Deep conservation of cis-element variants regulating plant hormonal responses

Michal Lieberman-Lazarovich, Chen Yahav, Alon Israeli, Idan Efroni

Plant Cell. Advance Publication August 29, 2019; doi: 10.1105/tpc.19.00129

Corresponding author: Idan Efroni, idan.efroni@mail.huji.ac.il

Review timeline:

TPC2019-LSB-00129	Submission received:	Feb. 23, 2019
	1st Decision:	April 6, 2019 *revision requested*
TPC2019-LSB-00129R1	1st Revision received:	June 29, 2019
	2nd Decision:	July 17, 2019 *accept with minor revision*
TPC2019-LSB-00129R2	2nd Revision received:	July 27, 2019
	3rd Decision:	Aug. 2, 2019 *acceptance pending, sent to science editor*
	Final acceptance:	Aug. 27, 2019
	Advance publication:	Aug. 29, 2019

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.)

TPC2019-LSB-00129 1st Editorial decision – *revision requested* April 6, 2019

We have received reviews of your manuscript entitled "Deep conservation of cis-element variants regulating plant hormonal responses." Thank you for submitting your best work to *The Plant Cell*. The editorial board agrees that the work you describe is substantive, falls within the scope of the journal, and may become acceptable for publication pending revision, and potential re-review.

We ask you to pay attention to the following points in preparing your revision.

1. While the CoMoVa approach is original and powerful, more information about its assumptions, robustness and limitations is required for making this a widely applicable tool.

2. The promoter length is limited to 500 bp, which is perhaps representative for Arabidopsis. However, as a generic tool is presented, it would be important to show application of the approach with extended promoter length, being more suitable for species with less compact genomes.

3. The analysis of the types of genes associated with different AuxREs should be extended.

In addition, please take note of all reviewers' comments and address each point. If there are good reasons why points cannot be addressed, please explain this clearly in your point-by-point response.

---------- Reviewer comments:

[Reviewer comments shown below along with author responses]

TPC2019-LSB-00129R1 1st Revision received June 29, 2019

Reviewer comments and **author responses**:

We thank the reviewers for their constructive and helpful suggestions. We have revised the manuscript, expanded the description of the algorithm and its limitation, and tested its ability to detect RE conservation using varying promoter

lengths. As a result of this analysis, we have found that 1-kb promoter sequence provides better results than 500 bp and have revised the analysis accordingly.

Reviewer #1:

Overall, I found the manuscript interesting and important for a wide audience, those who studies transcription factor binding sites.

We thank the reviewer for the positive comments.

But I have two concerns:
1) Authors did not describe clearly the limitations of CoMoVa method. The assumptions and the limitations are very important to interpret the results.

a. The method considers that the TFBS core sequence is more important for TF binding than the clumping tendency of the TFBS. However, this is not a problem for most TFBS, AuxRE is an unfortunate example, because the spacing between the AuxRE repeats might play more important role compared with the exact AuxRE sequence (Boer et al., 2014; Stigliani et al., 2018). The small number of the genes with a highly conserved AuxRE actually supports the idea that there is another factor, more important than the degenerate sequence of AuxRE variants. I suggest to discuss it both in results and in the discussion.

We agree with the reviewer that TF specificity is govern by many sequence features apart from the core sequence identity, including spacing from other RE. Indeed, we identify broad sequence conservation outside the core motif. However, we think that at this moment, there is insufficient data to quantitatively determine the relative contribution of the different sequence features to promoter activity. To clarify this we added this paragraph to the discussion:

"How promoters encode specific transcriptional responses is still far from being fully understood. The presence of particular DNA motifs contributes to TF binding but cannot explain the complex transcriptional response of the gene. Many sequence features contribute to TF-DNA binding specificity and ability to activate transcription of downstream genes. ARF proteins, for example, form complexes, and preferentially bind RE pairs with specific spacing (Boer et al., 2014; Galli et al., 2018; O'Malley et al., 2016). In this work, we highlight the contribution of specific variant nucleotides in vicinity of the core motif to the promoter transcriptional response profile."

b. It is not clear if the method applicable for more degenerated TFBS, which is difficult to describe by a consensus sequence.

This is indeed a limitation of the technique. We added the following note to the discussion explaining this limitation and its underlying cause:

"Identification of regulatory sequences in promoters is statistically challenging due to the large size of promoters. Methods that can reduce the sequence search space are required to address this problem. By confining the search just to two or three nucleotides in specific sequences contexts, CoMoVa was able to identify significant conservation of many REs even in relatively long sequences. The ability to identify such conservation is increased with larger number of genomes and is reduced by larger sequence search space (longer promoters) or by high-degeneracy of core motif. With the current number of genomes, the most effective promoter length for the method is ~1kb, although we could identify significant conservation even when 2kb and 3kb promoters were used. Further, while the method may not currently be well suited for TF with highly-degenerate binding sites, an increase in the number of sequenced plant genomes can compensate for the low information in the motif and expand the usability of the method."

c. Authors limited the search by upstream regions 500 bp in length. However many AuxRE are known to be placed outside this region. Thus the study on upstream regions of different lengths should be given (as a supplementary materials).

As the CoMoVa algorithm is alignment free, the use of additional sequences increases background levels and makes its hard to identify a statistically significant enrichment. To test this effect, we ran the analysis for promoter lengths of 1 kb, 2 kb and 3 kb. We now show that even at 3 kb, CoMoVa is able to identify motif conservation, however, since the algorithm is alignment-free, additional sequence means high background level and the number

of detected genes with a conserved motif is reduced. Thus, there is a trade-off between the ability to detect conservation and the length of the sequences. We describe this analysis in the text and in Supplemental Figure 2:

"Promoter sequences vary in length between species and RE may be found further than 1kb upstream to the TSS. However, a limitation of alignment-free algorithms like CoMoVa is that longer sequences reduces the ability to identify signal in the data due to higher background. To test the ability to detect of conserved variants at different distances from the TSS, we applied CoMoVa to identify auxRE conservation in sequences of 500bp, 2kb and 3kb upstream to the TSS. Consistent with the higher background, there was a reduction in the number of genes with significantly conserved auxRE for longer promoters. Additionally, analysis of the shorter 500bp promoter resulted in the identification of fewer conserved REs, probably due to loss of data (n=127, 215 and 157 for the 500bp, 2kb and 3kb regions, respectively; Supplemental Fig. 2A). Regardless of the reduced number of genes with conserved RE, significant enrichment was still detected for the TGTCCC/GG/TC/AC variants (Supplemental Fig. 2B; Supplemental Table S4) and there was a large overlap between the genes with conserved auxRE between the different promoter lengths, although some were unique to specific promoter lengths (Supplemental Fig. 2C). GO enrichment test for genes with conserved auxRE in longer promoters resulted in enrichment for similar terms as those detected for the 1kb promoters (Supplemental Table S4), suggesting that there are functional conserved auxRE located at a longer distance from the TSS. However, it should be noted that we cannot rule out that this apparent conservation of distant RE is due to the mis-annotation of the TSS in the available angiosperms genomes. "We also make this limitation of technique explicit in the discussion section:

"With the current number of genomes, the most effective promoter length for the method is ~1kb, although we could identify significant conservation even when 2-kb and 3-kb promoters were used."

As a result of this analysis, we chose to utilize an optimal promoter length of 1 kb instead of the 500 bp used in the original submission. We have updated the manuscript accordingly.

2) The rationale to use the same dataset for foreground and background is clear. However the choice for NNnnnnnnnnTGTC is unclear. AuxRes have a tendency to locate closely to each other (O'Malley et al., 2016, Stigliani et al., 2018), would not it give a bias to your analysis? Why not permute the foreground for each species with the preservation of dinucleotide distribution? Also it is unclear if you use the same rule to calculate background for cytRE and ABRE?

We agree that the choice of dinucleotides at position -8 bp for the background rates was not sufficiently explained.

We added in the text a more detailed description of the rationale behind our computation of sequence conservation background:

"In order to determine the significance of the conservation, the neutral variation rate at promoter positions had to be calculated. Defining a theoretical model for neutral substitution rates at promoters is difficult, as GC content and the distribution of dinucleotides is genome-specific (Gentles and Karlin, 2001) (Supplemental Fig. 1A) and the available genomes are not equally spaced in term of evolutionary distance. Thus, some of the observed conservation may be due to kinship. To correct for both these factors, background rates were determined empirically by calculating the conservation score distribution for the two variant nucleotides 8bp upstream to the core sequence (NNnnnnnnnnTGTC for the auxRE; Supplemental Fig. 1B). auxRE are sometimes found in close proximity to one another (O'Malley et al., 2016; Stigliani et al., 2018), which could affect the distribution of nucleotides at these positions. However, the overall distribution of dinucleotides at position -8 to the auxRE was found to be similar to the overall distribution in promoters, indicating that, genome-wide, this effect is negligible (Supplemental Fig. 1B).

The permutation approach suggested by the reviewer is a possible solution for the problem of computing background rates, but, in our opinion, the use of dinucleotide from promoters of orthologs has the advantage of accounting for kinship and is therefore preferable in this context.

Background rates of cytRE and abRE were computed using similar method: BnnnnnnnnACGTGnnnnnnnnK for abRE and DnnnnnnnnGATCnnnnnnnnYN for cytRE. We now make this explicit and added to the text:

For cytRE enrichment: "Similarly to the auxRE, we used matching adjacent sequence (DnnnnnnnnGATCnnnnnnnnYN) to compute the background conservation rates."

For abRE conservation: "(motif: BACGTGK, background: BnnnnnnnnACGTGnnnnnnnnK). "

Reviewer #2:

The manuscript is generally very well written. There are a number of points, highlighted below, which would improve the clarity and impact of the work.

We thank the review for the positive assessment of our paper.

1. Line 50-51 - This claim should be referenced.

We added the reference (Boer et al., 2014) which provides a well described example for this general claim.

2. Line 105 - which occurrence of the RE (if on average 3.5 auxRE per gene) was used in the remaining analysis in the first section of the MS?

Figure 1 - This is a really interesting figure, but as for my comment above, which of the 3.5 REs, per gene were used in this analysis?

The CoMoVa algorithm treats multiple occurrences of auxRE as a single one. We added this clarification in the text:

"auxRE variants in the promoters of all genes were extracted. Multiple instances of an auxRE variant were treated as one. Thus, the entire promoter sequence was converted into a vector of 16 Boolean values, representing the presence or absence of a specific motif variant in the promoter of a given ortholog. Following ortholog assignment, these presence\absence vectors were arranged according to a predefined species tree representing known phylogenetic relationships, resulting in a Boolean matrix (Fig. 1A-D)."

4. Line 127 - this claim must be clarified or expanded.

The claim "AuxREs are extremely common in the genome, but…" was updated. First, we list the number of auxRE found in the genome (updated to 1-kb promoters):

"A mean 91±9% of genes carried at least one degenerate TGTCNN auxRE, and among those, there were 6.1±0.8 auxREs per gene.".

Second, we clarify the significance:

" Using CoMoVa, we detected highly significant conservation for a 252 genes as compared to the background conservation of the neutral nucleotides at position -8 to the motif."

5. Line 154 - Typo, I believe that "complied" should be "compiled"

Typo was corrected.

6. Line 162 - What was the background data set for this enrichment?

The background dataset was the conservation of nucleotides at positions -8 to the core motif. We added a more in-depth explanation of the background calculation to the results. See reply to comment #2 of Reviewer #1.

7. Line 165-168 - The claim about "guiding negative feedback and cell wall modifications" should be expanded. It isn't clear to me based on this section of the manuscript what evidence supports this claim.

This claim was based on the identification of a conserved motif in the promoter of multiple AUX/IAA genes, which act to provide the negative feedback on the auxin response. Furthermore, multiple cell wall modifying enzymes were shown to have a conserved motif. We tried to clarify this point better in the text:

"Overall, the consistent biological role of genes with conserved auxRE variants, which included well known factors such as the negative auxin feedback genes, AUX/IAA and GH3 genes (Bargmann and Estelle, 2014), suggests that CoMoVa identified a conserved regulatory module acting downstream of auxin. "

8. Line 188 is this a "control" or a "test"?

We have corrected the line to: "As an additional test, we computed the conservation of the abRE motif, which mediates the ABA response and appears in 29±7% of all genes in all species…"

9. Line 199: This is a really interesting claim and should be expanded.

We agree this claim, about the usability of CoMoVa to identify targets for other, non-hormone related transcription factors, is rather terse. We removed it from the result section and now expand on this claim in the discussion:

"Conservation of the variant nucleotides in the core binding sequence was common in genes making up the core feedback response for the hormone tested. This raises the hypothesis that motif conservation can be used to identify TF targets. While this is limited only to conserved targets of the TF, it may help uncover the principle activities of different factors. For example, cytokinins have long been tightly associated with promotion of cell division, but the mechanism by which cytokinins mediate their control of the cell cycle has been elusive (Schaller et al., 2014). Interestingly, apart from the A-class RR, which are known downstream factors of cytokinin, we also identified deep conservation of specific cytREs in promoters of key histones and cell-cycle genes, suggesting that RRs may have direct transcriptional control on the cell cycle machinery. While the focus of this work was on transcriptional response to hormonal transcriptional responses, a degenerate RE was defined to many TFs (Franco-Zorrilla et al., 2014; O'Malley et al., 2016) and a similar approach can be applied to them."

10. Line 203: What does "active" mean in this context?

Meaning bound by transcription factor *in vivo.* We added an explanation in the text: "TF binding sites are widely distributed in the genome but only a subset of them is bound by TF *in vivo*. These TF-bound sites are characterized by sequence conservation in regions flanking the core binding site (Dror et al., 2015)."

11. Line 225: How did you determine that conservation declined with distance from the motif? Is this quantified?

We added a graph (Fig 5B) quantifying the overall conservation of nucleotides relative to the position of the core motif, showing the decline in conservation away from the motif.

12. Line 232: Could this be the result of kinship rather than selection?

Indeed, some of the sequence conservation is due to kinship. However, the fact the nucleotides closer to the core motif tend to be more highly conserved than nucleotides further away from it is an indication for selection. We added an explanation in the text:

"Conservation beyond the motif was more pronounced within plant families and while part of this apparent conservation may be due to kinship, the fact the nucleotides closer the to the core motif are more highly conserved than nucleotides further away from it, suggests that these sequences are selected to serve family-specific functionality."

13. Section beginning Line 282: This section is not sufficient to explain what was found. I'd suggest removing or expanding it.

This section describes the creation of a new auxin reporter based on the knowledge gained in this study. We expanded this section to better describe the results. We hope this better clarified the experiment.

14. Since this is the first release of CoMoVa algorithm (I believe), it would be useful to have more details about how it works.

We significantly expanded the description of the algorithm, its underlying assumptions and limitations across the manuscript, both in the results and discussion sections. We hope that this better clarifies the algorithm.

15. Figure captions of all figures should be expanded for clarity. The figures are for the most part excellent, but with short captions, they took some work to figure out.

Figure 2 needs great explanation:
a. 2A is unclear. What is the y-axis? What is the colour scale?
b. 2B is unclear. Why are only certain motifs shown on each row?
c. 2C is unclear. How are the left and right panels to be interpreted together?

We expanded the description of these panels in the legend.

17. Figure 3: How were the genes selected for panels B and D - what was the cut-off for being "most highly conserved?

We have removed these panels in this version. The list of genes with conserved motifs is provided in Supplemental Tables S3,S5 and S6, and key genes are highlighted in panels 2A, 3A and 4A.

18. Figure 6: This figure is very dense. I'd recommend breaking out the I - K panels to a separate figure.

We have split the figure, as suggested.

Reviewer #3:

The analysis of conservation in TF binding element variants across different plant species is a valuable strategy for the identification of core genes in hormonal pathways, and there is a lot of interest in the plant research community in identifying and functionally validating cis-regulatory elements in plant promoters for gene editing approaches.

We thank the reviewer for the positive assessment of our paper.

In general, the manuscript is clear and concise. However, it needs revisions as highlighted below.

Major revisions and questions:
- Line 293: there is no comparison presented in the paper between pIAAmotif and Dr5v2, so the conclusion is not justified.

We have removed the statement regarding the direct comparison between these promoters.

- When discussing GO terms, I would strongly recommend matching the GO terms presented in Figures. For example, lines 181-183 do not match what is presented in Suppl Fig 3A. This would help readers of your manuscript.

We have corrected the references in the text to match the GO terms in the table.

- Supplemental Fig 3 lacks panel C and I believe Supplemental Figure 3B is the wrong panel (does not match line 187).

We have updated the supplemental figures to comply with TPC regulation. The data previously provided in Supplemental Fig 3 was moved to the main figures with panel 3C.

- Figure 6G: match the names of the construct with those presented in the text. Is cac, c2a?

This error has been corrected and the construct is now named c2a throughout the manuscript.

- The TGTCAC variant is conserved (line 238) but it attenuates the response to auxin more than a non-conserved GC tail. This result on AuxRE analysis is quite unexpected, as the authors also point out. It would be interesting if the authors analyzed and highlighted which genes have this conserved variant and if there is an enrichment for known biological processes/pathways.

The small number of genes with conserved TGTCAC motifs precludes the use of functional enrichment test. By looking at the gene list, we could not identify an obvious distinguishing feature. We include a description of the genes containing a conserved TGTCAC motif in the text:

"Despite TGTCAC motif producing weak response to auxin, 18 genes were shown to have this conserved motif in their promoters, including the auxin transporter ABCB1 (Bailly et al., 2008) and two MYB genes (Supplemental Table S3). However, we could not identify an obvious common functionality for these genes."

- The use of the CoMoVa algorithm is restricted to the analysis of the first 500 bp from TSS and 100 bp downstream. In species with large genomes, it is very unlikely that all relevant regulatory elements are included in these proximal regions and there are many examples of enhancer elements tens of kb away from TSS having a functional role. How

does CoMoVa perform when analyzing longer distances from TSS? Is conservation observed also for more distant REs? If so, it would make it a more valuable tool and should be highlighted.

We have included conservation tests for promoters of 1 kb, 2 kb and 3kb and detail the results of this analysis in the text. Please see answer to Reviewer #1, section 1c.

Minor comments and corrections:

- Line 141: is "TGTCC" supposed to read "TGTCCC"?

Corrected.

- Lines 141-143: It is not possible to see whether the genes mentioned have the TGTCC (TGTCCC?) variant in Suppl Fig 2B. Can the authors present this data in a different way? Or simply add another panel to the figure?

We have updated the supplemental figure to comply with TPC instructions. We now provide this data as Supplemental Table 2.

- Line 147: "...as well as two EXPANSIN two mannose..." please fix

Corrected.

- Line 154" correct "complied" with compiled

Corrected.

- Line 162-163: should add something like "in agreement with previous findings in Arabidopsis and maize (O'Malley et al, Galli et al)"

We have corrected the text to read: "Despite large differences in the length of intergenic regions between species, and in agreement with previous studies (O'Malley et al., 2016; Galli et al., 2018; Zemlyanskaya et al., 2016), auxREs were enriched in the +50 to -250bp region relative to the TSS across all tested angiosperms (Fig. 2D-E)."

- Line 178: if this refers to ten A-class RRs, why mention ARR-B class throughout the manuscript rather than ARR-A? (see lines 39, 69 etc.)

B-class ARR are activated by cytokinin to induce the transcription of A-class ARR. We added a clarification to the text: "Amongst these were highly conserved AGATTT and TGATTT variants in seven of the ten A-class RRs (Fig. 3A). These negative feedback regulators of the cytokinin response are known targets of the B-class ARRs, as shown for multiple species (Kieber and Schaller, 2018)."

- Lines 218 and 270: correct Brassicaceae

Corrected.

- Line 279: is it supposed to read "Figure 6B"?

Indeed. We corrected the error.

- Supplemental Fig.6 qPCR data relative to what control?

Control was *ACTIN*. We added a description of the qPCR to the M&M:

"Arabidopsis seedlings at 10 days after transfer to light were transferred to agar plants containing 1μM 2,4-D, or mock, and incubated for 6h. RNA was extracted from ~100 seedlings using TRI REAGENT (MRC Inc.) according to the manufacturer's instructions, treated with DNAase (TURBO DNA-free kit, Invitrogen) to remove DNA remains, and cDNA was synthesized using SuperScript II reverse transcriptase (Invitrogen) following manufacturer's instructions. qRT-PCR was performed using Absolute Blue qPCR SYBR Green (AB-4162, Thermo Fisher Scientific) in a Rotor-Gene 6000 cycler (Corbett Research) with three technical repeats for each sample. Four biological replicates, from independently grown seedlings, were used. *ACTIN* was used as a reference gene."

- Correct phytozone to Phytozome throughout the manuscript.

Corrected.

- Line 429, cite reference related to dataset used.

The citation (Galli et al., 2018) was added.

TPC2019-LSB-00129R1 2nd Editorial decision – *accept with minor revision* July 17, 2019

We have received reviews of your manuscript entitled "Deep conservation of cis-element variants regulating plant hormonal responses." On the basis of the advice received, the board of reviewing editors would like to accept your manuscript for publication in *The Plant Cell.* This acceptance is contingent on revision based on the few remaining comments of our reviewers. As you will see, these are minor, involve mostly textual revision and perhaps inclusion of a small extension of your analysis.

---------- Reviewer comments:

[Reviewer comments shown below along with author responses]

TPC2019-LSB-00129R2 2nd Revision received July 27, 2019

Reviewer comments and **author responses**:

Reviewer #1:

There is one unresolved issue which, in my opinion, might spoil some of the results. Or might not.
I do understand why the background was determined empirically, and this is a good solution. But please demonstrate that the rule that you use for it (NNnnnnnnnnTGTC) is more or less invariant. I suggest to do the same analysis for at least one more rule (e.g., TGTCnnnnnnnnnNN, or NNnnnnnnnnnnnnnTGTC) and compare the results. If they match - no doubts, if they did not match, suggest and test another rule.
I believe that this option will make CoMoVa more trustworthy and invariant for TFs, which binding sites have a clumping tendency.

We ran the analysis with 4 different rules for background distribution using nucleotides as position -8, -10, -14 and -15. 89% of the significantly conserved motifs were identified as conserved by all four analyses. We think that this demonstrates that the algorithm is relatively robust in term of selection for background distribution. Regardless, the definition of background distribution is customizable and users may choose any distance from the core motif they see fit for the particular TF in question.

We added this analysis as Supplementary Fig. 2 and added the following text:

"To verify that this arbitrary selection for calculation of background nucleotides does not significantly affect the results, we also run the analysis using nucleotides at position -10bp, -14bp and -15bp upstream to the core motif for calculation of neutral rates. The number of significantly detected genes varied only slightly (Supplemental Fig. 2A) and of the 252 genes with significantly conserved motif, 224 (89%) were identified as significantly conserved using all four calculations of background distributions (Supplemental Fig. 2B). "

One more problem is with the logic about experimental testing of TGTCAC (lines 285-291). When you performed the analysis with 1 kb promoters, TGTCAC became not conserved (fig. 2A, lines 142-144). Thus you need to introduce the need to test TGTCAC in a different way (and in the lines 194-195 as well). It is not a big problem, though, as you still have it conserved within 500 bp length promoters.

While TGTCAC is not the most highly conserved motif, it still is conserved even with 1-kb promoters and 18 of the 252 genes have a conserved TGTCAC motif in their promoter. We made this modification to the text to clarify this:

"There was a marked bias in the conserved variants, with CC, GG, GA, and TC being the most highly conserved motifs, but we also detected conservation for AT, AA, AC, AG, CT, and TG."

Lines 259-261: another explanation, there are multiple RE in one promoter, and you mixed functional and not functional RE.

That is a possibility. In this case, conservation is expected to drop sharply outside the hexamer binding site rather than gradually decline, as we observe. It is true that for some genes we do not identify any conservation outside the motif, however, this is the minority and we don't think this is a major factor in our analysis.

Reviewer #3:

I have reviewed a previous version of this manuscript. The authors adequately addressed my main concerns. Below are minor corrections:

Several mistakes are found throughout the manuscript, i.e.:
Lines 121-122:"...gained or lost and between those that are highly conserved."
Line 140: "...for a 252 genes"
Line 187: "...the ability to detect of conserved..."
Line 251: "...also show signs of…"
Lines 475-476: I believe some text is missing in this section.

These errors were corrected.

TPC2019-LSB-00129R2 3rd Editorial decision – *acceptance pending* **Aug. 2, 2019**

We are pleased to inform you that your paper entitled "Deep conservation of cis-element variants regulating plant hormonal responses" has been accepted for publication in *The Plant Cell*, pending a final minor editorial review by journal staff.

Final acceptance from Science Editor **Aug. 27, 2019**