Peer Review File

Manuscript Title: A pan-grass transcriptome reveals patterns of cellular divergence in crops

Reviewer Comments & Author Rebuttals

Reviewer Reports on the Initial Version:

Referees' comments:

Referee #1 (Remarks to the Author):

This manuscript by Guillotin and co-workers describes a useful and novel resource for comparative transcriptomic analyses of root cells across three models species within the panciod grasses — including two major crops and one wild relative — and asks useful questions about the evolution of cell type specific expression across species and across whole genome duplications which were not previously addressable. I should add that my background and expertise encompasses plant genomics and evolution but not single cell/nuclei RNA-seq and so my review focuses on the portions of the manuscript where my own knowledge is relevant.

The manuscript is clearly written. With one exception below I found the analyses and interpretations appropriate. I have one significant concern which should be addressable by the authors without the need to conduct new wetlab experiments and a number of more minor questions or suggestions.

Major:

1. A substantial portion of the manuscript focuses on the evolution of cell type specific expression patterns of retained maize duplicated following the maize WGD (Figures 2 & 3 and associated text). When duplicate maize gene pairs have differences in expression pattern, the authors employ sorghum data to classify these differences as subfunctionalization (if the expression domains are within the expression domain of the maize gene pair's ortholog in sorghum) or neofunctionalization (if one or both maize gene copies is expressed in cell types in which the sorghum ortholog is not).

Given the inability to go back in time and observe the true ancestoral expression pattern immediately prior to WGD, referencing expression in an outgroup is a reasonable approach, but it will result in an over estimate of the frequency of neofunctionalization. This is because losses of expression domains in sorghum will be classified as gains of expression domains for one or both maize gene copies.

Because the authors also generated data from setaria, a further outgroup to maize and sorghum, they can estimate the frequency of this misclassification by asking how many of the cases of apparent maize neofunctionalization they observed (Figure 3d) are actually consistent with patterns of cell type specific expression of the setaria ortholog. If both sorghum AND setaria lack expression in a cell type where a maize gene is expressed, they have much stronger evidence of neofunctionalization. If maize and setaria genes are both expressed in a given cell type, but the sorghum ortholog is not, the evidence for neofunctionalization is of much lower confidence.

Particularly given the very surprising results described on lines 216-219, using the data the authors have already collected to increase the confidence of the authors' classifications of expression patterns as likely true neofunctionalization would significantly strengthen the paper.

Minor:

2. Figure 3a: I'm not sure I understand the partial subfunctionalization case. To me partial subfunctionalization implies some tissues/cell types in which both genes are still expressed and some tissues/cell types in which each gene is expressed independently. So for example copy one is expressed in cells A, B, C and copy two in cells C and D. The partial subfunctionalization case show here still has completely exclusive expression domains, just ones that are less balanced that the subfunctionalization case.
3. Line 23: It would be very helpful to clarify this refers to genes gaining novel (non-ancestral) expression domain. My first time reading the abstract I interpreted this as gain of novel protein domains which would also make some sense in context.
4. Line 105: inconsistent capitalization. If you're using title case it is appropriate to capitalize the names of maize, sorghum
5. 269: should this be "columella cells"? Similarly in the legend of Figure 4, should it be "the columella" or "columella cells"?
6. Lines 230-232: Authors should have the freedom to define terms however they would like but I would strongly encourage the authors not to try redefining subfunctionalization — which is typically understood to involve genes retaining different parts of their ancestral function — to include the case where both genes retain the same enzymatic function and pattern of expression. My suggestion to them is that this will make things needlessly confusing for readers.
7. Extended Data Figure legends: purely stylistic suggestion: it isn't necessary to define the type of plot in the legend. "Violin plot distribution of the number of UMI detected among cells vs. 453 nuclei in…" can become "Distribution of the number of UMI detected among cells vs. 453 nuclei in…" and so on.
8. No change needed, but Extended Data Figs 7 and 8 represent a LOT of work and adds a lot of confidence to the author's datasets and analysis. Nicely done.
9. Line 556. Please also include the common name for this accession: A10.1. Setaria folks will recognize that as the reference genotype for the S. viridis genome, but likely won't know the PI number without looking it up.
10. Lines 695-697: This sees like a floating fragment of text that should be somewhere else in the manuscript. Begins and ends with a sentence fragment.
11. Lines 60-61. Is there a good agreed upon definition of the complete set of major cell identities in a plant root? Might avoid some confusion is there is space to define or at least list the set of cell identities identified in this study.

Referee #2 (Remarks to the Author):

In their manuscript “Gene duplication and cellular divergence in crops”, Guillotin et al. report comparisons of single-cell and single-nucleus transcriptomics in the roots of maize, sorghum and Setaria. They identify cell types that evolved particularly rapidly at the transcript level, presenting the maize root cap mucilage-associated genes as a case study. Acknowledging the whole genome duplication event of maize, the authors describe if the function of duplicated genes may have diverged based on their expression domains. To assess the evolutionary fate of the duplicated genes, they score the number of cells in domains where each maize homeolog is dominant, both are co-expressed, or partition the domains where the ancestral sorghum ortholog is expressed. Based on these calculations, full dominance appears the most common category, however homeologs that partition their expression among cell types also display the highest rate of gaining a non-ancestral domain. Thus it is suggested that neo- and subfunctionalization are linked through transcriptional regulation of dosage compensation, in a manner that retains homeolog domains for extended periods of time until diverging into a new function.

The current work strikes as an impressive application of the single-cell RNA sequencing technique to understand evolution of cells and cell types in plants. Also, the established single-cell and nuclei transcriptomes of the roots of the three grass species will be a good resource for additional studies. It is useful to learn that one needs twice as many nuclei for the same level of information as protoplasts, and the two datasets may not differ greatly. However, although the authors report interesting findings of different ways that duplicated genes may have evolved in terms of their transcription at a cellular level for dosage compensation, the manuscript remains relatively descriptive. Experimental support is limited especially for the second part of the manuscript.

Major comments:
- The authors are focusing on “neofunctionalization” of gene expression patterns. In principle, gene expression patterns can be quite dynamic during evolution. The authors are sometimes referring to “expansion” of ancestral pattern of gene expression. How much of this “neofunctionalization” is simply a slight “expansion” of the expression domain to some of the surrounding cells, and how much it is jumping to totally different cell types?

- Related to the above, it would be informative to understand how much there is intra-species variation in the single-cell transcriptome. For example, I wonder if the authors could conduct another round of analysis in another distantly related maize cultivar?

-In the last part of the paper, the authors have chosen a case study of gene expression related to mucilage biosynthesis in root cap, and they describe a differential expression of a number of genes between maize and sorghum. I am not sure which dataset was used (single-cell or single-nuclei or combined). I wonder if it makes a difference whether mucilage (or its intermediate product) is produced in the cortex or the root cap? In other words, if producing mucilage in the root cap is an agronomically favorable trait.

- I would like the authors to analyze the neofunctionalization processes more widely. Are most of the neofunctionalized events concentrated on the root cap/cortex? If not, which other spatial domains

are often affected? Are there any specific cell types or is it just a random event? In addition, how much are the cells expressing neofunctionalized homeologs different from the sorghum cells expressing the same genes? Can the authors also try to find whether there are any specific changes in the cis-element of the genes with neofunctionalization?

- Two different categories, dominant and co-expressed homeologs, are described in maize. Can the authors identify any trends in changes of evolution of cis-acting elements that might explain to which category a gene belongs?

Minor comments:
-It is interesting how different the stress response genes are regulated in single-cell RNAseq. Although stated otherwise (lines 86-88) I think the change might not be subtle, as the enrichment analysis indicates the opposite (Fig 1f). I'd suggest the authors to discuss whether the amount of stress-induced genes affects their downstream analysis. Are there differences in stress response between studied species? I am also surprised to see the cell-type specific over- and underrepresentation of genes between experiments (Fig 1d). How does this happen and what can we do about this?

- The images of root cross sections with phloem reporters are not of high enough resolution to assess what cell type is expressing that gene in Fig 2b.

-The authors do the separate clusters and compare single-cell (sc) and single-nuclei (sn) RNA sequencing with whole root samples (Fig S1), which is good. However, it is not clear how sc and sn datasets are combined to improve the clustering. I suggest the authors elaborate the description of these methods.

-The authors are using a different ratio (~1:12) of cells:nuclei in the outgroup Setaria viridis while other species have a ratio of ~1:2 (line 109 and lines 253-294). What might be the implications of this, if any?

-Why are the bars in Fig3 b next to each other and in d on top of each other? This does not seem consistent. The data as such is interesting, however I think the text related to Fig 3 could be condensed.

-I suggest the authors improve the legend of Fig 2c. To me it is not clear what was done until reading the methods.

Referee #3 (Remarks to the Author):

In this manuscript, the authors describe an analysis of single-nucleus, single-cell and organ-system level transcriptome data in three grass species aimed at understanding expression divergence of genes in various cell types and associate it with the evolution of important agronomic traits. There

are several interesting and novel findings (to the best of my knowledge) presented in this manuscript, which would be of great interest to the broader community, including 1. comparison (and guidelines) of single-nucleus and single-cell transcriptome datasets to identify major cell identities, 2. a single-cell "pan-transcriptome" of these three species will serve as an important new community resource, 3. identification of genes with cell-type specific expression in two important agronomic species, and 4. discovery of genes that may have contributed to root cap "slime" - a key agricultural trait. I find the above aspects of the manuscript exciting and anticipate it would be highly cited for those reasons.

However, the findings associated with the analysis of homeolog expression patterns in Maize has been previously reported by multiple publications over the past decade. They cite Schnable et al., 2011 (PNAS) but miss the rich body of literature that further evaluates the fate of duplicate genes in Maize (e.g. Zhao et al., 2017; Renny-Byfield, et al., 2017) and many other polyploid crops.

I have a major concerns regarding the methods used to determine the "ancestral state" of expression for both Maize homoeologs. The authors used the expression patterns of the orthologous gene in Sorghum as a proxy for the ancestral state. This is heavily flawed assumption - both Maize and Sorghum genes have all evolved over the past 12 million years. Sorghum is not the ancestral state. Multiple species with a robust phylogeny would be needed to obtain more reliable estimates. Why was Setaria not included in this analysis (line 159)? The authors also assumed that the parental gene copies that contributed to the Maize polyploid event were similarly expressed (as noted by the authors on line 200-201). However, the current model suggests that Maize is an allo-polyploid (McKain et al., 2018), formed by species from two divergent lineages, thus it's quite likely that many homoeologs would not be similarly expressed (as would be the case for an autopolyploid). This likely contributed to the degree of subgenome expression dominance observed in Maize, as noted in Lines 163-165, and by previous studies including Schnable et al. (2011). The estimate of the ancestral state will greatly impact the analyses presented by the authors.

Also, subgenome differences in selective constraints (lines 171-172, 214-215) has also been described in these previous studies in Maize as well as other polyploids.

The model highlighted by the authors on how genomic-balance may lead to increased sub- and neo-functionation has similarly been previously reported by dozens of previous studies in a wide variety of polyploids. The authors provide very nice examples of this at cell type level in roots. However, determining the actual rate of sub- vs neo-functionalization is dependent on addressing points raised above. However, they should cite previous studies that have shown this in both plant and non-plant systems (e.g. Cotton - Peng et al., 2020). Several reviews have been published on this topic over the past decade.

My recommendation to the authors would be to focus on the truly novel findings presented in this manuscript. The impressive dataset presented in this manuscript alone can be used to identify genes that have diverged in expression, without needing to classify it as sub/neo-functionalization, and can be used to identify distinct cell-type specific expression in a particular species.

Additional comments:
Line 29: Suggestion "... comparing gene expression ... "
Line 110: How were orthologs identified between these three species?
Line 147: Suggestion "... to quantify expression patterns ... "
Line 155-156: If maintaining dosage among interacting genes is critical in newly-formed polyploids, why not ancient polyploids? Additional details are needed in the paragraph -- help guide the average reader.
Line 254: Subgenome dominance has been shown to be associated with transposable elements (TE) differences. Given that TE may also contain regulatory elements, and that TE abundance is high and variable in gasses, are TEs contributing to the observed expression differences?

Mckain et al., 2018 - https://www.biorxiv.org/content/10.1101/352351v1
Peng et al., 2020 - https://academic.oup.com/genetics/article/182/2/503/6062852
Renny-Byfield., 2017 - https://academic.oup.com/mbe/article/34/8/1825/3742530
Zhao et al., 2017 - https://pubmed.ncbi.nlm.nih.gov/29180596/

Referee #4 (Remarks to the Author):

Here, Guillotin et al. compare all cell transcriptomes across three different types of grass (maize, sorghum, and Setaria). Using a combination of single-cell and single-nucleus sequencing, the authors show (via comparative cellular analysis) the more rapid divergence of some cell-type transcriptomes, likely by the recruitment of gene modules from other cell types. By investigating maize whole genome duplication, the authors demonstrate dosage compensation in surviving co-expressed homeologs – speculating that this effectively maintains a stoichiometric balance to allow for homeolog retention while new functions arise. My following comments focus on the single-cell analysis, which is performed rigorously.

- The authors use a Seurat-based integration method to assess cell type across the key crop species. One concern is that this method can over-integrate closely related cell types. The authors should support their initial cell-type mapping by using additional integration methods.
- MetaNeighbor is adapted to quantify the similarity of cell clusters across datasets using a given marker set of genes and their orthologs. Again, I am concerned about using a single method for this analysis and that this may be heavily influenced by the integration method used. Can the authors also train a cell-type classifier to support these data?
- The data presented show that transcriptomes of columella, phloem, cortex subcluster 3, endodermis, pericycle, and stele cell types are most divergent compared to Setaria. Again, it would be essential to demonstrate the true biological differences instead of potential batch effects.
- The authors use a measure of co-expression conservation to examine divergent expression patterns across species in co-expression networks. Here, the authors could consider (if possible) using an algorithm such as SCENIC to infer gene regulatory networks from their single-cell profiling. This analysis may give some additional insight.

Author Rebuttals to Initial Comments:

We resubmit our manuscript, generating a new dataset that addresses reviewers' concerns about the need to use *Setaria* as well as sorghum in the ancestral state inference. We re-analyzed all data using that new dataset and performed other analyses suggested by reviewers. Most of the changes in the ms are in the whole genome duplication section, as we focused that analysis on how duplication events affected cell type-specific expression patterns, as per reviewer comments. The other sections contain additions based on suggested analyses by reviewers, in particular, network analysis. We believe these changes have made the ms much stronger, and we appreciate the thoughtful and constructive set of comments. All our answers to reviewer comments are blue, with the full original reviewer comments in black. When quoting excerpts from the ms, we add author names to the numbered citations.

Referees' comments:

Referee #1 (Remarks to the Author):

This manuscript by Guillotin and co-workers describes a useful and novel resource for comparative transcriptomic analyses of root cells across three models species within the panciod grasses — including two major crops and one wild relative — and asks useful questions about the evolution of cell type specific expression across species and across whole genome duplications which were not previously addressable. I should add that my background and expertise encompasses plant genomics and evolution but not single cell/nuclei RNA-seq and so my review focuses on the portions of the manuscript where my own knowledge is relevant.

The manuscript is clearly written. With one exception below I found the analyses and interpretations appropriate. I have one significant concern which should be addressable by the authors without the need to conduct new wetlab experiments and a number of more minor questions or suggestions.

Major:

1. A substantial portion of the manuscript focuses on the evolution of cell type specific expression patterns of retained maize duplicated following the maize WGD (Figures 2 & 3 and associated text). When duplicate maize gene pairs have differences in expression pattern, the authors employ sorghum data to classify these differences as subfunctionalization (if the expression domains are within the expression domain of the maize gene pair's ortholog in sorghum) or neofunctionalization (if one or both maize gene copies is expressed in cell types in which the sorghum ortholog is not).

Given the inability to go back in time and observe the true ancestoral expression pattern immediately prior to WGD, referencing expression in an outgroup is a reasonable approach, but it will result in an over estimate of the frequency of neofunctionalization. This is because losses of expression domains in sorghum will be classified as gains of expression domains for one or both maize gene copies.

Because the authors also generated data from setaria, a further outgroup to maize and sorghum, they can estimate the frequency of this misclassification by asking how many of the cases of apparent maize neofunctionalization they observed (Figure 3d) are actually consistent with patterns of cell type specific expression of the setaria ortholog. If both sorghum AND setaria lack expression in a cell type where a maize gene is expressed, they have much stronger evidence of neofunctionalization. If maize and setaria genes are both expressed in a

given cell type, but the sorghum ortholog is not, the evidence for neofunctionalization is of much lower confidence.

Particularly given the very surprising results described on lines 216-219, using the data the authors have already collected to increase the confidence of the authors' classifications of expression patterns as likely true neofunctionalization would significantly strengthen the paper.

The reviewer makes an important point that we needed to rectify with more data generation. In the first submission, the Setaria dataset was smaller than the datasets for the other species analyzed, so we could only spot check agreement with sorghum expression for the more highly expressed genes. To carry out a more complete analysis of ancestral expression, we generated more single-cell RNA-seq data for Setaria, and the manuscript now analyzes 10,613 cells and 12,192 nuclei, numbers comparable to those for the other species. In the revised manuscript, we now consider regulatory neofunctionalization, which we term domain expansion in the manuscript, when maize is expressed in a cell type where neither sorghum nor Setaria orthologs are expressed. The new data removed about a third of homeologous duplicate genes, but it did not change any trends. We point out the following sections, where the new data is used.

"Given the comprehensive coverage of a combined analysis, we pursued both whole cell and nucleus profiling to investigate cellular evolution in the maize-sorghum-*Setaria* clade. Thus, we generated profiles for sorghum (3,510 cells and 7,620 nuclei) and *Setaria* (10,613 cells and 12,192 nuclei, Supplementary Table 1)."

"We used concordance between sorghum and *Setaria* to infer ancestral expression domains for each duplicate gene pair."

Minor:

2. Figure 3a: I'm not sure I understand the partial subfunctionalization case. To me partial subfunctionalization implies some tissues/cell types in which both genes are still expressed and some tissues/cell types in which each gene is expressed independently. So for example copy one is expressed in cells A, B, C and copy two in cells C and D. The partial subfunctionalization case show here still has completely exclusive expression domains, just ones that are less balanced that the subfunctionalization case.

Thank you for pointing this out. The schematic could have represented the concept of partial subfunctionalizaton better. We revised it in the new Fig. 2d (pasted below). We consider subfunctionalization when both homeologs are expressed in distinct cell types but they can share some expression as well.

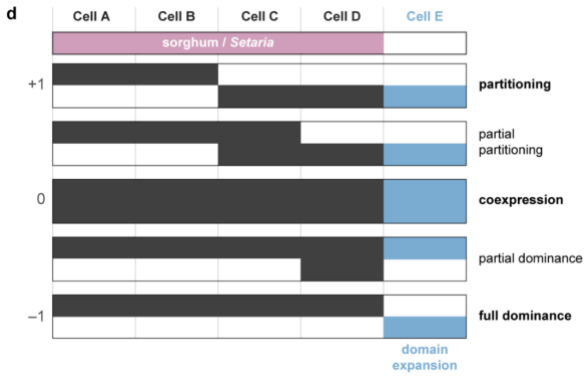


Figure 2d legend: "**d** Conceptual schematic of hypothetical expression patterns between duplicate gene pairs following a metric with a scale ranging from full dominance (-1) to equal co-expression (0) to regulatory subfunctionalization (1). Example intermediate states are also shown. Blue shows regulatory neofunctionalization."

3. Line 23: It would be very helpful to clarify this refers to genes gaining novel (non-ancestral) expression domain. My first time reading the abstract I interpreted this as gain of novel protein domains which would also make some sense in context.

We now use the term "expression domains" in the abstract and elsewhere now to make it clear we are talking about transcript localization and not protein domains.

4. Line 105: inconsistent capitalization. If you're using title case it is appropriate to capitalize the names of maize, sorghum

This has been corrected

5. 269: should this be "columella cells"? Similarly in the legend of Figure 4, should it be "the columella" or "columella cells"?

This has been corrected

6. Lines 230-232: Authors should have the freedom to define terms however they would like but I would strongly encourage the authors not to try redefining subfunctionalization — which is typically understood to involve genes retaining different parts of their ancestral function — to include the case where both genes retain the same enzymatic function and pattern of expression. My suggestion to them is that this will make things needlessly confusing for readers.

For the sake of focusing on the effects of WGD on cellular identity, we eliminated the phrase stating that dosage compensation was a form of subfunctionalization. The section is now more focused on gene duplication as it contributes to new cellular expression domains. The dosage compensation observation is stated without drawing the parallel to the case of regulatory subfunctionalization by transcript dosage.

"In keeping with Genome Balance models, we observed that co-expressed WGD homeologs showed expression patterns indicative of dosage compensation (Birchler & Veitia, 2012; Muyle *et al.*, 2022)[51,52], while this pattern was weaker or non-existent in other duplicate classes (Fig. 3a, Extended Data Fig. 10a-c)."

7. Extended Data Figure legends: purely stylistic suggestion: it isn't necessary to define the type of plot in the legend. "Violin plot distribution of the number of UMI detected among cells vs. 453 nuclei in…" can become "Distribution of the number of UMI detected among cells vs. 453 nuclei in…" and so on.

It is a bit clunky to refer to type of plot each time in the legend, but we did feel this was important for a reader who is not familiar with the plot type to know the technique being used. We also wanted to stay consistent with the main figures. To address the concern, we gave the extended figure legends conceptual titles and eliminated repetition of plot types when possible.

8. No change needed, but Extended Data Figs 7 and 8 represent a LOT of work and adds a lot of confidence to the author's datasets and analysis. Nicely done.

Thank you for recognizing the extensive corroboration of the data. We hope this will be a unique resource for the community.

9. Line 556. Please also include the common name for this accession: A10.1. Setaria folks will recognize that as the reference genotype for the S. viridis genome, but likely won't know the PI number without looking it up.

Thank you. We included that information in the methods.

10. Lines 695-697: This sees like a floating fragment of text that should be somewhere else in the manuscript. Begins and ends with a sentence fragment.

Corrected, apologies for not catching that in proofing the final version.

11. Lines 60-61. Is there a good agreed upon definition of the complete set of major cell identities in a plant root? Might avoid some confusion is there is space to define or at least list the set of cell identities identified in this study.

To address the point, we give a modern anatomy reference that defines the major cell types of the root when we first bring up the topic, and, in that place in the text, we refer to Figure Legend 1a,b,c, that lists the cell types. We also now point out that these are the major cell types in the figure legend.

"In both Arabidopsis and maize, cells and nuclei generated UMAP clusters corresponding to all the major cell identities (Ray F. Evert, 2006)[28] (Fig. 1a-c; Extended Data Fig. 2, 3)."

From Figure legend 2c: "**c** Dot plots of Arabidopsis marker genes in cells (blue) or nuclei (red), showing all the cell types defined from clusters in this study."

Referee #2 (Remarks to the Author):

In their manuscript "Gene duplication and cellular divergence in crops", Guillotin et al. report comparisons of single-cell and single-nucleus transcriptomics in the roots of maize, sorghum and Setaria. They identify cell types that evolved particularly rapidly at the transcript level, presenting the maize root cap mucilage-associated genes as a case study. Acknowledging the whole genome duplication event of maize, the authors describe if the function of duplicated genes may have diverged based on their expression domains. To assess the evolutionary fate

of the duplicated genes, they score the number of cells in domains where each maize homeolog is dominant, both are co-expressed, or partition the domains where the ancestral sorghum ortholog is expressed. Based on these calculations, full dominance appears the most common category, however homeologs that partition their expression among cell types also display the highest rate of gaining a non-ancestral domain. Thus it is
suggested that neo- and subfunctionalization are linked through transcriptional regulation of dosage compensation, in a manner that retains homeolog domains for extended periods of time until diverging into a new function.

The current work strikes as an impressive application of the single-cell RNA sequencing technique to understand evolution of cells and cell types in plants. Also, the established single-cell and nuclei transcriptomes of the roots of the three grass species will be a good resource for additional studies. It is useful to learn that one needs twice as many nuclei for the same level of information as protoplasts, and the two datasets may not differ greatly. However, although the authors report interesting findings of different ways that duplicated genes may have evolved in terms of their transcription at a cellular level for dosage compensation, the manuscript remains relatively descriptive. Experimental support is limited especially for the second part of the manuscript.

Major comments:
- The authors are focusing on “neofunctionalization” of gene expression patterns. In principle, gene expression patterns can be quite dynamic during evolution. The authors are sometimes referring to “expansion” of ancestral pattern of gene expression. How much of this “neofunctionalization” is simply a slight “expansion” of the expression domain to some of the surrounding cells, and how much it is jumping to totally different cell types?

That is an important point. To address it, we re-analyzed the regulatory neofunctionalized genes separating cases that could be considered slight expansion within a tissue (subtype to subtype, e.g., from Epidermis1 to Epidermis 2), from a more dramatic expansion (cell type to cell type, e.g., from Epidermis to Xylem). The new analysis showed a fairly even split between the two categories (55% slight expansion and 45% dramatic expansion). In the new analysis, we display, in Fig. 3f and Supplementary Fig 9i-o, the dramatic expression domain expansions. Overall, we found that certain cell types, such as columella cells and also xylem/phloem, appear to be the primary targets of expansion domains. This was an excellent question that led to an analysis that now appears in the main text and in a figure panel.

“Finally, certain cell types appeared to be more likely domain expansion destinations than others (Fig. 3f). The trends were similar for all duplicate classes, with the specialized vascular cells and root cap cells most frequently comprising the new expression domains. Cortex was the least frequent sink for new domains, although one of the most frequent source domains (Fig. 3f, Extended data Fig. 10e-h). Overall, the data shows how gene duplication, particularly WGD, frequently provides genetic material for the transcriptional divergence of specific cell types.”

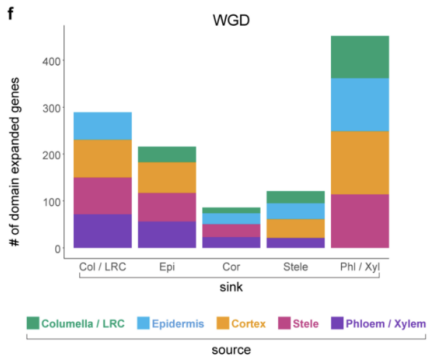
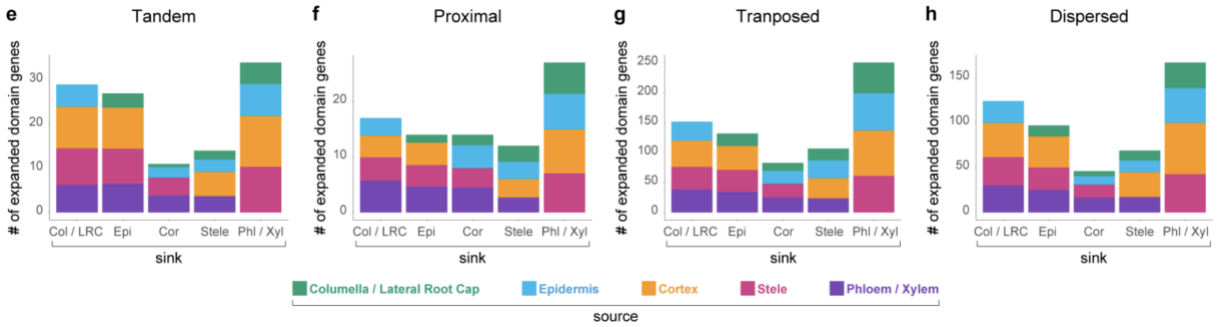


Fig 3f

Figure 3f legend: "**f** Regulatory neofunctionalized genes categorized by their new expression domains. Colors within a bar graph show their ancestral cell-type domain (Methods)."



Extended Data Fig.10e-h legend: "**e-h** Distribution of maize genes displaying regulatory neofunctionalization of expression into new cell types. Colors signify the cell type of origin."

- Related to the above, it would be informative to understand how much there is intra-species variation in the single-cell transcriptome. For example, I wonder if the authors could conduct another round of analysis in another distantly related maize cultivar?

We agree this would be interesting, but, respectively, we feel it is out of the scope of the current study. It was difficult to fit all the current analysis in this paper, including the relevant new suggestions from the reviewers. We feel this interesting question will have to be a next step.

-In the last part of the paper, the authors have chosen a case study of gene expression related to mucilage biosynthesis in root cap, and they describe a differential expression of a number of genes between maize and sorghum. I am not sure which dataset was used (single-cell or single-nuclei or combined). I wonder if it makes a difference whether mucilage (or its intermediate product) is produced in the cortex or the root cap? In other words, if producing mucilage in the root cap is an agronomically favorable trait.

We have now clarified in the text that the comparison of cell types was done on our combined Cells+Nuclei dataset.

"To ask about cellular divergence more broadly, we next examined the entire transcriptome of each cell cluster to determine which cell types changed most dramatically in maize and sorghum compared to the outgroup *Setaria*. For all comparative analysis, we used the combined cell and nuclei dataset. To compare cell identity across species, we used MetaNeighbor on our combined dataset (Fig. 4a)."

On the second part of the question, there is not much in the literature on the functional consequences of where mucilage is produced. However, we do now include an analysis of genes that appear to have been under selection during domestication (Wang et al., 2020). That analysis revealed that the genes found to be divergent between maize and sorghum columella also showed a significant enrichment for genes involved in maize domestication. We refer to this analysis in the last section of the main text.

"The most parsimonious model is that the recruitment of the mucilage module occurred before the maize WGD, as both maize homeologs in the mucilage-annotated genes tended to share expression in the columella. However, the set of mucilage genes showed a significant overlap with genes previously defined as under selection during domestication (Wang et al., 2020)[62] (Supplementary Table 8), suggesting they play a role in agricultural traits."

- I would like the authors to analyze the neofunctionalization processes more widely. Are most of the neofunctionalized events concentrated on the root cap/cortex? If not, which other spatial domains are often affected? Are there any specific cell types or is it just a random event?

Indeed, the re-analysis we performed to answer this question showed that specific cell types, such as columella and some vascular cells, were frequent "destinations" for regulatory neofunctionalized duplicates, as per Figure 3f above. This was a great suggestion as the regulatory neofunctionalized gene analysis concurred with the analysis on "fast evolving" cell types in the next section. We make that connection in the section on whole genome duplications.

"The analysis showed that, in both maize and *Setaria*, transcriptomes of columella, phloem, cortex subcluster 3, endodermis, pericycle, and stele cell types are the most divergent compared to *Setaria* (Fig. 4a), suggesting that the function of these tissues diverged from *Setaria* before the maize-sorghum split. In addition, certain cell types—such as cortex subcluster 1 and 4, and several stele clusters—were significantly diverged between maize and sorghum, implying additional divergence after the maize-sorghum split. We note that the fast-evolving cell types were largely consistent with the sink tissues favored for regulatory neofunctionalization by duplicate genes (compare Fig. 4a with 3f). Interestingly, in maize, columella was among the most divergent cell type relative to *Setaria* (Fig. 4a)."

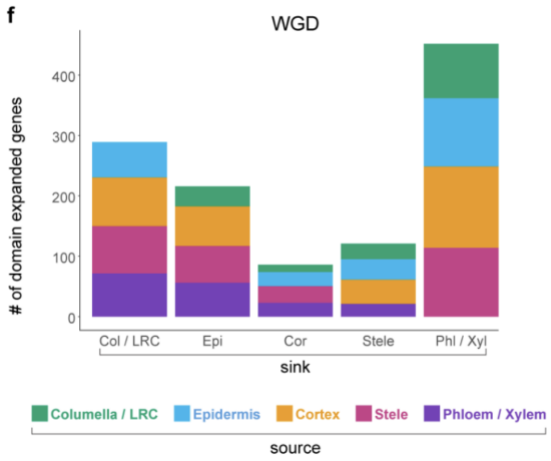


Figure 3f legend: "**f** Regulatory neofunctionalized genes categorized by their new expression domains. Colors within a bar graph show their ancestral cell-type domain (Methods)."

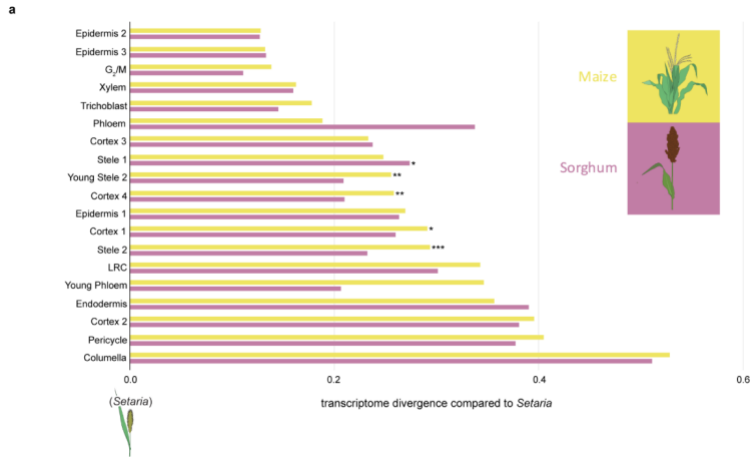


Figure 4a legend: **"a** MetaNeighbor analysis showing a quantification of transcriptome divergence among cell types in maize and sorghum compared to the outgroup *Setaria*. Statistical significance between maize and sorghum was performed using the Hanley McNeil test (see method, p *<0.05,**<0.01,***<0.001)."

In addition, how much are the cells expressing neofunctionalized homeologs different from the sorghum cells expressing the same genes? Can the authors also try to find whether there are any specific changes in the cis-element of the genes with neofunctionalization?

The first part of the question was addressed in response to the previous question. The cells that were the frequent destination of regulatory neofunctionalization were also the most divergent in the wider analysis.

For the second part of the question, we performed cis-regulatory analysis using available tools (MEME suite, FIMO) guided or not guided by open chromatin analysis in maize roots (Marand et al., 2020). We did not find any signals (e.g., gain of cis-regulatory elements) for the regulatory neofunctionalized genes (but see response to next question for strong trends in dominant genes introns). The cis-regulatory changes involved in domain expansion may be more subtle than is detectable with current cis-regulatory information and methods.

- Two different categories, dominant and co-expressed homeologs, are described in maize. Can the authors identify any trends in changes of evolution of cis-acting elements that might explain to which category a gene belongs?

This was an excellent suggestion. We did find a strong trend in the comparison of dominant vs. non-dominant homologs, showing a much lower conservation of cis-elements in their introns (and no other genic region) than co-expressed homologs (Fig. 3d and Supplementary Fig. 9h). We surmised in the ms that this might be due to loss of intron-mediated enhancement. The analysis led to additions to the main text and a new figure panel.

"In addition, we found that homeologs in the WGD class showed a dramatic decrease in the conservation of intronic *cis*-regulatory sites between the dominant and non-dominant homeolog compared to homeologs in the co-expressed class—a feature not observed in other duplicate

classes, nor in promotors (Fig. 3d; Extended Data Fig. 10d; Supplemental Table 6). This could represent a possible loss of intron-mediated expression enhancement in the non-dominant homeolog. These two genomic features are consistent with prior findings that suggest dominant homeologs may have retained ancestral gene function (Renny-Byfield *et al.*, 2017; Zhao *et al.*, 2017)[49,54], while non-dominant homeologs may adopt new functions or become pseudogenes."

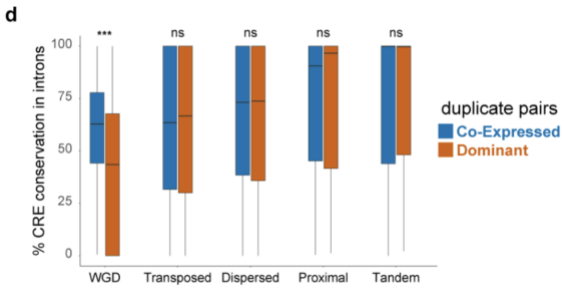
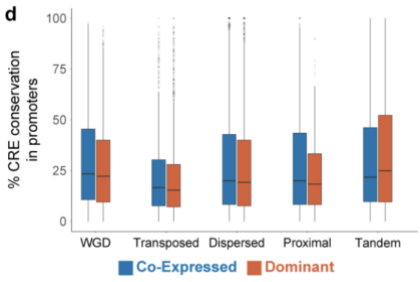


Figure legend 3d: "**d** Cis-regulatory element conservation rate between duplicate pairs in introns split into co-expressed and dominant categories… in d, Wilcoxon signed-rank test, with p *** < 0.001."



Extended Data Figure legend 10d: "**d** Conservation rate of *cis*-regulatory elements between WGD homeolog pairs in promoters. The plot shows no major differences between co-expressed and dominant gene pairs, and no major differences among the different classes of duplication."

Minor comments:
-It is interesting how different the stress response genes are regulated in single-cell RNAseq. Although stated otherwise (lines 86-88) I think the change might not be subtle, as the enrichment analysis indicates the opposite (Fig 1f). I'd suggest the authors to discuss whether the amount of stress-induced genes affects their downstream analysis. Are there differences in stress response between studied species? I am also surprised to see the cell-type specific over- and underrepresentation of genes between experiments (Fig 1d). How does this happen and what can we do about this?

To address this point, we performed all the downstream analyses with and without the stress related genes implicated in cell profile enrichment. All the trends remained the same, and, in general, these genes were a minor fraction of the genes involved in each analysis. We list the potential protoplasting induced stress genes in all the species now in the Supplemental Table 3.

On the issue of why the stress-related genes appear to be cell-type specific, we strongly suspect that, even though cells experience the same uniform protoplast-generation stress, protoplasts themselves still have the capacity to respond in a cell type-specific manner. We saw some hints of this in a previous study on nitrogen responses among cell types in which we found that nitrogen treatment during protoplast generation enhanced the cell type-specific responses we found when treating only intact roots. We will have to save this thought for another paper when we gather more data on the phenomenon. In the meantime, we now added a sentence in the paper, pointing out that dual cell and nucleus profiling can identify such cell-specific stress responses.

- The images of root cross sections with phloem reporters are not of high enough resolution to assess what cell type is expressing that gene in Fig 2b.

We substituted an improved image.

-The authors do the separate clusters and compare single-cell (sc) and single-nuclei (sn) RNA sequencing with whole root samples (Fig S1), which is good. However, it is not clear how sc and sn datasets are combined to improve the clustering. I suggest the authors elaborate the description of these methods.

We updated the methods section to elaborate on the techniques used in Seurat that combine datasets from different sources.

From Methods: “For the co-clustering of cells and nuclei, either dataset were treated similarly, all replicates were integrated at once using the seurat 'SCT' approach (Hafemeister & Satija, 2019)[31]. First raw reads were normalized using the *SCTransform* function, then *SelectIntegrationFeatures* was used to identify anchors between the datasets, using 3000 features. “

-The authors are using a different ratio (~1:12) of cells:nuclei in the outgroup Setaria viridis while other species have a ratio of ~1:2 (line 109 and lines 253-294). What might be the implications of this, if any?

This issue is related to comments by other reviewers that Setaria should be used in the wider analysis of ancestral gene expression, so, indeed, it did have important implications, as the reviewer pointed out. To address this issue, we generated more single-cell data for Setaria. The new dataset now contains almost equivalent numbers of celsl vs nuclei (10,613 cells and 12,192 nuclei). It also strengthens the inference of the ancestral gene expression states.

-Why are the bars in Fig3 b next to each other and in d on top of each other? This does not seem consistent. The data as such is interesting, however I think the text related to Fig 3 could be condensed.

We modified the figure, shortening the discussion of this analysis in the text and simplifying the figure. We explain the components of these graphs in question in Fig. 2.

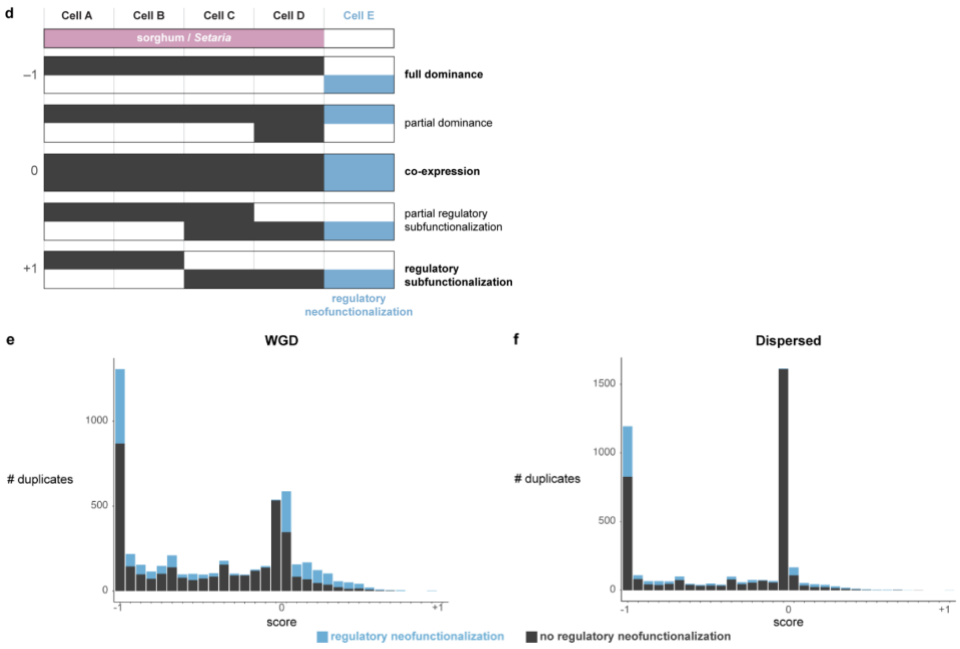


Figure legend 2d-f. "**d** Conceptual schematic of hypothetical expression patterns between duplicate gene pairs following a metric with a scale ranging from full dominance (-1) to equal co-expression (0) to regulatory subfunctionalization (1). Example intermediate states are also shown. Blue shows regulatory neofunctionalization. **e-f** Distribution of duplicate gene expression patterns using the metric described in (d) for WGD homeologs (e) and dispersed duplicate (f) pairs having similar with median Ks. Number of genes: 5,052 (WGD homeologs); 4,140 (dispersed duplicates)."

-I suggest the authors improve the legend of Fig 2c. To me it is not clear what was done until reading the methods.

We have modified the text in Figure 2 to describe the overall methods more clearly, as per above.

Referee #3 (Remarks to the Author):

In this manuscript, the authors describe an analysis of single-nucleus, single-cell and organ-system level transcriptome data in three grass species aimed at understanding expression divergence of genes in various cell types and associate it with the evolution of important agronomic traits. There are several interesting and novel findings (to the best of my knowledge) presented in this manuscript, which would be of great interest to the broader community, including 1. comparison (and guidelines) of single-nucleus and single-cell transcriptome datasets to identify major cell identities, 2. a single-cell "pan-transcriptome" of these three species will serve as an important new community resource, 3. identification of genes with cell-type specific expression in two important agronomic species, and 4. discovery of genes that may have contributed to root cap "slime" - a key agricultural trait. I find the above aspects of the manuscript exciting and anticipate it would be highly cited
for those reasons.

However, the findings associated with the analysis of homeolog expression patterns in Maize

has been previously reported by multiple publications over the past decade. They cite Schnable et al., 2011 (PNAS) but miss the rich body of literature that further evaluates the fate of duplicate genes in Maize (e.g. Zhao et al., 2017; Renny-Byfield, et al., 2017) and many other polyploid crops.

We agree that we did not do an adequate job of putting our results in the context of prior work, as our intent was to draw focus on the novel aspects of the ms concerning the cell-type specific comparisons of duplicate genes. To address this concern, we made major modifications to the duplication section, **The Maize WGD Makes a Major Contribution to Cellular Divergence."**

First, we have revised the section to better focus the homeolog analysis on what is learned about the way the WGD contributed to changes in cell identities –i.e., the unique perspective of our analysis. We are also more explicit now in pointing out prior results and what our analysis adds to prior knowledge. This whole section now reflects that emphasis.

I have a major concerns regarding the methods used to determine the "ancestral state" of expression for both Maize homoeologs. The authors used the expression patterns of the orthologous gene in Sorghum as a proxy for the ancestral state. This is heavily flawed assumption – both Maize and Sorghum genes have all evolved over the past 12 million years. Sorghum is not the ancestral state. Multiple species with a robust phylogeny would be needed to obtain more reliable estimates. Why was Setaria not included in this analysis (line 159)?

To address this important issue, we generated more single-cell profiles for Setaria. In the first version of the ms, we had fewer cells of Setaria so we could only spot check agreement with sorghum gene expression patterns for agreement in Setaria. We did not explain this in the original manuscript. To rectify the issue in the ancestral state inference, our updated analysis now uses 10,613 cells and 12,192 nuclei, comparable to the other species, enabling us to include Setaria in the ancestral expression state analysis. We now infer an ancestral state by concurrence of cell-type specific expression in both sorghum and Setaria. This eliminated about a third of the genes we had used previously, although it did not change the major trends concerning domain expansion or partitioning. All the analysis was redone with the sorghum/Setaria concurring genes forming the basis to infer ancestral domains.

"Given the comprehensive coverage of a combined analysis, we pursued both whole cell and nucleus profiling to investigate cellular evolution in the maize-sorghum-*Setaria* clade. Thus, we generated profiles for sorghum (3,510 cells and 7,620 nuclei) and *Setaria* (10,613 cells and 12,192 nuclei, Supplementary Table 1)."

From the section that starts analysis of of ancestral state:

"We used concordance between sorghum and *Setaria* to infer ancestral expression domains for each duplicate."

The authors also assumed that the parental gene copies that contributed to the Maize polyploid event were similarly expressed (as noted by the authors on line 200-201). However, the current model suggests that Maize is an allo-polyploid (McKain et al., 2018), formed by species from two divergent lineages, thus it's quite likely that many homoeologs would not be similarly expressed (as would be the case for an autopolyploid). This likely contributed to the degree of subgenome expression dominance observed in Maize, as noted in Lines 163-165, and by previous studies including Schnable et al. (2011). The estimate of the ancestral state will greatly impact the analyses presented by the authors.

We have changed several sections to address the alloploid origin of maize and the consequences for our analysis and conclusions. We first state clearly the likely allopolyploid origins of maize.

“In addition, since sharing a common ancestor with sorghum, maize underwent a whole genome duplication (WGD) 5 to 12 million years ago, likely following a hybridization (allopolypoidy)(Swigonova, 2004; Estep *et al.*, 2014; McKain *et al.*, 2018).”

We also state in the ms that the likely allopolyploid origin of maize means that some of the gene expression differences we observed in homeologs could have originated in the parental genomes before hybridization.

“We note that we cannot determine if differences in gene expression between duplicated genes occurred in the parent genomes or, more likely (Hughes *et al.*, 2014)[48], after WGD (Chaudhary *et al.*, 2009; Zhao *et al.*, 2017; McKain *et al.*, 2018)[14,46,49].”

Also, subgenome differences in selective constraints (lines 171-172, 214-215) has also been described in these previous studies in Maize as well as other polyploids.

We now make a clear distinction about the points that have been shown in prior studies, and we cite the previous studies when those observations are made in our data.

For example, “As found in previous studies (Renny-Byfield *et al.*, 2017; Zhao *et al.*, 2017)[49,54], dominant members of a homeolog pair showed greater purifying selection (Fig. 3c).”

We cite previous studies that found M1 expression bias and cite previous studies and reviews on the issue of expression bias and dominance.

The model highlighted by the authors on how genomic-balance may lead to increased sub- and neo-functionation has similarly been previously reported by dozens of previous studies in a wide variety of polyploids. The authors provide very nice examples of this at cell type level in roots. However, determining the actual rate of sub- vs neo-functionalization is dependent on addressing points raised above. However, they should cite previous studies that have shown this in both plant and non-plant systems (e.g. Cotton - Peng et al., 2020). Several reviews have been published on this topic over the past decade.

We now cite Peng et al. and other studies on this topic and more carefully put the novel aspects of our analysis in the context. The addition of the new Setaria data, as above, now improves the inference of ancestral state and classification of neo and subfunctionalization.

We cite what we found to be the most specific reference on link between regulatory subfunctionalization and neofunctionalization, which is one point we draw in the subfunctionization analysis.

“In addition, while regulatory subfunctionalization between homeologs was a minor outcome, this category of homeologs showed the highest rate of regulatory neofunctionalization (59%) compared to any other duplicate class (e.g., Fig. 2e,f, Extended data Fig. 9b-d). The trend is consistent with models in which regulatory subfunctionalization is a transitory state that facilitates regulatory neofunctionalization (Rastogi & Liberles, 2005)[56].”

My recommendation to the authors would be to focus on the truly novel findings presented in this manuscript. The impressive dataset presented in this manuscript alone can be used to identify genes that have diverged in expression, without needing to classify it as sub/neo-functionalization, and can be used to identify distinct cell-type specific expression in a particular species.

We have streamlined the WGD section of the manuscript and addressed the reviewer's valid critiques, clarifying the novel aspects of ms. In the revision, there were aspects of the gene duplication analysis that reinforced the analysis of overall cellular divergence, which is a unique perspective of the ms. For this and other reasons, we felt the duplicate analysis was important and left it in the ms with some key revisions and new analysis. We appreciate the frank critique and we believe the reviewer has helped to improve the ms by focusing it on the cell-type specific insights. On the issue of the use of the data in the wider community, we also now provide a comprehensive flat file that will allow access to the processed data, with average expression of every genes accross cell type for each species (Supplementary Table 1). The raw data, including the integrated Seurat objects, will be available in SRA/GEO.

Additional comments:
Line 29: Suggestion "… comparing gene expression .". "

This has been modified

Line 110: How were orthologs identified between these three species?

We had cited these two publications:

Zhang, Y. et al. Differentially regulated orthologs in sorghum and the subgenomes of maize. Plant Cell 29, 1938–1951 (2017) and Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. U. S. A. 108, 4069–4074 (2011).

We clarify the source of the data in the text.

"We took advantage of prior comparative genomic sequence analyses in maize, sorghum, and *Setaria* that mapped orthologs among the three species, including the homeologs created by WGD in maize(Schnable *et al.*, 2011; Zhang *et al.*, 2017)[11,15] (hereafter subgenome M1 and M2)."

From Methods, "We used prior studies to obtain a list of WGD homeologs in the maize1 and maize2 genomes(Schnable *et al.*, 2011; Zhang *et al.*, 2017)[11,15].

Line 147: Suggestion "... to quantify expression patterns ... "

The particular sentence was removed in the new version of the text, but we made sure to use "expression pattern" instead of pattern alone in the manuscript.

Line 155-156: If maintaining dosage among interacting genes is critical in newly-formed polyploids, why not ancient polyploids? Additional details are needed in the paragraph -- help guide the average reader.

This particular paragraph has been revised. We streamlined the discussion of Genome Balance and dosage compensation in order to focus on the cell-type specific aspects of gene duplication.

Line 254: Subgenome dominance has been shown to be associated with transposable elements (TE) differences. Given that TE may also contain regulatory elements, and that TE abundance is high and variable in gasses, are TEs contributing to the observed expression differences?

We obtained TE genomic mappings from Stitzer et al., 2021, as well as a curated list of genes having TEs (mostly gypsy and copia TE) in their promoters from Amachandran et al. 2020 and performed several analyses to ask whether TEs were associated with dominance vs. subdominance categories or regulatory neofunctionalized homeologs. In brief, we queried whether distance from the homeolog start site (as found in Yin et al., 2022) or presence in proximal promoter, intron, or UTRs showed any signal. While we could recapitulate the relationship between TE distance and subgenome dominance found in Yin et al., 2022, we could not detect any relationship between TE location and any of our categories of differentially expressed homeologs.

Stitzer et al., 2021 -
https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009768#sec018
Amachandran et al., 2020- https://academic.oup.com/g3journal/article/10/12/4387/6048670.
Yin et al., 2022
https://academic.oup.com/mbe/article/39/10/msac198/6709529

We note that the dominant and subdominant categories in cells showed only a marginal enrichment for maize1 and maize2, respectively. i.e., the TE relationship for the subgenomes was not strong enough to emerge in our dominant vs. subdominant classes or in regulatory neofunctionalization. We did not have the space to elaborate on the negative result so these analyses do not appear in the manuscript.

The reviewer provided these references below in relation to comments above.

Mckain et al., 2018 - https://www.biorxiv.org/content/10.1101/352351v1
Peng et al., 2020 - https://academic.oup.com/genetics/article/182/2/503/6062852
Renny-Byfield., 2017 - https://academic.oup.com/mbe/article/34/8/1825/3742530
Zhao et al., 2017 - https://pubmed.ncbi.nlm.nih.gov/29180596/

We incorporated the references provided by the reviewer in the appropriate points of the ms.

We thank the reviewer for the constructive critiques.

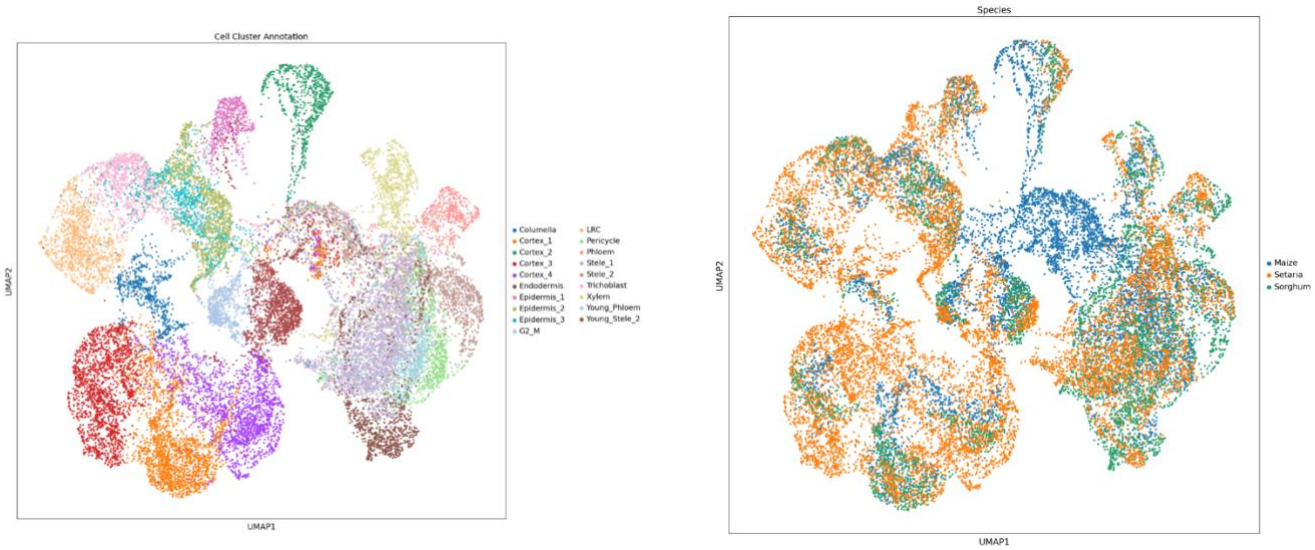
Response to Reviewer 4
Referee #4 (Remarks to the Author):

Here, Guillotin et al. compare all cell transcriptomes across three different types of grass (maize, sorghum, and Setaria). Using a combination of single-cell and single-nucleus sequencing, the authors show (via comparative cellular analysis) the more rapid divergence of some cell-type transcriptomes, likely by the recruitment of gene modules from other cell types. By investigating maize whole genome duplication, the authors demonstrate dosage compensation in surviving co-expressed homeologs – speculating that this effectively maintains

a stoichiometric balance to allow for homeolog retention while new functions arise. My following comments focus on the single-cell analysis, which is performed rigorously.

- The authors use a Seurat-based integration method to assess cell type across the key crop species. One concern is that this method can over-integrate closely related cell types. The authors should support their initial cell-type mapping by using additional integration methods.

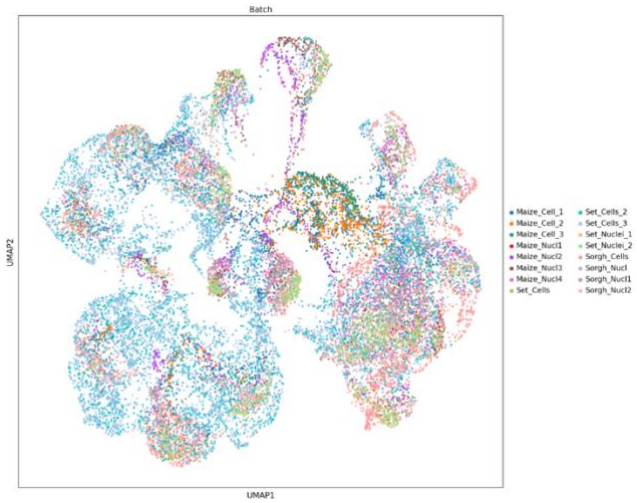
The reviewer raises a good point that it is important to confirm integration with multiple methods, as Seurat can potentially over integrate clusters, mixing cell types that should be distinct. To confirm that the Seurat integration did not over cluster and that the clustering was replicable, we instead used the Python supervised integration method scGen. This method uses a variational autoencoder to learn the underlying latent space for the cell types, and then batch corrects using this learned model. In our experience, supervised integration methods are more effective for cross species integration in plant data, which is a challenging integration scenario. We show that we are able to successfully integrate the cross-species nuclei and cells following filtering of nuclei for sufficient read depth. Following integration, we place 88% of cells into a cluster that contains multiple species, failing only to integrate a small group of maize cells. Additionally, we show that our integration is accurate, integrating 79% of cells into a cluster primarily of cells of the same cell type. Cells inaccurately placed are primarily in a cluster composed of young stele, stele, pericycle and young phloem, showing that the issue may be slight *under*-integration of closely related cell types, not over-integration. We now elaborate on this cross check of integration methods in Supplementary Figure 6b,c and the methods.



Extended Data Figure legend 6b-c: “b-c UMAPs generated by additional integration of the dataset using a Python supervised integration method scGen. This method uses a variational autoencoder to learn the underlying latent space for the cell types. b Different colors represent the clusters identified by the Seurat integration mapped onto the new scGen integration, showing Seurat classification was in relative agreement with the scGen classification. i.e., scGEN clusters

have relatively homogenous coloration. c The same UMAP as in (b), this time showing the species distribution. Overall, each cluster has cells from each of the three species."

We did not include the figure below but it shows cell types from the different species mapping to the same clusters using the alternative integration method.



-

From Methods:

"Validation of Integration using scGEN

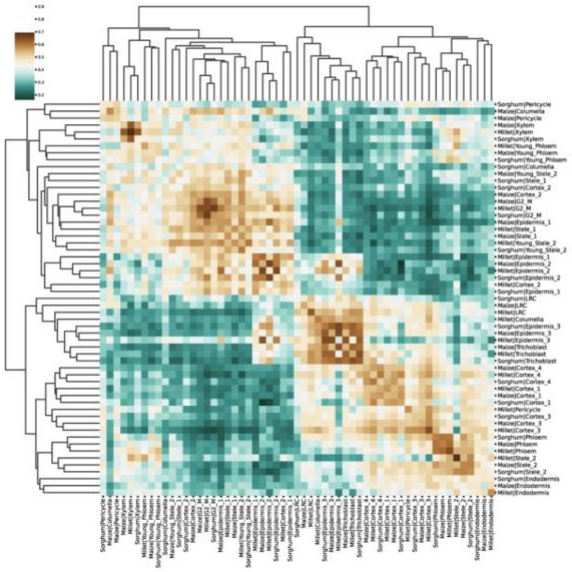
To evaluate the integration of nuclei and cells across three plant species, we repeated the integration using the supervised integration method scGEN (Lotfollahi *et al.*, 2019)[33]. We utilized scGEN version 2.1 to train a model using the scgen.train function, and utilized the scgen.model.batch_removal function to correct our data. Following correction, we utilized the Scanpy (Wolf *et al.*, 2018)[70] calculate the nearest neighbors using scanpy.pp.neighbors, and generated a 2D projection using UMAP, via sc.tl.umap. We then used sc.tl.leiden clustering algorithm at a .6 resolution to identify clusters, which we evaluated for mixing and accuracy of integration."

MetaNeighbor is adapted to quantify the similarity of cell clusters across datasets using a given marker set of genes and their orthologs. Again, I am concerned about using a single method for this analysis and that this may be heavily influenced by the integration method used. Can the authors also train a cell-type classifier to support these data?

We understand the reviewer's concern about the replicability of our results given the integration method, and that MetaNeighbor may be influencing the results. To investigate the impact of the integration method on the results, we first evaluated the unintegrated data using MetaNeighbor. Metaneighbor is designed to work on unintegrated data by stratifying comparisons across the datasets. We found that the results on unintegrated data recapitulated what we saw in the integrated data (Extended Figure 6a). While most cell types were most similar to their homologous type across species as we expected, several cell types did not cluster with their homologous types. These types included columella, pericycle, cortex subcluster 2, LRC,

epidermis subcluster 1, supporting our post integration conclusion of divergence in several of these cell types.

a



Additionally, we trained a different cell classifier, Cell Typist, on our Setaria data, and attempted to classify our maize and sorghum data using this Setaria-based model. We used our original cell labels, as the classifier was unable to successfully learn finer cell subclusters using the more refined cell labels. We then calculated the percentage of each cell type we were able to correctly classify using this model. The percentage of successfully classified cells was moderately correlated with our MetaNeighbor results, (Pearson's r = 0.52 for Setaria -> sorghum, Pearson's r = 0.41 for Setaria -> maize). This performance difference was likely a result of MetaNeighbor's design being uniquely suited for the task. Unlike other, more complex methods, MetaNeighbor evaluates the transcriptional profile of each cell type without training a complex model, allowing it to function well on much smaller sample sizes. Because of this, we are able to evaluate cross-species data for which we are underpowered to train a deep learning-based classifier. It is these "failure-to-learn" cases that bring the correlation down to ~0.5, but our observations of divergence are still confirmed; i.e., the new method was not able to learn a model where MetaNeighbor also observed divergence. For example, the model struggles with columella, several of the stele cell types, and cortex subcluster 3 – cell types we identified as diverging rapidly in maize and sorghum in comparison to Setaria. We point out that the duplicate gene analysis focusing on domain expansion concurs with these results, showing columella and vascular cells have diverged most dramatically. We did not include this analysis in the manuscript.

- The data presented show that transcriptomes of columella, phloem, cortex subcluster 3, endodermis, pericycle, and stele cell types are most divergent compared to Setaria. Again, it would be essential to demonstrate the true biological differences instead of potential batch effects.

To address this question, we performed a more thorough analysis of potential batch effects: first, we generated additional data, adding Setaria cells so that each species had a similar number of cells and nuclei. With this new data, we compared the cell type similarity between each individual batch in our unintegrated data to see if batch effects were driving clustering in

the data prior to correction. We find that our data is primarily clustered by cell type, and that there are not strong batch effects present in the data driving clustering. Overall, the results did not change upon adding several new batches and, as demonstrated above, we can show robustness across integration methods and the lack of strong batch effects in the raw, unintegrated data. We point out that we also checked the inferred expression patterns of genes across maize and sorghum using extensive *in-situ* and spatial transcriptomics analyses in Supplemental Fig. 7 and 8. Given these analyses, we are confidence in our results demonstrating biological differences.

- The authors use a measure of co-expression conservation to examine divergent expression patterns across species in co-expression networks. Here, the authors could consider (if possible) using an algorithm such as SCENIC to infer gene regulatory networks from their single-cell profiling. This analysis may give some additional insight.

This was an excellent suggestion. We utilized the MINI-EX pipeline to infer gene regulatory networks in each of our 3 species. As MINI-EX requires a list of transcription factors and their motifs for each species, we used our orthologs to transfer the maize transcription factor list to Sorghum and Setaria. Running MINI-EX, we found 15 regulons that were highly conserved across species, and biologically plausible for the cell type they were identified in. This analysis also showed that regulons were often "swapped" between cell types in different species. This fit the pattern we had already observed in the swapping of the mucilage module between sorghum/Setaria and maize. The regulon analysis reinforced a major theme in the ms that cell type evolution on this scale involved swapping gene expression modules between cells. The new regulon analysis appears in the main text.

In the section: "**A Pan-Transcriptome Map Identifies Conserved Markers and Gene Modules"**

"One potential use of cell type-specific pan-transcriptome data is to search for highly localized and conserved gene expression modules. We used MINI-EX to identify cell type-specific networks across the three grass species (Ferrari *et al.*, 2022)[34]. The analysis revealed 15 transcription factors (TFs) and putative targets (regulons) conserved in specific cell types across all three species (Extended Data Fig. 9a, Supplementary Table 5). In five of the fifteen cases, mutants in predicted TFs or direct Arabidopsis orthologs have been shown to exhibit cell type-specific phenotypes corresponding to the conserved regulon localization (Ingram *et al.*, 1999; Donner *et al.*, 2009; Guo *et al.*, 2020; Li *et al.*, 2020; Wang *et al.*, 2022)[35-39]. These results highlight the ability of comparative cell type analyses to reveal conserved cellular mechanisms across species and connect specific genes to specific cellular functions."

In the last section: "**Root Cap "Slime" Drives a Case of Rapid Cell Type Divergence"**

"Prior studies in animals have identified cooption of gene modules from one cell type to another as a mechanism of cellular diversification (Arendt, 2008)[63]. We asked how frequently gene expression modules, such as the mucilage group, switched cellular localization by focusing on regulons that have different cell type-specific expression patterns in maize compared to sorghum and/or *Setaria* (swapped regulons). Although annotated regulons comprise just a subset of all potential TF-downstream targets, we identified more than 50 swapped modules across cell types. The swapped modules are prime candidates for genes that could confer differences in cellular traits between maize and related species (Supplementary Table 5)."

Reviewer Reports on the First Revision:

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have done an outstanding job of addressing my concerns. I particularly want to highlight the expansion of the number of cells in the Setaria dataset which was not something I asked for or felt was necessary to address my concerns, but which significantly strengthens the paper and the authors ability to infer true gain of function regulatory events on duplicated maize paralogs vs loss of function regulatory events on the equivalent sorghum co-orthologous gene.

I have no further concerns regarding the methodology, significance, or robustness of this manuscript.

Referee #2 (Remarks to the Author):

The authors opted out of expanding their analyses to another distantly related maize cultivar, however addressed all other concerns of mine. I think the manuscript is clearly more informative and compelling in its revised form. Many of the descriptions were changed and they elaborated how and why they did what, which makes it much better to understand. Overall, I think the manuscript is potentially valuable, although still rather descriptive.

Referee #3 (Remarks to the Author):

The reviewers adequately addressed my previous set of comments and concerns, including now having more robust estimates of ancestral states.The revised manuscript puts their findings in the context of previous research in polyploid for maize. No additional comments or concerns. I anticipate that this will be a highly cited and impactful to the broader plant science community.

- Patrick Edger, Michigan State University

Referee #4 (Remarks to the Author):

The authors have addressed all my previous concerns.

Author Rebuttals to First Revision:

Referee #1 (Remarks to the Author):

The authors have done an outstanding job of addressing my concerns. I particularly want to highlight the expansion of the number of cells in the Setaria dataset which was not something I asked for or felt was necessary to address my concerns, but which significantly strengthens the paper and the authors ability to infer true gain of function regulatory events on duplicated maize paralogs vs loss of function regulatory events on the equivalent sorghum co-orthologous gene.

I have no further concerns regarding the methodology, significance, or robustness of this manuscript.

Thank you.

Referee #2 (Remarks to the Author):

The authors opted out of expanding their analyses to another distantly related maize cultivar, however addressed all other concerns of mine. I think the manuscript is clearly more informative and compelling in its revised form. Many of the descriptions were changed and they elaborated how and why they did what, which makes it much better to understand. Overall, I think the manuscript is potentially valuable, although still rather descriptive.

We felt adding another distantly-related cultivar of maize is potentially valuable but another question that requires a separate treatment. On the descriptiveness of the ms, we did not feel a functional case study was the best way to present the new perspective on cellular evolution or validate the data. Rather our main goal was to first validate the unprecedent dataset with extensive in-situ and spatial transcriptomics, which sets up an array of functional analysis for the community, and then analyze the evolutionary trends that the validated dataset uncovered.

Referee #3 (Remarks to the Author):

The reviewers adequately addressed my previous set of comments and concerns, including now having more robust estimates of ancestral states.The revised manuscript puts their findings in the context of previous research in polyploid for maize. No additional comments or concerns. I anticipate that this will be a highly cited and impactful to the broader plant science community.

- Patrick Edger, Michigan State University

Thank you.

Referee #4 (Remarks to the Author):

The authors have addressed all my previous concerns.

Thank you.