nature portfolio

Peer Review File

A sorghum pangenome reference improves global crop trait discovery

Corresponding Author: Dr John Lovell

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.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 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
This is an impressive and comprehensive study on sorghum genomics. The authors present a new reference assembly (BTx623 V5), a 33-member pangenome, and whole genome resequencing of nearly 2000 genotypes. Together, these resources constitute a long-term asset for the sorghum community. The manuscript is generally well written, the data are of high quality, and the analyses are state of the art. The highlights include the discovery of major structural variants at the domestication locus SH1 and the use of diagnostic kmers to bridge between pangenome assemblies and diversity panels. Overall, this is a highly valuable contribution to crop genomics. The authors should, however, improve coherence between sections, provide missing methodological details, clarify the interpretation of certain analyses, and address the above points to strengthen the manuscript.

Major strengths
• Careful selection of pangenome members and extensive phenotypic characterization.
• Thorough discussion of pangenome graphs and alternative approaches.
• Clear presentation of SH1 haplotypes and their geographic distribution.
• Use of advanced landscape genomics approaches to study migration and adaptation.

Weaknesses
1. Coherence of the manuscript: The section on climate adaptation and migration reads like a standalone mini-paper and is only loosely connected to the pangenome theme. The dhurrin section is more closely related, but the link to the pangenome could be emphasized more clearly.
2. Methodological details: The new BTx623 assembly is compared to version 3, but not to version 4. The rationale should be explained. The RNA-seq experiments are insufficiently described; Supplementary Data 6 lists only accessions and read counts. Growth conditions and tissue sampling must be reported for reproducibility. The use of different assemblers for different genomes also needs justification.
3. Analyses and interpretation: The comparison with Arabidopsis thaliana is not well motivated. A cereal crop would have been a more appropriate reference point. The ADMIXTURE analysis does not report the optimal K. The interpretation of the landscape genomics results should more clearly acknowledge that subpopulation structure and farmer choices, in addition to natural selection, shape patterns of differentiation. Candidate genes for drought adaptation should be presented more prominently, but with the caveat that their functional role remains hypothetical. Extending the kmer-based genotyping beyond SH1 and the dhurrin BGC would strengthen the case and illustrate the broader utility of this method.Additional analyses, such as GO term enrichment for low-drought regions and visualization of extended haplotypes, would add depth.

Specific comments
• It would be useful to clarify how many loci were queried with the kmer-based approach.
• For the dhurrin analysis, a kmer-based GWAS could provide a stronger demonstration of the usefulness of the pangenome resources.
• The overlap of drought boundaries with subpopulation boundaries should be highlighted.
• For the migration surfaces, explain why separate E-W and N-S analyses were performed, or provide a combined result.

Minor points
• Supplementary Fig. 1A: clarify why telomeric sequences appear in mid-chromosome.
• Supplementary Fig. 1B: clarify meaning of black density bars.
• Use “striga” or “witchweed” consistently.
• Lines 58 and 96: homogenous -> homogeneous
• Line 70: the dhurrin BGC is detected in GWAS for dhurrin content, not in climate-gene associations.
• Lines 129-132: examples cited are not clearly visible in Suppl. Data 1-2.
• Suppl. Note 1 appears to be missing.
• Lines 163-167: cross-referencing between Suppl. Data 4 and Suppl. Table 1 could be improved.
• Line 420: justify use of different assemblers.
• Line 434: clarify what “Illumina correction” entails.
• Line 440: threshold of 90% identity for gene retention seems low.
• Line 497: state explicitly that pangenome-based genotyping was limited to SH1 and dhurrin.
• Line 505: incomplete sentence.
• Line 570: results for optimal K missing.
• Line 596: GO analysis results not referenced in main text.
• Line 607: clarify whether Cycles model can capture management practices such as fertilization of landraces.
• Supplementary Fig. 8: caption should read “resequenced short variant” rather than “resequencing short variant”.

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)
Summary: This is a significant amount of work which involved improving sorghum BTX623 V3 reference genome to version 5, generating a sorghum pangenome using a diverse and representative set of 33 genotypes, characterizing two significant adaptation traits - SH1-1 and dhurrin.
The data quality and the figures used for illustration are of extremely high quality.

Here are a few things that could be improved in the paper:
1. Given the availability of other sorghum pangenomes, it would have been great to indicate what gaps the new pangenome will be filling. It will also be good to explain why the current datasets were not rather used to improve previously generated pangenomes. This is especially because the analysis done could have also been done successfully with the existing pangenomes
2. The title may need to be changed to focus on adaptation, which is the main subject of the manuscript. It is difficult to see how the work done in the manuscript has anything to do with future resilience. Unless some modeling was done to capture future scenarios, it might be better to avoid making statements on the future.
3. There needs to be a clear objective of the work that runs through the whole manuscript to give a clear message on the value that the paper is adding. This is lacking.
4. Sub-titles may need to be revised. For example, Under the topic, "Integrating single-reference and pangenome approaches to better understand multiple domestication events", the section simply discusses the 3 different alleles/haplotypes and not necessarily confirming whether these were independent domestication events or not.
5. The title "exploring the balance between climate adaptation and migration in sorghum germplasm" also specifically explored drought, so the title can be changed to capture drought.
5. What was the purpose of high throughout phenotypic characterisation of the 33 genotypes since the number was not optimal to make conclusions on haplotype or trait variations?
6. Under future targets, only dhurrin biosynthesis is explored, perhaps as a proof of concept. But there is room to explore several other obvious targets using advanced machine learning tools that would add more value to the manuscript than was done with the current analysis.
7. There is absolutely no link between what appears in the conclusion and the results obtained. The conclusion talks about consumer preferences, something that was not part of the study. The authors also talk about introducing beneficial alleles without really providing a strategy for doing so in the manuscript.

Referee #4

(Remarks to the Author)
Morris and co-authors developed a pangenome for sorghum from a diverse core of 33 accessions. This advancement represents a significant achievement for plant sciences, which will have broad applications for sorghum improvement for stakeholders ranging from small-holder farmers to large multinational companies. Overall, the figures, text and tables superbly summarized the volumes of data contained within this presentation. The text is well-written. In summary this manuscript artfully compiles years and volumes of data in making a monumental contribution to the sorghum field.
Specific areas
Lines 75-87. Although this topic is a very important one societally, this paragraph appears out of place here, because it does not directly relate to the development of this sorghum pangenome. Consider moving it to the discussion.
Lines 258-272. For readers who may be unfamiliar with these analyses, please briefly explain this concept “random mating

at geographic scales". Clearly, plants such as Arabidopsis or sorghum are self-pollinating at the plant scale and pollen flow from either species does not occur at any measurable levels beyond 1 km.
Lines 297-298. More appropriate references demonstrating the importance of lignin deposition for drought tolerance should be cited. 62. Jeon et al 2023 Plant deals with salt tolerance and does not mention lignin deposition or monolignol biosynthesis in the text or figure. 63. Li et al. 2019 deals with overexpression of a sorghum Erecta (receptor kinase) homolog in maize, which confer drought tolerance and lignin concentrations may be increased, but statistical analyses were not performed.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have thoroughly revised their manuscript, and their response letter is convincing. My co-reviewer and I recommend acceptance.

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)
Summary: The generation of the pangenome is certainly a great development for sorghum, and the added attempt in this revision to integrate previous versions is commendable.
However, this manuscript still lacks originality. A simple reporting of structural variations without a validation of the phenotypes leaves most of the results speculative and hypothetical at best. Here are just a few further concerns:
1). it is not clear from the write-up, the significant improvements to the current BTX623 V3 other than clarity in the positions of genes. Do you need a pangenome to get this clarity? Usually, this can be achieved by generating a better assembly of the same genome. The objective of a pangenome generation should go well beyond clarity of positions of genes.
2). Other than providing a baseline for relating selection, what was the purpose of adding the 33 diverse sets in the sequencing panel? Did the addition of SRN39 for example, add to the discovery of more resilience genes or more Striga resistance genes? The authors say it stood out for strong performance under drought, but how did that feature add value to the pangenome information?

Dear Dr. ,

Thank you for a thorough and thoughtful review of our manuscript now titled “A sorghum pangenome reference improves global crop trait discovery”. The reviewers brought up a number of good points, many of which we were able to address with improvements to analysis, visualization, and text changes. Several of the comments asked for clarification, which when we probed deeper, we realized we could answer with additional data that we had on hand but had opted previously to not include. To this end, we have included several additional field and high-throughput phenotyping datasets, a more detailed analysis of climate-SNP associations, and an analysis of botanical type-subpopulation assignment-SHATTERING1 haplotype associations. These new datasets required the inclusion of three new authors, several supplementary files, and methods sections. Combined, we believe that these additions expand the impact and scope of this article without substantially expanding the text length or distracting from the main message.

Below, we have copied full comments from the three reviewers and editorial notes. We provide a response for all comments and line numbers (if appropriate) where associated changes have been made. Please don’t hesitate to contact us if any questions or concerns arise.

Geoff, Nadia, and John

Referee #1 (Remarks to the Author):

This is an impressive and comprehensive study on sorghum genomics. The authors present a new reference assembly (BTx623 V5), a 33-member pangenome, and whole genome resequencing of nearly 2000 genotypes. Together, these resources constitute a long-term asset for the sorghum community. The manuscript is generally well written, the data are of high quality, and the analyses are state of the art. The highlights include the discovery of major structural variants at the domestication locus SH1 and the use of diagnostic kmers to bridge between pangenome assemblies and diversity panels. Overall, this is a highly valuable contribution to crop genomics. The authors should, however, improve coherence between sections, provide missing methodological details, clarify the interpretation of certain analyses, and address the above points to strengthen the manuscript.

Major strengths: • Careful selection of pangenome members and extensive phenotypic characterization. • Thorough discussion of pangenome graphs and alternative approaches. • Clear presentation of SH1 haplotypes and their geographic distribution. • Use of advanced landscape genomics approaches to study migration and adaptation.

[1.1] Coherence of the manuscript: The section on climate adaptation and migration reads like a standalone mini-paper and is only loosely connected to the pangenome theme. The dhurrin section is more closely related, but the link to the pangenome could be emphasized more clearly.

> **Thank you for bringing this to our attention. We agree with the reviewer’s perspective (one also shared with reviewer 3 [3.3] and reviewer 4 [4.1]). Taken together, these comments note that there are many places where different analyses and approaches presented could be better integrated with a more obvious analysis and biological question that connects all sections. We have taken these suggestions seriously and have now updated all sections of the paper to be better tied to the common line of evidence. For brevity, we will not detail all places where the text was better integrated, but below we highlight a few important ones.**
>
> **Specifically, we now better integrate the climate section in several ways: (1) we preface the need for climate-aware breeding with more nuance in the introduction [lines 77-79] and specify why a pangenome helps to bridge local and global improvement goals [lines 91-96]; (2) we now present the phenotypic variation in our panel with far more nuance (see note to other reviewers [comments 3.2, 3.6, 4.3]) and connect these traits to climates and sorghum diversity [lines 138-149, 267-270 , and elsewhere]; (3) we discuss SHATTERING1 in the context of both climate and spatial structure and have conducted several new analyses to directly link SH1 with previous results of structure and upcoming detailed climate analyses (results are presented in the main text [lines 200-207], and new figures to this effect are included in extended data figure 3); and (4) we have re-written the lead in to**

the climate section [lines 213-220] to both connect to population stratification discussed above and the landscape processes in this section.

We have also worked to integrate the dhurrin vignette with the larger narrative, including modifying the text in the lead into the dhurrin results to reference the new results about the phenotypes in our reference and diversity panel [lines 275-279]. We also more directly connect pangenome-based trait discovery to breeding efforts in the completely restructured conclusion [lines 325-329].

[1.2] Methodological details: The new BTx623 assembly is compared to version 3, but not to version 4. The rationale should be explained.

BTx623 v4 was never publicly released. It is fairly common to have public releases skip numbers when the internal releases are vetted, then improved upon (e.g. human genome hg8 preceded hg10). We now state this in the main text [line 106] and provide the full V4 release notes as part of a new Supplementary Note 1.

[1.3] The RNA-seq experiments are insufficiently described; Supplementary Data 6 lists only accessions and read counts. Growth conditions and tissue sampling must be reported for reproducibility.

We have corrected this omission and now include a full methods section related to plant growth and tissue collection [lines 509-531]

[1.4] The use of different assemblers for different genomes also needs justification.

We have added a statement that sequencing methods and coverage necessitated different assembly methods [lines 532-533], and now include Supplementary Note 2, which has detailed assembly methods and justification for all 33 genomes.

[1.5] Analyses and interpretation: The comparison with Arabidopsis thaliana is not well motivated. A cereal crop would have been a more appropriate reference point.

Agreed - we have added a cereal (pearl millet) and now treat Arabidopsis as the baseline for a highly differentiated inbred locally adapted species and pearl millet as a panmictic outbreeding species with an overlapping range to sorghum [lines 227-233].

[1.4] The ADMIXTURE analysis does not report the optimal K.

Like with many population genetic analyses, the optimal *k* in our dataset is not obvious. Previously, we had opted for *k* = 8 to enhance backwards compatibility to previous studies. Partially motivated by this comment and the need to re-run the analyses after excluding four duplicate samples (*n* of unique genotypes in the resequencing diversity set now is 1,984), we now accomplish a full exploration of *k* = 2-13 (Supplementary Note 5). There is evidence for both low and high k values, and we opt for *k* = 10 both because of empirical evidence and to allow for better integration with previous studies that have *k* = 8-10 [lines 215, 697-702].

[1.5] The interpretation of the landscape genomics results should more clearly acknowledge that subpopulation structure and farmer choices, in addition to natural selection, shape patterns of differentiation.

It was our intention to have this line of evidence as one of the primary drivers of variation. We have now made a conscious effort to connect farmer/breeder choices/preferences to the background of sorghum diversity [lines 81-82, 87-91, 121-122, 208-210, 216-218], and mention it as a non-climatic source of variation including as it relates to our new analysis of botanical type variation [line 210].

[1.6] Candidate genes for drought adaptation should be presented more prominently, but with the caveat that their functional role remains hypothetical.

The total amount of text we can spend on candidate genes was a point of discussion during revisions and original manuscript preparation. In the end, we opted to keep the actual candidate gene list as

part of the supplementary information (new supplementary data 8), but now report this full list and reference functional annotations of particular candidates therein [lines 264-267].

[1.7] Extending the kmer-based genotyping beyond SH1 and the dhurrin BGC would strengthen the case and illustrate the broader utility of this method.

We now mention that the method has been used to great effect in other studies in the conclusions [lines 325-329]. Specifically, We have cited two additional works, VanGessel et al. 2025 and Maina et al. 2025 (in review), where the kmer-based genotyping was also applied. In both cases, the method was used to genotype loci that harbor structural variation and are of interest for sorghum breeding for either aphid resistance (RMES1, VanGessel) or *Striga hermonthica* resistance (LGS1, Maina). The differences in the structure of both of these loci, as characterized by the kmer-based genotyping, have been linked to functional, breeding-relevant variation and also were used to accelerate breeder decisions and for marker development. We agree that extending the kmer-based genotyping to additional loci demonstrates the utility of the method, and plan to do so in future papers where we will have more space. In the end we do not explore other regions here. This is mostly because we'd prefer to dig deeper into a few loci and leave larger exploration for future work.

[1.8] Additional analyses, such as GO term enrichment for low-drought regions and visualization of extended haplotypes, would add depth.

We had performed GO term enrichments for low-drought regions, but the extended haplotype outliers contained too many genes for interpretable GO enrichments. We have added Extended Data Figure 4 with visualization of extended haplotypes for accessions assigned to high and low drought prevalence.

[1.9] It would be useful to clarify how many loci were queried with the kmer-based approach.

The number of loci typable by kmers in SH1 (n = 6,533) and the dhurrin BCG (n = 139,866) are now reported in the methods [lines 635-636].

[1.10] For the dhurrin analysis, a kmer-based GWAS could provide a stronger demonstration of the usefulness of the pangenome resources.

We agree that a kmer-based GWAS is a powerful tool for exploration of associations between sequence diversity and phenotypes in a pangenomic context. Our kmer-genotyping method is analogous to a kmer-GWAS, but specifically to aggregate across an a priori-defined region, like the LD block defined in the dhurrin analysis or the SH1 locus. Conceptually, our method is similar to that described in *The barley pan-genome reveals the hidden legacy of mutation breeding* (Jayakodi et al. 2020) where single copy kmers were extracted from structural variants within 20 assemblies and counted within illumina short read libraries. The resulting matrix was used to display the global pattern of structural variation (which generally matched SNPs) via PCA. Where our methodology differs is: rather than extracting kmers from global SVs across the genome and trying to scan for large-scale effects from a kmerGWAS scan, we instead leverage the local haplotype structure at pre-defined regions of interest (which simultaneously combines variation derived from SNPs, indels and large SVs) and are able to confidently assign individual libraries to the observed pangenome haplotype structure. This allows fine-grain comparisons that would otherwise be lost by whole genome scans. We have provided additional context in the methods section [lines 620-622] on the similarity between our method and a kmer-based GWAS.

[1.11] The overlap of drought boundaries with subpopulation boundaries should be highlighted.

We have added the text [lines 242-245] "Drought status was broadly different among admixture (*k* = 10) subpopulations ($\chi^2_{9, N = 431}$ = 82.664, *P* < 0.001), and we observed different patterns of allele sharing within drought classes. For example, accessions from high drought prevalence populations had lower allelic dissimilarity than those from less drought prone populations (Fig. 4D), suggesting that alleles were more likely to move between spatially disjunct populations experiencing drought."

to acknowledge the overlap of subpopulation boundaries and water availability gradients, relative to the "high" and "low" drought prevalence assignments.

[1.12] For the migration surfaces, explain why separate E-W and N-S analyses were performed, or provide a combined result.

Our presentation of the separate axes is to present estimations of gene flow relative to axes where crops are cultivated in relatively uniform (E-W) and contrasting (N-S) photoperiod regimes. We have clarified this in the text [lines 465-466, 234-235]

[1.13] Supplementary Fig. 1A: clarify why telomeric sequences appear in mid-chromosome.

In Sorghum, several chromosomes have true blocks of interstitial telomere repeats. This is a commonly observed phenomenon in many species; however, most telomere finding efforts focus on chromosome termini and do not seek out interstitial repeat blocks (and thus these go unnoticed). Since the point of this plot (now part of Supplementary Note 1) is to highlight assembly completeness and not to discuss genomic complexities, the presence of interstitial telomere repeats will likely confuse the readers. As such, we have removed them from the plot and now only flag telomere sequences within 10kb of the chromosome termini. The figure caption now reflects this change.

[1.14] Supplementary Fig. 1B: clarify meaning of black density bars. **Updated to note that black denotes un-annotated sequence (now Supplementary note 1).**

[1.15] Use "striga" or "witchweed" consistently. **Updated - only use 'striga' now after the initial definition**

[1.16] Lines 58 and 96: homogenous -> homogeneous **Updated here and elsewhere**

[1.17] Line 70: the dhurrin BGC is detected in GWAS for dhurrin content, not in climate-gene associations. **Text removed**

[1.18] Lines 129-132: examples cited are not clearly visible in Suppl. Data 1-2. **We have moved this citation to the supplement, so that it now references all genes and positions of all the variants.**

[1.19] Suppl. Note 1 appears to be missing. **This was an upload error. It now is Supplementary Note 3.**

[1.20] Lines 163-167: cross-referencing between Suppl. Data 4 and Suppl. Table 1 could be improved. **Our analysis of phenotype data is now completely restructured.**

[1.21] Line 420: justify use of different assemblers. **See [comment 1.4]**

[1.22] Line 434: clarify what "Illumina correction" entails. **Updated**

[1.23] Line 440: threshold of 90% identity for gene retention seems low. **This is an unnecessary detail that we have dropped (since it was only for internal QC of the assemblies). 90% is used in the case of high divergence between assemblies.**

[1.23] Line 497: state explicitly that pangenome-based genotyping was limited to SH1 and dhurrin. **Updated**

[1.24] Line 505: incomplete sentence. **Updated**

[1.25] Line 570: results for optimal K missing. **Updated**

[1.26] Line 596: GO analysis results not referenced in main text. **We have added additional text to clarify that the "wavelet outliers" [lines 261-264] were used in the GO analysis.**

[1.27] Line 607: clarify whether Cycles model can capture management practices such as fertilization of landraces. **Updated**

[1.28] Supplementary Fig. 8: caption should read “resequenced short variant” rather than “resequencing short variant”. **The supplementary material is now completely revised and this error has been corrected.**

Referee #3 (Remarks to the Author)

Summary: This is a significant amount of work which involved improving sorghum BTX623 V3 reference genome to version 5, generating a sorghum pangenome using a diverse and representative set of 33 genotypes, characterizing two significant adaptation traits - SH1-1 and dhurrin. The data quality and the figures used for illustration are of extremely high quality. Here are a few things that could be improved in the paper:

[3.1] Given the availability of other sorghum pangenomes, it would have been great to indicate what gaps the new pangenome will be filling. It will also be good to explain why the current datasets were not rather used to improve previously generated pangenomes. This is especially because the analysis done could have also been done successfully with the existing pangenomes

> **We worked hard to vet existing sorghum resources and spent considerable time working to integrate sorghum genomes from other sources into our pangenome. In the end, we were able to bring in Wray, BTx642 and RTx430 into our diversity pangenome, but were not able to bring others in without degrading the quality of our resource. We have no intention of defacing the other resources in the manuscript, and it is not possible to describe why we did not incorporate existing resources without bringing up data quality issues. Very quickly, the publicly reported 2021 genomes are far too fragmented to include here; the 10 genomes reported in 2023 (Volker) were scaffolded against BTx623 V3 and likely contain the assembly errors that made us need to build a new genome; at the time of the assembly of this pangenome, none of the assemblies on sorghumbase had high enough per-base quality to effectively annotate (and we did try). This is necessary since annotations hosted on 3rd party sites cannot be integrated due to methodologically derived PAV (see our review on this matter: https://www.biorxiv.org/content/10.1101/2025.08.14.670405v1.abstract).**

[3.2] The title may need to be changed to focus on adaptation, which is the main subject of the manuscript. It is difficult to see how the work done in the manuscript has anything to do with future resilience. Unless some modeling was done to capture future scenarios, it might be better to avoid making statements on the future.

> **We have revised the manuscript title to “A sorghum pangenome reference improves global crop trait discovery” to more accurately reflect the manuscript narrative.**

[3.3] There needs to be a clear objective of the work that runs through the whole manuscript to give a clear message on the value that the paper is adding. This is lacking.

> **We have addressed this and several other comments to clarify the objective of the paper and detail these in our responses to reviewer #1 [comment 1.1]. In short, we now hope that the thread of trait discovery via pangenomics is more clearly the objective of the paper. We have added statements to connect back to this theme in the title, abstract [lines 69-71], introduction [lines 89-94], results [lines 171-173, 180-182, 266-270] and conclusions [lines 328-336].**

[3.4] Sub-titles may need to be revised. For example, Under the topic, "Integrating single-reference and pangenome approaches to better understand multiple domestication events", the section simply discusses the 3 different alleles/haplotypes and not necessarily confirming whether these were independent domestication events or not.

> **We have revisited all of the sub-titles to reflect the contents of the section and to fit within guidelines of subheading title length.**

[3.5] The title "exploring the balance between climate adaptation and migration in sorghum germplasm" also specifically explored drought, so the title can be changed to capture drought.

See above comment. This subtitle has been revised to "Dhurrin biosynthesis pangenomic variants".

[3.5] What was the purpose of high throughout phenotypic characterisation of the 33 genotypes since the number was not optimal to make conclusions on haplotype or trait variations?

The phenotype data was not adequately incorporated into the narrative in previous drafts. We have now re-written this section [lines 138-149], incorporated additional context for how the phenotypic data aids in interpretation of diversity in the phangenome reference members and diversity panel [lines 117, 267-270, 277-279] and include the new Extended Data Figure 2 related to trait variation.

[3.6] Under future targets, only dhurrin biosynthesis is explored, perhaps as a proof of concept. But there is room to explore several other obvious targets using advanced machine learning tools that would add more value to the manuscript than was done with the current analysis.

We agree with the reviewer that integration of machine learning tools is a valuable exercise and could provide valuable insights; however, we believe that it is outside of the scope of the current paper. Here, we aim to introduce our sorghum pangenomic resource, contextualize the forces that have shaped the diversity present in the pangenome resource, and demonstrate the use of tools which connect diversity in reference assemblies to broader diversity panels. We see this suggestion as an interesting direction for a future paper.

[3.7] There is absolutely no link between what appears in the conclusion and the results obtained. The conclusion talks about consumer preferences, something that was not part of the study. The authors also talk about introducing beneficial alleles without really providing a strategy for doing so in the manuscript.

We hope that the previously tenuous link has been solidified in our completely rewritten conclusion [lines 321-338]. This text now provides two concrete examples where this approach to trait discovery has led to new discoveries and implementation in real world small holder breeding programs.

Referee #4 (Remarks to the Author):

Morris and co-authors developed a pangenome for sorghum from a diverse core of 33 accessions. This advancement represents a significant achievement for plant sciences, which will have broad applications for sorghum improvement for stakeholders ranging from small-holder farmers to large multinational companies. Overall, the figures, text and tables superbly summarized the volumes of data contained within this presentation. The text is well-written. In summary this manuscript artfully compiles years and volumes of data in making a monumental contribution to the sorghum field.

Specific areas

[4.1] Lines 75-87. Although this topic is a very important one societally, this paragraph appears out of place here, because it does not directly relate to the development of this sorghum pangenome. Consider moving it to the discussion.

We appreciate the reviewer's suggestion and agree that this information is better placed elsewhere. We have restructured the presentation of this text in both the introduction [lines 80-86] and conclusion [lines 321-338] sections.

[4.2] Lines 258-272. For readers who may be unfamiliar with these analyses, please briefly explain this concept "random mating at geographic scales". Clearly, plants such as Arabidopsis or sorghum are self-pollinating at the plant scale and pollen flow from either species does not occur at any measurable levels beyond 1 km.

We have revised our terminology to consistently use "panmixia" in place of "random mating"[lines 224-232]. We additionally have edited the text to describe this method [line 224] "'multilocus wavelet genetic dissimilarity', which models changes in allele frequencies and identifies spatial scales where

genetic similarity is higher or lower than expected under panmixia (random mating)" to clarify the null expectation of this methodology. We now refer to results from wavelets as deviations from expected allelic dissimilarity [lines 228-232 and elsewhere]. See our response to reviewer #1 [comment 1.3] for details about why we included A*rabidopsis thaliana* and how we have updated this analysis.

[4.3] Lines 297-298. More appropriate references demonstrating the importance of lignin deposition for drought tolerance should be cited. 62. Jeon et al 2023 Plant deals with salt tolerance and does not mention lignin deposition or monolignol biosynthesis in the text or figure. 63. Li et al. 2019 deals with overexpression of a sorghum Erecta (receptor kinase) homolog in maize, which confer drought tolerance and lignin concentrations may be increased, but statistical analyses were not performed.

We thank the reviewer for their attention to the references cited and have incorporated these suggestions. We have corrected these references and systematically reviewed all others to ensure accuracy and appropriateness.

Editorial comments

STATISTICS: When revising your manuscript, you should ensure that any statistical analysis used is sound and that it conforms to Nature's guidelines). A collection of articles explaining the basics of statistical analysis and advice on how to best present it can be found here.

Checked

REPRODUCIBILITY: All of the checklists provided with the current submission (Reporting summary, and Code and software checklist (if applicable)) should be updated to reflect the revisions made and submitted with the revised manuscript.

We have completed these where appropriate

LENGTH: In print, biological sciences papers do not normally exceed 8 pages on average; the final print length, however, is at the editor's discretion. The typical length of an 8-page article with 5 modest (quarter-page) display items is 4300 words. If a composite figure (with multiple panels) must occupy at least half a page in order for all the elements to be visible, the text length may need to be reduced accordingly to accommodate such figures. Essential but technical details can be moved into the Methods or Supplementary Information (see below).

Our total main text length is 4191 words. Two of our figures are larger than ¼ page, so we targeted <4,300 word limit. We can adjust as needed during copy editing. If further reductions are needed, we can move Fig. 2 to the extended data.

TITLE: Titles cannot exceed 75 characters (including spaces); they must not contain punctuation.

The title of the initial submission "Developing future resilience from signatures of adaptation across the sorghum pangenome" was at 87 characters. In response to the reviewers' comments and the length requirements, we have updated to "A sorghum pangenome reference improves global crop trait discovery" (66 characters).

SUMMARY PARAGRAPH: All Nature papers begin with a fully referenced paragraph, typically no longer than 200 words, aimed at readers in other disciplines. This paragraph starts with a 2- to 3-sentence, basic introduction to the field; continues with a 1-sentence statement of the main findings starting 'Here we show' or an equivalent phrase; and finally, concludes with 2 to 3 sentences putting the main findings into general context so it is clear how the results described in the paper have moved the field forward. A downloadable, annotated example is available here. In some cases it may be necessary to exceed this limit modestly in order to explain complex material for readers in other fields. The extra length, however, is for introduction and context, and not for additional technical information.

The previous summary paragraph conformed to these specifications except for length. It has now been shortened to 198 words.

MAIN TEXT: If further introductory material is necessary, the main text can begin with up to 500 words of introduction expanding on the background to the work (some overlap with the summary is acceptable), before proceeding to a concise, focused account of the findings, and ending with 1 or 2 short paragraphs of discussion. Sections are separated with subheadings (up to 40 characters including spaces) to aid navigation.

The introduction is at 431 words with as little overlap to the summary paragraph as possible. Previous headings all exceeded 40 characters. These have all been shortened. Discussion (labeled 'conclusions) is a single 256 word paragraph.

REFERENCES: As a guideline, most papers should include no more than 50 main text references; additional references can be cited in (and listed after) the Methods section, as detailed below.

We have reduced redundancy of the citations so that there are now 53 main text references. The other 64 can be found in the methods.

FIGURE LEGENDS: These should be listed sequentially after the main text references and not in the figure files; they should not exceed 300 words each.

Checked no legends are > 300 words.

Each legend should begin with a brief title for the whole figure and continue with a short description of each panel and the symbols used.

Checked and no additional information was required.

Each figure legend should contain, for each panel where relevant, the following information:

* the exact sample size (n) for each experimental group/condition, given as a number, not a range;

Checked and n is reported where necessary

* a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc);

Checked and no additional information was required.

* a statement of how many times the experiment shown was replicated;

Temporal replication has been added where appropriate

* definitions of statistical methods and measures:

This information was added where required

* very common tests (e.g., t-test, simple Chi-square tests, Wilcoxon and Mann-Whitney tests) can be identified by name only, but more complex techniques should be described in the Methods;

Checked and no additional information was required.

* whether tests are one-sided or two-sided;

Checked and no additional information was required.

* whether there are adjustments for multiple comparisons;

Checked and no additional information was required.

* the statistical test results (e.g., P values);

Checked and no additional information was required.

* the definition of 'center values' as median or average;

Checked and no additional information was required.

* the definition of error bars as s.d. or s.e.m.

Checked and no additional information was required.

Any descriptions too long for the figure legend should be included in the Methods section; see here for further explanation.

We have opted to include all necessary information in the figure legends, except where it would exceed 300 words. We can alter the legends as needed during copy editing.

METHODS: After the main text figure legends there should be a section entitled "Methods", which provides the full, step-by-step instructions that would allow other researchers to replicate the results. The Methods section will not appear in print but will appear online in the full-text HTML and PDF versions. The Methods section should be written as concisely as possible but should contain all elements necessary to allow interpretation and reproduction of the results. If there are additional references in the Methods section, their numbering should continue from the last reference in the main text, and they should be listed following the Methods section. Specialized methods that require chemical structures, figures or tables, or methods requiring equations, cannot be accommodated in the Methods section of the main text file. If such information is part of the Methods, the entire Methods section must instead be included within a Supplementary Information text file.

Checked and added more detail where necessary in the main text. We have now added supplementary notes where tables or extensive details were needed.

ETHICS STATEMENT: For research involving human research participants, the Methods section must include an ethics statement. This statement should provide the name of the committee that approved the study; confirm that the research was performed in accordance with all relevant guidelines and regulations; and confirm that informed consent was obtained from all participants. If the study was granted an exemption from requiring ethics approval, details of the committee granting the exemption must be included.

N/A

For research involving human embryos or gametes, or human stem cells in contexts requiring ethical oversight, the Methods section must include an ethics statement. This statement should provide the name of the committee that approved the study and confirm that the research was performed in accordance with all relevant guidelines and regulations. We encourage authors to follow the principles laid out in the 2021 ISSCR Guidelines for Stem Cell Research and Clinical Translation: https://www.isscr.org/guidelines. Where necessary, the ethics statement should also describe the conditions of donation of materials, such as human embryos or gametes, and confirm that informed consent was obtained from all donors of cells or tissues.

N/A

MAIN TEXT STATEMENTS: Several statements (which will not appear in print but will appear online in the full-text HTML and PDF) are required after the Methods, and before the Extended Data legends. First, there should be an Acknowledgements section, listing grant/financial support.

Updated accordingly

Next, we require a detailed Author Contribution statement; the specific contributions of each author, particularly in terms of which authors performed which specific experiments, must be listed.

Updated accordingly

This is followed by a Competing Interest statement. Financial or non-financial interests should be noted here, as well as any patents; patent information should include at a minimum what is covered by the patent and who submitted the patent application.

Updated accordingly

Finally, an Additional Information statement should include information regarding reprints and permissions and name the author(s) to whom correspondence and requests for materials should be addressed. For details of "end note" style and an example see here.

Updated accordingly

DATA AND CODE AVAILABILITY STATEMENTS: All original research manuscripts published in Nature Portfolio journals must include a Data availability statement (DAS). This statement must make the conditions of access to the “minimum dataset” that is necessary to interpret, verify and extend the research in the article, transparent to readers. This minimum dataset may be provided through deposition in public community/discipline-

specific repositories, custom proprietary repositories for certain types of datasets, or general repositories like Figshare, Zenodo and Dryad. Providing large datasets in supplementary information is strongly discouraged and the preferred approach is to make data available in repositories. More information on Nature Portfolio's reporting standards and preparing your Data availability statement can be found here.

We have updated the data availability section with new information. All raw reads have been deposited, but the genome assemblies and annotation submissions have been halted with the following note "Due to the absence of either a FY 2026 appropriation or Continuing Resolution for the Department of Health and Human Services, [responses will be delayed] until normal government operations resume." Nonetheless, all data that will eventually be made available there is also already available on Phytozome.

For all studies using custom code or mathematical algorithms that are deemed central to the conclusions, a Code availability statement (CAS) must be included, indicating whether and how the code or algorithm can be accessed, including any restrictions to access. the CAS should be provided as a separate section after the DAS but before the references. Code should be deposited in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cited in the reference list. We encourage you to manage subsequent code versions and to use a license approved by the open source initiative. Additional details can be found here.

All methods presented here used publicly available software, assimilated into pipelines. As such the descriptions of the pipelines provided in the methods and supplementary information should suffice. The one exception is the kmer-based genotyping which used novel unpublished code to accelerate speed of sequence matching — we do not intend to make this code publicly available. Since the basic operations of this method are generic (string search and counting) and our code only increases speed, we believe inclusion is not necessary. Please contact John Lovell if this response is not satisfactory.

Wherever possible the data used in the paper should be placed into a public data repository or presented as Supplementary Information. If data can only be shared on request, please explain why in the DAS and in the cover letter. The DAS must list which data are included (e.g. by figure panels and data types) and mention any restrictions on availability. If a dataset generated or analysed during the study is publicly available and has a unique DOI, please include it in the reference list and cite the dataset in the Methods.

All data used in this paper are available through the urls provided in the data availability statement or in the supplementary datasets.

Accession numbers for any newly determined sequences, structures, microarray or zoobank data; project IDs for MG-RAST data; accession numbers for X-ray crystallographic coordinates and structure factor files; or comparable NMR or cryoEM data, should be included only in the DAS.

N/A

DISPLAY ITEMS: We ask that you take stock of all the data that have been generated throughout the review process and ensure that only the data most central to the conclusions are presented in the main text figures. Although figures can be presented within the text file during the review process, they must be removed from the final text file and uploaded as separate individual files.

Updated accordingly

Figures should be comprehensible to readers in other or related disciplines and assist their understanding of the paper. Figures should be as small and simple as is compatible with clarity. Main text figures (but not Extended Data) must be provided in an editable format. Acceptable formats include .ai, .cmx, .cdr, .doc, .eps, .pdf, .ppt, .ps, .psd, .svg and .xls; unacceptable formats include .cvs, .gif, .jpg, .png and .tif. All panels of a figure should be logically connected; each panel of a multipart figure should be sized so that the whole figure can be reduced by the same amount and reproduced on the printed page at the smallest size at which essential details are visible.

Checked and all figures conform to these specifications. To reduce file size, gwas figure panels were converted to 600dpi rasters and placed into the otherwise vector graphics. We are happy to provide the originals, but the file sizes are each > 500Mb.

For guidance, Nature's standard figure sizes are 9, 12, or 18 cm wide; the maximum permitted height of a figure is 17 cm. All panels of figures must be on a single page and assembled into a rectangular shape for publication; any essential alignments (parts horizontal, vertical, spacings of stereo pairs, etc) should be indicated.

All figures are constructed to be full-width exactly 18cm. At this full width display size, all fonts conform to the below specifications.

Nature requests that authors of accepted manuscripts contribute towards the total cost of reproduction of colour figures in print. You will be charged a fee for the print reproduction of colour figures (see here for cost information).

We will revisit figure color and number once costs are quoted (during editorial process)

Tables should be prepared using the Table menu in Microsoft Word.

N/A

FIGURE FORMATTING: Lettering in all figures (e.g., labelling of axes) should be in uniform, sans-serif font, in lower-case type, and large enough to permit substantial reduction for publication (the minimum font size permitted is 5 pt).

All fonts are helvetica > 5pt. Some use of upper case letters is necessary for accuracy.

Separate parts of a figure are labelled a, b, etc.

All figure panel titles have been updated to lower-case. We prefer to also have a short title accompany the panel. If this is unacceptable, please contact John Lovell.

Units have a single space between the number and the unit, and follow SI nomenclature or the nomenclature common to a particular field.

Checked and all figures conform to these specifications

Thousands are separated by commas (1,000).

Checked and all figures conform to these specifications

Unusual units or abbreviations are defined in the legend.

Checked and all figures conform to these specifications

IMAGE PRESENTATION: Data figures are integrated into the main paper online. An overview of the key features of this presentation may be found here. We strongly encourage you to consult the guidelines. A discussion of our standards regarding how images should be prepared and presented can be found here.

Photos of plants (Fig. 1) have had their background removed as noted in the figure caption and raw un-modified images are found in Supplementary Note 2. Predicted molecule structures (Ext. Data Fig 5) are exactly as the stated software produced them without any modifications other than labels.

EXTENDED DATA: Extended Data do not appear in print but are included online within the full-text HTML and integrated in the downloadable PDF. Extended Data are an integral part of the paper, and only data that directly contribute to the main message should be included. All Extended Data must be referred to in the main text, and their legends should be listed sequentially at the end of the main text, not in the Extended Data files.

Extended data figures are formatted exactly as the main text figures are and conform to all specifications

The Extended Data should be assembled into a maximum of 10 A4 size, multi-panelled display items, submitted as individual files in .jpg, .tif or .eps format only. They should be of the same quality as print figures, but there are important differences in their formatting. More specific instructions can be found here. We encourage authors who are describing complex processes to include a schematic of the main finding as part of the Extended Data to aid readers unfamiliar with the immediate discipline.

Checked and all figures conform to these specifications

SUPPLEMENTARY INFORMATION: Supplementary Information (SI) is online-only, peer-reviewed material that is essential background to the study (e.g., large data sets, more complex methods, and calculations), but which is too large or impractical, or of interest only to a few specialists, to justify inclusion in the print version of the paper (see here for further details). While SI should not typically contain data figures (any figures additional to those appearing in the main text should be formatted as Extended Data), we require that the raw, uncropped data for gels be presented as an SI figure (see below).

There are three analyses that require detailed responses, which are provided in Supplementary Notes 1-6. In these cases, the associated figures apply primarily to the text in the supplementary notes, and are therefore better suited to inclusion therein and not as extended data figures.

Tables may be included in SI, but only if they are unsuitable for formatting as Extended Data (e.g., tables containing large data sets or raw data tables that are best suited to Excel files).

Three tables are required for results interpretation but not suitable for inclusion in extended data figures. These are included in the supplementary information along with the six notes.

If a manuscript has SI, each discrete item of the SI (e.g., videos, tables) must be referred to at an appropriate point in the main manuscript.

Checked and all SI elements are referenced in the main text or methods

We ask that you provide a word file entitled “SI Guide”, containing a cover page with manuscript title and author information; a table of contents (preferably with page numbers); and then any SI text, notes, figures, and titles and legends for any separate SI files. Additional information can be found here. Please pay careful attention to the formatting of the SI because it is not subedited. After the paper has been accepted, SI files can only be amended for critical changes to the scientific content, not for style.

The supplementary material now conforms to the guidelines here: https://www.nature.com/nature/for-authors/supp-info

SOURCE DATA (GRAPHS): To increase transparency, we strongly encourage you to provide, in spreadsheet form, the data underlying the graphical representations used in the figures. In the case of all experiments presenting data from animal models, this is a requirement and is not optional. This is in addition to our well-established data-deposition policy for specific types of experiments and large datasets. Online readers of the manuscript will be able to access the graphical source data directly from the figure legend. Spreadsheets can only be submitted in .xls, .xlsx or .csv formats. One file per figure is permitted. If there is a multi-panelled figure, the source data for each panel should be clearly labeled in the file; alternatively the source data for a figure can be included in multiple, clearly labeled sheets within an Excel file. File sizes of up to 30 MB are permitted, but it is expected that the vast majority of graphical source data files will be considerably smaller than this. When submitting these files with your manuscript, you should select the "Source Data" file type and use the title field in the file description tab to indicate the figure(s) to which the source data pertain.

Source data for each panel of each figure (or, when necessary, each data type within each panel) and named accordingly as separate tabs in a single xlsx spreadsheet.

RAW DATA (GELS): You must provide the original source images for all data obtained by electrophoretic separation (e.g., EMSA, northern/Southern/western blots, etc). The raw images must be assembled into a single .pdf or .tif file (multiple gels on a single page is encouraged). The file should be uploaded as Supplementary Figure 1. The full scanned images must be in uncropped form and contain size/molecular weight markers and loading controls. There should be an accurate indication of how the gels were cropped for the final figure. Please clarify in the figure legends and raw data files whether controls (such as beta-actin) were run on the same gel as loading controls, or on separate gels as sample processing controls (see here). While the data can be displayed in a relatively informal style, there must be a correspondence between each source data image and a specific main text or Extended Data figure. The main text or Extended Data figure legends should refer to the uncropped scans explicitly (e.g., "For gel source data, see Supplementary Figure 1."). For examples, see here or here.

N/A

THIRD PARTY RIGHTS: Please identify any content used in your article, whether in the main text, extended data or supplementary information, that comes from a third party. This could include figures, tables, images, videos or text boxes that are reproductions or adaptations of items that have previously been published elsewhere and/or are owned by a third party. It also encompasses pictures taken by professional photographers, maps and images downloaded from the internet. You must obtain the right to use each of these items and provide evidence that you have these rights for your paper. You will also need to give proper attribution to the copyright holders in your paper. We ask that you fill out a Third party rights table and upload this with your manuscript. You can find more information about third party rights here.

N/A

ORCID: As part of our efforts in improving transparency in authorship, we now request that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on Nature's manuscript tracking system prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. More information and instructions on how to associate your ORCID and MTS accounts can be found here. If you have any issues attaching an ORCID identifier to your MTS account, please contact the Platform Support Helpdesk.

We have ORCIDs for coauthors who have one. Please direct us for the best way to include these. Please note that Todd Mockler is deceased and does not have contact information.

Referee #3 (Remarks to the Author):

Summary: The generation of the pangenome is certainly a great development for sorghum, and the added attempt in this revision to integrate previous versions is commendable.

However, this manuscript still lacks originality. A simple reporting of structural variations without a validation of the phenotypes leaves most of the results speculative and hypothetical at best. Here are just a few further concerns:

> **We hope this is a minority opinion. The data generated and many of the analyses accomplished are completely novel. It is true that many of the phenotypic associations remain uncomplemented by genome editing; however, we are not sure how this would increase originality.**

1). it is not clear from the write-up, the significant improvements to the current BTX623 V3 other than clarity in the positions of genes. Do you need a pangenome to get this clarity? Usually, this can be achieved by generating a better assembly of the same genome. The objective of a pangenome generation should go well beyond clarity of positions of genes.

> **The need for a pangenome and the need to update an inaccurate existing reference genome do not rely on the same lines of evidence. The former is evidenced by Figure 2, which shows structural, gene presence-absence, and sequence variation. This diversity clearly shows that sorghum is a model for why we need pangenomics. The need for a new reference genome is evidenced in supplementary note 1. It should be pretty clear - the old genome has numerous large gaps and false structural variants. This issue has been projected onto many existing sorghum genomes, since they used the erroneous v3 coordinates as a synteny scaffold. Our updates resolve these issues.**

2). Other than providing a baseline for relating selection, what was the purpose of adding the 33 diverse sets in the sequencing panel? Did the addition of SRN39 for example, add to the discovery of more resilience genes or more Striga resistance genes? The authors say it stood out for strong performance under drought, but how did that feature add value to the pangenome information?

> **Among other purposes, the genomes form the foundation from which variants are found and to which they are referenced. Without them, none of the pangenome informed variant analyses would be possible**