Differentially Regulated Orthologs in Sorghum and the Subgenomes of Maize

Yang Zhang, Daniel W. Ngu, Daniel Carvalho, Zhikai Liang, Yumou Qiu, Rebecca L. Roston, and James C. Schnable


Corresponding author: James C. Schnable schnable@unl.edu

Review timeline:

<table>
<thead>
<tr>
<th>Manuscript Number</th>
<th>Submission received</th>
<th>1st Decision</th>
<th>2nd Decision</th>
<th>Final acceptance</th>
<th>Advance publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPC2017-00354-RA</td>
<td>May 5, 2017</td>
<td>June 2, 2017</td>
<td>revision requested</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPC2017-00354-RAR1</td>
<td>July 5, 2017</td>
<td>July 6, 2017</td>
<td>acceptance pending, sent to science editor</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>July 18, 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>July 21, 2017</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

REPORT: (The report shows the major requests for revision and author responses. Minor comments for revision and miscellaneous correspondence are not included. The original format may not be reflected in this compilation, but the reviewer comments and author responses are not edited, except to correct minor typographical or spelling errors that could be a source of ambiguity.)


Each of the editors appreciated the cross species comparison. The major criticism is that the outcome is descriptive based on a very limited RNA Seq dataset, which yields rather weak inferences. The single RNaseq time point for cold stress was viewed as a major flaw. It is possible that the lack of correlations may be because one or two of the species up- or down-regulate the orthologous gene at a different time, and one cannot get a sense of this in the absence of a time course. The mix of genomic and transcriptomic data sets complicated the analysis. The editors concluded that these shortcomings limited the biological inferences, which is a major criterion for publication in The Plant Cell.

TPC2017-00354-RA  1st Editorial decision – revision requested June 2, 2017

The reviewers liked the novel approach to comparing gene expression between species. You should note in the revision that this study serves as a proof of concept for such studies.

Note that the sampling and nature of "biological replicates" should be described precisely (i.e. different plants, parts of plants, pooled tissue, independent pools of tissue, sampled at different times, etc.). The reader should know exactly what was sampled; what forms the basis of the calculation of any means and statistical parameters reported. This is also necessary to ensure that proper statistical analysis was conducted.

The figures are not very attractive and need attention from a professional illustrator. Please see the attached file for some examples of specific issues. You will need to download the file and view in comment mode in Acrobat. Apply these comments to all figures, including the supplemental figure. Supplemental Figure 7 is egregiously poor in terms of reader accessibility. Note that all supplemental materials must be in Arial or Helvetica except for nucleotide or protein sequences, which may be in Courier or other non-proportional font. You should consider using the freeze
pane option in Excel for the supplemental table. The tables are not presently acceptable in PDF format, but we assume that they will be available in .xls format for the reader. The journal requires that all new data be made available in the main figures. The supplemental figures should be reserved only for large datasets, reagent, genotype or method validation, replicate experiments or evaluations, etc. Please ensure that this is the case because you have a number of supplemental figures. The figure titles could be improved as well. For instance, it is not necessary to indicate in the title "Box plot..." or "Scatter plot...". You may consider titles that are declarative of the take home message from the figure. As a young investigator, we hope that the suggestions on the figures will be useful to you for all future submissions. It is important that effort be put into presentation of the research in proportion to the effort you and your coauthors have put into its execution.

Reviewer comments:

TPC2017-00354-RAR1 1st Revision received July 5, 2017

Reviewer comments and author responses:

Reviewer #1

Point 1. The Introduction needs to do a much better job of putting this work in context. Davidson et al. (2012) is not mentioned here (though it is elsewhere), nor are other seemingly relevant papers such as Wang et al. (2014).

RESPONSE: These are both very important papers in the field and we apologize for their omission from the previous draft of the introduction. Both have been added in this draft (Line 96-97, page 4.)

Point 2. The authors present alignment to a common reference genome as a problem, but then do not really address that in this paper. Where does this paper fit in to the existing literature? What are the authors doing that is unique or special?

RESPONSE: In discussions prior to submission Burkart-Waco et al.’s work, using species closely related enough that all reads could be aligned to a common reference genome was brought up as potentially relevant to our work with grass species that are sufficiently diverged that alignment to a common reference genome was impractical. We agree with the reviewer that as currently written, this section is confusing and potentially misleading to the reader and have revised it substantially to highlight the differences between the analytical problem we seek to address and previous work. (line 85-134, page 3-5).

Point 3. How sensitive are comparisons of ka/ks and open chromatin to expression level? The authors do a minimum cutoff of RPKM>2 in the supplement, which to my eye seems to have an effect. Can this be quantified? I worry that there may be more power to detect DRO/DE2 for genes that are highly expressed in both species, and genes that are more highly expressed are expected to have lower ka/ks and more open chromatin. Efforts to tease these apart (for example, perhaps make comparisons of ka/ks and MNASE in bins of expression?) would be good.

RESPONSE: The reviewer brings up an important point. The new Figure 1B shows the distribution of expression levels in maize and sorghum among DE0, DE1 and DE2 genes. We do not observe an obvious bias towards DE2 genes being more highly expressed in either or both species. However, to test whether the pattern of open chromatin accessibility we observe is the result of DE2 genes or DROs tending to be genes with higher expression levels and/or lower Ka/Ks ratios, we divided our gene set into three tertiles based on both expression and Ka/Ks ratio (Supplemental Figure 6), 9 total bins. While the resolution of this analysis is much lower, as the number of genes per category per bin can be quite small, the link between conserved differential expression and open chromatin in the promoter is visible in both the first and last tertiles of gene expression. (Line 149, pages5; line 583, page 20)

Point 4. I’m not totally clear how much the estimates of shared regulation (18% etc.) depend on the statistical power to identify DE genes within each species. If the power is low, won’t that reduce the amount of shared genes? Some
RESPONSE: Again, this is an important point. In the revised manuscript, we took two approaches to addressing the concerns with regard to the power to detect differential expression. In the first we used simulated data from 1000 randomly selected maize genes from our dataset, with 10% (100 genes) simulated to have true changes in expression. The results are included in a new supplementary table 2, and reproduced below.

In addition, based on Schurch et al.’s work with yeast RNA-seq datasets (citation below and added to the main text), the statistical power of DESeq2 to detect differentially expressed genes with a minimum 2 fold change in expression log₂(Treat/Control) cutoff and 3 biological replicates was estimated to be in the range of 0.65-0.90.

Based on both the observed power in yeast using DESeq2 with the same fold change cut off and number of replicates, as well as the simulation study above based on our own data, it is possible to calculate the expected proportion of the set of gene pairs classified as differentially expressed in either species, which would be identified as differentially expressed in both under the assumption that all cold-responsive expression truly is conserved between the two species.

The expectation for this proportion (DE in both/DE in either) was calculated using the formula Expected proportion = power*²/1-(1-power)*². Based on our simulated results below, DESeq2 does a good job of controlling FDR; however, if the false discovery rate were high, the formula above would need to be revised to include the frequency of false positives, which would increase the value of the denominator.

In the lowest power scenario (0.628 based on a ratio of 2, much lower than our observed median ratio of 3.3) we calculate that, if every gene pair had conserved cold responsive gene regulation between sorghum and maize, the expected proportion of DE2/(DE2 + DE1 maize + DE1 sorghum) would be 46%, which is significantly higher than our observed value of 25%.

We have revised the section of the Results focused on estimating the proportion of genes exhibiting conserved regulation to a discussion of the implications of false negatives in differential gene expression analysis on our estimates. (Page 9, lines 279-291).

Point 5. Why is it necessary to invoke neutrality? Why not stabilizing selection on cold response? If it's highly polygenic, individual loci could change in frequency while response could stay the same. Naively, I might expect under such a model to see greater conservation for genes of large effect. Perhaps worth considering in the Discussion at least other models that might explain the data, especially if these lead to testable predictions with the given data.

RESPONSE: Our initial assessment was that neutrality was likely to be a more parsimonious explanation than something like stabilizing selection. However, we agree with the reviewer that stabilizing selection on a large population of genes where a transcriptional response to cold of any individual gene has some fitness impact is also consistent with our results. We have updated (Page 21, line 610) the Discussion to incorporate this as an additional model.

Point 6. Why analyze CO₂ assimilation for 6 species? Just to make the point that maize and sorghum are both cold sensitive (which seems evident from the citations in lines 145-146)? It might be clearer to drop the other species and focus on just maize/sorghum.

RESPONSE: We apologize that the justification for the inclusion of the CO₂ assimilation curves for multiple species was unclear in the previous version of the manuscript. Our goal for this figure was to demonstrate both that the protocol used to apply cold stress in this paper was successful at distinguishing cold-sensitive and cold-tolerant panicoid grass species and to place the degree of similarity in phenotype between maize and sorghum in the context of the degree of variation present within the clade of panicoid grasses. We have revised the text and legend to state the purpose of this
graphic more explicitly, and revised the formatting of this figure to improve readability. The revision version of this figure is Fig. 2D (Page 6, line 173-183).

Point 7. Line 169: If maize2 genes are expressed at a lower level, shouldn't that subtract from the DE1-maize and DE2 categories, but not the DE1-sorghum category? Why are there fewer sorghum-specific genes for maize2?

RESPONSE: The previous draft did not clearly enough explain that the results on Line 169 and Figure 1 (now Fig. 3) incorporate data from maize1-sorghum or maize2-sorghum gene pairs including those cases where only one gene copy (either maize1 or maize2) is present in maize. So the number of DE1-sorghum genes is lower for the sorghum-maize2 comparison because the total number of gene pairs is also lower. We have revised the text, including moving additional material into primary figures in an attempt to ameliorate this potential source of confusion (Page 11, line 294-304).

Point 8. The authors raise the concern of cross-species comparisons when the two species have different baseline levels of expression. I think DESeq2 solves this by including species as an effect in the model; this should be clarified. If so, I don't understand why different baselines is a novel problem and/or applying a standard linear modeling approach a novel solution.

RESPONSE: I'm guessing this is the key point that we failed to communicate well. The problem is that there are different models for incorporating a species-specific effect into a differential expression model (see new Fig. 5B). Testing for differential expression with a species specific correction factor in the model using ANOVA will classify the example gene calculated under the multiplicative expectation as differentially regulated between species and the example gene calculated under the additive expectation as not significantly differentially regulated between species. DESeq2, using a Generalized Linear Model, will make the opposite classification of these two example genes (see also Supplemental Table 3 in this revised version).

Previous work looking at different genotypes of the same species have used either only one of these two models (such as Lovell 2016, now cited in the main text) or used both (such at Waters 2017, now cited in the main text) but have not attempted to test which of these two approaches is a more accurate reflection of the underlying biological reality of how genes are regulated.

We’ve re-written text, introduced additional supplemental data (Supplemental Table 3), and revised Fig. 3 (now Fig. 5) (page 14-15, line 403-414).

Reviewer #2:

The manuscript from Zhang and colleagues addresses a really important question - how do we compare gene expression patterns across related species. The authors do a really nice job of describing this problem and ways to think about it. They find evidence for shared and species-specific responses to cold stress in maize and sorghum. They also find some evidence for differences in selective pressure and chromatin state for genes with conserved and variable responses. In the end, this manuscript does a nice job of developing some tools for assessing and describing patterns across species but lacks a strong biological take-home message. One issue is that the authors provide figures that have relatively little biological data. A number of the supplemental figures are nice and could be worked into the main text (detailed below).

Point 1. The authors do a good job of identifying syntenic gene pairs (a strength of this group) for these analyses. They demonstrate that there is similarity in expression between species (Figure S1 is nice and could go into the main text). The authors do a nice job of providing phenotypic data documenting the effects of prolonged cold stress on maize, sorghum and other grass species.

RESPONSE: Thanks for the comment; the majority of Supplemental figure 1 has been moved to Fig. 1 in this revised version.

Point 2. The authors then proceed to call DE genes in cold vs. control for each species. They provide evidence for more species-specific DE than shared responses. However, the shared responses are found more often than expected by chance. I would have been quite interested in seeing a better way to visualize this data. A scatter plot of
cold/control in sorghum vs. maize or a clustered heat-map of the cold/control ratios for all genes that are DE in at least one species might provide better visualizations of this data. Currently it is not clear whether the genes are responding in both but are only significant in one species or whether responses are truly only happening in one species.

RESPONSE: Thank you for this helpful suggestion. A scatter plot of all DE genes of cold/control ratios following the format proposed by the reviewer has been added to the revision as Fig. 3B. We have merged the following two comments together, as we feel they both speak to the same issue with the previous draft.

Point 3. The authors proceed to generate data for a time-course series in both species and to analyze DE at multiple time points. This analysis shows that the level of shared responses increase at later time points. This data could have been presented in more depth. One weakness of this section is that the authors make substantial efforts to describe the logic and the test (which they do nicely) but then only show a very brief summary of the results. It is important for the authors to have a better description of these findings and the interpretations.

RESPONSE: We would like to thank the reviewer for pointing out this gap in the previous version of the manuscript. We have revised the figures to provide more information to the reader on the results of the time course analysis. In addition, we have included an additional paragraph in the Results looking at the biological pathways that show consistent patterns of regulation between maize and sorghum and the absence of function or pathway enrichment among the genes that respond to cold only in one of the two species examined (Page18-19, line 484-514).

Point 4. The authors then look at altered expression for orthologs in maize. I was not convinced that this was the best way to evaluate the relative effects on expression in these pairs. I might be missing something, but one could imagine using an ANOVA in which you assess the species and treatment (and species x treatment interaction) as a more powerful way to identify the cases described in Figure 3B, but I might be missing something here. Lovell et al., 2016 Genome Research provided a model for doing this. While they were studying genotype variation, you could replace genotypes with species in this model.

RESPONSE: Lovell 2016 is an important reference and we apologize for its omission from the previous draft. In addition to Lovell 2016, we also identified Waters 2017 based on the terminology used above. Waters 2017 discusses both the DESeq2 interaction-based approach and the ANOVA-based approach. We tested both approaches using simulated data and found that where genes were regulated in an additive fashion, an ANOVA-based approach would be more suitable for identifying genes with significant changes in gene regulation, the opposite pattern from that of DESeq2 interaction model (See Supplemental Table 3). We have added a discussion of ANOVA as an alternative approach to our Results section. (Line 403-414, page 14-15)

Point 5. The last portions of the Results provide some comparisons of properties of the genes with conserved response relative to species-specific responses. They find lower Ka/Ks for genes with conserved responses. They also assessed CNSs and predicted TF binding sites for these sets of genes (Figures S5 and S6). These were interesting and I would encourage the authors to find a condensed version of these supplemental figures to include in the main text.

RESPONSE: Thank you for this suggestion. We have added a subset of the CNS results to the new main Fig. 6C in this revision.

Point 6. They also look at chromatin accessibility for these sets of genes. The authors note that conserved responsive genes have more open chromatin than other genes. I was not fully convinced by this interpretation. The plots in Fig 4C-D are quite interesting. However, it was hard for me to see that this was a statistically significant difference in the DE2 versus the other classes of genes. To me, it seems that all sets of genes have fairly similar profiles.

RESPONSE: Thank you for bringing up this concern. We have made two modifications as a result. The first is to plot the open chromatin data (now in Figure 7) on a log scale to increase the visibility of
different patterns. The second is to mark two standard deviations of uncertainty around the DE2 MNase result (95% confidence interval), showing that at variation locations relative to the TSS, this confidence interval is non-overlapping with the observed values for DE1 maize and DE1 sorghum.

Reviewer #3:
Zhang and colleagues present an analysis of gene expression variation for orthologous genes between maize and sorghum under cold treatment. By taking seedling plants from both species under control and cold-stress time points, the authors are able to investigate not only which genes show differential expression within a single species (DE1), but the more complex view of conserved differential expression across species (DE2). This level of investigation may be of greater importance for identifying major stress response genes beyond just a single species.

The authors provide useful data and theory as to the hurdles involved in DE analysis when investigating multiple species and how to overcome them. This includes the variety of pitfalls that can arise due to different underlying assumptions when examining differential expression across species (species x treatment effect). The authors promote a 'multiplicative' model based on predictions available via the unique subgenome dataset available within maize. Overall, the most conserved DE2 genes are found rapidly after stress, display evidence of purifying selection, and often have more open chromatin states in their promoters based on external MNase maize data.

I find this to be a useful paper on an interesting topic that is sure to be of interest as the opportunity for pan-species DE analysis increases. As the authors note, the ability to compare expression levels between syntenic genes may help to identify common regulators and genetic factors important to specific stresses.

Point 1. My main concern is tied to the selection of a multiplicative model as a likely method to separate conserved regulatory orthologs (CROs) compared to those differentially regulated (DROs). In part, why must there be only one possible outcome for the comparison of orthologs? The prediction results indicate that the multiplicative model is better for predicting, but the results still look quite noisy for both. Could orthologs be separated based on the observations alone and look for other molecular factors to more naturally separate the classes? Given, this is no easy task and the authors note the caveats of this type of analysis within their Discussion. If nothing else, the authors do not overstep their data and provide a useful base to build from for future analyses of this complex situation.

RESPONSE: The reviewer raises an excellent point that different types of genes may tend to follow different models of gene regulation. In response to this point, we tested whether maize gene pairs that show models more comparable to the predictions of the additive or multiplicative models tend to be enriched in particular GO terms or display different distributions of Ka/Ks ratios or log2 transformed expression from two time points control. The results of these analyses are now reported in (page 15-16 and line 440-450, Supplemental Figure 4). Overall, we did not find significant differences between the two populations of gene pairs. However, we think the reviewer's idea is a really interesting one and the current set of GO annotations doesn't have a lot of statistical power. We have also added some text to the Discussion (page 23 and line 620-627) describing this point (that it may be possible to separate these classes of genes based on additional molecular traits measured from the genome in the future).

Point 2. The authors also note that in the classification of DE0/1/2 ortholog pairs, the CRO possibilities are complex. Specifically, examples such as CRO DE1 (Fig 3B, SF3) are only just missing classification as DE2. Of the thresholds used for differential expression, how many DE1 pairs are 'close' to being DE2?

RESPONSE: That is a good question. To address it, we used the set of 850 maize1/sorghum gene pairs classified as DE1 Maize in Fig. 3A and the set of 1,507 maize1/sorghum gene pairs classified as DE1 Sorghum in the same dataset. By relaxing the adjusted p-value cut off from 0.05 to 0.1 for sorghum, 22 of the DE1 maize gene pairs were reclassified as DE2 gene pairs. By relaxing the adjusted p-value cut of from 0.05 to 0.1 for maize, 39 DE1 sorghum gene pairs were reclassified as DE2 gene pairs. In both cases, this works out to approximately 2.7% of genes classified as DE1.
Point 3. Given the curious result of MNase data in maize, would it be worthwhile to look through other chromatin datasets to confirm and or expand on the open chromatin observation? Methylation? Histones? Small RNAs? As it stands, it feels a bit tacked on right at the end.

RESPONSE: We chose to use the MNase data because of previous results showing significant correlations between conserved noncoding sequences and open chromatin (Zhang 2012 doi: 10.1101/gr.131342.111, Vera 2014 doi: 10.1105/tpc.114.130609, Lai 2017 doi: 10.1016/j.molp.2017.05.010). We have added citations to these manuscripts to explain the reasoning for our interest in the MNase open chromatin data. We agree that investigating other epigenetic/chromatin state marks would be an interesting path to pursue going forward, however, this rapidly becomes a large open-ended investigation without a clear conclusion. (Line 562-569, page 20)

Point 4. Within the time course results, it is noted that early time points have fewer, but more conserved DE2 observations. As more genes become DE through the time course, are most of the previously identified genes at earlier time points maintained in their DE2 state in subsequent time points? I think the time course could have a bit more in depth analysis in regards to the patterns observed through time, as it is a very interesting dataset. If some DE2 genes drop off over time, is the drop off itself conserved between species?

RESPONSE: Thanks for this question. We have revised this portion of the analysis to incorporate a Sankey Diagram, which should make it possible for readers to get a sense of the patterns and questions the reviewer is asking about (Fig. 4A). Generally, genes tended to remain in the same categories between related time points, with more DE0 genes moving to all three cold-responsive categories in later time points. However, there are certainly significant numbers of genes exhibiting every possible pattern of movement between groups across time points.

Point 5. You note that baseline can differ significantly, but you begin the Results (ln 125) noting how similar they are. Seems a bit of a contradiction.

RESPONSE: We have revised the statement on line 225 (line 396-401, page 14)) and we apologize for the apparent contradiction. Overall expression levels are going to be correlated (genes involved in encoding components of photosystem complexes will be highly expressed in seedlings, most genes encoding transcription factors will be expressed at relatively low levels). Variation on the order to 2x between syntenic orthologs in maize and sorghum would be considered well correlated between comparing expression levels across all the genes in the genome, but not when comparing the same gene across multiple treatments. We have revised our text in both the Introduction (line 91-95, page 3-4; line 104-114, page 4) and Results (line 396-401, page 14) to describe the distinction between these two different standards of comparison.

TPC2017-00354-RAR1  2nd Editorial decision – acceptance pending  July 6, 2017

We are pleased to inform you that your paper entitled "Title" has been accepted for publication in The Plant Cell, pending a final minor editorial review by journal staff. At this stage, your manuscript will be evaluated by a Science Editor with respect to scientific content presentation, compliance with journal policies, and presentation for a broad readership.

Final acceptance from Science Editor  July 18, 2017