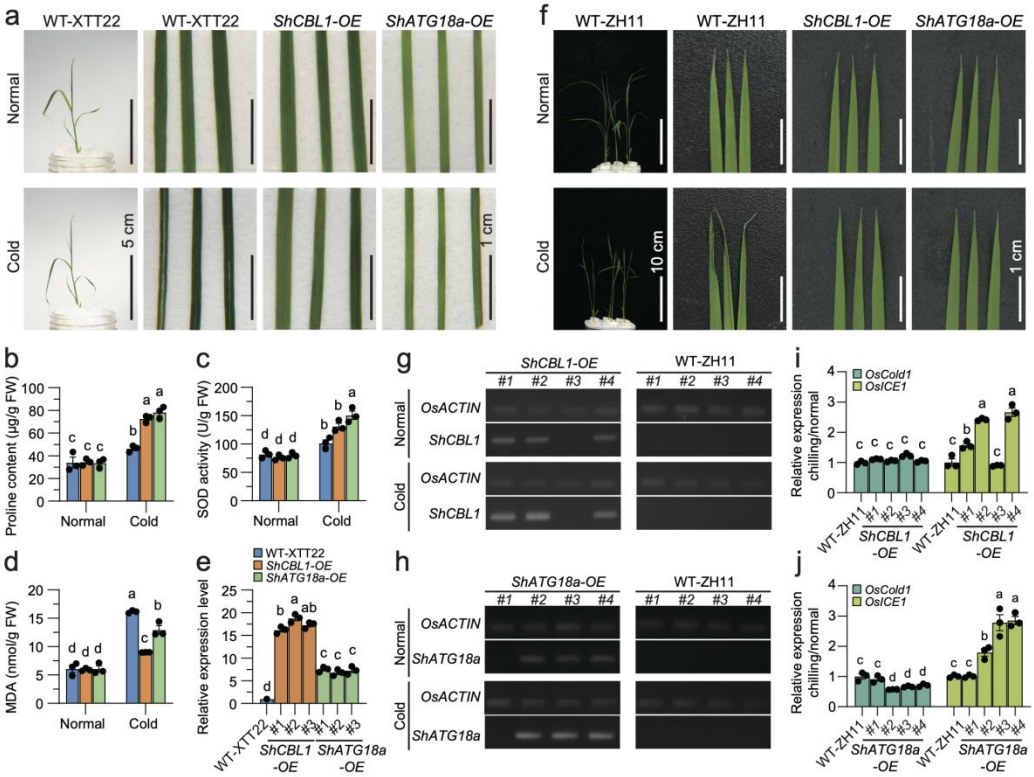
***Experimental Evidence from Transgenic Functional Validation.*** To directly validate the functional interpretations in our study, we conducted transgenic experiments for four key genes (*ShCBL1*, *ShATG18a*, *ShTIP1* and *ShSUT2*). These experiments provide direct evidence of gene function in a heterologous system (rice) and, most importantly, in sugarcane itself.

To validate the functions of genes identified through selective sweep analyses, we conducted overexpression studies on *ShCBL1* and *ShATG18a*. Overexpression of these genes in rice and sugarcane resulted in significantly increased proline content, superoxide dismutase (SOD) activity, and malondialdehyde (MDA) levels (**Extended Data Fig. 6**). These findings indicate that *ShCBL1* and *ShATG18a* function as regulators of the cold stress response, thereby validating predictions from our population genomics analysis.

The selective sweep analysis also identified an aquaporin protein (TIP1) that may be involved in vacuole development. Although functional validation of this gene's role in vacuole development requires extended time beyond the revision timeline, our new experiments demonstrate that overexpression of *ShTIP1* alters leaf width in both sugarcane (cultivar XTT22) and rice (ZH11) by reducing mesophyll cell size and decreasing the distance between veins (**Extended Data Fig. 7**). These results provide direct experimental evidence that ShTIP1 contributes to the regulation of cell size, supporting our hypothesis that vacuole-related genes were subject to selection due to their contributions to cellular architecture.

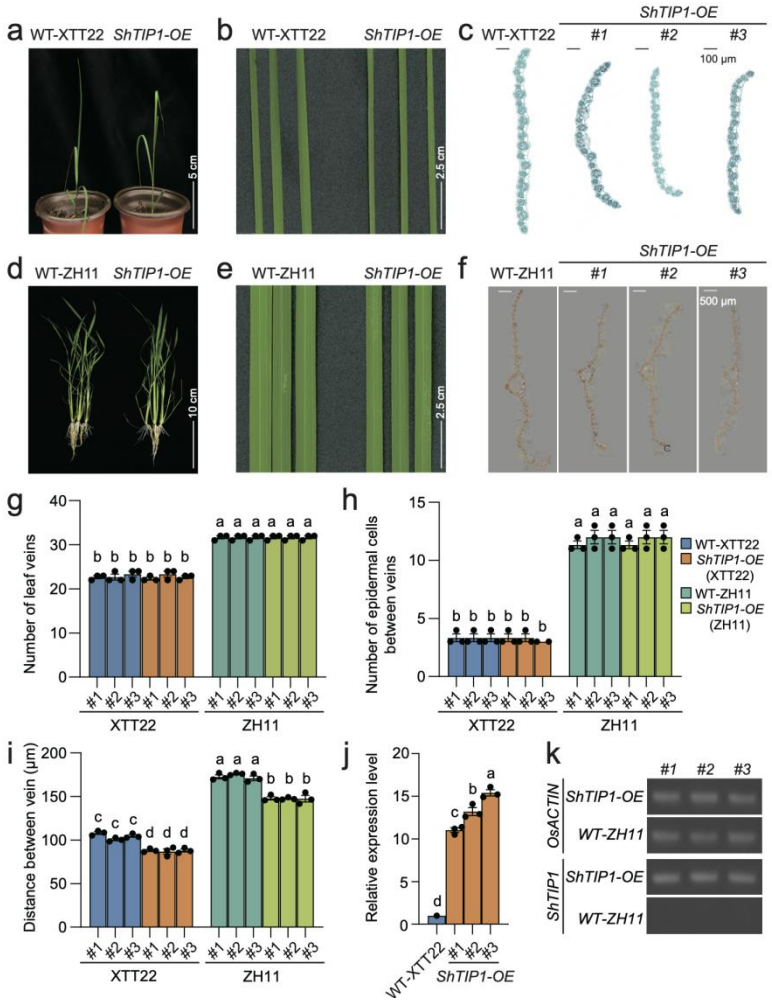
In addition to the three newly validated genes included in the revised manuscript, our original submission also reported that overexpression of the *ShSUT2* gene in cultivated sugarcane significantly increased parenchyma cell

size (**Extended Data Fig. 9**). Collectively, these experimental results elevated several of our key findings from predictive to experimentally validated conclusions, substantially strengthening the manuscript.

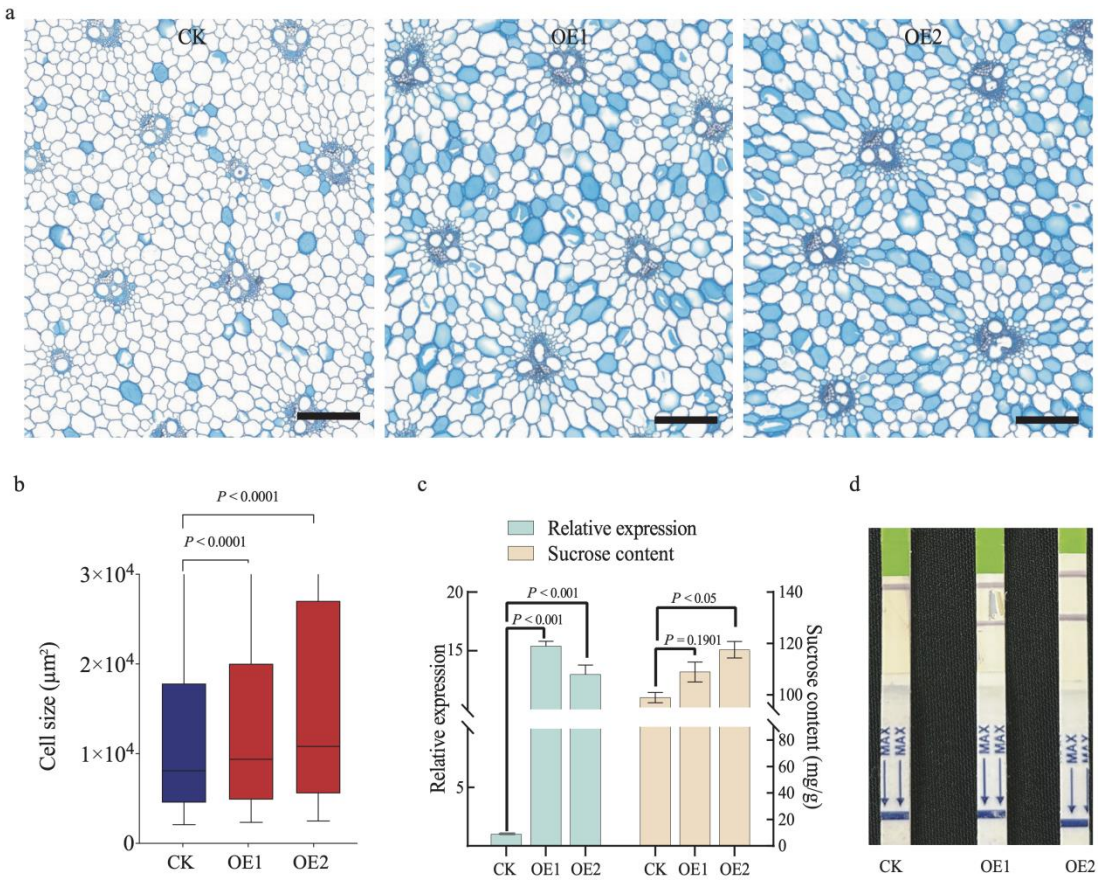



**Extended Data Fig. 6. Overexpression of *ShCBL1* and *ShATG18a* enhances cold tolerance in sugarcane and rice.**

(a) Phenotypes of wild-type (WT-XTT22) and transgenic sugarcane lines overexpressing *ShCBL1* (ShCBL1-OE) or *ShATG18a* (*ShATG18a*-OE) under normal (26 °C) and cold (-5 °C) stress conditions for 7 hours. Scale bars = 5 cm (left) and 1 cm (right). (b-d) Measurement of proline content (b), superoxide dismutase (SOD) activity (c), and malondialdehyde (MDA) content (d) in the leaves of WT-XTT22, *ShCBL1*-OE, and *ShATG18a*-OE sugarcane lines under normal and cold conditions. Data are presented as mean ± SD (n = 3). Different letters indicate statistically significant differences ($P < 0.05$, one-way ANOVA). (e) Relative expression levels of *ShCBL1* and *ShATG18a* in WT-XTT22 and transgenic sugarcane lines as determined by qRT-PCR. (f) Phenotypes of wild-type rice (WT-ZH11) and transgenic rice lines overexpressing *ShCBL1* or *ShATG18a* under normal and cold (4 °C) stress conditions for 5 days. Scale bars = 10 cm (left) and 1 cm (right). (g) Semi-quantitative RT-PCR analysis of *ShCBL1* and *ShATG18a* expression in transgenic rice lines under normal and cold conditions. *OsACTIN* was used as an internal control (For gel source data, see **Supplementary Figure 30**). (h, i) Relative expression levels of cold-responsive genes *COLD1* (h) and *ICE1* (i) in WT-ZH11, *ShCBL1*-OE, and *ShATG18a*-OE rice lines under normal and cold conditions. Data are presented as mean ± SD (n = 3). Different letters indicate statistically significant differences ($P < 0.05$, one-way ANOVA); n.s. indicates no significant difference.




**Extended Data Fig. 7. Overexpression of *ShTIP1* alters leaf morphology in sugarcane and rice.** (a) Phenotypes of 30-day-old wild-type (WT-XTT22) and *ShTIP1*-overexpressing (*ShTIP1*-OE) sugarcane plants. Scale bar, 5 cm. (b) Leaf morphology of WT-XTT22 and *ShTIP1*-OE sugarcane. Scale bar, 2.5 cm. (c) Transverse sections from the middle of leaves of WT-XTT22 and three independent *ShTIP1*-OE transgenic sugarcane lines (#1, #2, #3). Scale bar, 100 μm. (d) Phenotypes of 60-day-old WT (WT-ZH11) and *ShTIP1*-OE rice plants. Scale bar, 10 cm. (e) Leaf morphology of WT-ZH11 and *ShTIP1*-OE rice. Scale bar, 5 cm. (f) Transverse sections from the middle of leaves of WT-ZH11 and three independent *ShTIP1*-OE transgenic rice lines (#1, #2, #3). Scale bar, 500 μm. (g-i) Quantification of leaf anatomical traits in WT and *ShTIP1*-OE lines of sugarcane (XTT22) and rice (ZH11), including (g) number of leaf veins, (h) number of epidermal cells between veins, and (i) distance between veins. Data are presented as mean ± SD (n = 10). (j) Relative expression analysis of *ShCBL1*, in both WT and *ShTIP1*-OE lines of sugarcane and rice. In panels (g-j), different letters (a, b, c, d) above the bars indicate statistically significant differences between the lines, as determined by a one-way ANOVA test ($P < 0.05$). (k) Semi-quantitative RT-PCR analysis of *ShTIP1* expression in transgenic rice lines with *OsACTIN* as an internal control (For gel source data, see **Supplementary Figure 31**).




**Extended Data Fig. 9. Phenotypic analysis of *ShSUT2*-overexpressed sugarcane lines.** (a) Cross-sectional views of the 12th internode from 6-month-old CK (ROC22), OE1 (overexpression line 1), and OE2 (overexpression line 2) plants. Scale bar = 500 μm. (b) Statistical comparison of parenchyma cell size between CK and overexpressed lines ($P < 0.0001$, Wilcoxon test). The lowest 10% and largest 10% outliers were removed from the comparison. (c) Relative expression levels (qRT-PCR) of *ShSUT2* and stem sucrose content (HPLC analysis) in CK, OE1, and OE2. (d) Immunoblot verification of *ShSUT2* transgenic case using Basta antibodies.

***Integration of Existing Literature for Previously Validated Genes***. In response to the reviewer's feedback, we conducted a thorough literature search and have now integrated citations for published studies that had already provided functional validation for two of the genes in sugarcane: TB1 (Teosinte Branched 1)[3] and SPS (Sucrose Phosphate Synthase)[4].

2. Assembly Validation: Could the authors provide additional validation beyond sequencing data? At a minimum, validation of several large structural variants would be helpful. Additionally, can any small HE (homoeologous exchange) regions be experimentally validated, for example by FISH?

Response: Recognizing the important suggestion of supporting our findings

with experimental evidence, we conducted validation experiments using haplotype-specific oligo-FISH probes, focusing on both large SVs and smaller HE regions.

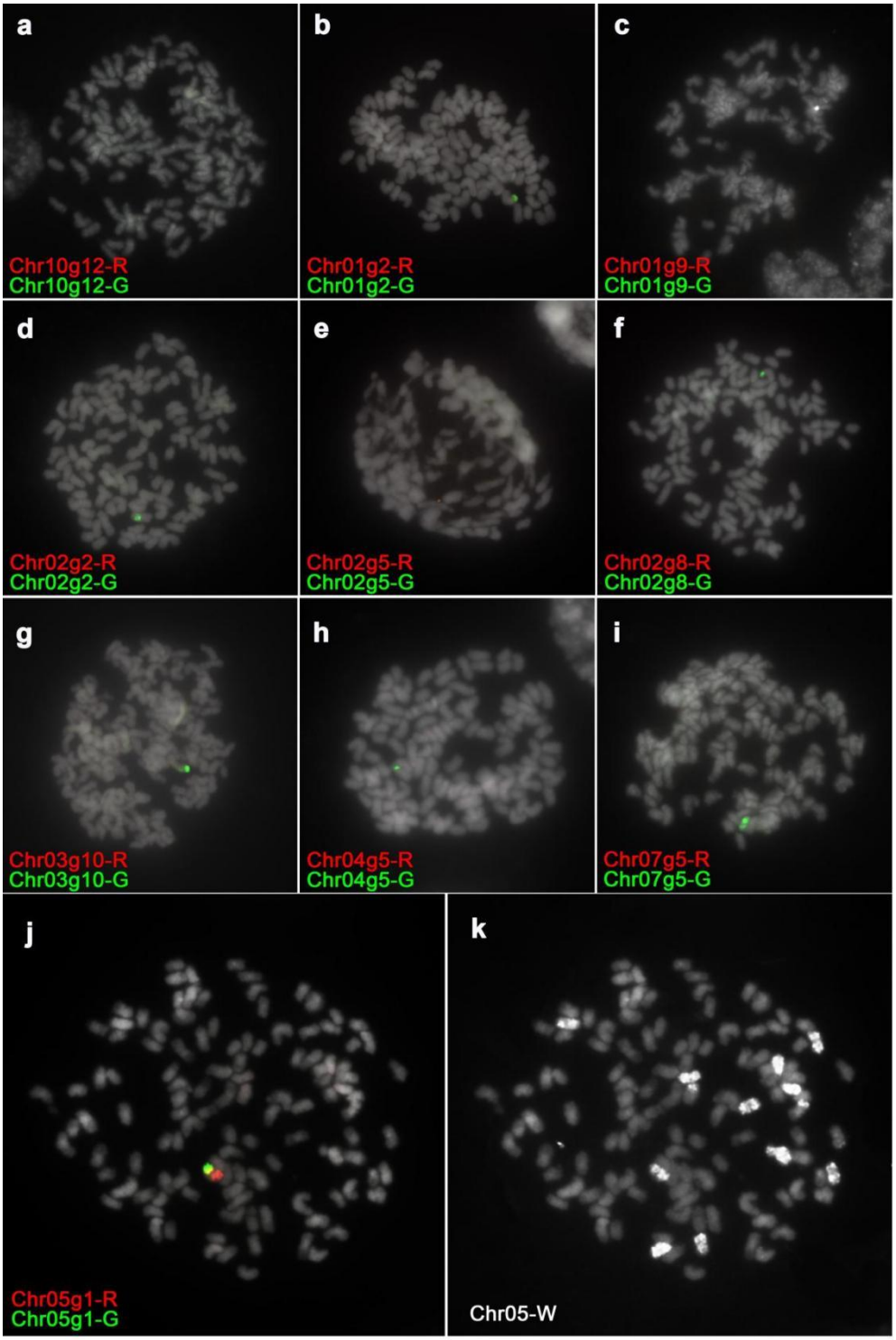
To validate large SVs, we designed and synthesized ten sets of haplotype-specific oligo-FISH probes derived from ten different chromosomes (**Supplementary Table 13**). Each set comprised two pairs of probe clusters: one targeting sequences from the SS subgenome and the other from the SO subgenome. FISH experiments on metaphase chromosomes demonstrated that seven of ten probe sets generated clear, chromosome specific fluorescent signals (**Extended Data Fig. 1**). Among these, Chr05g1 was the only chromosome for which both SS- and SO-derived probe clusters successfully hybridized.

For Chr05g1, the haplotype-specific probes targeting the SS-derived region (Chr05g1-R) spanned from 1.56 Mb to 42 Mb, while probes targeting the SO-derived region (Chr05g1-G) covered 42 Mb to 81.72 Mb. FISH results confirmed that both probe clusters co-localized onto the same chromosome (**Extended Data Fig. 1**). Additionally, chromosome painting probes previously developed for Chr05[5] independently validated the identity of this recombinant chromosome. These results conclusively demonstrated that the first 42 Mb of Chr05g1 originates from the SS subgenome, while the remainder derives from the SO subgenome, providing strong experimental support for the accuracy of our assembly in resolving large SVs.

In response to the reviewer's request to validate smaller HE regions, we attempted to use oligo-FISH probes designed for these regions. However, the sensitivity of oligo-FISH for detecting small regions presented significant technical challenges. Successful detection requires two critical conditions: (1) a sufficiently large number of probes within each cluster (greater than 5000 probes per cluster, based on our experience) and (2) a high probe density (greater than 0.12). When either condition is not met, the resulting fluorescent signal is too weak for reliable microscopic detection.

Despite these challenges, it is important to emphasize that the inability to validate certain smaller or low-complexity regions does not indicate errors in the assembly. In fact, all successfully hybridized haplotype-specific probe clusters consistently produced clear signals localized to single chromosomes, further supporting the accuracy of our assembly. Importantly, no evidence of switched errors was observed.

We have added the relevant Results (**Lines 221-228, Page 9**) and Methods (**Lines 1280-1299, Page 49**) in our revised manuscript.




**Extended Data Fig. 1. Chromosome painting on metaphase chromosomes of POJ2878 using haplotype-specific oligonucleotide probe sets.** (a-i) Nine haplotype-specific oligonucleotide probe sets were used, with the suffixes 'R' and 'G' denoting two distinct target regions on a single chromosome, labeled with red and green fluorescent probes, respectively. (j) Two sets of Chr05g1-specific oligonucleotide probes, Chr05g1-R and Chr05g1-G, were used to paint the SS-derived (red, positions 1.56–42 Mb) and SO-derived (green, positions 42–81.72 Mb) chromosomal segments,

respectively. (k) The Chr05-W probes (white) are previously developed chromosome painting probes designed to label all homologous copies of Chr05. Scale bars: 5 μm.

3. Population Analysis: The manuscript discusses three groups—S. hybrid, S. spontaneum (S.s), and S. officinarum (S.o). Based on the mapping method described (line 1290), all accessions were aligned to the POJ2878 reference genome. Theoretically, there should be no shared SNPs between S.s and S.o. This casts doubt on the interpretation of some of the population structure in Figure 2. A more detailed explanation and methodological clarification are needed.

Response: We acknowledge the concern that shared SNPs between *S. spontaneum* (S.s) and *S. officinarum* (S.o) might seem counter-intuitive when aligned to the POJ2878 reference genome. We offer the following clarification to address this point and demonstrate the robustness of our findings.

First, the use of the POJ2878 reference genome is integral to our analytical framework. As an auto-allo-aneuploid resulting from the hybridization of S.o and S.s, the POJ2878 genome contains not only species-specific regions but also a large number of evolutionary conserved loci that are shared across the *Saccharum* genus. These shared loci are essential for capturing genetic variations at the same sites across the three major species: hybrid sugarcane, S.o, and S.s.

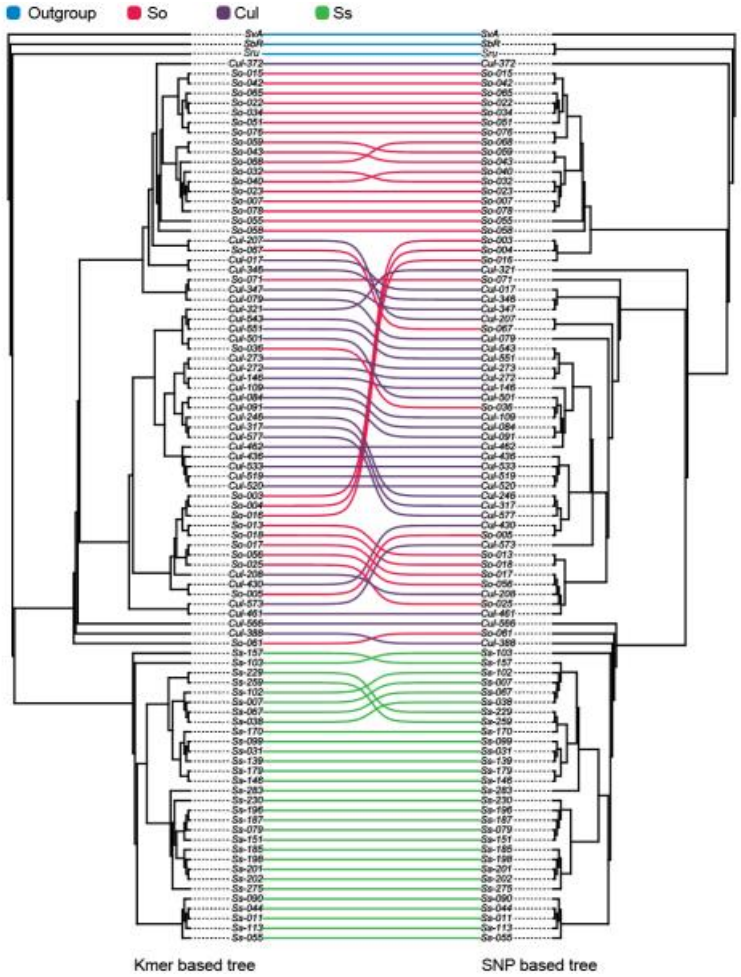
To further validate the population structure shown in Figure 2, we constructed a phylogenetic tree using a k-mer-based approach[6], which is entirely reference-free and therefore unaffected by the quality or characteristics of any single genome assembly. For this analysis, we randomly selected 30 representative samples from each of the three populations and included *Setaria viridis* (SvA), *Sorghum bicolor* (SbR), and *Erianthus rufipilus* (Sru) as outgroups.

As shown in **Response Figure 1**, the two methods yield highly consistent results for the major clades. Both the reference-based and k-mer-based trees clearly separate most *S. officinarum* (So) and *S. spontaneum* (Ss) into distinct, well-supported clades. The primary topological incongruence is observed among the group of cultivated hybrids (Cul), whose admixed and heterogeneous genetic backgrounds are expected to be more sensitive to the choice of reference. This result strengthens our confidence that the main conclusions regarding the population structure of the ancestral species are sound.

To classify this concern, we provided an explanation in the Method section (**Lines 1450-1454, Page 54**), which reads:

“The SNPs shared between *S. spontaneum* and *S. officinarum* in the POJ2878 genome represent a substantial number of evolutionary conserved loci shared across the *Saccharum* genus. These conserved loci are critical for capturing genetic variations at homologous sites across hybrid sugarcane, *S.*

*spontaneum* and *S. officinarum*."




**Response Figure 1. Comparative phylogenetic analysis of randomly selected sugarcane accessions using reference-free and reference-based methods.** The left-hand tree was generated using a reference-free method based on k-mer distances, while the right-hand tree was constructed using a reference-based approach from single nucleotide polymorphisms (SNPs) called against the modern cultivar ('POJ2878') reference genome. Lines connect the identical accessions between the two trees to visualize topological concordance and incongruence. Accessions are color-coded by species or group: *Saccharum officinarum* (So) are shown in red, *Saccharum spontaneum* (Ss) in green, modern cultivated hybrids (Cul) in purple, and outgroup taxa in blue.

4. Domestication Analysis: According to Healey et al. (https://www.nature.com/articles/s41586-024-07231-4#Sec1), domestication should be discussed in the context of S. officinarum and S. robustum, rather than modern cultivars. The breeding process involving current cultivars should be classified as crop improvement, similar to how rye was used to enhance

disease resistance in wheat. Thus, the interpretation and analysis of domestication here appear to be inaccurate.

Response: We agree that the domestication of *S. officinarum* should be contextualized through comparisons with its direct ancestor, *S. robustum*. During our revision process, we incorporated data from a recent genomics study on wild *Saccharum* species[7], which included 43 whole-genome re-sequencing datasets of *S. robustum*. These datasets provided a valuable foundation for investigating the domestication of *S. officinarum*. Through comparative analysis of the population re-sequencing data, we identified 42.50 Mb of selective sweeps under artificial selection in *S. officinarum*, encompassing 1,788 protein-coding genes. This new analysis has been incorporated into the revised manuscript (**Lines 460-483, Page 18**) and is presented in Figure 4b.

Additionally, we fully agree with the reviewer's point that breeding efforts involving modern sugarcane cultivars should be categorized as crop improvement rather than domestication. To address this, we have revised the manuscript (**Lines 484-486, Page 19**) to clearly distinguish between domestication and crop improvement, ensuring there is no ambiguity in our interpretation and analysis.

5. Genome Assembly Focus: The assembly is arguably the paper's highlight. While tools like Allele-Express and KMERIA are valuable for polyploid genome analysis, the manuscript lacks a clear connection between these tools and the assembly itself. There is insufficient explanation of how results from these tools feed back into genome construction or interpretation.

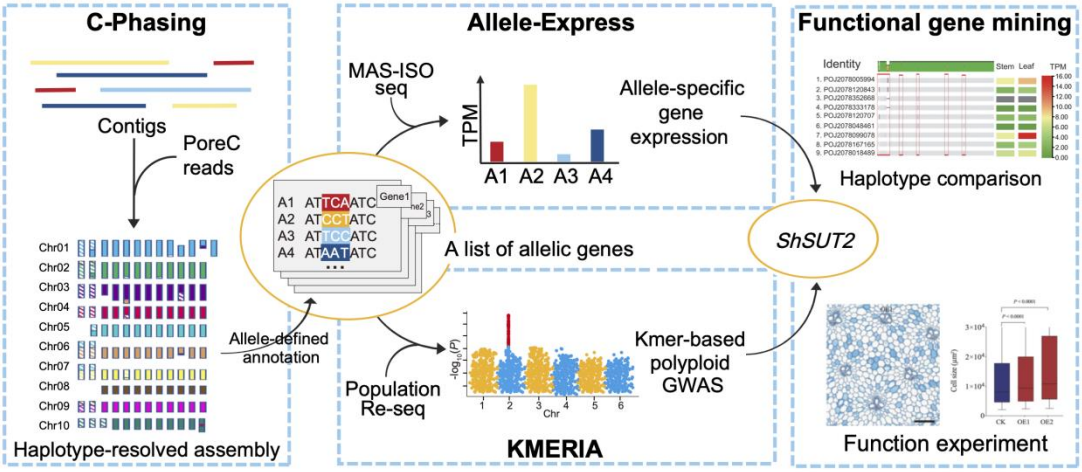
Response: Our study establishes an integrated analytical framework in which the three tools function as interconnected components of a comprehensive polyploid genomics pipeline:

***C-Phasing as the Foundation***: C-Phasing leverages Pore-C sequencing data to achieve haplotype-resolved assembly at chromosomal scale. This tool addresses the fundamental challenge in polyploid genomics—accurately separating homologous and homeologous sequences. The resulting POJ2878 assembly provides 118 individually phased chromosomes across 10 homoeologous groups, serving as the essential reference framework for all downstream analyses.

***KMERIA and Allele-Express as Functional Interpreters***: *KMERIA* performs k-mer-based GWAS, which avoids the errors inherent in complex polyploid genotyping. This approach further leverages a high-quality reference genome to facilitate the biological interpretation of significant k-mer associations. Allele-Express directly depends on the haplotype-resolved assembly and allele-defined annotation for accurate allele-specific expression (ASE) quantification for these genes identified by *KMERIA* or population genetics. Without precisely phased chromosomes, distinguishing transcripts from different alleles would be impossible in such a complex polyploid.

For example, k-mers associated with parenchyma cell traits were mapped to the POJ2878 assembly, leading to the identification of candidate genes such as *ShSUT2*. Using the phased genome, we identified allelic variants across different haplotypes, with one specific allele showing high expression in leaf tissue. This provides a valuable opportunity to target key alleles for further functional validation or for use in breeding programs.

To address the reviewer's concern, we have added **Extended Data Fig. 10**, which illustrates this integrated workflow, showing how: (1) C-Phasing outputs enable precise allele-defined gene annotation, (2) this annotation facilitates accurate ASE quantification by Allele-Express, and (3) KMERIA results are biological interpretation through mapping back to the assembly. Additionally, we have revised the Discussion section (**Lines 587-622, Pages 22-23**) to emphasize how the integration of these tools forms a synergistic analytical framework that exceeds the capabilities of each tool individually.




**Extended Data Fig. 10: Integrated Analytical Framework for Polyploid Genomics.** This figure illustrates the integrated workflow of three computational tools for comprehensive polyploid genome analysis. C-Phasing uses Pore-C sequencing to generate haplotype-resolved assembly with allele-defined gene annotation. Allele-Express leverages this assembly to quantify allele-specific expression from MAS-ISO seq data, identifying functionally important haplotypes. KMERIA performs k-mer-based GWAS on population re-sequencing data, with significant associations mapped back to the reference assembly for biological interpretation. The workflow demonstrates functional gene mining, exemplified by the identification and validation of *ShSUT2* through haplotype comparison and functional experiments. This integrated approach creates a synergistic framework where high-quality assembly enables precise functional analysis and meaningful interpretation of population-scale genetic data.

Minor Comments and Suggestions:

6. Please provide a coverage distribution plot for the whole-genome assembly. Additionally, indicate the proportion of IBD (identity-by-descent) regions across the haploid genome. How many regions are shared among haplotypes? What is the maximum and average number of haplotypes sharing a region?

Response: The coverage distribution plot for the whole-genome assembly is provided in **Supplementary Figure 4**. Our analysis indicates that 98.14% of genomic regions exhibit uniform coverage, with only 0.82% showing potential collapsed regions.

Due to the aneuploid nature of the sugarcane genome, identifying haploid genomes and distinguishing combinations of non-homologous chromosomes remains unfeasible. However, the POJ2878 phased genome exhibits a high proportion of identity-by-descent (IBD) regions (98.15%), suggesting that a similar level of IBD regions would likely be observed in a hypothetical haploid genome.

Given this high proportion (98.15%) of IBD regions in the POJ2878 genome, we anticipate extensive sharing of IBD regions across haplotypes. This suggests that quantifying IBD regions shared among haplotypes in the POJ2878 genome alone may do not yield meaningful insights. To address this, we conducted a focused analysis by randomly selecting five cultivars and examining the IBD regions introgressed from the POJ2878 genome. The results, summarized in **Response Table 1**, reveal variability in the number and size of IBD regions across cultivars. The number of IBD regions shared among haplotypes ranges from 764 to 9,163, with sizes varying between 284.63 Mb and 1,973.98 Mb. The maximum number of haplotypes sharing a single IBD region ranges from 6 to 11 across individuals, while the average ranges from 1.06 to 1.40.

Calculating the IBD statistics for all the 613 cultivars is computationally intensive and not directly relevant to the primary focus of this study. While informative, this analysis does not substantially contribute to the overarching discussion or conclusions of our work and has therefore not been included in the revised manuscript.

Response Table 1. Statistics of IBD regions introgressed from the POJ2878 genome in five randomly selected individuals

| Sample ID | No. of IBD introgressed from POJ2878 | Size of IBD introgressed from POJ2878 (Mb) | No. of shared IBD among haplotypes | Size of shared IBD among haplotypes (Mb) | Max No. of haplotypes sharing an IBD | Average No. of haplotypes sharing an IBD |
|---|---|---|---|---|---|---|
| Cul_024 | 16653 | 3521.86 | 4516 | 744.45 | 9 | 1.40 |
| Cul_267 | 8247 | 1585.96 | 764 | 284.63 | 11 | 1.16 |
| Cul_387 | 18847 | 4670.84 | 946 | 593.96 | 6 | 1.06 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Cul_401 | 23182 | 3322.88 | 3861 | 721.15 | 9 | 1.23 |
| Cul_424 | 36608 | 5052.37 | 9163 | 1973.98 | 10 | 1.37 |
| Cul_476 | 24399 | 4048.97 | 4366 | 1073.32 | 10 | 1.29 |

7. Provide an example demonstrating how methylation data helped resolve complex regions during genome assembly.

Response: The role of methylation-aware haplotype phasing in resolving complex genomic regions is thoroughly discussed in our companion paper on C-Phasing[8], and therefore is not included in this study. Below, we provide a brief summary with examples to address the reviewer's concern.

Polyploid genomes, such as those of sweet potato and sugarcane, are characterized by highly similar or identical-by-descent regions that often cause ambiguous read mappings, resulting in fragmented assemblies and unanchored contigs. To overcome this challenge, the C-Phasing pipeline incorporates the Methalign module, which utilizes native 5-methylcytosine (5mC) methylation signals from Pore-C reads as a "fifth base." This approach refines alignments by penalizing mismatches in methylation patterns and prioritizing mappings that match allele-specific methylation profiles derived from HiFi or ONT reads.

For example, during the assembly of the hexaploid sweet potato genome (2n=6x=90) using ~50× Pore-C data and HiFi contigs, standard alignment methods without methylation data left ~15% of the genome unanchored, mapping only ~2.39 Gb (85.14% of the estimated genome) and leaving gaps in low-contact regions. By integrating methylation data, valid alignments increased by ~9.5%, and contact counts rose from 37.77 million to 46.14 million. This resulted in the anchoring of ~2.86 Gb (102.79% of the estimated genome) and enabled the resolution of haplotypes such as Chr01g5/g6 and Chr10g5/g6. Incorporating methylation data rescued misaligned reads, improved assembly completeness by ~5%, and demonstrated the utility of epigenetic signals in resolving sequence ambiguities without requiring additional data types such as trio or pollen genomes.

Details and relevant figures are available in the C-Phasing paper, which has been posted as a preprint (https://doi.org/10.21203/rs.3.rs-7343323/v1).

8. How are Pore-C connections allocated among shared haplotypes?

Response: Allocating Pore-C connections among shared haplotypes in highly complex polyploid genomes is challenging due to the high similarity of homologous sequences. Our approach, implemented in the C-Phasing pipeline, uses a multi-step algorithmic strategy to address this.

First, the *Methalign* module leverages native 5mC methylation signals as an additional alignment feature, enabling more precise, haplotype-specific mapping of Pore-C reads that would otherwise map ambiguously due to near-identical sequences.

Next, the *Hyperpartition* module utilizes Pore-C's high-order concatemers—representing multi-fragment chromatin interactions—to cluster contigs and their Pore-C connections into distinct shared haplotypes.

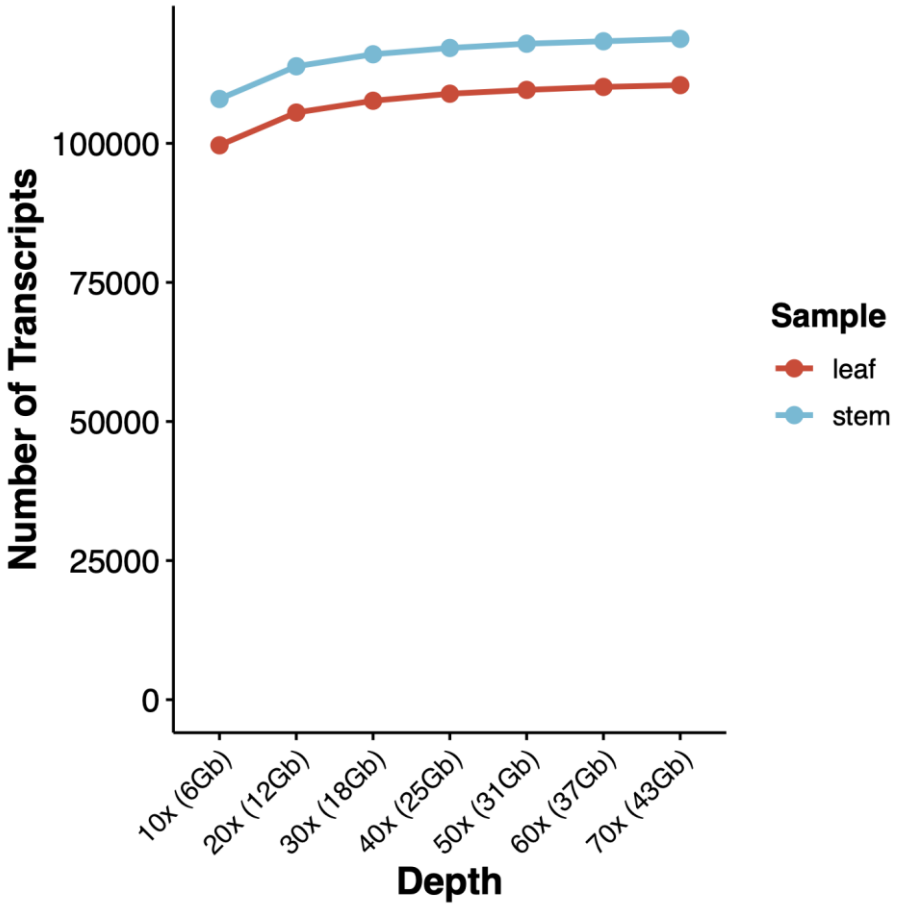
Finally, the *Kprune* module filters out spurious connections caused by residual mapping ambiguity, ensuring that only genuine haplotype-specific Pore-C connections remain.

In summary, the allocation follows a clear "map → cluster → filter → complete" workflow. This method has been validated in ultra-complex genomes such as hexaploid sweet potato and aneuploid sugarcane, improving haplotype-resolved assembly completeness by up to 23% and reducing misallocated connections to below 0.2%. Full methodological details are provided in our C-Phasing companion paper.

9. For Iso-Seq data, include a saturation curve showing the relationship between sequencing depth and transcript discovery. This is necessary to confirm whether the data are sufficient for reliable quantification.

Response: A saturation curve illustrating the relationship between sequencing depth and transcript discovery has been included in **Supplementary Figure 11**. The curve demonstrates that the number of identified transcripts increases with sequencing depth but gradually reaches a plateau, indicating that sequencing beyond 30x depth provides minimal additional benefit in identifying novel transcripts.

In this study, we generated MAS-ISO-seq reads for both leaves and stems, with three biological replicates for each tissue. The data size for each replicate exceeds 19 Gb, corresponding to approximately 32x sequencing depth (estimated transcriptome size: 616.78 Mb). This level of sequencing depth ensures that the Iso-Seq data produced in our study is sufficient for reliable transcript quantification. The analysis has also been described in the main text (**Lines 293-296, Page 12**).




**Supplementary Figure 11. Saturation analysis of MAS-ISO-seq transcript discovery.** The saturation curves illustrate the number of unique transcripts identified (y-axis) as a function of sequencing depth (x-axis) in leaf (red) and stem (blue) samples. The plateauing of both curves indicates that the sequencing depth was adequate for comprehensive transcriptome coverage, capturing the majority of expressed transcripts in both tissues.

10. Please describe in detail the method used for identifying favorable haplotypes.

Response: We have provided the details for identifying favorable haplotypes in the METHOD section (**Lines 1463-1475, Page 55**), which is also attached below:

***Identification of Shared IBD Regions and Favorable Haplotypes.*** To investigate shared identical-by-descent (IBD) regions across sugarcane accessions, we utilized IBDseq[9] with its default settings. We quantified the contribution of each IBD region by introducing a metric, termed IBD density, which measures the proportion of varieties sharing IBD segments with POJ2878. The IBD density for a given genomic window is calculated as:

$$\text{IBD density} = \frac{\text{Number of varieties sharing IBD segments with POJ2878}}{\text{Total number of varieties examined in the population}}$$

The genome was analyzed in 50-kb non-overlapping windows, and regions within the top 5% of IBD density were designated as hotspots of

POJ2878 contributions, representing favorable haplotypes. These hotspots mark genomic regions with substantial genetic contributions from POJ2878, and likely harbor key alleles associated with important agronomic traits in sugarcane.

11. Figure 3d contains confusing labels. The relationship between genome haplotypes and labels such as so_hap and ss_hap needs to be clarified. Additionally, clarify what determines haplotype count—are they based on genome haplotypes, isoforms, or copy number variation?
Response: We apologize for any confusion caused by the original labels. The figure legend of **Figure 3d** has been revised as follows for clarity:

"The x-axis represents the genes analyzed, with genome haplotypes (e.g., _10g5) denoting the chromosome ID where favorable alleles were identified. The y-axis labels indicate the progenitor origin of these haplotypes, along with their randomly assigned numbers. Haplotype counts were determined using allele-aware annotations derived from the haplotype-resolved genome assembly. Although multiple copies of a gene may be annotated, only haplotypes containing at least one mutation, compared to alternative forms, were included in the analysis. For instance, while a total of 12 copies of the *MNS2* gene were annotated, only eight haplotypes were retained for further analysis: three allelic genes from SS and five from SO."

12. Of the three companion papers mentioned for the major methods, only the paper on C-Phasing is cited. The other two are not provided and should be included.
Response: Thank you for bringing this to our attention. In the revised manuscript, we have included the details of the Allele-Express method as **Supplementary Note 2**. Additionally, the methodologies for C-Phasing and KMERIA are fully described in two companion papers, which have been released as preprints[8,10]. We have now included citations for both the C-Phasing and KMERIA papers in the revised manuscript.

13. The Allele-Express GitHub repository lacks English documentation. Please add an English-language explanation of the tool.
Response: The English documentation for this tool has been added in the Github repository.

14. A workflow diagram for the Allele-Express method should be provided. Also, explain how homologous gene groups are defined and what filtering parameters are used (e.g., mapping quality, edit distance). How does Allele-Express associate transcripts with complex genomic regions? How are alleles assigned to homologous gene groups and genomic loci?

Response: We have provided a workflow diagram (**Extended Data Fig. 5**) to visually outline the step-by-step process of the Allele-Express method. A detailed description of the algorithm, along with its validation, is included in **Supplementary Note 2**.

Homologous gene groups are defined as allelic genes located on homologous chromosomes that originate from the same ancestral species. Homoeologous gene groups, on the other hand, are orthologous genes found between subgenomes derived from different progenitor species.

To ensure high-quality transcription quantification, Allele-Express employs a hierarchical ranking of alignment results based on several metrics. Alignments are first prioritized by MAPQ (Mapping Quality Score) in descending order, as higher scores indicate unique, reliable alignments with minimal ambiguity. Next, alignments are sorted by edit distance in ascending order, where lower values indicate greater sequence similarity between the read and reference transcript. This is followed by the AS tag (Alignment Score) in descending order, which rewards matches while penalizing mismatches. Finally, alignments are ranked by the DE tag (Deletion/Insertion Penalty Score) in ascending order, minimizing fragmented or error-prone alignments caused by insertions or deletions. The alignment that ranks highest across these metrics is selected for downstream analyses, including allele-specific expression quantification. This multi-tiered approach ensures that only the most reliable alignments are used, characterized by minimal edit distances and optimal alignment scores.

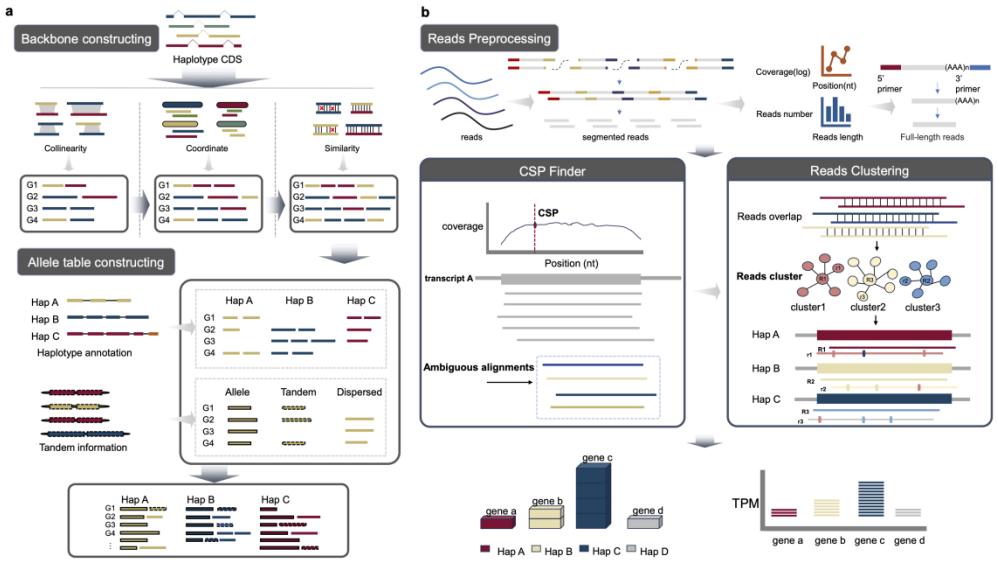
To address the challenge of assigning reads that align to multiple genomic regions, we developed two complementary strategies: the read-depth stable point (CSP)-based approach and a read clustering-based approach. The CSP method identifies a stabilization point within each transcript, beyond which reads are classified as ambiguous. Meanwhile, the clustering approach utilizes many-against-many sequence comparisons of full-length MAS-ISO-seq reads, followed by a greedy incremental clustering algorithm based on overlapping read similarity. Ambiguous reads are confidently assigned to specific transcripts if other reads within the same cluster align uniquely to those transcripts. Together, these methods enable accurate assignment of MAS-ISO-seq reads to their respective transcripts, with only a small fraction of reads remaining unassigned.

Alleles are assigned to homologous gene groups and genomic loci using a multi-step process that integrates synteny analysis, coordinate-based mapping, and sequence homology searches:

- Synteny Analysis: Using MCScanX, conserved genomic blocks are identified based on collinearity among homologous chromosomes, requiring a minimum of five gene pairs to define synteny blocks.
- Coordinate-Based Mapping: CDS sequences are aligned to a monoploid reference genome using GMAP, grouping genes with ≥80% overlap in genomic coordinates.

- BLAST Homology Search: Additional homologous genes are identified through BLASTN, requiring ≥80% sequence identity.

This comprehensive approach ensures robust identification and assignment of alleles to homologous gene groups and genomic loci. Details regarding the algorithm and validation of Allele-Express are provided in **Supplementary Note 2**.

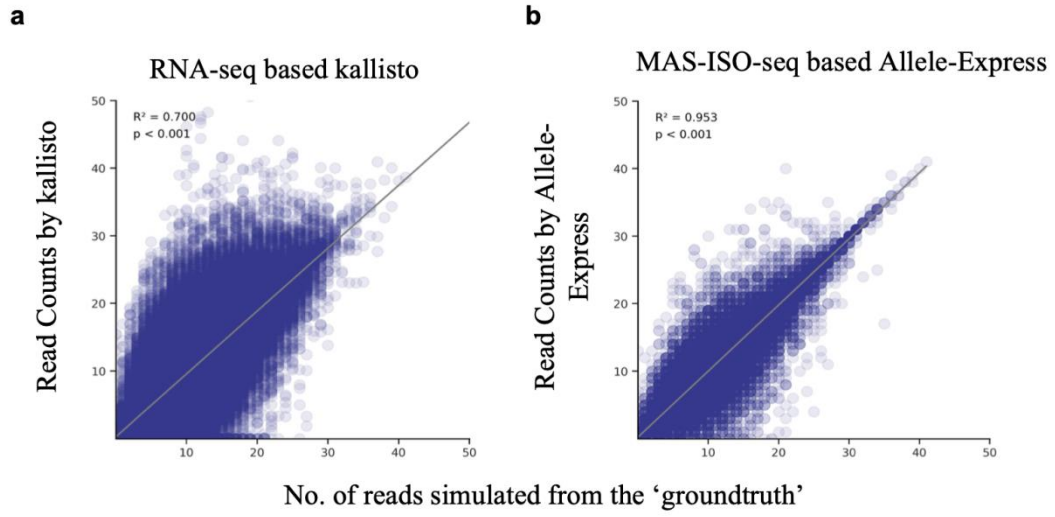



**Extended Data Fig. 5. Overview of the Allele-Express framework.** (a) The *Allele* module for identifying allelic genes in haplotype-resolved genomes. The process begins with the construction of a haplotype backbone from Coding Sequences (CDS). An allele table is subsequently constructed by integrating three criteria: collinearity, genomic coordinates, and sequence similarity. This multi-faceted approach classifies genes into allelic, tandem, or dispersed groups, leveraging tandem gene annotations and haplotype information. (b) The Expression module for measuring allele-specific gene expression. This module first preprocesses raw MAS-ISO-seq reads. To resolve ambiguous alignments, two complementary methods are employed: the CSP (coverage stable point) finder, which identifies a stable point in read coverage to trim ambiguous read ends, and a read clustering approach, which groups reads based on sequence similarity to confidently assign ambiguous reads. Finally, the expression of each allele is quantified and normalized to Transcripts Per Kilobase Million (TPM).

15. Can Allele-Express and Iso-Seq quantification results be validated—either experimentally or using RNA-seq data? If so, please provide supporting evidence. Is this method reliable enough to replace traditional RNA-seq for stable quantification?

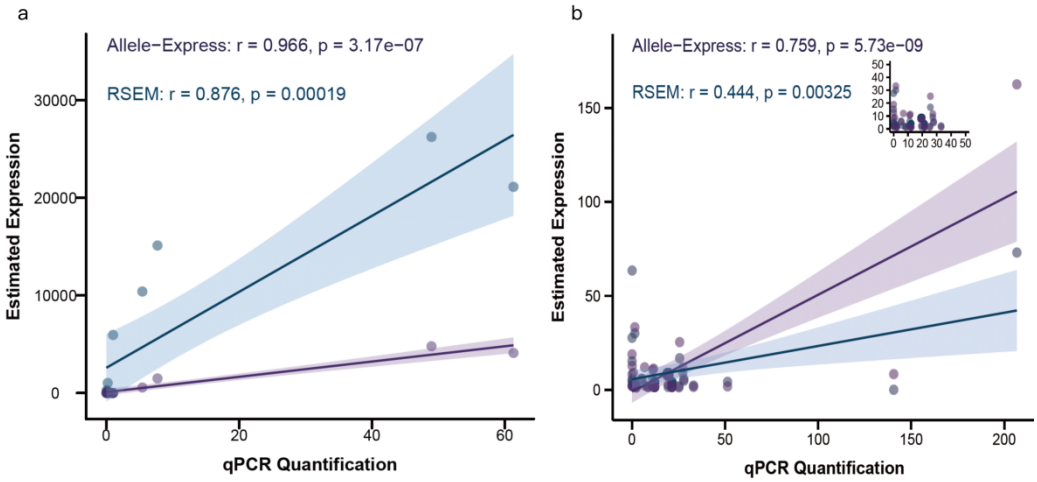
Response: Yes, the quantification results from Allele-Express and Iso-Seq have been validated experimentally and through comparative analysis with RNA-seq data. Supporting evidence includes:

***RNA-seq Comparative Analysis*** (**Supplementary Figure 28**). Benchmarking against RNA-seq quantification tools (e.g., Kallisto) revealed substantial improvements in accuracy for Allele-Express. In polyploid sugarcane, Allele-Express achieved an $R^2 = 0.95$, compared to RNA-seq algorithms like Kallisto ($R^2 = 0.70$). Scatter plots for MAS-ISO-seq based Allele-Express results showed tighter clustering of Allele-Express predictions along the regression line, indicating consistent accuracy across all expression levels.




**Supplementary Figure 28. Benchmarking of different gene quantification approaches in the polyploid sugarcane genome.** The x-axis represents the number of reads simulated from the 'ground truth' based on the POJ2878 genome. The y-axis shows the calculated read counts from RNA-seq using the Kallisto algorithm (a) and MAS-ISO-Seq using the Allele-Express algorithm (b).

***Quantitative RT-PCR Validation*** (**Supplementary Figure 29**): Validation experiments were conducted using qRT-PCR for both diploid *Arabidopsis thaliana* and polyploid hybrid sugarcane (POJ2878). In *A. thaliana*, Allele-Express achieved a Pearson correlation coefficient of $r = 0.966$ with qRT-PCR measurements, explaining approximately 93% of variance ($r^2 \approx 0.93$). In polyploid sugarcane, Allele-Express demonstrated robust performance with $r = 0.759$, significantly outperforming traditional RNA-seq methods (e.g., RSEM, $r = 0.444$).




**Supplementary Figure 29. Experimental validation of Allele-Express performance using quantitative RT-PCR.** Correlation analysis between computational gene expression estimates and experimental qRT-PCR measurements across different genomic complexities. (a) Validation in diploid *Arabidopsis thaliana* showing correlation between qRT-PCR quantification (x-axis) and estimated expression levels (y-axis) for multiple selected genes. (b) Validation in polyploid sugarcane genome showing the same comparison. Shaded areas represent 95% confidence intervals for the regression lines. The inset in panel (b) shows a zoomed view of the low-expression region. Statistical significance was determined using Pearson correlation analysis.

The above results show that the MAS-ISO-seq based Allele-Express represents a significant advancement in gene expression quantification, particularly for polyploid genomes, and offers a reliable alternative to traditional RNA-seq methods. This information has been included in **Supplementary Note 2**.

16. In Figure 2 (panels a–d), the same colors are used for the three species, but the color assignments differ in panels e–f. Please use consistent color schemes across all subfigures.

Response: Thank you for pointing out the inconsistency in the color scheme of Figure 2.In the revised figure, we have implemented a unified color scheme across all panels (a–f), ensuring that each species is represented by the same color throughout. Additionally, the So/Cul labels in the phylogenetic tree have been updated to match this color scheme, eliminating any potential confusion.

17. What is the reference genome used in Figure 4a?

Response: The T2T monoploid genome of Ss82-114 was employed as the reference genome. The detailed information of Ss82-114 genome assembly

can be found in Supplementary Note 1. We have revised the legend of Figure 4a for clarity.

18. In line 667, label (e) appears to be incorrect.
Response: The error has been corrected.

19. Lines 64-6: To provide a balanced discussion, it should be acknowledged that the easy availability of sugar has contributed to a global health crisis. The phrase “reshaped human migration patterns” is also too euphemistic a reference to the transatlantic slave trade.
Response: We have revised the introduction accordingly (**Lines 67-75, Page 4**).

20. The terms “noble” and “nobilization” should be explicitly defined upon first use. These are technical terms specific to sugarcane breeding and may not be familiar even to geneticists working with other plant species.
Response: In the revised manuscript, we have provided clear definitions of the terms "noble breeding" and "nobilization" upon their first mention (**Lines 78-81, Page 4**) to ensure clarity for readers, who may not be familiar with these technical terms specific to sugarcane breeding.

Referee #1 (Remarks on code availability):

The code is accessible and properly documented.

Referee #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.
Response: Thanks for your valuable comments.

Referee #3 (Remarks to the Author):

This manuscript describes an impressive advancement in sugar cane genomics, really the last, biggest, and hardest agriculturally significant plant genome out there. Overall I was impressed with both how well the new genome assembly worked and how clearly the paper was written. I think this study would be of wide interest and the genome assembly is highly impactful with a lot of potential for reuse. I have a number of minor concerns and suggestions, primarily focused on the downstream analyses presented using this genome assembly and one modern concern regarding the novelty of the k-mer based association method.

Response: Thank you for your recognition of our work and for your valuable suggestions. Please see our point-by-point responses below.

Moderate:

1. Lines 437-446 (and 497): Kmer based GWAS is a technique that has been used in a number of previous contexts. Is KMERIA significantly different from the method of He et al 2024 which was also using kmers to link traits to copy number variation via GWAS? (He, C., Washburn, J. D., Schleif, N., Hao, Y., Kaeppler, H., Kaeppler, S. M., ... & Liu, S. (2024). Trait association and prediction through integrative k- mer analysis. The Plant Journal, 120(2), 833-850

Note: I don't question the biological results obtained via the author's method, I just don't believe this can correctly be framed as a novel solution, let alone a paradigm change

Response: We appreciate your insights and agree that the use of the term "novel" or "new paradigm" in our manuscript may not have been entirely accurate. We have revised this terminology in the updated manuscript to better reflect the contributions of our study.

Regarding the k-mer GWAS method described by He et al. (2024)[11], we are grateful for bringing this important work to our attention. Their study represents a significant contribution to the field, leveraging k-mer copy numbers for association analyses in diploid maize. We also recognize the value of their approach in integrating k-mers into genomic selection, which offers valuable insights and strategies worth learning from.

That said, we would like to emphasize that KMERIA differs from the method of He et al. (2024) in several critical aspects, particularly regarding its design and application to polyploid genomes:

**(1) Focus on Polyploid Genomics:**

While He et al.'s method is tailored to diploid maize, KMERIA is specifically designed for polyploid species, where traditional SNP-based genotyping often encounters great challenges. KMERIA uses k-mer copy numbers as proxies for genotypes, specifically modeling allelic dosage in polyploids. In contrast, He et al.'s framework is not directly applicable to polyploid systems due to the additional complexities of ploidy levels and allele dosage.

**(2) Statistical Framework:**

He et al.'s approach relies on simple regression for identifying trait-associated k-mers. While effective in their context, this method does not incorporate a linear mixed model (LMM) to account for kinship, which could lead to inflated false-positive rates. KMERIA, on the other hand, integrates an LMM to ensure statistical robustness in polyploid GWAS.

**(3) Computational Efficiency:**

We attempted to evaluate He et al.'s framework in a polyploid context but encountered great computational challenges. The k-mer matrix construction step in their method lacks optimization, resulting in excessive CPU and

memory usage. This restricts its applicability to species with small genomes and limited sample sizes, making it impractical for complex polyploid genomes. Furthermore, their method requires auxiliary SNP genotyping information to compute covariates such as population structure. In polyploids, this would necessitate additional adjustments for ploidy levels, further escalating computational demands and limiting feasibility for most polyploid populations.

In contrast, KMERIA is implemented in C/C++ with multiple algorithmic optimizations to enhance computational efficiency. These improvements allow for scalable and robust analyses of complex polyploid genomes, addressing the limitations of existing k-mer GWAS methods.

To address the reviewer's concerns and further improve the manuscript, we have made the following revisions:

**(1) Title Revision:**

We have replaced the phrase "a new paradigm" with more measured language to better reflect the contribution of our study. The revised title now reads:

"An integrated genomic framework for polyploid crops reveals the genome organization of sugarcane as a leading sugar crop."

**(2) Acknowledgment of Prior Work:**

We have explicitly acknowledged the contributions of He et al. (2024) and other k-mer GWAS studies in the Discussion (**Line 603, Page 23**), ensuring proper recognition of prior advancements in the field.

**(3) Clarification of KMERIA's Contribution:**

We have clearly articulated that the innovation of KMERIA lies in extending k-mer GWAS to polyploid allele dosage modeling, rather than in the invention of k-mer GWAS itself by stating "KMERIA extends these approaches through allele-dosage modeling and integrates a linear mixed model (LMM) to correct for population structure and kinship, improving both robustness and precision".

Minor:

2. Line 212: "exceeding" rather than "exeeds"? Something is not quite right in this sentence.

Response: Thank you for highlighting this issue. We have corrected the sentence for clarity and accuracy (**Lines 243-244, Page 10**). The revised sentence now reads:

"In the So-subgenome, LTRs account for 61.37%, while in the Ss-subgenome, they constitute 52.17% (**Supplementary Table 7** and **Supplementary Figure 7a**), exceeding the proportions observed in the progenitor lineage genomes, *S. spontaneum* 82-114 and *S. officinarum* XZ. "

3. 215-218: these dates would place the transposon bursts prior the hybridization events that produced POJ2878 which occurred less than 200

years ago. Does that suggest that the true progenitor lineages of used in producing POJ2878 were significantly different from those generated in this study?

Response: We agree with your observation. The dating results suggest that the true progenitor lineages used in producing POJ2878 are likely different from the two genomes of *S. spontaneum* 82-114 and *S. officinarum* XZ, which were collected in China. To address this, we have added a sentence to clarify why the LTR bursts in the POJ2878 subgenome predate the hybridization events that occurred less than 120 years ago. The revised sentence now reads:

"This increase in LTR content is attributed to distinct LTR bursts, which are estimated to have occurred approximately 0.16 million years ago in the So subgenome and 0.2 million years ago in the Ss subgenome (**Supplementary Figure 7b**), predating the hybridization events that produced POJ2878 less than 120 years ago. This dating analysis suggests that the true progenitor lineages used in producing POJ2878 may differ from those analyzed in this study."

4. 281-283: If it is necessary to use long read iso-seq technologies to measure gene expression, can short reads be confidently used for SNP/indel discovery?

Response: Genes are typically located in evolutionarily conserved regions with high sequence similarity, which makes it difficult for short-read RNA-seq to distinguish between alleles accurately. Long-read iso-seq addresses this limitation by capturing full-length transcripts, enabling precise allele-specific expression analysis.

In contrast, SNP and indel discovery relies on mapping short reads to the entire genome, including genic and intergenic regions. Although genic regions are conserved, they represent only a small fraction of the genome, accounting for 15.90% of the POJ2878 genome. The majority of the genome, particularly intergenic regions, displays much higher levels of heterozygosity than that in genic regions. Our analysis indicates that 89.32% of the identified SNPs originate from these intergenic regions, while only 10.68% from genic regions. This suggests that while long-read iso-seq is indispensable for resolving the complexities of gene expression, short-read sequencing remains a reliable and cost-effective approach for genome-wide SNP and indel discovery currently.

5. 304-306: 30 Mb out of a 10.37 Gb genome is actually not very much. I don't think this really does underscore the widespread use of POJ2878 in breeding.

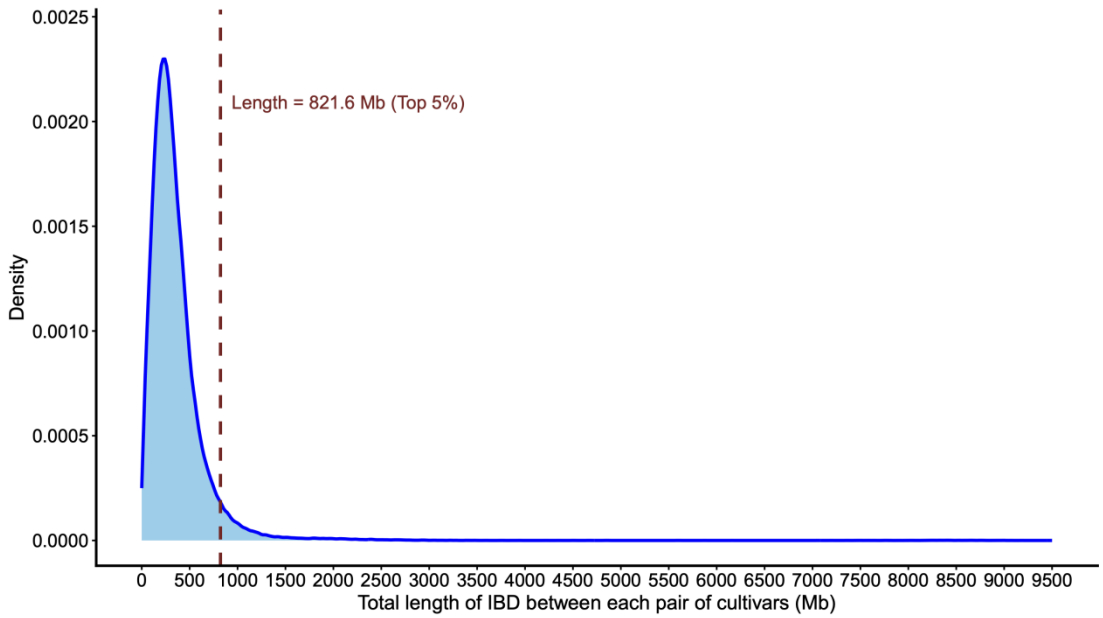
Response: We apologize for the confusion caused by an earlier misstatement in the manuscript and appreciate the reviewer's attention to this detail. The reference to "30 Mb of IBD sequences" was incorrect: this value does not represent a genome-wide threshold for POJ2878-derived introgression, but rather the minimum IBD segment size identified on a single chromosome

(Chr01g1) in initial analyses. This error arose from a misinterpretation of the IBDseq output, which we have since corrected.

To address this issue and establish a robust genome-wide threshold for POJ2878 introgression, we refined our approach by analyzing the distribution of shared IBD sizes across various cultivars. Through this analysis, we determined a minimum IBD size of 821.6 Mb for POJ2878-derived introgression across the entire genome, representing the top 5% of the longest IBD regions shared among cultivars (**Response Figure 2**).

Using the revised 821.6 Mb threshold, we reanalyzed a dataset of 573 hybrid sugarcane cultivars with confirmed geographical distribution. The updated results revealed that 546 out of 573 cultivars (95.3%) share over 821.6 Mb of IBD sequences with POJ2878—a threshold corresponding to the top 5% of the longest IBD regions shared among cultivars. This finding strongly highlights POJ2878's extensive use in global breeding programs.

All relevant content in the revised manuscript has been updated to reflect these corrections, including corresponding text (**Lines 368-373, Pages 14-15**) and updates to **Figure 3**.




**Response Figure 2. Density Distribution of Shared Identity-by-Descent (IBD) Total Length Between Cultivar Pairs.** The x-axis represents the total length of IBD shared between each pair of cultivars, while the y-axis indicates the probability density. A key feature highlighted in the plot is the vertical dashed line, which marks the threshold for the top 5% of the distribution. This line is positioned at a total IBD length of 821.6 Mb, signifying that 5% of all cultivar pairs in this study share a total IBD length of 821.6 Mb or greater.

6. 323-326: Typically optimal alleles for fitness/yield will not be those with the most extreme expression levels. See Kremling et al 2018. (Kremling, K. A., Chen, S. Y., Su, M. H., Lepak, N. K., Romay, M. C., Swarts, K. L., ... & Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature, 555(7697), 520-523.)

Response: We agree with your observation and have cited the referenced publication (Kremling et al., 2018) to support our conclusion in the revised manuscript.

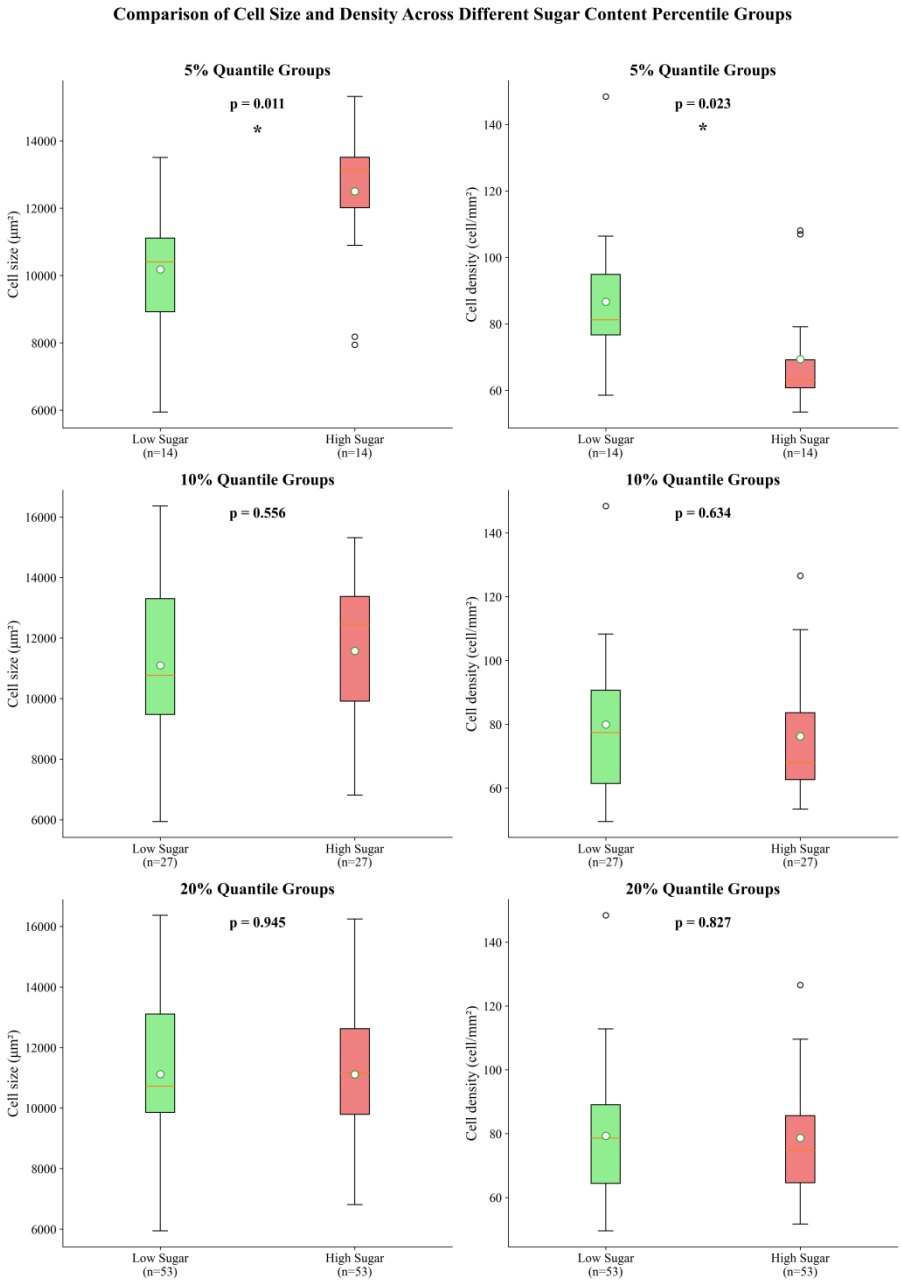
7. 425-436: If there isn't a significant correlation across the entire dataset it is fair to try a few other ways of slicing and dicing the data, but it does introduce a multiple testing problem. P values of only 0.04 and 0.01 aren't high enough to be convincing in the context of likely running a number of different tests with different cutoffs and approaches.

Response: We sincerely appreciate the reviewer's observation regarding the multiple testing problem and the need for more stringent statistical thresholds when conducting exploratory analyses with different data subsets.

To directly address this concern, we conducted a systematic analysis using three predefined grouping strategies to minimize multiple testing bias. (1) the extreme 5% of samples with highest versus lowest sugar content, (2) the top and bottom 10%, and (3) the top and bottom 20%. For each comparison, we evaluated both parenchyma cell size and density using the Mann-Whitney U test, as Shapiro-Wilk testing revealed non-normal distributions in our extreme value groups.

Our systematic approach yielded clear results (**Supplementary Figure 24**): significant differences emerged only when comparing the most extreme 5% groups for both cell size (12,483 μm² vs. 10,157 μm²; $P$ = 0.011) and cell density (69.86 vs. 87.13 cells/mm²; $P$ = 0.023). Importantly, broader thresholds (10% and 20% comparisons) showed no statistical significance.

These findings highlight that sugar accumulation is a complex, multifactorial process. While parenchyma cell characteristics contribute to sugar storage capacity, they represent only one component of a broader regulatory network. We have updated the Results section (**Lines 545-554, Page 21**) to reflect this nuanced interpretation.




**Supplementary Figure 24. Parenchyma Cell Characteristics in Samples with Extreme Sugar Content.** Boxplots comparing parenchyma cell size (left panels, in μm²) and cell density (right panels, in cells/mm²) between samples with the highest (High Sugar) and lowest (Low Sugar) sugar content. The comparisons are shown for the extreme 5% (top, n=14 per group), 10% (middle, n=27 per group), and 20% (bottom, n=53 per group) of samples based on sugar content. Statistical significance was determined using the Mann-Whitney U test, with p-values displayed above each plot. Asterisks (*) denote a statistically significant difference ($P < 0.05$). Within each boxplot, the white circle indicates the mean, the center line represents the median, the box boundaries mark the 25th and 75th percentiles, and the whiskers extend to the data range, with outliers shown as individual points.

8. For an international journal, describing populations of sugar cane as Chinese/non-Chinese will aid intelligibility rather than national/international which are terms whose meaning will vary depending on the country of the reader.
Response: Agree. We have revised the terminology in the manuscript to describe the populations as "Chinese" and "non-Chinese" instead of "national" and "international."

9. Typo in selective on line 388.
Response: The typo has been corrected. Thanks!

Referee #4 (Remarks to the Author):

Wang et al. present an impressive study of sugarcane genomics, offering valuable insights into its complex polyploid genome through haplotype-resolved chromosomal-level assemblies. The approaches used are novel and powerful. The results shown are solid and engaging. This work represents a state-of-the-art in resolving highly complex polyploid genomic studies using a combination of long-read and short-read methods. To further push the boundaries, this study would benefit from stronger integration across novel methods, particularly in leveraging the phased genome to investigate subgenome and homoeolog-specific dynamics, which represents one of its most novel contributions. Additionally, greater emphasis on biological interpretation and synthesis, especially in the context of evolutionary and functional genomics, would enhance its overall impact. Nonetheless, this is a substantial and exciting contribution to the field that, with some refinement, could serve as a foundational resource for sugarcane biology and comparative genomics. Detailed comments are shown below.
Response: We sincerely thank the reviewer for this constructive feedback and the kind words about our work. Please see our point-by-point responses below.

The first paragraph of the introduction is beautifully written. The second paragraph presents some technical challenges, but mostly focuses on sequencing technologies. They should have mentioned the progress of using long reads in sugarcane genome assemblies, such as Healay et al. 2024, Nature. The third paragraph is largely a duplication of the abstract. Overall, the introduction brought about the importance of sugarcane research and describes challenges in sugarcane genomics, but missed the opportunities to introduce the current progress and challenges in sugarcane biology and evolutionary studies, which are the major selling points of this study.
Response: We fully agree that the Introduction requires substantial revision to

better contextualize our work within the broader landscape of sugarcane biology and evolutionary genomics, rather than focusing primarily on technical aspects. In response, we have revised the Introduction to better situate our study within the broader context of recent advancements and challenges in sugarcane research. Specifically, we have incorporated key findings from recent landmark studies, including Healey et al. (2024, Nature)[12], Zhang et al. (2024, Nature Genetics)[13], and Olivier et al. (2025, Cell)[7]. These studies highlight important progresses in sugarcane genomics, particularly through the use of long-read sequencing technologies, as well as insights into the evolutionary processes shaping sugarcane's complex genome. Details of these updates can be found in the revised Introduction (**Lines 88-95, Pages 4-5**).

This study proposed an approach to identify the subgenome origin of hybrid sugarcane using T2T monoploid genome assemblies and resequenced hybrid sugarcane genotypes. How does the current method and finding represent an improvement over an existing study from Zhang et al. 2018, Nature Genetics?
Response: While the study from Zhang et. 2018[14] was a foundational step in sugarcane genomics, our work addresses a more complex system and provides significant improvements in scope, methodology, and biological insights. Below, we summarize the key differences:

***Research Focus:*** Zhang et al. (2018) assembled the haploid *S. spontaneum* genome (AP85-441, 1n=4x=32) and analyzed 64 *S. spontaneum* accessions through resequencing, with minimal focus on hybrid sugarcane. In contrast, our study not only assembled the highly complex hybrid sugarcane genome but also generated telomere-to-telomere (T2T) monoploid genomes for both progenitor species (*S. spontaneum* and *S. officinarum*). Additionally, we resequenced 78 *S. officinarum*, 290 *S. spontaneum*, and 613 hybrid sugarcane accessions, providing a comprehensive dataset to explore subgenome origins and evolutionary dynamics.

***Genome Assembly Quality:*** Zhang et al. (2018) used PacBio RSII sequencing, achieving a contig N50 of only 45 kb, with chromosomal assignment accuracy of 89%. In contrast, we utilized PacBio Revio HiFi and ONT Ultra-long sequencing, achieving a contig N50 of 15.60 Mb for the hybrid sugarcane genome—a 350-fold improvement. The monoploid genomes of *S. spontaneum* and *S. officinarum* were assembled to a telomere-to-telomere (T2T) level, providing unmatched resolution and accuracy.

***Methodological Advancements:*** Zhang et al. (2018) applied the ALLHiC algorithm for scaffolding autopolyploid genomes using Hi-C data. However, ALLHiC is limited by short-read sequencing, leading to chimeric errors in more complex polyploids[15]. Our study leveraged the long-read advantages of Pore-C technology and introduced the novel C-Phasing algorithm developed in our companion study[8], specifically designed to resolve the intricate haplotype structure of hybrid sugarcane. Furthermore, we developed an integrated

framework for polyploid genomics, combining C-Phasing, KMERIA (a k-mer-based GWAS tool), and Allele-Express (for allele-specific expression analysis).

***Biological Insights:*** Zhang et al. (2018) focused on the structural evolution of the *S. spontaneum* genome. In contrast, our study explores the complex genome structure and dynamics of hybrid sugarcane, population evolution, and the genetic basis of domestication and breeding as a primary sugar crop. For example, we identified extensive subgenome recombination, non-homologous rearrangements, and key genes under artificial selection.

In conclusion, while Zhang et al. (2018) laid the groundwork for sugarcane genomics, our study provides a significant leap forward by addressing the subgenome origin, domestication, and breeding of hybrid sugarcane with cutting-edge technologies and novel methodologies. This work offers new perspectives on the genetic architecture and evolutionary dynamics of sugarcane, advancing both fundamental science and breeding applications.

ePore-C seems an improved version as described by the companion paper, but only Pore-C was mentioned in the manuscript. It is not very clear which one was used.

Response: We confirm that this study utilized ePore-C, an enhanced version of the Pore-C technology that offers improved data yield. We have revised the method section to explicitly state this and have included a citation to the companion paper[8] where ePore-C is described in detail.

C-Phasing was developed and introduced as a new method in this manuscript, yet the authors also submit a companion manuscript describing the C-Phasing algorithm and benchmarks. This is double-dipping. They should either report the C-Phasing method as a stand-alone work and USE/cite it in this study, or describe the development of C-Phasing in this study without a companion paper.

Response: We sincerely thank the reviewer for raising this important point regarding the presentation of our methodologies in this study. To address the concerns, we have carefully revised our manuscript to ensure that the scope and novelty of this work are appropriately framed, avoiding any overlap with the companion manuscripts describing the development and benchmarking of C-Phasing and KMERIA. Below are the specific changes we have made:

**(1) C-Phasing[8] and KMERIA[10] as Companion Papers:**

We acknowledge that the detailed development and benchmarking of C-Phasing and KMERIA are more appropriately addressed in separate companion manuscripts. In the revised manuscript, we have removed the claim that C-Phasing and KMERIA represent new contributions of this work. Instead, we now treat these methods as established tools, with proper citations

to the companion manuscripts where their development and benchmarking are comprehensively described. The focus of this study has been shifted to the biological insights and applications enabled by their use in sugarcane genomics. The two companion manuscripts have now been posted as preprints with the following DOIs.

C-Phasing: https://doi.org/10.21203/rs.3.rs-7343323/v1

KMERIA: https://doi.org/10.21203/rs.3.rs-7347406/v1

**(2) Allele-Express in Supplementary Note 2:**

The development and validation of the Allele-Express algorithm are detailed in Supplementary Note 2 and acknowledged as a contribution to this study.

**(3) Revised Abstract and relevant text:**

We have rewritten the Abstract and the relevant sections to emphasize the biological and genomic insights gained in this study, rather than the novelty of the computational methods. While we acknowledge the use of C-Phasing and KMERIA, we now present these as supporting methodologies rather than primary contributions of this work.

In the revised manuscript, we clearly state that the development of C-Phasing and KMERIA is reported in separate manuscripts, while Allele-Express is a contribution in this study. We ensure that this work focuses on the application of these tools to generate biological insights into sugarcane genomics.

Similarly, based on the kmeria github and L1380-84, there is a manuscript in progress for this method as well. So, claiming the development of KMERIA as the novelty in this study seems improper. I want to point out that this work still represents great progress in polyploid genomics and population and quantitative genetics without describing C-Phasing, KMERIA, and Allele-Express as part of the work.

Response: Please refer to our previous reply for the C-Phasing companion paper.

L154, intra-chromosomal interaction is almost always the strongest. The authors may refer to intra-subgenome interactions that look stronger than inter-subgenome interactions.

Response: Agree. We have revised the sentence, which now reads:

“The analysis of chromatin contacts, derived from the alignment of Pore-C reads, reveals stronger intra-subgenome interactions than inter-subgenome interactions (**Figure 1c**)”

L201-204, is 25% enriched or depleted when compared to the genome background? If not comparing to the genome background or random expectation, you may say: “X% of the Ss-So recombined chromosomes were telocentric and acrocentric.”

Response: We acknowledge that the 25% figure was not a comparison to the genome background or random expectation. Following your suggestion, we have revised the statement to clearly report the proportion of recombined chromosomes that are telocentric and acrocentric. The updated sentence now reads:
“40% (4 out of 10) of the Ss-So recombined chromosomes were telocentric and acrocentric.”

L229-231, The difference in allele counts between So- Ss- is interesting. In Fig. 1b, the hybrid sugarcane genome has about two copies of the Ss subgenome and ~11 copies of the So subgenome. It seems that homoelogous genes are mostly retained in the Ss subgenome but not in the So subgenome. What may be the underlying mechanism causing this? Is it pseudogenization or biased fractionation? If so, which mechanisms prevail (frame shifts, premature stops, TE disruption, etc.)?
Response: We appreciate the opportunity to clarify the underlying mechanism behind the observed difference in allele counts between the So- and Ss-subgenomes.

The reviewer's interpretation that "homoeologous genes are mostly retained in the Ss subgenome but not in the So subgenome" is understandable. However, the observed difference in allele counts is not due to differential gene retention, pseudogenization, or biased fractionation. Instead, it arises from the distinct definitions of gene copy number versus allelic genes, combined with the differing ploidy levels of the two subgenomes in the hybrid genome.

To elaborate, gene copy number refers to the total number of homologous gene copies present across all haplotypes within a subgenome. For instance, if a gene is present on all 11 So-derived haplotypes, its copy number would be 11. In contrast, allelic genes are defined as homologous gene copies that differ by at least one nucleotide variant. Identical gene copies across multiple haplotypes are counted as a single allele.

For example, consider a gene with 11 copies (denoted as a-k) where copies a-d are identical in sequence, and copies e-k are also identical but differ from a-d by at least one nucleotide. In this case, the gene would be classified as having two alleles, despite having 11 total copies.

To address this distinction, we have revised the manuscript to clarify the definition of alleles (**Lines 262-264, Page 11**). The updated text now reads:

"These allele-defined genes represent instances where multiple allelic variants are present within the genome, with each allele differing by at least one nucleotide variant in coding sequences."

Table S4 shows over 99% short read mapping rate, which contradicts the results in L238 – 247, claiming mapping issues of short reads. Maybe Table S4 did not show the unique mapping rate, if so, please also show the unique

mapping rate.

Response: Yes, the previous Table S4 did not include the unique mapping rate, which has now been added. Additionally, we have revised the text in the Results section to highlight the low proportion of unique alignments in this genome (**Lines 166-168, Page 7**), which now reads: “The alignment of MGI-seq reads to the genome assembly achieved a mapping rate of over 99.93%, with **31.38%** of the reads uniquely aligned.”

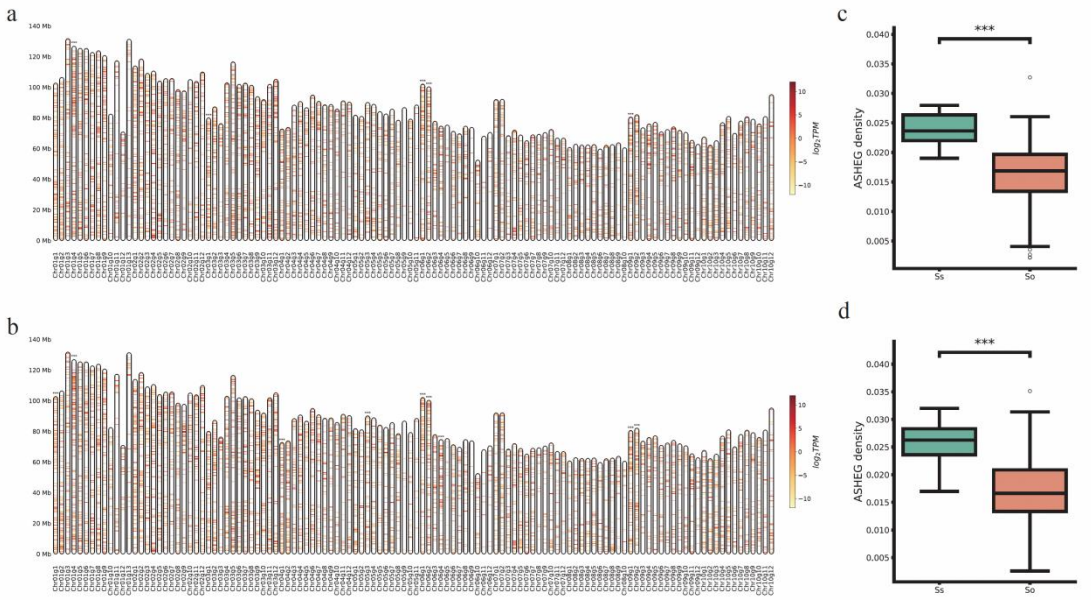
You have phased chromosomes and the allele-specific highly expressed genes (ASHEG) gives you power to quantify homo(eo)log/subgenome expression. What is the distribution of ASHEGs across homo(eo)logous chromosomes/subgenome? Are there any biased activations of chromosomes for expression (chromosomes with enriched ASHEG)?

Response: We analyzed the distribution of ASHEGs identified in mature leaf and stem tissues across all 118 chromosomes of POJ2878 (**Supplementary Figure 12a-b**). Visual inspection revealed uneven distribution, which we quantified by calculating ASHEG density (the proportion of ASHEGs relative to total genes) for each chromosome. To identify chromosomes with significant ASHEG enrichment, we applied a binomial test followed by Bonferroni correction for multiple testing (adjusted $P < 0.01$). This analysis identified five chromosomes with significant ASHEG enrichment in stem tissue (Chr01g4, Chr03g1, Chr06g1, Chr06g2, and Chr09g1) and nine chromosomes in leaf tissue (Chr01g1, Chr01g4, Chr04g1, Chr05g3, Chr06g1, Chr06g2, Chr06g4, Chr09g1, and Chr09g2), demonstrating tissue-specific biased activation patterns.

Notably, among the 10 chromosomes showing ASHEG enrichment across both tissues, seven are derived from the Ss subgenome while only three originate from the So subgenome. This prompted us to test for systematic expression bias at the subgenome level. Comparing ASHEG density between subgenomes revealed that the Ss subgenome exhibits significantly higher ASHEG density than the So subgenome in both tissues (stem: 0.024 vs. 0.016; leaf: 0.026 vs. 0.017; Mann-Whitney U test, $P < 0.001$; **Supplementary Figure 12c-d**).

This Ss-biased expression pattern is particularly noteworthy given that the So subgenome contributes ~80% of the POJ2878 genome content while Ss contributes only ~20%. The disproportionate contribution of Ss-derived chromosomes to high-level allele-specific expression suggests that Ss-derived alleles may play an outsized role in gene regulation and adaptive traits, despite their minority genomic representation.

We have added this analysis to the Results section (**Lines 301-320, Pages 12-13**) and included a new supplementary figure (**Supplementary Figure 12**) showing the chromosome-level distribution and subgenome comparison of ASHEG density.




**Supplementary Figure 12. Chromosome-level distribution and subgenome-biased expression of ASHEGs.** (a-b) Distribution of ASHEGs across all 118 chromosomes of POJ2878 in stem tissue (a) and leaf tissue (b). Each vertical bar represents an individual chromosome, with the height indicating chromosome length (Mb). The colored segments within each bar represent the distribution of ASHEGs along the chromosome, with color intensity reflecting the $\log_2$ TPM expression level. Chromosome labels on the x-axis indicate chromosome identity. ASHEG density (the proportion of ASHEGs relative to total genes per chromosome) was calculated for each chromosome, and chromosomes with significant ASHEG enrichment were identified using a binomial test followed by Bonferroni correction for multiple testing (adjusted $P < 0.01$). c, d Comparison of ASHEG density between Ss and So subgenomes in stem tissue (c) and leaf tissue (d). Box plots display median (center line), interquartile range (box), 1.5× interquartile range (whiskers), and outliers (circles). Statistical analysis was performed using Mann-Whitney U test with *** representing $P < 0.001$.

How are branches colored in Figure 2d? There are “78 S. officinarum (SO), 290 S. spontaneum (SS), and 613 hybrid sugarcane cultivars”. I thought these were red, green, and purple, respectively, but those counts do not appear to match branch counts. For example, samples located on the right of the cladogram with a high proportion of So-genomic components were colored purple on the branch.

Response: The color coding has been corrected. Please see the revised

Figure 2d.

Fig3a legend references red and blue, but the figure appears red and green.
Response: The legend in Fig3a has been corrected. It should be green rather than blue.

Also, I find this figure confusing. How are edge colors determined? How are nodes spatially oriented? Does the distance from POJ2878 tell us anything?
Response: Thanks for the opportunity to clarify the figure legend and address these concerns.

The color of each edge corresponds to the node with the higher contribution between the two connected cultivars.

The spatial positioning of the nodes was determined using the ForceAtlas2 force-directed algorithm[16], implemented in the Gephi software package. This algorithm arranges nodes to minimize visual overlap and enhances the readability of the network structure.

Regarding the distance from POJ2878, it does not hold a specific quantitative meaning in this layout. Since we did not apply edge weights in the analysis, the spatial arrangement reflects the ForceAtlas2 algorithm's optimization for visual clarity rather than a scaled measure of genetic relatedness. As such, proximity in the graph should not be interpreted as a direct indicator of genetic correlation.

L304-306 reported that 312/573 cultivars shared over 30 Mb of IBD sequences with POJ2878, which seems to be a very small number given the over 10 Gb genome size, the relatively short breeding history, and the lack of sexual recombination in sugarcane production. How is this not due to random noise?
Response: We apologize for the confusion caused by an earlier misstatement in the manuscript and appreciate the reviewer's attention to this detail. The reference to "30 Mb of IBD sequences" was incorrect: this value does not represent a genome-wide threshold for POJ2878-derived introgression, but rather the minimum IBD segment size identified on a single chromosome (Chr01g1) in initial analyses. This error arose from a misinterpretation of the IBDseq output, which we have since corrected.

To address this issue and establish a robust genome-wide threshold for POJ2878 introgression, we refined our approach by analyzing the distribution of shared IBD sizes across various cultivars. Through this analysis, we determined a minimum IBD size of 821.6 Mb for POJ2878-derived introgression across the entire genome, representing the top 5% of the longest IBD regions shared among cultivars (**Response Figure 2**).

Using the revised 821.6 Mb threshold, we reanalyzed a dataset of 573 hybrid sugarcane cultivars with confirmed geographical distribution. The updated results revealed that 546 out of 573 cultivars (95.3%) share over

821.6 Mb of IBD sequences with POJ2878—a threshold corresponding to the top 5% of the longest IBD regions shared among cultivars. This finding strongly highlights POJ2878's extensive use in global breeding programs.

All relevant content in the revised manuscript has been updated to reflect these corrections, including adjustments to the text (**Lines 368-373, Pages 14-15**) and updates to Figure 3.

Fig3c, is there a way to indicate which subgenome these are? Consider highlighting SUS2 so that it's more easily referenced in e-f.

Response: We have clarified the subgenomic origins of each chromosome in Figure 3c (x-axis) by using distinct colors: green for the Ss subgenome, red for the So subgenome, and black for recombined chromosomes. Furthermore, we have highlighted *SUS2* in red within Figure 3c to ensure it is easily identifiable and directly referenced in Figures 3e and 3f. Please see the revised Figure 3 for details.

L282, are SNPs called from unique read alignments?

Response: Yes, these SNPs were called from unique read alignments. We have modified the sentences to clarify this:

"Analysis of these uniquely aligned reads to the POJ2878 reference genome yielded 11.43 million SNPs and 7.36 million indels."

L376-381. Fig 4b is XPCLR not Fst.

Response: Thank you for your reminder. It is indeed XPCLR. We have made this correction.

Fig4a has peaks and labels, but many of the most distinct peaks are not labeled? Are they not containing genes?

Response: The distinct unlabeled peaks in Figure 4a do contain genes; however, they were not labeled for two reasons: (1) their functions remain unclear based on GO or KEGG annotations, and (2) they are unrelated to the traits being investigated in this study, such as those associated with adaptive evolution, natural selection, or traits targeted by human-driven artificial selection. To maintain clarity and ensure the figure remains focused on our core research objectives, we chose to highlight only the genes directly relevant to these topics.

Fig 4efgh, lack of data to further support selection over random drifting

Response: To address the question regarding selection versus random drift, we provide the following points:

**(1) XP-CLR Analysis to Account for Random Drift:**

We used the XP-CLR (Cross-Population Composite Likelihood Ratio)

method to analyze selective sweeps. This method is specifically designed to minimize the confounding effects of random drift and population history through its statistical framework. XP-CLR directly compares two competing models: the null hypothesis (H0), which attributes observed allele frequency differences to genetic drift under population divergence, and the alternative hypothesis (H1), which suggests that recent directional selection is required to explain these differences. When allele frequency differences (Δp) between populations exhibit spatial patterns consistent with linkage decay, and exceed what can be explained by random drift, it indicates the presence of positive selection rather than random drift[17].

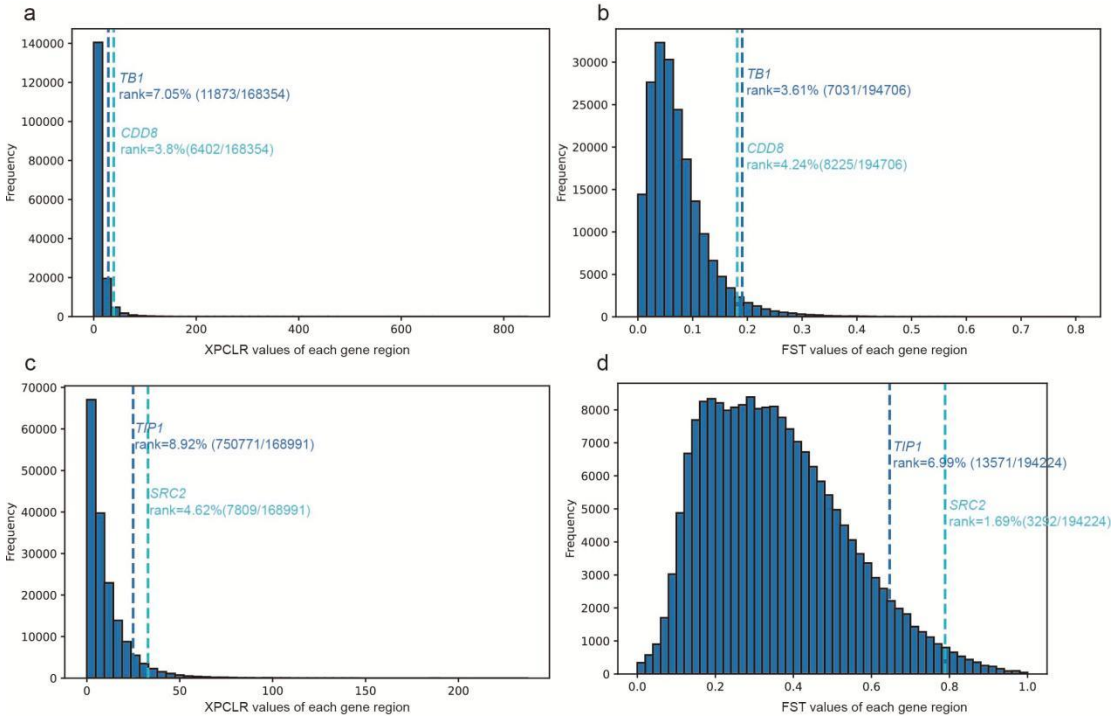
**(2) Selection in Key Adaptive Traits:**

Genes under selection are often associated with traits linked to environmental adaptation. In the case of cultivated sugarcane (Cul), traits such as sugar accumulation and tillering capacity are key advantages over wild species (*S. spontaneum*) and noble species (*S. officinarum*), directly reflecting adaptive evolution under selective pressure. Cross-population selection analyses (Cul vs. So; Figures 4c, 4e, 4g) revealed that tillering-related genes, such as *TB1* and *CCD8*, are under selection. Enhanced functionality of these genes likely contributes to higher yield, improving survival and reproductive efficiency in cultivated sugarcane. Similarly, cross-population analyses (Cul vs. Ss; Figures 4d, 4f, 4h) identified selection on vacuole development-related genes, such as *TIP1* and *SRC2*. These genes likely enhance vacuole function, which is crucial for sucrose storage. Their regulation optimizes sugar accumulation and storage, directly improving the economic traits of cultivated sugarcane and aligning with human demands for high sugar content.

**(3) Empirical Validation of Selection Signals**

To confirm that the observed selection signals (**Figure 4e-h**) exceed the expectations of random drift, we compared the XP-CLR and $F_{ST}$ values of candidate genes against genome-wide empirical distributions derived from all gene regions (**Supplementary Figure 22**). These distributions represent the expected range of genetic differentiation under random drift.

Candidate genes (TB1, CCD8, TIP1, and SRC2) were identified as significant outliers, with their XP-CLR and $F_{ST}$values falling in the extreme tails of the empirical distributions. To further quantify the combined significance of these metrics, we applied Fisher's combined probability test to merge the p-values of XP-CLR and $F_{ST}$. The resulting combined p-values were 0.0175 (TB1), 0.0118 (CCD8), 0.0382 (TIP1), and 0.0064 (SRC2), confirming that the differentiation signals for these genes significantly exceed what can be explained by random drift.




**Supplementary Figure 22: Empirical distributions of XP-CLR and $F_{ST}$ values for each gene region.** (a-b) Rankings of TB1 and CCD8 (Cul vs. So). (c-d) Rankings of TIP1 and SRC2 (Cul vs. Ss).

Fig 5efg are stretched and have a distorted ratio.

Response: Figure 5 has been revised.

How to use kmeria-identified genomic regions to find genes? How was fine mapping done? Line 1380-1384 are not detailed.

Response: We apologize that this section lacked sufficient detail to clarify our fine-mapping workflow. We have revised the manuscript (**Lines 1529-1541, Page 57**) to elaborate on the process as follows:

"To achieve precise genomic localization, KMERIA employs a read-based refinement strategy. Instead of mapping k-mers directly, we first retrieve sequencing reads that contain the significant k-mers. These longer, context-rich reads are then aligned to the monoploid reference genome using BWA (default parameters) or minimap2 (option -x sr). This approach substantially improves alignment accuracy and enables unique localization of the associated genomic regions with high confidence. Once significant k-mers are anchored through this read-based mapping, KMERIA defines a candidate genomic interval for gene discovery. By default, a 50-kb window is surveyed on both sides of the mapped site, though this can be adjusted according to the linkage disequilibrium (LD) decay of the studied population. All annotated genes within this interval are considered putative candidates underlying the

associated trait."

L462-467, OE varies too much (7.2% vs 28.92%). Why?

Response: We agree with this observation regarding the differences in cell size increase between the *SUT2* overexpression lines OE1 (7.2%) and OE2 (28.9%), and believe it reflects a common phenomenon in gene function studies. Below is our analysis and explanation for this phenomenon:

In plant trait biology, the complexity of the genetic background, physiological redundancy, and microenvironmental variability often causes the overexpression of a single gene to produce phenotypic effects of differing magnitudes across independent transgenic events. For example, in the study on TaCol-B5[18], yield increase among overexpression lines ranged from 7.8% to 19.8%. Therefore, the phenotypic variation we observed between OE1 and OE2 aligns with the general principles governing the genetic regulation of complex traits.

In addition, the regulation of cell size is a complex biological process, controlled by multi-layered mechanisms. While *SUT2* overexpression is confirmed as an important factor influencing this trait, the final phenotype is the result of the interaction between the gene product and the endogenous regulatory network. Beyond *SUT2* itself, other unknown mechanisms within the system may also modulate the final cell size.

This further underscores the importance of analyzing multiple independent transgenic lines in functional studies to avoid drawing conclusions based on a single, potentially atypical line.

The approach of the current manuscript has been focusing on finding genomic regions with varying methods (IBD/PAV/introgression/expression/selection), then performing GO/KEGG analysis to identify patterns. The integration between MAS-ISO-seq, GWAS, selection, and phased genome is weak. How biological questions can be answered by combining these data? The paper so far presents results from only one or two methods without organically combining them to strengthen the conclusion.

Response: Our study provided an integrated framework in which three tools function as interdependent components, enabling us to move from genome assembly to association mapping to functional validation in a unified way.

***C-Phasing as the foundation.*** Using Pore-C data, C-Phasing produces a chromosome-scale, haplotype-resolved assembly that separates homologous and homeologous sequences—a central challenge in polyploid genomes. The resulting POJ2878 reference comprises 118 individually phased chromosomes across 10 homoeologous groups. This assembly is not only a reference backbone; it is the prerequisite for allele-defined gene models and all downstream analyses.

***KMERIA and Allele-Express as functional interpreters***. *KMERIA*

performs k-mer-based GWAS, which avoids the errors inherent in complex polyploid genotyping. Crucially, the biological meaning of significant k-mers comes from mapping them back to the phased POJ2878 reference, where they can be assigned to specific haplotypes and alleles. Allele-Express then leverages the same phased reference and allele-level annotations to quantify allele-specific expression for candidate genes highlighted by KMERIA or population-genetic signals. In a genome as complex as sugarcane, accurate ASE depends on correct phasing; otherwise, transcripts from different alleles cannot be reliably distinguished. Our ASE analyses identify functionally relevant haplotypes (for example, aquaporin TIP2/3, linked to vacuolar water flow and cell expansion), illustrating how the assembly enables precise molecular dissection rather than serving as a static reference.

***Example of end-to-end integration.*** K-mers associated with parenchyma cell traits were mapped to POJ2878, pinpointing candidate genes including *ShSUT2*. Using the phased genome, we resolved allele variants across haplotypes and found one allele with elevated expression in leaf tissue. This chain—from association to allele-level mapping to ASE—illustrates how integrating the three tools generates testable biological hypotheses and concrete breeding targets.

To address the reviewer's concern, we added **Extended Data Fig. 10** to depict the workflow: (i) C-Phasing enables allele-defined gene annotation; (ii) Allele-Express provides accurate ASE based on those annotations; and (iii) KMERIA's signals gain biological interpretation through mapping to the phased assembly. We also revised the Discussion (**Lines 587-622, Pages 22-23**) to emphasize that the framework is synergistic: each tool increases the resolution and interpretability of the others, collectively enabling biological inference that would not be possible with any single method.

Referee #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Response: Thanks so much for your valuable comments.

References

1. Ma, Y. *et al.* COLD1 Confers Chilling Tolerance in Rice. *Cell* **160**, 1209–1221 (2015).
2. Chinnusamy, V. *et al.* ICE1: a regulator of cold-induced transcriptome and freezing tolerance in *Arabidopsis*. *Genes Dev.* **17**, 1043–1054 (2003).
3. Pribil, M. ALTERING SUGARCANE SHOOT ARCHITECTURE THROUGH GENETIC ENGINEERING: PROSPECTS FOR INCREASING CANE AND SUGAR YIELD. **29**, (2007).

4. Anur, R. M., Mufithah, N., Sawitri, W. D., Sakakibara, H. & Sugiharto, B. Overexpression of Sucrose Phosphate Synthase Enhanced Sucrose Content and Biomass Production in Transgenic Sugarcane. *Plants* **9**, 200 (2020).
5. Yu, F. *et al.* Chromosome□specific painting unveils chromosomal fusions and distinct allopolyploid species in the *Saccharum* complex. *New Phytologist* **233**, 1953–1965 (2022).
6. Wang, F. *et al.* MIKE: an ultrafast, assembly-, and alignment-free approach for phylogenetic tree construction. *Bioinformatics* **40**, btae154 (2024).
7. Garsmeur, O. *et al.* The genomic footprints of wild Saccharum species trace domestication, diversification, and modern breeding of sugarcane. *Cell* S0092867425010852 (2025) doi:10.1016/j.cell.2025.09.017.
8. Wang, Y. *et al.* Enhanced Pore-C with C-Phasing Enables Chromosomal-Scale, Haplotype-Resolved Assembly of Ultra-Complex Genomes. *PREPRINT (Version 1) available at Research Square* https://doi.org/10.21203/rs.3.rs-7343323/v1 (2025) doi:https://doi.org/10.21203/rs.3.rs-7343323/v1.
9. Browning, B. L. & Browning, S. R. Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data. *The American Journal of Human Genetics* **93**, 840–851 (2013).
10. Chen, S. *et al.* A k-mer-based GWAS approach empowering gene mining in polyploids. *PREPRINT (Version 1) available at Research Square* https://doi.org/10.21203/rs.3.rs-7347406/v1 (2025) doi:https://doi.org/10.21203/rs.3.rs-7347406/v1.
11. He, C. *et al.* Trait association and prediction through integrative *k* □mer analysis. *The Plant Journal* **120**, 833–850 (2024).
12. Healey, A. L. *et al.* The complex polyploid genome architecture of sugarcane. *Nature* **628**, 804–810 (2024).
13. Zhang, J. *et al.* The highly allo-autopolyploid modern sugarcane genome and very recent allopolyploidization in Saccharum. *Nat Genet* **57**, 242–253 (2025).
14. Zhang, J. *et al.* Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nature Genetics* **50**, 1565–1573 (2018).
15. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature Plants* **5**, 833–845 (2019).
16. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* **9**, e98679 (2014).
17. Chen, H., Patterson, N. & Reich, D. Population differentiation as a test for selective sweeps. *Genome Res.* **20**, 393–402 (2010).
18. Zhang, X. *et al. TaCol-B5* modifies spike architecture and enhances grain yield in wheat. *Science* **376**, 180–183 (2022).

# A point-by-point response to Reviewers' comments

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have thoroughly revised their manuscript and provided a detailed response letter. Overall, we are satisfied with most of the points, but two issues remain.
(1) The oligo-FISH analysis does not appear to have worked as intended, although the authors do not state this explicitly. Several panels show no signal or only a single colour, and the red probe frequently failed to yield a signal. Where signals are present, arrows should be added to indicate them clearly.
We appreciate the authors' willingness to present negative results; however, Extended Data Fig. 1 may confuse readers who do not also consult the Peer Review file. The presentation should therefore be improved, possibly by omitting negative results. Since Nature uses an open peer review policy, these results are already appropriately documented. We would also be curious to know why so many probes failed to produce a signal.

Response: We thank the reviewer for this thoughtful suggestion. We agree that including panels with negative results (no signal) could be confusing for readers who may not consult the peer review history.

Following your advice to improve the presentation, we have removed the panels showing negative results. We have retained only the panel (Chr05g) where the oligo-FISH probes successfully hybridized and generated clear signals in the revised **Figure 1b**. As requested, we have also added arrows to these panels to clearly indicate the fluorescent signals.

Regarding the question of why certain probes failed to produce signals, as briefly noted in our previous response, the primary limiting factors are probe quantity and density relative to the size of the target region. Specifically, successful oligo-FISH signal detection in sugarcane typically requires:

- A sufficient number of probes per cluster (empirically >5,000 oligos).
- A high probe density (typically >0.12 probes/kb).

However, designing **haplotype-specific probes** for the cultivated sugarcane genome is particularly challenging due to its high ploidy and structural complexity. In regions where no signals were detected, the amount of available unique sequences were insufficient to meet the requirements for effective probe density and coverage. Consequently, the fluorescence intensity fell below the detection threshold of the microscopy system. It is important to note that this is a known technical limitation of applying oligo-FISH to highly polyploid and repetitive genomes, rather than evidence of genome assembly errors.

(2) The section of the population analysis referring to "a large number of evolutionarily conserved loci" remains unclear. Specifically, we ask the authors to provide additional information:

How many of these conserved loci show overlap between S. s. and S. o. in the POJ2878 genome, and where are these overlapping regions located?
Why these signals cannot be explained by ambiguous read mapping rather than true evolutionary conservation.
Additionally, we wonder whether this could be related to aneuploidy—i.e., S. s. and S. o. reads both mapping to the same regions of the POJ2878 genome. Theoretically, the proportion of such shared regions should not be very high, and it is difficult to see how this would represent genome-wide differences.
Without these clarifications, it remains challenging to fully understand the biological significance of the population analysis in this context.

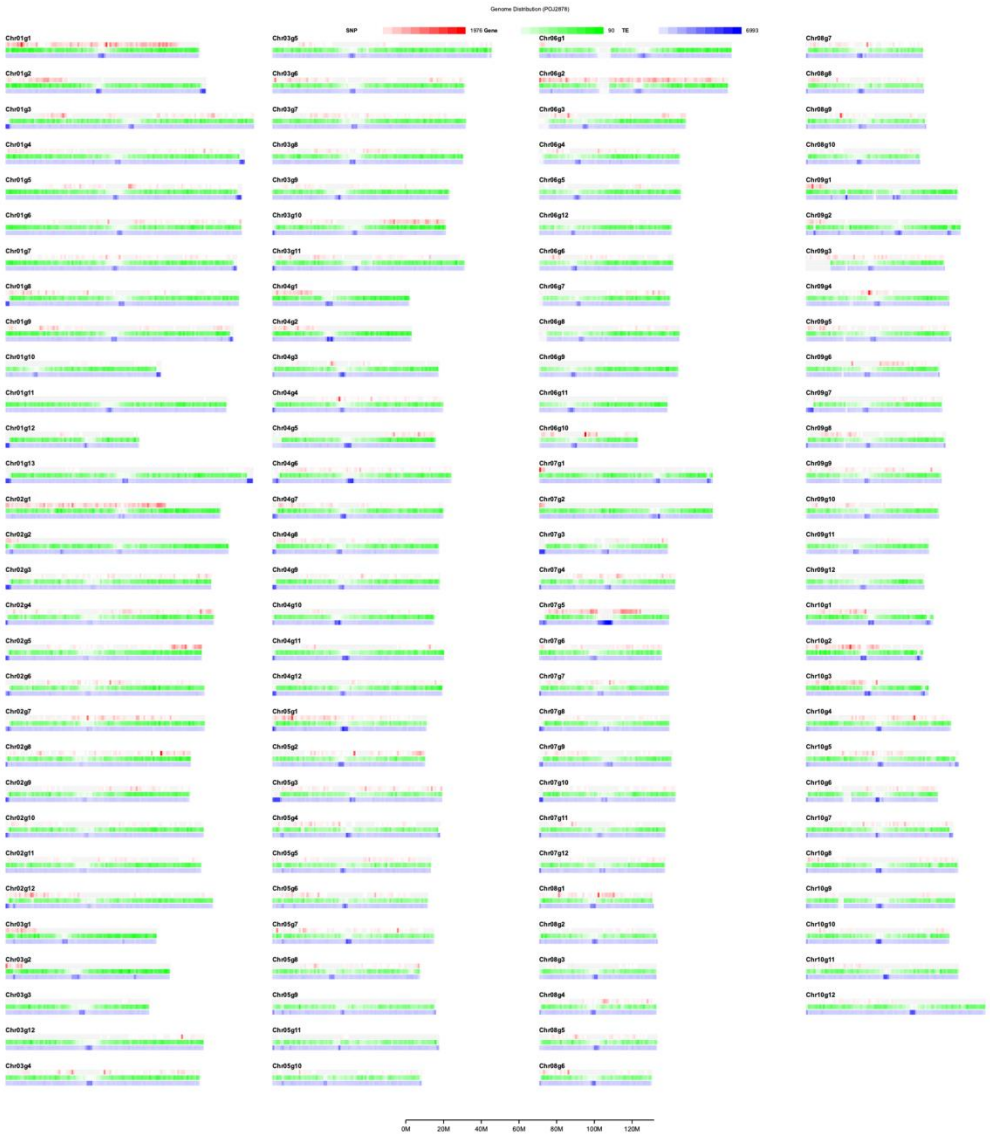
Response: We thank the reviewer for highlighting that the statement referring to a "large number of evolutionarily conserved loci" was insufficiently precise. In response, we: (i) quantify the extent of these shared SNP overlapped between *S. spontaneum* (Ss) and *S. officinarum* (So) on the POJ2878 coordinate system and describe where these loci are located, and (ii) provide multiple lines of evidence demonstrated that the signal is not attributable to ambiguous read mapping, aneuploidy, or sequencing depth artifacts.

For clarity, we refer to these sites as **high-confidence homologous anchor loci** (i.e., reliably mappable homologous coordinates shared between Ss and So in the POJ2878 reference), rather than implying that every site is invariant across species.

**(1) How many shared ("overlapping") loci are detected between Ss and So in the POJ2878 genome, and where are they located?**

Among the 11.43 million high-quality SNPs identified in the population, we defined shared SNP loci as sites with a call rate >50% in both populations (i.e., >50% of Ss samples and >50% of So samples had non-missing genotypes calls). Using this criterion, we identified 656,089 shared SNP loci, representing 5.74% of all high-quality SNPs. Additionally, we calculated the genomic regions that can be mapped by both So and Ss reads. By analyzing 50 Ss and 50 So individuals with mappability rates exceeding 50% in both populations, we identified a total of 412 Mb of shared genomic sequences that can be confidently mapped in both progenitor species. These shared regions comprise 3.97% of the POJ2878 genome, a proportion comparable to that observed for shared SNP loci.

We mapped these loci onto the 118 POJ2878 chromosomes (**Response Figure 1**). The shared loci are distributed across 117 chromosomes, with some forming distinct high-density clusters that predominantly co-localize with gene-rich regions. In contrast, transposable elements (TEs) display a complementary distribution pattern and are enriched in regions with lower gene densities and reduced SNP abundance.



**Response Figure 1. Distribution of high-confidence homologous anchor SNP loci, gene and TEs across the 118 chromosomes in POJ2878 genome.**

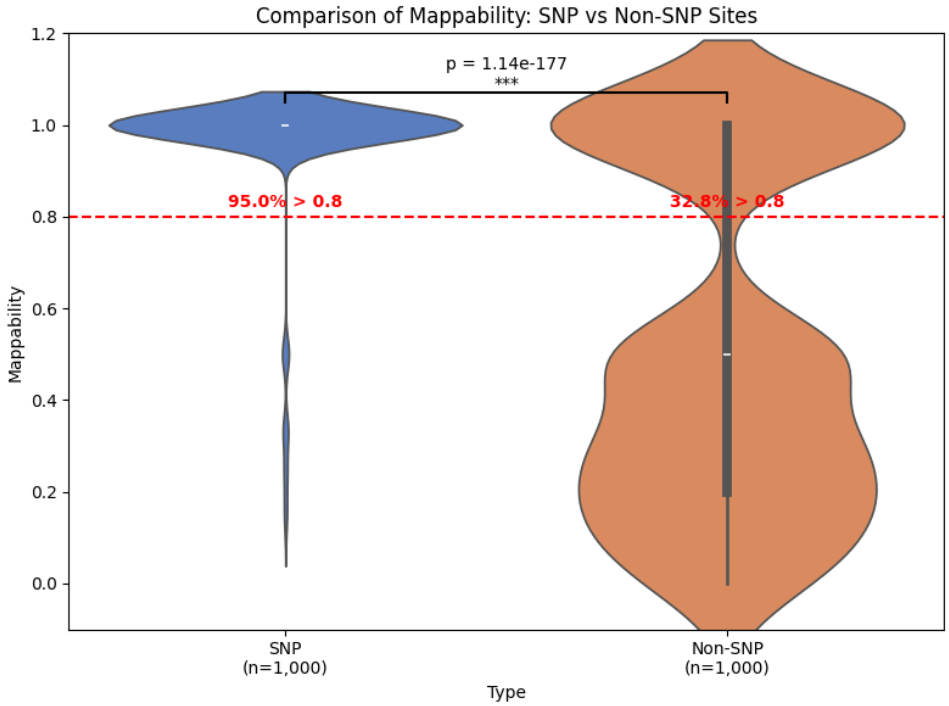
**(2) Why the shared loci are unlikely to be explained by ambiguous read mapping**

We fully recognize that ambiguous mapping is a major concern for short-read analyses in highly polyploid sugarcane genomes. To minimize this risk, our SNP calling pipeline incorporated stringent alignment filters to retain only high-confidence, uniquely aligned reads. Specifically, we excluded reads flagged for multiple alignments (e.g., SA/XA tags) and retained only reads meeting the criteria AS:i > XS:i and MAPQ > 20.

To evaluate whether this filtering strategy effectively reduces mis-mapping, we simulated 50 million pairs of PE150 reads (i.e., 100 million reads) from the POJ2878 reference sequence using *wgsim* (default settings). The simulated reads were aligned them back to the POJ2878 genome, and the same alignment filters were applied. After filtering, 36,067,470 reads remained, of which 59,276 were incorrectly mapped, corresponding to **a mis-mapping rate of only 0.16%**. These results indicate that our

alignment and filtering settings achieve high mapping specificity under this reference framework.

We next examined whether shared SNP loci are preferentially located in uniquely mappable genomic regions. Genome-wide mappability was calculated using *GenMap* (-E 0), where higher values indicate higher sequence uniqueness and lower ambiguity for short-read alignment. We compared 1,000 randomly sampled shared SNP loci with 1,000 random genomic positions, and found a strong enrichment of high mappability among shared loci: 95.0% of shared SNP loci show mappability >0.8, whereas only 32.8% of random sites exceeded this threshold (**Response Figure 2**). These results demonstrated that shared loci are concentrated in uniquely mappable regions, arguing against ambiguous read mapping as the primary driver of the observed signals.




**Response Figure 2. Comparison of genome mappability between SNP sites and random genomic sites (i.e., Non-SNP).** Mappability scores were calculated for 1,000 randomly selected SNP sites (left) and 1,000 randomly selected non-SNP genomic sites (right). The violin plots display the distribution of mappability scores (ranging from 0 to 1) for each group, with internal box plots indicating the median and interquartile ranges. The red dashed line marks a mappability threshold of 0.8, with the percentage of sites exceeding this threshold annotated above each group (SNP: 95.0%; Non-SNP: 32.8%). Statistical significance was determined using a two-sided Mann-Whitney U test (***p < 0.001).

### (3) Why aneuploidy/depth abnormalities are unlikely to account for the genome-wide shared-loci pattern

We agree that aneuploidy or sequencing depth irregularities could, in principle, bias read mapping by generating abnormal copy-number targets and artificially inflating

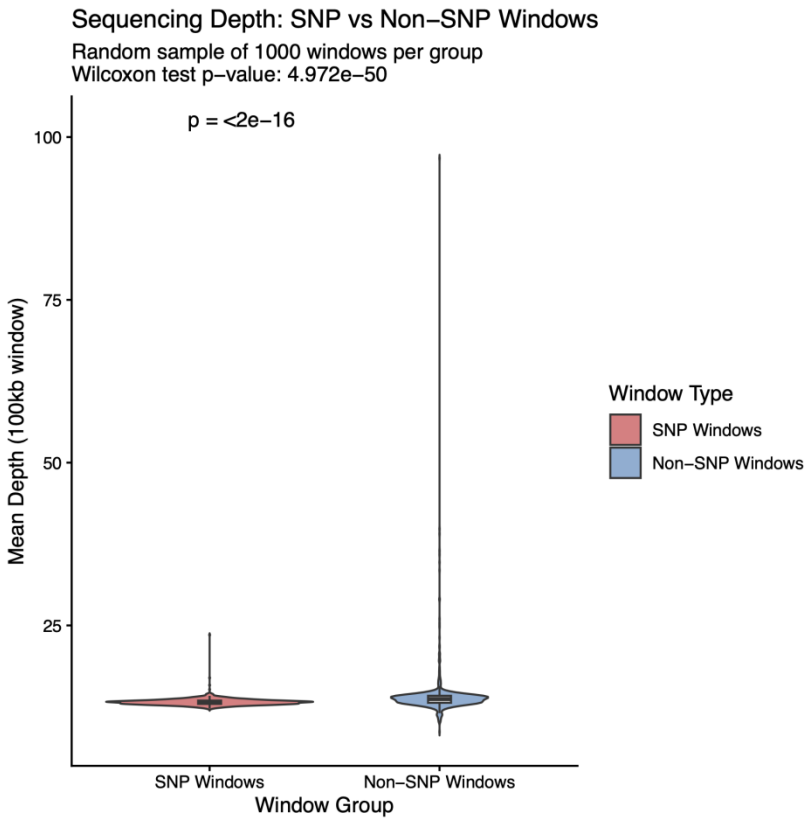
apparent overlap. We therefore explicitly examined whether depth abnormalities co-localize with shared loci.

First, assembly quality evaluation indicates a very low collapsed error rate (0.82%) as reported in the main text. This suggests limited large-scale assembly collapse and reduces the likelihood that widespread copy-number artifacts are driving the observed patterns.

Second, we compared the distribution patterns of these shared loci with the genome-wide sequencing depth profile (**Response Figure 3**). Visualization across the genome shows that regions with high densities of shared SNPs (dark blue heatmap) consistently coincide with stable sequencing depth (red track), indicating that these loci are are primarily located in reliable, single-copy genomic regions. In contrast, regions exhibiting abnormal depth variation display a pronounced depletion of shared SNPs, further supporting the effectiveness of the filtering strategy.

We also compared the sequencing depth distribution of windows containing these SNPs against a random set of genomic windows (**Response Figure 4**). The analysis revealed a statistically significant difference in sequencing depth between the two groups (Wilcoxon rank-sum test, $p < 2.2e\text{-}16$). Notably, windows containing shared SNPs exhibited substantially lower depth variation ($SD \approx 0.54$) compared to non-SNP windows ($SD \approx 3.51$), indicating that SNPs are preferentially located in stable, single-copy genomic regions. In contrast, non-SNP windows exhibited higher depth variability and greater susceptibility to copy number expansion, further confirming the effectiveness of our filtering strategy.

**Response Figure 3. Genome-wide landscape of sequencing depth and shared SNP density.** The red tracks represent the average sequencing depth calculated in 100kb non-overlapping windows across 118 chromosomes. The underlying chromosomal heatmaps display the density of shared SNPs within the same windows, where darker blue indicates higher density.



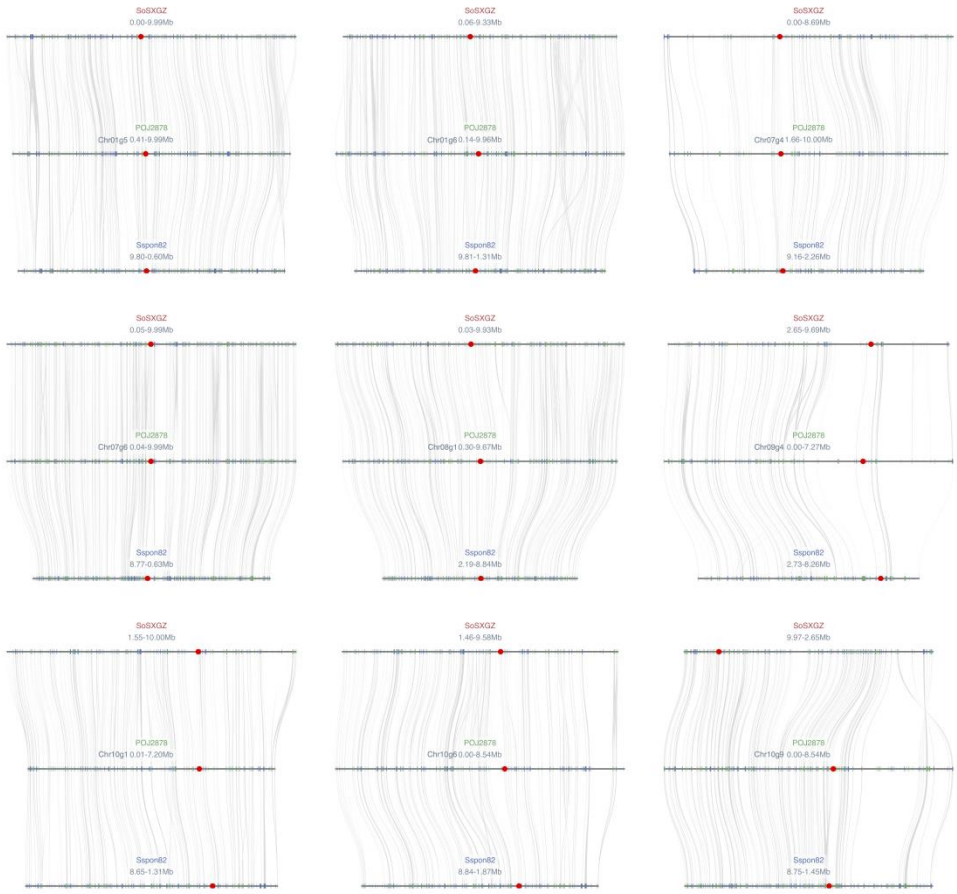

**Response Figure 4. Comparison of sequencing depth between SNP-containing and non-SNP windows (i.e., a set of randomly selected genomic loci).** Violin plots showing the distribution of mean sequencing depth (100kb windows) for windows containing shared SNPs (red, n=1,000) versus windows without shared SNPs (blue, n=1,000). The internal box plots indicate the median (thick line) and interquartile range. Statistical significance was determined using the Wilcoxon rank-sum test ($p < 2.2e-16$).

### (4) Evidence that shared loci represent homologous anchor loci conserved across Saccharum genomes

To evaluate whether shared loci reside within conserved homologous sequence contexts across progenitor genomes, we randomly sampled 500 loci from the 656,089 shared set and examined their syntenic context using the T2T monoploid assemblies of Ss82-114 and SoXZ generated in this study. We classified a locus as conserved across the three genomes if the SNP coordinate and its ±200 bp flanking sequence could be identified in POJ2878, Ss82-114, and SoXZ within syntenic regions. Under this criterion, 57.8% (289/500) of sampled loci met the conservation requirement. In addition, 27.2% (56 in So and 80 in Ss) were detected in only one of the progenitor genomes.

Representative examples are shown in **Response Figure 5**, and the full set of synteny plots can be downloaded at the following link (https://drive.google.com/file/d/1_QkNp6MOkWxR1E5GaouoGd2-B0q4hjvQ/view?usp=sharing). The remaining loci that could not be recovered in both progenitor assemblies likely reflect local assembly gaps, structural variation or presence–absence variation among genomes, and/or conservative flanking-sequence matching thresholds, rather than systematic mapping artifacts.

These results support that a substantial proportion of shared loci correspond to biologically meaningful homologous anchors conserved across Saccharum genomes, rather than mapping artifacts, and the overall conclusion is not affected by the minority of loci that could not be matched across all three genomes.



**Response Figure 5. Local synteny examples for nine SNP loci between POJ2878 and the two progenitor genomes (SoXZ and Ss82-114).** In each panel, tracks (top to bottom) show SoXZ (red label), POJ2878 (green), and Ss82-114 (blue); gray ribbons connect syntenic gene anchors. Red dots mark the SNP position mapped in each genome using the ±200 bp POJ flank (identity ≥ 80%, query coverage ≥ 0.80), and synteny support is evaluated using anchor blocks with a ±3 Mb padding around the block span.

**(5) Note on reference representation and implications for population analysis**

We agree that POJ2878 is a single cultivar and therefore cannot represent all genetic variations across the *Saccharum* genus. Variants absent from POJ2878 cannot be represented in a POJ2878-based coordinate system. Nevertheless, our population analyses are based on a large set of high-confidence homologous anchor loci that remain after stringent mapping and quality controls. The inferred population structure is robust to these filtering steps, indicating that our conclusions are not driven by artifacts associated with reference representation, ambiguous mapping, or depth abnormalities. Importantly, the resulting clustering and relationships are also consistent with the well-documented evolutionary and breeding history of *S. spontaneum*, *S. officinarum*, and modern hybrids, providing an independent line of support for the biological interpretability of the population analysis.

In the revised manuscript, we now provide genome-wide quantification of shared loci between Ss and So (656,089; 5.74%). We also added a synteny-based validation using the two progenitor T2T monoploid assemblies, along with an explanation that unmatched loci are most consistent with assembly gaps/structural or presence–absence variation, or threshold stringency, rather than systemic mapping artifacts. Changes can be found in our revised manuscript (**Lines 1,304-1,317, Pages 47-48**)

Referee #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.
Response: We sincerely thank the reviewer for your time and effort in reviewing our manuscript.

Referee #3 (Remarks to the Author):

I appreciate the substantial amount of work the authors have put into revising and updating the manuscript. They have fully addressed the concerns I raised in the previous round of review and I have no additional points to raise.
Response: We sincerely thank the reviewer for your time and effort in reviewing our manuscript.

Referee #4 (Remarks to the Author):

Wang et al. present a revised manuscript with greatly improved quality and clarity. The reviewer still has several concerns about the manuscript's conceptual framing, data interpretation, use of selection methods, and gene mapping. The selection analyses are improved, but may still suffer from the overly permissive cutoffs, which likely produced extensive false positives, undermining conclusions about selective sweeps. Concerns are also raised about the insufficient detail in fine mapping, inadequate discussions of biological insights, and several minor issues. Detailed comments are shown below.

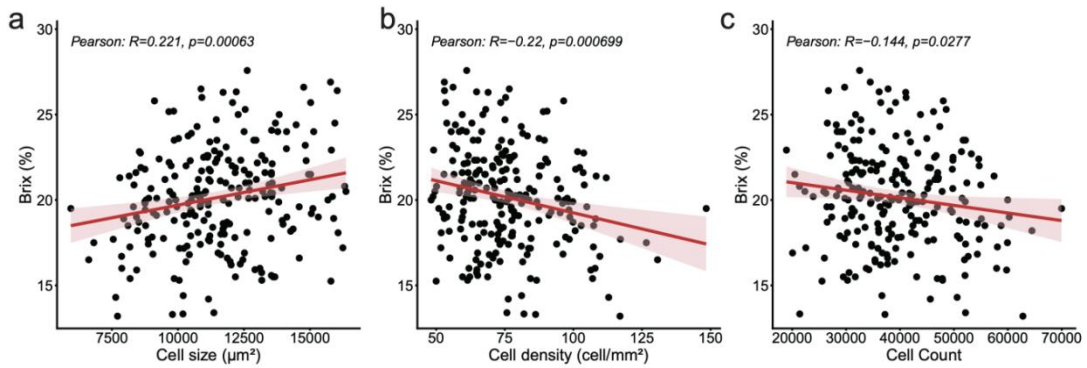
The introduction is still mostly focusing on sugarcane genomics, not sugarcane biology and evolutionary genetics. The reasoning behind GWAS studies of parenchyma cell size and sucrose storage capacity is unclear. What is the significance of focusing on these traits? Are these poorly studied traits? Are there any related genes reported? For the population genetics analyses, are there any genes reported previously as undergone domestication? Without these information, it is difficult to comprehend the significance of this study and the necessity of conducting the reported analyses.

Response: We thank the reviewer for this important comment. We agree that the previous version did not sufficiently articulate the biological and evolutionary-genetic motivations for (i) focusing on parenchyma cell size and sucrose storage traits in GWAS and (ii) performing population-genetic scans for domestication/improvement signatures. In the revised manuscript, we clarified the underlying biological questions, the specific knowledge gaps, and how our analyses address them.

**(1) Biological significance and rationale for parenchyma cell size and sucrose storage capacity**

Sugarcane's remarkable productivity is underpinned by its ability to accumulate sucrose at extraordinary concentration in stalk parenchyma cells. A central, yet poorly resolved, biological question persists: which cellular and genetic mechanisms determines the capacity and efficiency of sucrose storage in these cells? Evidence from other crops supports a mechanistic connection between parenchyma cell expansion/size and sugar accumulation (e.g., sugar beet *BvBZR1*[1]; strawberry tonoplast transporter *FvMATE51*[2]), and physiological studies in sugarcane suggest that parenchyma cell-wall properties help accommodate the high turgor pressure associated with massive sucrose storage[3]. We therefore revised the Introduction (**Lines 72-83, Page 4**) to explicitly frame parenchyma cell traits as biologically central and underexplored targets for understanding (and potentially improving) sucrose storage in sugarcane.

To strengthen agronomic relevance of our findings beyond early developmental stages, we collected additionally sugar-content measurements at the mature stage (300 days after planting) and tested their association with parenchyma cell traits. In the previous version, sucrose-related measurements were taken at 210 days, when stalks were not yet fully mature, which likely reduced our power to detect trait-sugar relationships at the final storage stage. Using the 300-day Brix data, we observed significant correlations between parenchyma cell traits and Brix ($P < 0.05$; **Supplementary Figure 22**), confirming that these cellular traits are relevant to sugar content at maturity and supporting their potential application in sugarcane breeding.




**Supplementary Figure 22. The correlation analysis between Brix and parenchyma cell traits.** Scatter plots showing the relationships between Brix content (%) and three parenchyma cell characteristics: (a) cell size (μm²), (b) cell density (cell/mm²), and (c) cell count. Each black dot represents an individual sample. The total sample size (n) is 265. Red lines indicate linear regression trends with shaded areas representing 95% confidence intervals. Pearson correlation coefficients (R) and p-values are displayed in each panel.

**(2) Are any related genes reported?**

We now address this at two levels. First, we cite prior evidence from other crops that links parenchyma cell expansion/tonoplast transport to sugar accumulation (as described above), highlighting plausible biological pathways despite the limited gene-level validation in sugarcane. Second, our GWAS identified candidate genes with established roles in cell/organ growth or transport in other species (e.g., homologs of *AtMob1A*, a MATE-family transporter, and a WD40-domain organ-size regulator). Notably, we found that *SUT2* is significantly associated with parenchyma cell size. Functional validation further demonstrated that overexpression of *ShSUT2* significantly increases parenchyma cell size (**Extended Data Fig. 9**; $P < 0.0001$), providing direct gene-level support that these traits are connected to interpretable biological mechanisms.

**(3) Population genetics: are there genes previously reported as having undergone domestication?**

Yes. A key goal of our population-genetic analyses was to identify genome-wide signatures of selection across three stages (Ss natural selection, So domestication, and hybrid improvement) and connect them to interpretable trait biology. In the revised text, we emphasize that several genes identified in our selective sweep analyses are involved in biological pathways widely recognized as targets of domestication and crop improvement in grasses, including plant architecture, flowering time, stress adaptation, and sugar transport/metabolism. Importantly, our candidate genes include orthologs of well-established domestication and adaptation targets reported in other crops, such as *TB1*, a classic domestication gene underlying reduced branching/tillering in maize[4] and sugarcane[5], *FT* (flowering locus T)[6,7] and BX8[7], which involved in benzoxazinoid biosynthesis. These well-documented examples provide independent support that our population-genetic scans capture biologically meaningful targets of selection rather than technical artifacts.

For subgenome allelic counts, 1.96 / 5 ≈ 5.03 / 12, so allelic counts appear primarily driven by gene #. On the other hand, the Ss subgenome has an average of ~2 copies, while the So subgenome has an average of ~11 copies. The number of alleles per gene copy is 1.96/2 = 0.98 for Ss and 5.03/11 = 0.46 for So, suggesting higher per-copy allelic diversity in Ss, not So (Line 264 – 266). This observation contradicts the higher average sequence identity in So (~95.51%) than in Ss (~95.00%) (Lines 266 – 273).

Response: We thank the reviewer for noticing this error. We incorrectly use of the term "allelic diversity" in the original text when dividing the average allele by either 2 or 11. We deleted relevant text.

Table 1 and Suppl. Table 1 shows the N50 of ONT UL reads as 155.5 kb, but Suppl Figure 1b does not seem to have such a read distribution. Also, the longest reads from

ONT sequencing are usually at the megabase level, even for regular ONT sequencing, but in the Suppl. Table 1, the longest ONT UL read shows 812 kb. Please double-check your data and statistics.

Response: We sincerely thank the reviewer for their meticulous observation and valuable reminder. After a thorough re-examination of our raw data and statistical records, we identified an oversight that led to inconsistencies among the tables, figures, and text.

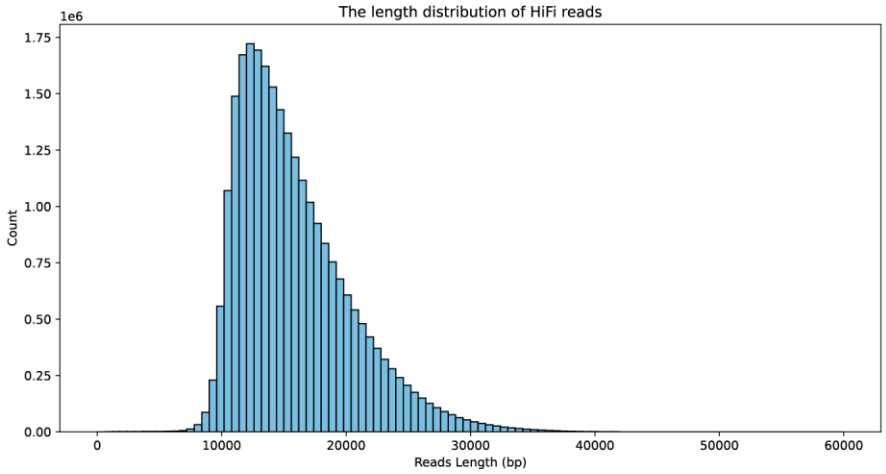
Specifically, two separate batches of ONT ultra-long sequencing were performed for the POJ2878 genome assembly.

1. The first batch yielded suboptimal read quality. For example，the maximum read length was only 812.15 Kb and did not reach the megabase scale, resulting an unsatisfactory genome assembly.
2. We therefore performed a second sequencing batch, which produced substantially improved data quality, with a maximum read length of 1.23 Mb. This second dataset was used for the final assembly reported in the manuscript.

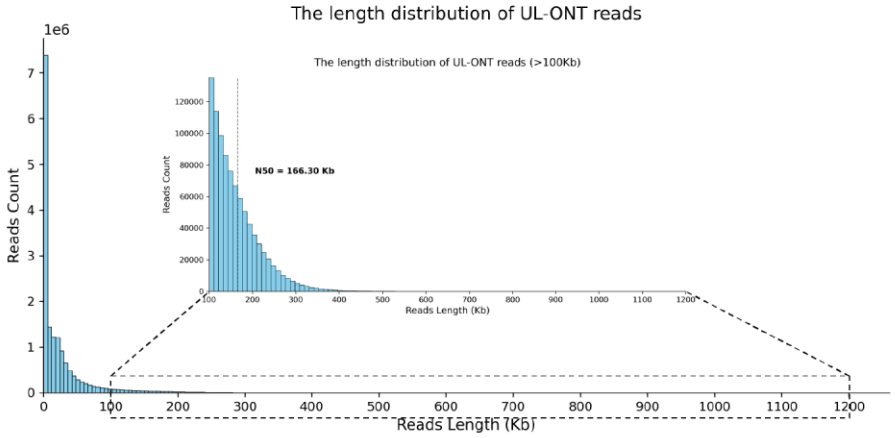
Regrettably, we inadvertently failed to update the statistical tables to reflect the second batch of data. We have now corrected these tables to present the statistics of the second batch, including the megabase-level maximum read length (1.23 Mb). We have also revised the corresponding descriptions in the manuscript (**Lines 124-125, Page 6**), **Extended Data Table 1** and **Supplementary Table 1** accordingly.

It is worth noting that only ONT ultra-long reads longer than 100kb were used for the genome assembly. However, the previous Supplementary Figure 1b showed the length distribution of all raw reads, which did not accurately reflect the reported N50 value. To address this issue, we have redrawn this figure (now **Supplementary Figure 2b**) to display the distribution of the filtered reads (>100 kb) and have clearly marked the N50 position. This revision ensures consistency with the data presented in **Extended Data Table 1** and **Supplementary Table 1**.

a




b




**Supplementary Figure 2. Length distribution of sequencing reads.** (a) Length distribution of PacBio HiFi reads. (b) Length distribution of ONT ultra-long reads. The distribution of ONT ultra-long reads exceeding 100 kb (used for assembly) is specifically magnified, with the N50 position explicitly indicated.

Figure 3a is still confusing and not straightforward, and the main text citing this result (L367 - 372) carries more information than the figure itself. The links between cultivars are almost all covered by nodes. The log2 transformation inappropriately enlarges smaller nodes.

Response: We realized that the original Figure 3a may have caused confusion for the reviewer and readers. To clarify and better support our statement—"IBD analysis revealed that 95.3% of cultivars shared over 821.6 Mb of IBD sequences with POJ2878, underscoring its extensive use in global sugarcane breeding programs"—we have replaced it with an empirical cumulative distribution plot. This revised figure provides a clearer and more intuitive representation of the data, thereby improving the accuracy of our presentation.

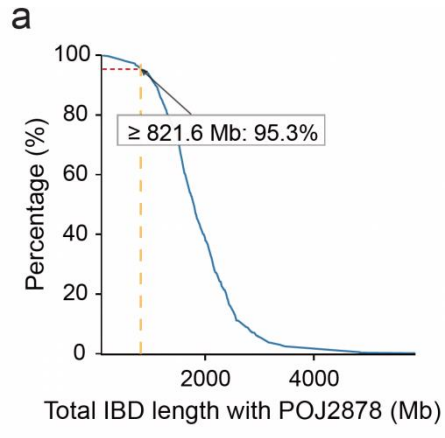



**Figure 3a. Empirical cumulative distribution of total identity-by-descent (IBD) length shared between 573 sugarcane cultivars and the founder cultivar POJ2878.** The x-axis shows total IBD length per cultivar (Mb), and the y-axis shows the percentage of examined cultivars. The dashed line marks 821.6 Mb, indicating that 95.3% of cultivars share ≥821.6 Mb of IBD sequence with POJ2878.

The selection analyses are not performed reasonably. For the genome assembly size of 10.37Gb, they identified 11.43 million SNPs, corresponding to an average of 907 bp per SNP. They used 10kb windows with 2kb sliding steps to calculate XP-CLR and Fst, effectively using 11 SNPs per window and 2 SNPs per step. Such a shallow number of data points in each calculation will result in extremely high false positives, despite these methods may be developed to account for random drifting. Furthermore, the use of 5% and 10% cutoff values to call selective sweeps are very loose criteria, which basically calls the whole genome undergone selective sweeps, as shown in their Figure 4a-d, where the majority of chromosomes are beyond the critical value(s).

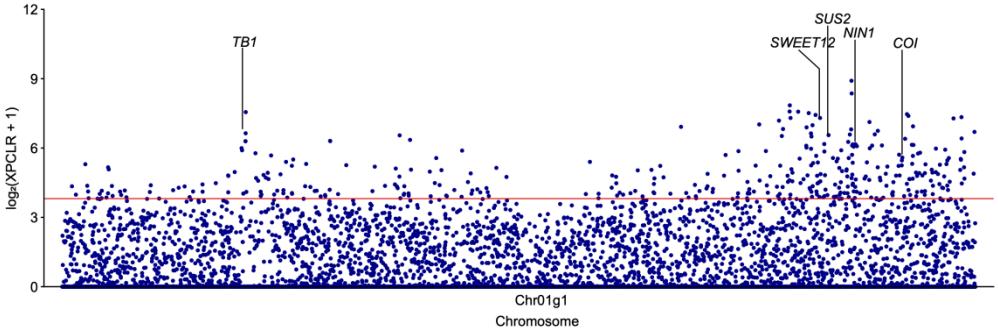
Response: We thank the reviewer for raising this important concern regarding marker density, window design, and the stringency of sweep calling.

Following the reviewer's suggestion, we re-ran the selection analyses using more stringent and statistically robust parameters. Specifically, we (i) applied an empirical top 5% cutoff, a commonly used threshold in genome-wide selection scans in plants, to define outlier windows (replacing the previous top 10%), and used 50-kb windows with 10-kb steps for window-based scans; and (ii) required multi-method support for each reported sweep window to minimize stochastic outliers. For natural selection in *S. spontaneum*, a window was retained only if it is supported by at least two of the three metrics (SeeD, RAiSD, and nucleotide diversity). For early domestication in *S. officinarum* and crop improvement in hybrids, a window was retained only if it is supported by both XP-CLR and $F_{ST}$ (i.e., outlier in both statistics under the same top 5% criterion). Under these updated settings, each window contains on average of 65.14 SNPs, substantially increasing marker density and the reliability of window-based statistics.

With these revised parameters, the overall evolutionary interpretation remains consistent with our previous conclusions (**Fig. 4j**). However, the more stringent

framework refines the candidate set and highlights additional biologically plausible genes within high-confidence signals, including candidates involved in carbohydrate/sucrose metabolism and transport (e.g., *ShFPB*, *ShINV*), flowering regulation (*RFT1*, *TFL4* and *AGLs*), and shoot and root branching (*TB1*, *NRT1.1*). The updated results are summarized in the main text (**Lines 335-383, Pages 13-15**) and **Figure 4**.

Regarding the reviewer's comment that "the majority of chromosomes are beyond the critical value(s)" in the original Fig. 4a-d, we note that chromosome-wide tracks in the overview panels were visually compressed, which may have made outlier windows appear densely clustered. To improve interpretability, we now provide an enlarged view of a representative chromosome (**Response Figure 6**), showing that the retained sweep windows are distributed in discrete intervals rather than spanning entire chromosomes.




**Response Figure 6. Distribution of selective sweeps across Chr01g1.**

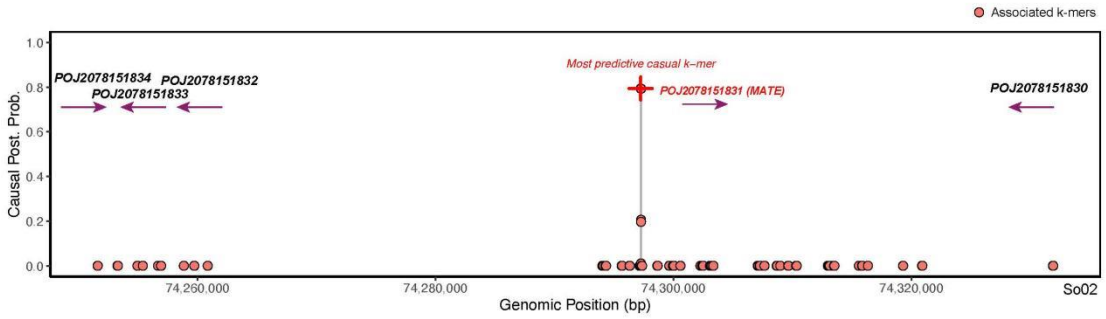
The revision provides more detail on how to use KMERIA identify candidate genomic intervals for gene discovery. They mentioned using LD to adjust the genomic interval. However, details regarding fine mapping are still missing. How were the reported genes fine-mapped from the at-least 100kb region? Without further details, these results may not be reproducible.

Response: We thank the reviewer for the insightful follow-up comment. We acknowledge that our previous response did not provide a sufficiently detailed description of the fine-mapping procedures applied to the significant loci identified by KMERIA. Further refinement of these intervals is essential for pinpointing the most likely causal genes. To address this issue, we have incorporated a comprehensive description of the fine-mapping strategy and the specific criteria in the revised **Methods** section (**Lines 1,402-1,419, Pages 51**):

"To distinguish the most likely causal genes from the initial candidates, we integrated statistical causal inference with functional evidence. Fine-mapping was performed within the ±50 kb genomic windows centered on the significantly associated k-mers. We employed the CAVIAR[8] to quantify the Causal Posterior Probability (CPP) for each associated k-mer within the interval by integrating GWAS *Z*-scores with a local LD correlation matrix calculated using PLINK2 with the

'*--r-unphased square*' option[9]. CAVIAR additionally identifies a minimal causal set, representing the smallest group of k-mers with a collective confidence level of ≥95% in capturing all true causal variants. Within this set, the top five k-mers with the highest CPP were designated as the primary statistical anchors for gene prioritization. To ensure the biological relevance of the statistically identified candidates, genes within the fine-mapped intervals were prioritized based on the following criteria: (i) Priority was given to genes that either overlapped with the k-mer exhibiting the highest CPP or were located within its immediate flanking regions. (ii) Candidates were referenced for functional orthologs with genes previously reported to be involved in biosynthesis or regulatory pathways in related species (e.g., *Sorghum bicolor*, *Oryza sativa*, and *Zea mays*). This integrated approach facilitated the transition from broad association intervals to high-confidence functional candidates."

Furthermore, we now provide fine-mapping results for the genes highlighted in the Manhattan plots, summarized in **Supplementary Table 13**. We have also included a visualization of the fine-mapping results for a representative locus in **Response Figure 7**. By calculating the CPP for all k-mers within a ±50 kb window surrounding the lead association signal, we prioritized the most likely causal gene from five initial candidates. Subsequent functional characterization, based on amino acid sequence homology alignment with related species, identified this prioritized candidate as a MATE (multi-drug and toxic compound extrusion) transporter gene.




**Response Figure 7. Statistical fine-mapping of GWAS candidate genes within the genomic intervals**. The plot illustrates the causal inference analysis for a representative 100 kb interval on chromosome So02. The X-axis represents the genomic position (bp), and the Y-axis denotes the Causal Posterior Probability (CPP) calculated using the CAVIAR. Colored circles represent k-mers within the ± 50 kb region of the lead association. The Y-axis reflect their statistical likelihood of being causal variants. Red cross indicates the most predictive causal k-mer, which possesses the highest CPP in the local region. Purple arrows represent the gene models annotated within this interval. Red gene ID highlights the prioritized functional candidate identified through the convergence of high CPP and functional orthologs with related species such as Sorghum, Rice and Maize.

The discussion still focuses heavily on the new methods and reads as though these are products of this paper.

Response: We appreciate the reviewer's valuable feedback and have revised the manuscript to address this concern.

We have removed the section "An integrated genomic framework for polyploid crops" from the main text and moved the technical details regarding the framework's construction to **Supplementary Note 4**. The revised Discussion now emphasizes: 1) The genetic innovations underlying sugarcane domestication and improvement, including the role of POJ2878 and key haplotypes, in shaping modern sugarcane cultivars; 2) The implications of our findings for breeding strategies, particularly the importance of balancing subgenome contributions and optimizing haplotype dosages in polyploid crops; 3) The link between cellular anatomy (e.g., parenchyma cell traits) and agronomic performance, highlighting actionable targets for enhancing sugar yield; and 4) Broader insights for genomics-assisted breeding in sugarcane and other polyploid crops, with an emphasis on the practical deployment of genomic tools and the regional tailoring of breeding strategies.

Please see our revised Discussion for details (**Lines 431-491, Pages 17-19**).

Minor comments:

Figures 1b,d,e are essentially the same. Some of them can become supplementary to reduce the size of this figure. Currently, the fonts are very small and impossible to read without zooming in.

Response: To improve clarity, we have moved Figures 1b and 1d to **Extended Data Figure 1** and enlarged the fonts for better visibility.

There is a typo on line 48 – "profling with by our newly"

Response: Thanks! The typo has been corrected.

Figure 4j, species names are not labeled.

Response: The species names have been labeled in the revised Figure 4j.

Referee #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Response: We sincerely thank the reviewer for your time and effort in reviewing our manuscript.

References

1. Li, N. *et al.* BvBZR1 improves parenchyma cell development and sucrose accumulation in sugar beet (Beta vulgaris L.) taproot. *Front. Plant Sci.* **16**, 1495161 (2025).

2. Wang, K. *et al.* Overexpression of Tonoplast Transporter FvMATE51

Simultaneously Increases Fruit Size and Sugar Accumulation in Strawberry. *Plant Biotechnol. J.* pbi.70478 (2025) doi:10.1111/pbi.70478.

3. Moore, P. H. & Cosgrove, D. J. Developmental Changes in Cell and Tissue Water Relations Parameters in Storage Parenchyma of Sugarcane. *Plant Physiol.* **96**, 794–801 (1991).
4. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**, 1160–1163 (2011).
5. Pribil, M. *et al.* ALTERING SUGARCANE SHOOT ARCHITECTURE THROUGH GENETIC ENGINEERING: PROSPECTS FOR INCREASING CANE AND SUGAR YIELD. *Proc Aust Soc Sugar Cane Technol* **29**, (2007).
6. Wang, S. *et al. FLOWERING LOCUS T* Improves Cucumber Adaptation to Higher Latitudes. *Plant Physiol.* **182**, 908–918 (2020).
7. Wang, X. *et al.* Genome-wide Analysis of Transcriptional Variability in a Large Maize-Teosinte Population. *Mol. Plant* **11**, 443–459 (2018).
8. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying Causal Variants at Loci with Multiple Signals of Association.
9. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015-0047–8 (2015).

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have satisfactorily addressed our outstanding concerns. Although we doubt that the approach followed by the authors is generally applicable to other polyploid systems, they provide sufficient empirical evidence that the detected SNP loci can be used to study population structure in Saccharum. We therefore recommend acceptance.
Response: We sincerely thank you for your positive recommendation and for your constructive feedback throughout the review process. The extreme genomic complexity of *Saccharum* presents unique challenges, and we acknowledge that applying this approach to other polyploid systems with different evolutionary histories would require further validation and optimization. Your rigorous review has significantly improved the quality of our manuscript.

Referee #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.
Response: We sincerely appreciate the time, effort, and valuable insights that both you and your co-reviewer have dedicated to evaluating our manuscript.

Referee #4 (Remarks to the Author):

We appreciate the authors continued improvements. Their manuscript represents a significant contribution to sugarcane genomics, and they have thoroughly addressed our comments.
Response: We sincerely appreciate the time, effort, and valuable insights that both you and your co-reviewer have dedicated to evaluating our manuscript.

Minor:
ASHEG was not defined (L228).
Response: "ASHEG" stands for "allele-specific highly expressed genes."
We have now provided the full definition where the acronym first appears in the revised manuscript (**Lines 243, Page 10**).

For non-sugarcane researchers, the manuscript should state more clearly that Brix is a measure of sugar content.
Response: We have now explicitly defined Brix in the revised text (**Line 419, Page 16**) as follows:

"Significant correlations were observed between these cell traits and Brix (a measure of sugar content; $P<0.05$; Supplementary Figure 22), supporting the potential breeding value of parenchyma cells."

Referee #5 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

Response: We sincerely appreciate the time, effort, and valuable insights that both you and your co-reviewer have dedicated to evaluating our manuscript.