Identification and classification of cis-regulatory elements in the amphipod crustacean Parhyale hawaiensis

Dennis August Sun, Jessen V. Bredeson, Heather S. Bruce and Nipam H. Patel

DOI: 10.1242/dev.200793

Editor: James Briscoe

Review timeline

Submission to Review Commons:	19 December 2021
Submission to Development:	21 March 2022
Accepted:	21 April 2022

Reviewer 1

Evidence, reproducibility and clarity

The contribution by Sun et al. describes a very deep and thorough analysis of an Omni- ATAC-seq approach to identifying cis-regulatory elements in the crustacean Parhyale. This is a resource paper, so it does not explicitly have a research question or conclusions. The findings are a detailed dataset of putative regulatory elements, tested and validated with a number of different computational approaches, and - to a lesser extent - with a number of experimental approaches.

The authors' work is very thorough, and while it may be possible to add more analyses and more validations, the work presented in the manuscript is impressive and stands on its own as a useful body of data. No additional work is needed to make this a complete contribution.

The text is very well written and clear. It is a bit arduous in some places, but that is understandable, given the technical nature of the paper. The figures are clear and many of them are very eye-catching (in a positive sense).

All in all, I have no criticism of this contribution. It is a very carefully executed and thorough analysis.

Significance

I am not aware of any other species outside of the main experimental model organisms for which there is data about putative regulatory elements that is as detailed as that presented in this manuscript. It is thus not only a fantastic resource for people working on Parhyale, but also a model for how such data can and should be generated for other species. The authors say this explicitly in their concluding paragraphs and I agree. The Parhyale community will pounce on this paper as a useful resource, whereas people working on other species might be inspired by it to generate equivalent data for their communities.

I am an evolutionary developmental biologist who has worked on a number of species that are not traditional model species (I avoid the term "non-model", since every species is a model for something). I for one, fall into the category of people who will be inspired to generate equivalent data, although I must confess that I do not have the bioinformatic expertise of the authors, and therefore I am not able to critically assess the specifics of the tools they have used to generate and validate their data.

Reviewer 2

Evidence, reproducibility and clarity

Sun et al used omni-ATAC sequencing that is a modified version of classical ATAc-seq to identify and characterise the cis-regulatory elements in the P. hawaiensis genome. They further use long and short reads to improve upon existing gene annotation for this organism. The in-depth analysis ensures the results and conclusions to be sound however few points below might be needed to be addressed before the acceptance of manuscript.

In Introductory paragraph 2, sentence one, authors suggest that gene regulation plays more important role in evolutionary process than genes. Although a significant amount of research has been dedicated to gene regulation based evolution still this field is in nascent form. For example evidence of inheritance of the gene regulation pattern across generation is scarce and requires more evidence. I suggest authors to modulate the claim that still gene based evolution is the main paradigm instead otherwise.

Authors have repeatedly used S21 and S22 throughout the manuscript to support their claims with clustering etc. May authors shed some light on the differences in replicates for these timepoints. Furthermore, I could not find Fig 3J, perhaps author would like to point out Fig 3H.

The majority of ATAC-seq peaks in the distal intergenic regions is a very surprising result. Authors defend this result by suggesting that this organism has big genome. May author perform a short analysis that shows that these peaks are indeed represent nearby genes or may point towards 3D genome organisation. For example, I see that this genome might have regions in the genomes that are densely organised in gene clusters, in those cases does the pattern remains same i.e he majority of the genes are very distant from each other and hence use vital regulatory elements?

Significance

The study by Sun et al is timely in nature and significantly improve the gene annotation of P. hawaiensis. It definitely advances the current knowledge for this organism regulatory elements. The comparison to other model organisms can be further improved by extending the discussion of the results especially in context of distal regulatory elements. The resource generated will be helpful for the researchers working in the field of developmental biology.

Reviewer 3

Evidence, reproducibility and clarity

In this study, Sun et al. use RNAseq and ATAC-seq in 15 stages of embryonic development of the amphipod crustacean Parhyale hawaiensis to analyse gene regulation genome-wide. They assess the data in multiple ways to provide a more complete genome annotation, understand temporal changes in gene regulation, and identify different classes of cis- regulatory elements including associated GO terms and putative transcription factor binding site enrichment. The authors have made a great effort to account for potential biases in their datasets (one impressive example is the comparison of multiple transcriptome assemblies and the following quality assessment) and I enjoyed reading this manuscript for its great explanations of method usage (i.e., what each bioinformatic package does, why it was used etc.) and the overall style.

I want to make a few suggestions that would make the study - in my opinion - even better:

- I felt that the first paragraph of the introduction is not necessary.
- Use of Genrich: I presume this was run on both duplicates simultaneously? This is not clear from the methods section. It might have implications for downstream analyses (e.g., differential accessibility between time points) because running on both sequencing library replicates simultaneously leads to a single "replicate" of peaks per time point, while running it individually leads to two. However, I have never tested if this actually does make a difference. Maybe the authors have and can comment on this?

- In general, I thought that the bioinformatic methods (i.e., the code or the options used for each program) would have been helpful for my understanding in some cases. The authors say that these will be published on an accompanying GitHub repository, which should be fine if this is sufficient for journal policy.
- The section on the IDE2 models (the paragraph at the end of page 4/beginning of page 5) was unclear to me but appears sound. (The only instance where I didn't quite understand what the program actually does.) Maybe this can be explained a bit easier?
- On page 7, Fig.3J needs changing to 3H. This figure should, in my opinion, also contain the absolute number of peaks for each time point to set the individual proportions into context.
- Last paragraph of the "Improving the Parhyale genome annotation" section: I think this needs to focus on those regions of the genome for which the location is known - after all, the "unknown" regions" could all be "distal transgenic", which would significantly change the relative proportions.
- On page 9, t-SNE is mentioned but doesn't seem to be cited.
- The third paragraph on page 9 ("We evaluated the differences...") should mention the fact that clusters 1 and 2 are the only ones with significant proportions of exonic and intronic peaks. In the accompanying figure (5C), the total number of peaks would again be helpful.
- In figure 5D, I can't quite make out at which stage the dip in the peak of Cluster 8 occurs. This is quite an unusual pattern of accessibility change, and I can't help but wonder if it has something to do with the quality of one of the libraries? Also, the fact that half of the peaks fall into unmapped regions of the genome is unusual, and I feel this deserves more discussion.
- On page 10, the abbreviation PFM appears, but it is only explained in the legend of Fig.4. This should appear in the text.
- The section on "Concordant and discordant expression and accessibility" is the one I disagree most with. The authors seem to suggest that a repressive cis-regulatory module should become less accessible when the gene is activated. However, they leave trans-acting factors completely out of their conceptualisation here. It is in general likely the availability of transcription factors that leads to repression, while the "silencer" can be well accessible in all cells. Moreover, it has become clear in recent years that CRMs are not just repressors or enhancers per se but can act as either depending on the availability of transcription factors. I think these facts could partially explain the weak correlation and should be discussed.

Significance

This manuscript will greatly advance research in the emerging model organism Parhyale through a more complete genome annotation and vast amounts of gene expression and chromatin accessibility data (and accompanying analyses) at various stages of development. However, the impact goes far beyond the Parhyale community, and I believe this paper can be seen as a blueprint for similar studies in other organisms. The excellent documentation and comparison of their bioinformatic methods makes their re-use straightforward and much of the authors' pipeline can be used for a "standard" ATAC-seq data analysis - I am likely to use many of their methods myself. Therefore, I think the audience can range from the "classic" evo-devo community to developmental biologists, scientists interested in gene regulation in general, and bioinformaticians.

My own expertise is in gene regulation through transcriptional control, and I use different seq approaches (ATAC, CUT&RUN, RNAseq) to study this process.

Author response to reviewers' comments

1. General Statements [optional]

We are grateful for the very kind, thoughtful, and detailed comments of the reviewers, which we have strived to fully integrate into the revised manuscript.

Of note are the concerns with the data from stages S21 and S22, which we acknowledge do appear to be qualitatively and quantitatively distinct from the other samples. While we are unable to completely disambiguate meaningful biological variation from technical or experimental noise using our data, we hope a few additional analyses and visualization tools we have included can provide greater confidence in the reliability of our findings.

Additionally, while attempting to evaluate Reviewer #2's suggestions about examining the distribution of intergenic peaks along the genome, we discovered an error in our code that resulted in the improper assignment of peak categories. The error resulted in the improper assignment of intronic and exonic peaks as intergenic peaks. While the largest group of peaks in our dataset remains distal intergenic peaks (30.2%), and distal intergenic peaks remain a larger proportion of our intergenic peaks than proximal intergenic peaks, many of the peaks originally assigned to the intergenic categories have been reclassified as exonic or intronic peaks. We have updated our code and figures upon reanalysis of our data and have revised our findings and discussion accordingly.

2. Description of the planned revisions

Reviewer #3, Comment #3 of 11

"In general, I thought that the bioinformatic methods (i.e., the code or the options used for each program) would have been helpful for my understanding in some cases. The authors say that these will be published on an accompanying GitHub repository, which should be fine if this is sufficient for journal policy."

We are still at work compiling the code for our analyses into a more reader-friendly form and setting up a GitHub repository to enable easy access to more detailed methods for interested readers. Some of the most important settings have been included in the Methods and Supplementary Methods sections, but we hope to include more thorough detailing of our pipelines in the GitHub repository. The raw data for portions of the RNA-Seq and all of the ATAC-Seq data have been uploaded to the Sequence Read Archive, and we are finalizing additional raw data submission. We are also in the process of determining what data to include in our Gene Expression Omnibus submission, which we hope to include all pertinent final data analysis files as well as any intermediate or accompanying datasets which would facilitate downstream analyses. The large size and number of our final analysis files has resulted in some challenges with data transfer and storage, which has delayed the upload and submission process.

We are also collating several of the data visualization scripts built for this manuscript into a Jupyter notebook. This tool will enable the visualization of ImpulseDE2 models and peak classifications for arbitrary genes and genome regions of a user's choice, alongside additional functions which are discussed in this revision plan.

3. Description of the revisions that have already been incorporated in the transferred manuscript

We have addressed the following substantive concerns with the manuscript:

Reviewer #2, Comment #2 of 3:

"Authors have repeatedly used S21 and S22 throughout the manuscript to support their claims with clustering etc. May authors shed some light on the differences in replicates for these timepoints. Furthermore, I could not find Fig 3J, perhaps author would like to point out Fig 3H."

Reviewer #3, Cross-comment #2 of 3:

"Focus on stages S21/S22: This might indeed be somewhat problematic. The libraries from these two stages (particularly S21) seem to be very different from those from the other stages. In the PCA (Fig. 1C), S21 doesn't cluster well with anything, and the difference between the two replicates is massive compared to other stages. The accessibility pattern (Fig. 1D) also looks odd. The libraries also have the lowest scores for % of mapped reads (Fig. S2B), fragment size distribution (S2E), and Spearman correlation (S2I). All this could be biologically sound and be due to a major developmental transition at this point, but maybe it justifies revisiting the data and testing whether leaving out S21 (and/or S22) makes a big difference for the clustering analyses."

1. Reviewers #2 and #3 discussed concerns with the outlying nature of libraries S21 and S22. We had also previously held concerns about these samples and had performed some analyses to examine whether the global properties of our dataset are dramatically changed upon removing those samples. We did not observe dramatic changes to the structure of our data in the absence of the S21/S22 samples.

a. Samples S21 and S22 appear to be highly separated from the rest of our data using Principal Components Analysis. We had also previously believed that this suggested that these samples might be problematic. However, a colleague indicated to us that researchers in microbiome ecology had observed similar phenomena, often caused by strong single axes of variation (or "linear gradients") in the datasets. In "Uncovering the Horseshoe Effect in Microbial Analyses" (mSystems, 2017) by Morton et al., the authors describe how a strong linear gradient can create a "horseshoe effect" or "Guttman effect", where PCA results in the two ends of a linear gradient appearing to come together in ordinal space. The authors also describe a similar "arch effect" which strongly resembles the general shape of our PCA curve. We suggest that the strong apparent "outlier" appearance of S21 and S22 may be exaggerated or induced by the technical "arch effect" phenomenon, and may be caused by a strong single biological gradient – a developmental timecourse – which our data aimed to capture.

NOTE: We have removed unpublished data that had been provided for the referees in confidence.

b. We also performed PCA on our dataset with the S21 and S22 time points removed prior to performing the analysis (see right panel, bottom). When we did so, we observed that the relative positions of the remaining libraries remains largely similar, with time points closer to the middle of development showing a positive loading in PC2, and time points closer to the beginning and end of development showing a negative loading. This suggests that the second major axis of variation in our dataset would remain a contrast between middle vs. terminal timepoints, even without the S21/S22 data, and that the relative positioning of the remaining data within PC-space is not entirely driven by S21/S22.

c. To further assess the degree of the S21/S22 samples' outlying effects, we also performed ImpulseDE2 analysis to generate model fits without S21/S22 data. Doing so allowed us to determine to what degree the S21/S22 stages are necessary for driving the accessibility trajectory of individual peaks, and of the data more broadly. We performed IDE2 with either all data, or the S21/S22 data removed prior to input into IDE2. This generated two sets of model fits to the "cloud" of accessibility vs. time measurements: one that included the S21/S22 data, and one without. We evaluated, for each peak in our dataset, the time point at which the IDE2 model achieved maximum accessibility (the "IDE2 max fit"), and plotted both the "all" and "noS21S22" data as a histogram (see right panel, top graph). The presence of peaks that achieve predicted maximum accessibility in the S21/S22 stages in the "no S21/S22" data is a result of how we calculate "max fit", which does not require that there is a known accessibility value at a given timepoint; only that the time point during which the model fit is maximum is closest to the timing of that developmental stage. Overall, we still observed early, middle, and late enrichment of IDE2 max fit even when the S21/S22 data are removed. We do see a rightward shift in the middle timepoint histogram in the direction of later stages, although this may be expected given the absence of concrete accessibility values at S21/S22 in the "no S21/S22" data. This indicates that our data globally retain the general trends of early, middle, and late enrichment of accessibility in the absence of the S21/S22 data.
Moreover, this suggests that, even without the S21/S22 data, the remaining data from early and late stages result in a model fit that still predicts maximum accessibility at middle developmental stages for many peaks.

NOTE: We have removed unpublished data that had been provided for the referees in confidence.

d. To further measure the influence of the S21/S22 data in IDE2 model fit, we also evaluated the degree of change in the global behavior of a peak when the S21/S22 stages were removed. This analysis aimed to assess whether removing S21/S22 data resulted in an IDE2 model with the same general trajectory as with all data, as opposed to the more stringent requirement of evaluating whether the exact developmental stage of the peak was changed. To perform this analysis, we grouped developmental stages into five quintiles, each representing three stages of development. We asked, for each peak in our dataset, whether that peak's IDE2 max fit was "stable" when the S21/S22 data were removed; that is, if the quintile of the IDE2 max fit was altered when the S21/S22 data were removed (i.e. if a peak moved more than 3 developmental stages away from its original position), a peak was considered "unstable". We observed that over 80% of peaks in each quintile remained "stable" after removing the S21/S22 data, suggesting that the vast majority peaks show the same general trajectory of accessibility even without the S21/S22 data. Peaks within the middle time points appeared to be more unstable than peaks at the terminal timepoints, which could be expected given that the S21/S22 timepoints constituted the middle-most timepoints in our dataset.

We acknowledge that the S21/S22 timepoints still appear to be qualitatively different in other ways. Moreover, we acknowledge that some of the peaks in our dataset are "dependent" on the S21/S22 stages, given that their accessibility trajectory changes when these stages are removed. It is difficult to determine whether a change in accessibility trajectory for a given peak caused by the removal of S21/S22 data is indicative of technical differences in sample preparation, such as batch effects; biological variation, such as a potentially unknown mutant or sick embryo; or due to genuine wildtype biological processes that occur at the S21/S22 stages.

These caveats acknowledged, a comparative analysis of the data in the absence of the S21/S22 stages suggests that much of the global picture of development remains the same. In the interest of providing the data we generated as a resource, we decided to include the S21/S22 data in the final manuscript we have prepared for submission.

We have included an additional supplementary figure (Supp. Fig. 2.2) highlighting these further analyses, which we hope future readers will consider when performing their own analyses with these timepoints, as well as a summary of the ways we evaluated this potential concern in the Supplementary Methods. To facilitate future users of this dataset, we will include the model parameters calculated from IDE2 using both the full dataset and the data with S21/S22 removed in the GEO accession data, as well as a Jupyter notebook (ParhyaleATACExplorer.ipynb) that allows users to plot the raw accessibility data and IDE2 model fits for individual peaks of interest (C, example on right panel), so that downstream experiments can consider the potential differences with the S21/S22 samples.

NOTE: We have removed unpublished data that had been provided for the referees in confidence.

Reviewer #2, Comment #3:
"The majority of ATAC-seq peaks in the distal intergenic regions is a very surprising result. Authors defend this result by suggesting that this organism has big genome. May author perform a short analysis that shows that these peaks are indeed represent nearby genes or may point towards 3D genome organisation. For example, I see that this genome might have regions in the genomes that are densely organised in gene clusters, in those cases does the pattern remains same i.e he majority of the genes are very distant from each other and hence use vital regulatory elements?"

Reviewer #3, Cross-comment #3 of 3:
Peaks in distal intergenic regions: I agree that this could be elaborated on. It might also be that >10 kb is not actually that distal for Parhyale. I would suggest to split the "distal peaks" further (e.g., in 10 kb or 2-log steps, or whatever makes most sense) and try to understand if >10 kb is mostly <20 kb, or if most of them are hundreds of kb from the nearest gene?

2. Reviewers #2 and #3 expressed interest in understanding the absolute distribution of distal intergenic peak distances from nearby genes in our dataset. In generating the analyses to address this question, we stumbled upon an error in our code that reveals that the true number of intergenic peaks is much lower than we had originally reported. We discuss the nature of the error below. Moreover, we address the previous question using the new data, which overall still indicates that distal intergenic peaks remain a large portion of the *Parhyale* genome.
   a. To address Reviewer #2's comments with respect to the presence of potential clusters of intergenic regions, we built a Python tool (included in ParhyaleATACExplorer.ipynb) enabling the visualization of different cis-regulatory element categories along a genomic

coordinate. Upon plotting our data with this tool, we observed problems with the categorization of the peaks – namely, that intronic and exonic peaks were erroneously classified as intergenic peaks (see right panel, top). We analyzed our script for classifying annotations more carefully and realized that we had erroneously used “bedtools closest” instead of “bedtools intersect” to try to identify all peaks overlapping with gene annotations in our genome. We corrected this error and observed the expected distribution and categories of peaks in our data (right panel, bottom).

NOTE: We have removed unpublished data that had been provided for the referees in confidence.

b. The revised peak categories have been added to the updated manuscript in Fig. 3H and Fig. 5C. The categories of peaks we observed differ substantially from our previous results, in that we observe a much higher representation of exonic and intronic peaks in our dataset, with intronic peaks now representing 28.2% of all peaks (increased from <1%), and distal intergenic peaks representing 30.2% (decreased from 51.2%). While distal intergenic peaks remain the largest category over time, the proportion is relatively equal to the fraction of intronic peaks. Intergenic peaks (distal and proximal combined) now make up only a slightly larger fraction of peaks (37.2%) than gene body peaks (exon, intron; total 34.4%). This updated result is a significant departure from our previous report, and we have updated the text of the manuscript to correct this mistake.

c. While intergenic and distal intergenic peaks constitute a much smaller portion of our data, we still wanted to address Reviewer #2 and #3’s questions about the distribution of distances between intergenic peaks and nearby genes. We generated a plot to illustrate the number of intergenic peaks at variable distances to the nearest gene (B, right panel). As illustrated in the plot, there are a very large number of distal intergenic peaks, including many peaks >100kb away from the nearest gene. The average distance of intergenic peaks from the nearest gene was 73,351bp. We neglected to mention in the original manuscript that one of the rationales for choosing a 10kb cutoff as “distal intergenic” was that peaks beyond this distance would be considerably more difficult to isolate as single fragments combined with a proximal promoter using PCR, agnostic of their orientation with respect to the promoter element. Such peaks could not have been easily identified using previous transgenic approaches, and are thus distinguished from “proximal” peaks by their necessary identification using techniques such as ATAC-Seq. We have updated the text to reflect this distinction.

NOTE: We have removed unpublished data that had been provided for the referees in confidence.

d. Given that both intergenic and gene body peaks appeared to comprise large fractions of our revised data, we also examined the relative enrichment of intergenic and gene body peaks with respect to time (after normalizing for the fraction of “unknown” peaks, as suggested by Reviewer #3). We observed that the proportion of peaks belonging to intergenic and promoter regions declined slightly as development progressed, while the proportion of gene body peaks increased (E, below). There appeared to be slightly more intergenic peaks than gene body peaks at all developmental time points, and the ratio of intergenic peaks to gene body peaks declined very slightly over time (F, below). These data indicate that intergenic and gene body peaks have different enrichment trajectories over time. As development progresses, gene body peaks are increasingly enriched, and may have a greater impact on gene regulation. We have added these additional observations to the text and to a new Supplementary Figure 2.3.

NOTE: We have removed unpublished data that had been provided for the referees in confidence.

We have also addressed the following textual and conceptual concerns with the manuscript:

Reviewer #3, Comment #1 of 11
I felt that the first paragraph of the introduction is not necessary.

1. We believe the introductory paragraph helps frame the paper in the context of the broader scope of advances in technologies for emerging research organisms – currently, it has become straightforward to both generate a genome sequence and to identify and manipulate coding genes of interest across diverse taxa, but the identification of gene regulatory mechanisms remains more difficult. We have edited the introduction to better reflect this perspective and to link the first paragraph to the rest of the paper.

Reviewer #2, Comment #1 of 3
"In Introductory paragraph 2, sentence one, authors suggest that gene regulation plays more important role in evolutionary process than genes. Although a significant amount of research has been dedicated to gene regulation based evolution still this field is in nascent form. For example evidence of inheritance of the gene regulation pattern across generation is scarce and requires more evidence. I suggest authors to modulate the claim that still gene based evolution is the main paradigm instead otherwise."

Reviewer #3, Cross-comment #1 of 3
Evolution via gene regulation vs. coding sequence: While (to my understanding) it is largely accepted in the field that changes to the CDS will often have more deleterious effects than changes to the expression of a gene, I agree that this could be elaborated on a bit.

2. As requested by Reviewers #2 and #3, we have clarified the language surrounding the debate between gene functional and gene regulatory evolution to indicate that both mechanisms appear to be important for evolutionary processes, with the importance of the latter more recently revealed.

Reviewer #3, Comment #2 of 11
Use of Genrich: I presume this was run on both duplicates simultaneously? This is not clear from the methods section. It might have implications for downstream analyses (e.g., differential accessibility between time points) because running on both sequencing library replicates simultaneously leads to a single "replicate" of peaks per time point, while running it individually leads to two. However, I have never tested if this actually does make a difference. Maybe the authors have and can comment on this?

3. In response to Reviewer #3's inquiry about Genrich, we have added additional clarifying information into the Methods section. "Genrich analysis was run on both duplicate libraries simultaneously; Genrich performs peak calling on each peak individually, and then merges the p-values of the replicates using Fisher's method to generate a q-value, obviating the need to calculate an Irreproducible Discovery Rate (IDR)." We did not test running Genrich on individual libraries, opting for the more conservative approach of using the combined q-value as a filtering score for peak quality. For further information, the reviewer can see the Genrich Github repository section here: < https://github.com/jsh58/Genrich#multiple-replicates>

Reviewer #3, Comment #4 of 11

The section on the IDE2 models (the paragraph at the end of page 4/beginning of page 5) was unclear to me but appears sound. (The only instance where I didn't quite understand what the program actually does.) Maybe this can be explained a bit easier?

4. As requested by Reviewer #3, we have attempted to explain the methods and logic of using ImpulseDE2 a bit more clearly:

   "To identify regions of dynamically accessible chromatin, we used the ImpulseDE2 (IDE2) pipeline (Fischer et al., 2018). IDE2 differs from other software for differential expression analysis in that it allows the investigation of trajectories of dynamic expression over large numbers of timepoints. It does so by modeling a gene expression trajectory as an “impulse” function that is the product of two sigmoid functions (Chechik and Koller, 2009; Yosef and Regev, 2011). This approach enables the modeling of a trajectory of gene expression in three parts: an initial value, a peak value, and a steady state value, thus summarizing an expression trajectory using a fixed number of parameters. With the ability to capture the differences between early, middle, and late expression values for each gene in a dataset, IDE2 also enables the detection of transient changes in gene expression or accessibility during a time course. Identifying differential expression over large numbers of timepoints is difficult for more categorical differential expression software such as edgeR and DESeq2, which generally use pairwise comparisons between timepoints to assess change over time (Love et al., 2014; Robinson et al., 2010).”

Reviewer #2, Comment #2 of 3

2-2) Authors have repeatedly used S21 and S22 throughout the manuscript to support their claims with clustering etc. May authors shed some light on the differences in replicates for these timepoints. Furthermore, I could not find Fig 3J, perhaps author would like to point out Fig 3H.

Reviewer #3, Comment #5 of 11

On page 7, Fig.3J needs changing to 3H. This figure should, in my opinion, also contain the absolute number of peaks for each time point to set the individual proportions into context.

5. As requested by Reviewer #3, we have added a bar charts representing the number of peaks found at each time point (Fig. 3H) and the number of peaks found in each cluster (Fig. 5C) to the peak type proportion plots. We have also fixed references to Fig. 3J to instead refer to Fig. 3H – we apologize for the confusion.

Reviewer #3, Comment #6 of 11

Last paragraph of the "Improving the Parhyale genome annotation" section: I think this needs to focus on those regions of the genome for which the location is known - after all, the "unknown" regions" could all be "distal transgenic", which would significantly change the relative proportions.

6. We have revised our analysis of this topic with our updated peak type proportions, as described above in point 2d above under “substantive concerns”.

Reviewer #3, Comment #7 of 11

“On page 9, t-SNE is mentioned but doesn't seem to be cited.”

7. As requested by Reviewer #3, we have added citations for the t-SNE method, as well as scikit-learn, the software we used for t-SNE visualization.

Reviewer #3, Comment #8 of 11

“The third paragraph on page 9 ("We evaluated the differences...") should mention the fact that clusters 1 and 2 are the only ones with significant proportions of exonic and intronic peaks. In the accompanying figure (5C), the total number of peaks would again be helpful.”

8. After identifying the error in our peak category classification pipeline, this observation was no longer true. However, upon examining the new distributions by cluster, we observed that in Clusters 3–7, for which we observed GO enrichment for developmental processes, there appeared to be slightly higher enrichment of intronic regulatory elements than distal intergenic regulatory elements. These results resemble the observation from recent work showing that tissue-specific enhancers are enriched in intronic regions in various human cell types (e.g. Borsari et al. 2021, *Genome Research*). We have noted this new observation in the text.

Reviewer #3, Comment #9 of 11

In figure 5D, I can't quite make out at which stage the dip in the peak of Cluster 8 occurs. This is quite an unusual pattern of accessibility change, and I can't help but wonder if it has something to do with the quality of one of the libraries? Also, the fact that half of the peaks fall into unmapped regions of the genome is unusual, and I feel this deserves more discussion.

9. In Figure 5D, Reviewer #3 asks about a dip in accessibility for Cluster 8 peaks. The dip in accessibility was actually observed for Cluster 9 peaks and is marked by the asterisk in that panel. We have updated the figure legend to clarify the significance of the asterisk and have referred readers to examine Supp. Fig. 5.1B, where the IDE2 model fits more clearly show a collective dip in accessibility for Cluster 9 peaks. Upon examining the size distribution of the clusters, we have also noticed that Cluster 8 is the smallest cluster. We have noted the small cluster size and high “unknown” peak enrichment for Cluster 8 in the text.

Reviewer #3, Comment #10 of 11
“On page 10, the abbreviation PFM appears, but it is only explained in the legend of Fig.4. This should appear in the text.”

10. Reviewer #3 mentions that on page 10, we use the abbreviation for position frequency matrices (PFMs) without previous reference. We first introduce the abbreviation on page 8, but given the repeated use of “PFM” on page 10, we have added an additional explanation of the abbreviation on page 10, for ease of reading.

Reviewer #3, Comment #11 of 11
“The section on "Concordant and discordant expression and accessibility" is the one I disagree most with. The authors seem to suggest that a repressive cis-regulatory module should become less accessible when the gene is activated. However, they leave trans-acting factors completely out of their conceptualisation here. It is in general likely the availability of transcription factors that leads to repression, while the "silencer" can be well accessible in all cells. Moreover, it has become clear in recent years that CRMs are not just repressors or enhancers per se but can act as either depending on the availability of transcription factors. I think these facts could partially explain the weak correlation and should be discussed.”

11. We appreciate the comments from Reviewer #3, which alerted us to the more recent literature around the bifunctional potential of regulatory elements. We have revised our claims to clarify that concordance and discordance analysis cannot be used to directly assign “enhancer” or “silencer” identity to given regulatory elements. Instead, we suggest that evaluating concordance and discordance can be useful for downstream users of our data, such as those aiming to build reporter constructs for a given gene of interest. To facilitate such tool development, we have built additional functions into a Jupyter notebook to enable the visualization of accessibility, gene expression, fold change of accessibility and gene expression, significance of fold change, and concordance/discordance assignment for arbitrary peak-gene pairs. An example of this visualization is shown on the following page. Panel A shows the region around the *Engrailed-1* and *Engrailed-2* loci in *Parhyale* (text labels within the plot region were added manually in Illustrator). Panel B shows visualization of the En1 promoter peak alongside En1 expression. Significant log fold changes (DESeq2 padj < 0.05) are marked by asterisks in the bar plots, and concordance/discordance assignment at each time point is indicated by the color of the comparison text (red = concordant, blue = discordant). Panels C and D show accessibility and expression visualization for a single peak (En1 peak5) compared to two nearby genes (En1 and En2). We hope to include sufficient documentation in our GitHub repository such that using these tools is accessible for most researchers, even with limited programming knowledge.

NOTE: We have removed unpublished data that had been provided for the referees in confidence.

4. Description of analyses that authors prefer not to carry out
We were unable to easily visualize the distribution of regulatory elements across the whole genome as suggested by Reviewer #2. One challenge of working with the *Parhyale* genome is the lack of complete chromosomes. The genome is distributed across ~290,000 contigs of variable size. We were unable to find any software that could be easily and quickly set up to visualize our data, although we will provide in a Jupyter notebook the tools for local visualization of peak types that we developed.

Original submission

First decision letter

MS ID#: DEVELOP/2022/200793

MS TITLE: Identification and classification of cis-regulatory elements in the amphipod crustacean Parhyale hawaiensis

AUTHORS: Dennis August Sun, Jessen V Bredeson, Heather S Bruce, and Nipam H Patel
ARTICLE TYPE: Research Article

I am happy to tell you that your manuscript has been accepted for publication in Development, pending our standard ethics checks. As you will see, both referees were satisfied with the revisions to the study that were made in response to the initial reviews provided through Review Commons.

Reviewer 1

Advance summary and potential significance to field

In this manuscript, Sun et al. analyse gene regulaion during development of the crustacean, Parhyale hawaiensis. Using ATAC-seq and RNAseq approaches in multiple stages during embryonic development, they illustrate transcriptional changes as well as accessibility trajectories of genomic elements and improve the genome annotation of this emerging model organism. The authors not only provide a thorough analysis of a multitude of genomic datasets that will, without a doubt, be the source of countless future scientific endeavours in Parhyale, but their work can also serve as a blueprint for the analysis of similar datasets in other organisms.

Comments for the author

I have reviewed this manuscript for Review Commons, and the authors have dealt with my previous comments admirably. I have no further comments.

Reviewer 2

Advance summary and potential significance to field

Sun et al used omni-ATAC sequencing that is a modified version of classical ATAc-seq to identify and characterise the cis-regulatory elements in the P. hawaiensis genome. They further use long and short reads to improve upon existing gene annotation for this organism. In the first version of the manuscript the in-depth analysis ensures the results and conclusions to be sound and revised version of the manuscript clarifies doubt raised by all the reviewers satisfactorily. I do not have further comments on the manuscript. I wish good luck to authors.

Comments for the author

None