Oviedo, January 15th, 2024

Dear Editor,

We thank the opportunity to submit the revised version of the manuscript “**PlantFUNCO: integrative functional genomics database reveals clues into duplicates divergence evolution (MS ID#: MBE-23-0764)”.**

In this revised version, we have addressed all suggestions and questions raised during the review process, contributing to an overall improvement in the quality of this work. According to the referees’ feedback, manuscript underwent a thorough grammatical and text review. Additional analyses were conducted, additional figures and supplementary information were included, and documentation for the PlantFUNCO website was added. The main text was modified in response to reviewers’ comments. Detailed responses to the reviewers are provided below in a point-by-point manner.

We believe that our manuscript is now well suited for publication in “Molecular Biology and Evolution”, and we hope that both the reviewers and the editorial committee share our opinion.

Sincerely yours,



**Dr. Mónica Meijón**

Associate Professor

Plant Physiology

Department of Organisms and Systems Biology

University of Oviedo

Spain

Telf.: +34 985 104 796

E-mail: [meijonmonica@uniovi.es](about:blank)

**MS ID#: MBE-23-0764**

**MS TITLE: PlantFUNCO: integrative functional genomics database reveals clues into duplicates divergence evolution**

***Associate Editor's comments to the author:***

The comments and recommendations two expert reviewers are now available for your manuscript. These reviewers had mixed opinions about your manuscript. They found the potential scientific impact of your work to be medium/high. They found that the manuscript needs a lot of improvement in text and a lot of additional data analysis. Editors generally agree with their concerns and recommendations, which led to a designation of medium priority.

Note that many manuscripts receiving medium priority based on reviewer comments are not accepted by the Board of Editors. I look forward to seeing the revised manuscript.

# We wish to thank the Board of Editors and the Associate Editor to give us this opportunity to submit our revised work and anonymous reviewers for their feedback, which helped us to improve our manuscript. We have modified the manuscript according to all the reviewers’ criticisms.

----------

***Reviewers' comments:***

***Reviewer: 1***

*Comments to the Author:*

The manuscript “PlantFUNCO: integrative functional genomics database reveals clues into duplicates divergence evolution” by Roces et al. presents a functional conservation database of chromatin states and conservation scores across three plant model species, Arabidopsis thaliana, Oryza sativa, and Zea mays. The authors main objective – to characterize functional divergence based on chromatin state among species – is an important goal in plant biology. I found the sections on paralog divergence and validation of mitochondrial AOX to be particularly intriguing. However, the countless grammatical errors and questionable statements made getting through this manuscript a challenge. A bigger concern is the lack of careful consideration in experimental design. The authors collected epigenomics data produced by multiple labs for a range of tissue/developmental stages. Chromatin is highly dynamic spatiotemporally. Some chromatin modifications are more stable (among tissues and dev. stages) than others, which will bias conservation estimates. None of these points are discussed or accounted for by the authors in the analysis. While I do appreciate the question posed by the authors, the execution requires revising to explicitly address the discrepancies among the collected data. That being said, I do think there is utility to the intra-specific comparisons as paralogs are essentially matched. This manuscript has potential to be of interest once the above and below concerns are addressed.

# First, we want to thank the reviewer for the time taken in reviewing our manuscript. The insightful comments provided will undoubtedly contribute to the enhancement of our work. We appreciate reviewer’s perspective on the utility of paralog functional divergence and intra-specific comparisons. We note that further clarification about the approach and the specific use of the resources generated would be really helpful and we again want to thank the reviewer for pointing this out. We genuinely hope that the comments and analyses presented below will address the concerns raised by the reviewer.

*Major points:*

There are numerous grammatical and nonsensical statements throughout the manuscript that made reading this a challenge. Rather than listing each offense (which are too numerous to cite), I would recommend that the authors get another set of eyes on the manuscript

# We appreciate the comment provided by the reviewer. The manuscript has undergone further scrutiny, and the final version, which incorporates all the new modifications, has been professionally proofreaded by an official scientific English service.

Line 102: Chromatin states vary significantly based on tissue, development, and environmental condition during samples. Epigenomic data sets were collected from a range of tissues and developmental stages that were developed by multiple labs. How did the authors account for these discrepancies in their model? What are the consequences of using unmatched data among species? In my opinion, failing to control for sampling is resulting in a fundamentally flawed apple to orange comparisons throughout the manuscript and limits the utility of this study. This is particularly critical in cases where the authors are comparing specific loci among species; because chromatin states are highly dynamic among tissues, of course there will be stark differences between rice seedling chromatin and seed chromatin in maize. A result of using multiple samples per species is a bias towards stable states - it is expected that genes that are constitutively expressed (“Active” chromatin state) will be more conserved among species; these are housekeeping genes. This calls into question the reliability of all the inter-specific analyses.

# We thank the reviewer for these questions. As the comment contains several questions, we have decided to address them in a point-by-point manner.

**Chromatin states vary significantly based on tissue, development, and environmental condition during samples. Epigenomic data sets were collected from a range of tissues and developmental stages that were developed by multiple labs.**

# In response to reviewer comment, we have added a new sheet titled “CS intraspecies variance\_corr” to **supplementary table S1**. In summary, we computed intra-species variance and correlation between log2(fold enrichment + 0.5) signal tracks for each species, chromatin modification and combination of conditions (tissue, age and study; see **supplementary table S1**). We used deepTools with a binSize of 5 kb, consistent with the bin size reported in metazoan inter-species chromatin states (Ho, J., et al. Comparative analysis of metazoan chromatin organization. *Nature* **512**, 449-452 (2014)).A single signal track for each combination of conditions (tissue, age and study) was used, ensuring that replicates from the same study did not affect the computation of intra-species correlation and variance.

# Example: for H3K27me3 in *Arabidopsis thaliana*,intra-species variance and correlation were computed using six signal tracks , representing six unique combinations of tissue, age and study. Pairwise correlations between the six signal tracks were calculated, and intra-species variance and correlation were determined as the standard deviation and mean, respectively, of the correlation coefficients from all pairwise comparisons.

# The chromatin modifications data in our collection presented strong intra-species positive correlations with relatively low intra-species variance, with very few exceptions. These results suggest that chromatin modifications across the conditions present in our collection tend to regulate the same core set of regions, thereby indicating consistency and low discrepancy in the tracks obtained from multiple labs and/or different range of tissues and developmental stages.

# We modified the methods section to introduce the new table added:

**[ Methods, Page 14, Lines 475-477]**

[Additional information detailing intra-species correlations and variance can be found in **supplementary table S1**]. To guarantee the reproducibility of the analysis, a docker was created and it is available at [https://hub.docker.com/r/rocesv/plantina-chiplike](about:blank).

**How did the authors account for these discrepancies in their model? What are the consequences of using unmatched data among species?**

# We computed inter-species chromatin states using data from multiple conditions. When employing this data design to compute chromatin states in a stacked fashion, the primary goal is to characterize the potential epigenomic regulation and chromatin state status of genomic regions, considering all the signals across the conditions in our collection. This approach allowed us to differentiate between constitutive active/repressive regions across the conditions collected in our data and those regions with potential bivalent regulation. This data design and approach have been used several times previously, with interesting applications reported (Liu, Y., et al. PCSD: a plant chromatin state database, *Nucleic Acids Research*, Volume 46, Issue D1, Pages D1157-D1167(2018); Hazarika, R.R., et al. Molecular properties of epimutation hotspots. *Nature Plants*, Volume 8, Pages 146-156 (2022); Vu, H., Ernst, J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biology*. Volume 23, 9 (2022); Horvath, S., et al. DNA methylation clocks for dogs and humans. *PNAS*. Volume 119 (2022); Arneson, A., et al. A mammalian methylation array for profiling methylation levels at conserved sequences. *Nature Communications*. Volume 13, 783 (2022)).

# It should be noted that the use of this data design and approach has several advantages and disadvantages.

# Disadvantages: We lose tissue- and condition-specific epigenetic regulation information. Additionally, comparisons between specific loci among species should be avoided only considering the inter-species chromatin state framework due to non-matching data use (as discussed in the next comment).

# Advantages: By leveraging data diversity, this approach is capable of detecting constitutive active/repressed elements. States computed from multiple conditions are more predictive of external annotations than tissue- and condition-specific annotations. For instance, the success of ChromImpute and other related approaches underscored the value of considering shared information across chromatin modifications and cell types. The flexibility of the hiHMM algorithm, applied to detect homogeneous chromatin modification genome segments across multiple species using a stacked-like approach, allowed us to discern which chromatin states favor or avoid the co-occurrence of specific chromatin modifications across species in all conditions collected. Consequently, we could extrapolate more general (broad but shallow) relations of chromatin states conservation and/or divergence. Since chromatin states conservation/divergence relations were determined based on chromatin modifications composition and genomic distribution (see minor comment below), it shares the same assumptions as the chromatin states themselves and should be understood as an additional layer of interpretation when performing intra-species analyses.

# Rather than comparing specific loci among species, the primary objective of our inter-species chromatin states is to establish a general annotation of chromatin states for intra-species characterization, aiding in the prediction of other genomic elements/mechanisms, such as paralog redundancy. Our approach added an additional layer of interpretation, offering insights into the relations of conservation or divergence of chromatin states definitions between species across the conditions collected in our data. Despite the mentioned disadvantages, the employed approach has previously provided valuable biological insights and is well-suited to address our specific research questions.

**In my opinion, failing to control for sampling is resulting in a fundamentally flawed apple to orange comparisons throughout the manuscript and limits the utility of this study. This is particularly critical in cases where the authors are comparing specific loci among species; because chromatin states are highly dynamic among tissues, of course there will be stark differences between rice seedling chromatin and seed chromatin in maize.**

# We apologize for any confusion. It should be noted that we intentionally avoided the comparison of specific loci among species in the inter-species chromatin states results section (“Characterization of shared and species-specific chromatin states”). As discussed above, this is because such comparison is not the main focus of the inter-species chromatin states framework, considering the discussed advantages and disadvantages of the data collection design. We were well aware that specific loci comparisons among species required matched data or a specific algorithm designed for dealing with non-matching data. Two important points should be highlighted:

# 1) The biological question addressed in the inter-species chromatin states results section (“Characterization of shared and species-specific chromatin states”) involved conducting intra-species region descriptive analyses with various annotation modules to further validate the definitions of the chromatin states. We leveraged an additional layer of interpretation, gaining insights into how conserved or divergent the chromatin state definitions were across species for the conditions in our data collection. We would like to emphasize that despite the involvement of evolutionary-related annotations such as orthologs, conserved non-coding elements, etc., in the chromatin states descriptive analyses, we were not directly comparing the chromatin states between equivalent loci in different species (a conscious decision due to non-matching data design). Instead, we reported intra-species chromatin state trends with evolutionary-related tracks, such as significant genomic overlaps or the proportion of genes covered by a specific state that had orthologs in the other species. In conclusion, the descriptive analyses conducted in the inter-species chromatin states results section focused solely on the genomic coordinates for one species at a time, without engaging in direct comparisons of equivalent loci across different species.

# 2) We exclusively compared specific loci among species using the LECIF algorithm in the results section titled “Defining functional- genomic conservation score and the database”. The LECIF algorithm is explicitly designed to handle non-matching data taking as input orthologous-like regions derived from whole genome alignments pipeline. We opted to stay at the region-level rather than gene-level to retain non-coding regulatory information. The LECIF algorithm integrated non-matching chromatin states, chromatin modifications and gene expression data to output a single functional genomics conservation score. Among the analyses performed with LECIF scores (“Defining functional- genomic conservation score and the database”) there were three using inter-species chromatin states information (**Fig. 6E-G; Chromatin States**). The state-specific LECIF scores distribution (violin plots) did not compare equivalent loci across species, as only the genomic coordinates for the query species were used at a time for describing LECIF score distribution. The analyses comparing chromatin state similarity between high/low and low/high LECIF/PhyloP genomics score regions (horizontal grouped barplot) and the cross-species chromatin state similarity between regions with low, medium and high LECIF score (line plots) were the only analyses in the manuscript comparing equivalent loci/regions across species. Cross-species chromatin states similarity in the original LECIF reference (Kwon, S.B., Ernst, J. Learning a genome-wide score of human-mouse conservation at the functional genomics level. *Nature Communications*. Volume 12, 2495 (2021)) was performed based on a 7 states model jointly obtained using a pseudo-genome concatenating human and mouse chromosomes and non-matching human and mouse chromatin modifications data (different conditions such as tissues, cell types, age, and laboratories in human vs mouse). Therefore, highlighting the urge of having similar chromatin states number and definitions between the species compared rather than matching data. The greater flexibility of the hiHMM algorithm to learn chromatin states across species coupled to the ability of LECIF algorithm scores to deal with non-matching data filled this gap in our case.

# We reviewed discussion section to clarify the specific purpose of each generated resource:

**[ Discussion, Page 11, Lines 355-356]**

While this flexible framework provides a consistent definition of CS across multiple genomes, [making the extrapolation of intra-species analyses between them easier], the stack approach allows for an understanding of the potential epigenomic regulation over several tissues/conditions such as differentiating constitutively active/repressive regions (Vu and Ernst, 2022).

**[ Discussion, Page 13, Lines 421-427]**

While we expect PlantFUNCO to be useful, we acknowledge certain limitations. [Owing to our data collection design, the main goal of inter-species CS resources is to conduct intra-species analyses while leveraging the advantage of having additional layers of interpretation, including direct correspondence between CS and conservation/divergence relationships established across species. Direct functional cross-species comparisons of equivalent loci or CSs should be only undertaken in conjunction with the plants’ LECIF scores, as this algorithm is explicitly designed to handle highly diverse datasets.]

**A result of using multiple samples per species is a bias towards stable states - it is expected that genes that are constitutively expressed (“Active” chromatin state) will be more conserved among species; these are housekeeping genes. This calls into question the reliability of all the inter-specific analyses.**

# We support the notion that using multiple samples/conditions per species could lead to a clearer differentiation between genomic regions with potential bivalent regulation and those with constitutive active/repressive (stable) status across all the samples/conditions. It should be noted that chromatin state models were initially designed to annotate the non-coding genome; thus, constitutive active states also encompassed other interesting genomic elements such as constitutive active promoters and enhancers. As later noted by the reviewer in a minor comment, finding housekeeping genes, which are suggested to be more conserved, in constitutive active states is totally expected. This was the primary reason why we incorporated conservation-related tracks to futher validate the chromatin state definitions in the results section titled “Characterization of shared and species-specific chromatin states”. That being said, rather than a bias, this is an inherent and desired feature when computing chromatin states with multiple samples/conditions and one of the main goals for adopting this approach in our study and those conducted previously (Liu, Y., et al. PCSD: a plant chromatin state database, *Nucleic Acids Research*, Volume 46, Issue D1, Pages D1157-D1167(2018); Vu, H., Ernst, J. Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biology*. Volume 23, 9 (2022)).

# In case the reviewer is referring to how housekeeping genes could influence conservation/divergence relations in the inter-species chromatin states, it is essential to clarify that the relations of conservation/divergence between chromatin states were determined based on chromatin modifications composition and distribution over genomic elements (see minor comment below). As the level of conservation of the genomic elements covered by each chromatin state was not taken into account for defining the aforementioned relationships, it should not introduce bias into those relationships. A concrete example of this scenario could be CS1 bivalent state, which represented a conserved definition across species, yet it did not imply an active state.

# Motivated by the concept of conservation in chromatin states, we further research the comparison between comparative genomics conservation and functional genomics conservation in the LECIF results section titled “Defining functional- genomic conservation score and the database”. Although constitutive active elements (including housekeeping genes) are suggested to be highly conserved, there are no simple tools to determine the extent of functional conservation between model species beyond manually cheking for orthology and gene expression similarity. In these scenarios, we expect that LECIF scores, which summarize the degree of functional conservation from multiple modalities and conditions, along with the tools provided in PlantFUNCO, could be useful for the molecular biology and evolution community.

*Minor points:*

Line 40: chromatin state includes more than just “epigenetic” marks (epigenetics should be used to define heritable non-DNA-encoded changes that influence transcription). A more specific term would be “chromatin modifications”. The use of “epigenetic marks” throughout the manuscript should be changed in the cases where the authors are referring to chromatin modifications.

# The reviewer is completely right. We have replaced the term “epigenetic mark” throughout the manuscript. Some examples are in the following sentences:

**[ Introduction, Page 2, Lines 40]**

**[ Results, Page 4, Lines 110]**

**[ Methods, Page 14, Lines 458]**

**[ Supplementary Material, Page 28, Lines 941]**

Please include a few genome browser panels from each species to support the chromatin state calls.

# As requested by the reviewer, we have included genome browser panels for each species to supplementary fig. S1. Supplementary fig. S1 legends was modified accordingly.

Figure 1: The % genome coverage for each state ranges widely among species. How do the authors reconcile the huge differences between species? Are the values observed in each species expected? If so, what is the justification?

# We appreciate the questions raised by the reviewer. To address these inquiries, we calculated the maximum coverage range for each chromatin state (maximum coverage value – minimum coverage value) in our plant inter-species chromatin states:

# Maximum coverage range values (%): CS1 = 1.2, CS2 = 0.9, CS3= 1.9, CS4 = 2.9, CS5 = 2.2, CS6 = 5.4, CS7 = 2.7, CS8 = 10.4, CS9 = 13.7, CS10 = 0.2, CS11 = 5.5, CS12 = 2, CS13 = 12.5, CS14 = 15.9, CS15 = 15.3, CS16 = 6.

# We checked the maximum coverage range values from metazoan inter-species chromatin states reported in Ho, J., et al. Comparative analysis of metazoan chromatin organization. *Nature* **512**, 449-452 (2014). Excluding quiescent, low signal and unmapped states, we found maximum coverage range values between 0.9-8.5 and 1.7-9.8 (0.9-24.9 and 1.7-9.8 including quiescent, low signal and unmapped states). In summary, most of our maximum coverage range values seem to fall within the expected range for an inter-species framework.

# Excluding quiescent, low signal and unmapped states (CS14-CS16), the chromatin states that could show substantial differences in genome coverage are CS8, CS9 and CS13 (10.4, 13.7 and 12.5, respectively). Considering the definitions of each chromatin state and the molecular biology background of each species under study, these observed values are expected. The genomes under study differ in size and transposon content: *Arabidopsis thaliana* has a genome size of ~ 135 Mb with a transposon content of ~ 23.41 %, *Oryza sativa* has a genome size of ~ 430 Mb with a transposon content of ~ 42 %, and *Zea mays* has a genome size of ~ 2.4 Gb with a transposon content of ~ 80 %.

# The lower coverage in *Z. mays* compared to *A. thaliana* and *O. sativa* in CS8 may be explained by the fact that protein-coding genes and transcription end sites constitute a smaller fraction of *Z. mays* genome considering its size and transposon content. CS9, defined as an active state in *A.thaliana* and *O.sativa* but representing a repressed intergenic state in *Z. mays*, showed higher coverage in *Z. mays* compared to *A. thaliana* and *O. sativa*, likely due to intergenic regions and transposons constituting a larger fraction of *Z. mays* genome. Furthermore, CS13’s higher coverage in the monocots (*O. sativa* and *Z. mays*) can be attributed to the larger genomes and higher percentage of transposons and intergenic regions in these species.

Lines 115-118: It is not clear how divergence between (or within) species is being determined.

# We sincerely apologize for any inconvenience. The divergence/conservation among inter-species chromatin states is determined based on the composition of chromatin modifications (**fig. 1**, **top panel**) and the distribution across various genomic elements (**fig. 1**, **bottom panel**). This methodology aligns with the approach established in the original hiHMM algorithm article (Sohn, KA., et al. hiHMM Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics*. Volume 31(13), 2066-74 (2015)). Although this is mentioned after line 118, where various degrees of divergence with examples are enumerated, we acknowledge the need for enhanced clarity in the initial statement. Consequently, we have revised the highlighted lines to address the reviewer’s comment.

**[ Results, Page 4, Lines 119-121]**

[The various degrees of CS divergence were determined based on chromatin modifications composition (**fig. 1**, **top panel**) and genomic distribution (**fig. 1**, **bottom panel**). Ranging] from less to more divergent:

Figure 2: The text is hard to read. This is a very busy figure that would benefit greatly from a bit of revising. I would recommend either breaking this figure into two separate figures or moving the least interesting parts to the supplemental. This also applies to Figure 3 which would not be legible on a standard 8.5”x11” US letter.

# The reviewer is correct. In the reviewed manuscript, we have split Figure 2 into two new main figures, Figure 3 into one main and one supplementary figure.

Lines 153-155: The decreasing trend reflects the % coverage of each state (with the exception of CS11 which is likely constitutive heterochromatin that is typically depleted of genes).

# We appreciate the reviewer for bringing up this important point. As highlighted by the reviewer, the decreasing trend in the chromatin states reported could be reflected by the coverage percentage of each state. To provide clarification, we have introduced a new sheet titled “CS summary” to supplementary table S1, summarizing the percentage of protein-coding genes (number of protein-coding genes / total number of genes) and the ratio of orthologs (number of protein-coding genes having orthologs in other species / number of protein-coding genes) in each chromatin state and species. As values are divided by the total number of genes and protein-coding genes in each state, we could compare CS with different coverage percentages.

# Since the definition of CS9 differed substantially between species, being active in *A. thaliana* and *O. sativa* but repressive in *Z. mays*, the CS9 values for *Z. mays* were excluded from the computation mentioned before. However, both options (with and without *Z. mays* CS9 value) are provided in the table for reference.

# With the new percentages and ratios provided, we observed that active and bivalent states exhibited higher percentages of protein-coding genes than heterochromatic states, as expected. It is important to note that there is no clear evidence supporting a specific order between bivalent and active states in terms of protein-coding genes. Thus, we have rectified the hierarchical order reported in the results section for the number of protein-coding genes.

# On the other hand, in terms of the ratio of orthologs, we found evidence supporting the notion that active states had greater ratio of orthologs than bivalent states and subsequently greater than heterochromatic states (active > bivalent > heterochromatic). Given that the divergent state CS10 had different definitions for each species (heterochromatic, bivalent and active for *A. thaliana*, *O. sativa* and *Z. mays*, respectively), we have rectified the results statement to only reflect the order between active, bivalent and heterochromatic states.

# To illustrate, examples indicating that the ratio of orthologs followed an active > bivalent > heterochromatic trend, independently of genome coverage and number of protein-coding genes, include: 1) *A. thaliana* active CS7 with 3.4 % genome coverage and 73.66 % ratio of orthologs, while bivalent CS4 had 4.7 % genome coverage and 66.34 % ratio of orthologs. 2) *O. sativa* active CS7 with 0.7 % genome coverage and 75.27 % ratio of orthologs, while bivalent CS5 had 3.3 % genome coverage and 46.98 % ratio of orthologs. 3) *Z. mays* repressed intergenic CS9 with 18.3 % genome coverage, 3571 protein-coding genes and 27.47 % ratio of orthologs, while bivalent CS3 had 1.1 % genome coverage, 2091 protein-coding genes (a lower number than heterochromatic state due to its genome coverage) and 81.97 % ratio of orthologs (a greater ratio despite the lower number of protein-coding genes).

# Lastly, it is crucial to note that functional enrichments should not be significantly influenced by genome coverage, which might otherwise lead to an increase in the gene set size. This is because, after testing all gene sets, the enrichment scores are normalized for the size of the gene set before the p-values are corrected. Furthermore, to ensure comparable gene set sizes, default settings establish minimum and maximum gene set sizes at 10 and 500, respectively. Functional enrichments were computed separately for each species taking into account species-specific backgrounds.

# Taking everything into consideration, we have modified the following lines to rectify the trend of protein-coding genes and to only reflect the order between active, bivalent and heterochromatic states:

**[ Results, Page 5, Lines 161-166]**

We observed a remarkable gradient across functional groups, excluding quiescent/no signal from the analysis due to the lack of epigenetic regulation (**fig. 3**; **supplementary fig. S3** **and supplementary fig. S4;** [**supplementary table S1**). A decreasing trend in gene functional convergence (KO and GO) and the proportion of orthologous relationships was identified, following the order active > bivalent > heterochromatin, illustrated by CS6 > CS1 > CS11, respectively (the first state of each functional group was selected for representation). CS10 represented a divergent state corresponding to heterochromatic, bivalent and active states in *A. thaliana*, *O. sativa* and *Z. mays*, respectively.]

**[ Supplementary Material, Page 29, Lines 965-966]**

**Supplementary table S1. Epigenomic data collection [and chromatin states summary of protein-coding genes and orthologs percentage.]**

Line 159-160: “… linking CS with high regulatory/transcriptional activity to evolutionary constraint patterns.” How did the authors come to this conclusion? It seems more likely that ChromHMM is identifying housekeeping genes as transcriptionally active (as they are transcribed ubiquitously among tissues) and these are the most conserved genes, which is totally expected. A direct comparison of chromatin states among orthologs would have been more informative.

# We agree with the reviewer’s observation. At that stage of the manuscript, the conclusion might have been premature. The results presented in this section, supporting this statement, indicated that active and bivalent states (CS with high regulatory/transcriptional activity) had higher ratios of orthologous genes, showed greater functional convergence (GO and KO enrichments) and were predominantly enriched in conserved PhastCons elements compared to heterochromatic states. We have rectified the sentence to address the reviewer’s concerns (**Results, Page 5, Lines** **166)**. As stated by the reviewer, identifying housekeeping genes, which are expected to be highly conserved, is totally expected. This was the rationale behind our inclusion of conservation-related track to further validate the proposed chromatin state definitions.

# As discussed above, while comparing chromatin states among orthologs would undoubtedly be interesting, we aimed to avoid the comparison of specific loci among species at that stage of the manuscript. This is because such comparison is not the primary focus of the inter-species chromatin states framework, considering the data collection design. We are well aware that ortholog comparisons require matched data or a specific algorithm designed for unmatched data.

# To fill the gap proposed by the reviewer in direct comparison of chromatin states in orthologs across species, we have applied the LECIF score analyses (**fig. 6**). The LECIF algorithm is specifically developed to deal with unmatched data and takes as input orthologous-like regions derived from whole genome alignments pipeline. We chose to stay at the region-level rather than the gene-level to retain non-coding regulatory information. In the LECIF results section (“Defining functional- genomic conservation score and the database”), the reviewer can find two analyses referring to a direct comparison of chromatin states among orthologous-like regions across species: 1) Dice similarity between chromatin states groups within regions with low, mid and high functional conservation scores (LECIF) (**fig. 6E-G**, **Chromatin States module**, **line plots**). 2) Dice similarity between chromatin states across functional conservation scores (LECIF) and comparative genomics conservation scores (PhyloP) (**fig. 6E-G**, **Chromatin States module**, **horizontal grouped bar plots**).

# Finally, to support the aforementioned statement about the similarities of chromatin states across orthologous-like regions without considering functional and comparative genomics conservation scores, the area under the lineplots (**fig. 6E-G**, **Chromatin States module**, **line plots**) and the mean between the three reported points consistently showed higher values for active and bivalent states than heterochromatic and the divergent states (quiescent, low signal and unmapped group excluded due to the absence of epigenetic regulation).

Lines 214-239: How the authors came to the conclusions in this section is not clear. Many of the sentences would benefit from rephrasing, which perhaps contributed to the lack of clarity in this section.

# We apologize for any inconvenience. The highlighted lines by the reviewer point to the results paragraph where Cusack et al., 2021 models are reproduced, introducing our new chromatin state metrics. There are two main conclusions in the underlined section:

# 1) The results from the Cusack et al., 2021 modelling are reproducible, ensuring accurate comparisons. This conclusion is supported by the number of variables and the relative importance of the six feature categories in the models RD4 and RD9, which do not include chromatin state information (**fig. 5F**). These results are very similar to the ones reported in the original reference, Cusack et al., 2021. Additionally, the intersection with the highest number of features was common to all models, both with (RD4C-RD9C) and without (RD4-RD9) chromatin state information (**fig. 5I**). This suggests that the core predicting power remained constant for all the models, securing that the observed differences in the performance of the models with the inclusion of the chromatin state information is rigorously accurate.

# 2) The inclusion of chromatin state information improved predictions for RD9 inclusive redundancy gene pairs (a general definition that also contained RD4 extreme redundancy gene pairs). However, this improvement was not observed for RD4 extreme redundancy gene pairs (where single-mutants have no abnormal phenotype, and the double mutant is lethal). This conclusion is substantiated by the higher and lower AUC-ROC and AU-PRC values (**fig. 5G-H**) detected for RD9C (inclusive redundancy gene pairs with chromatin state information; **AUC-ROC= 0.665; AU-PRC = 0.651**) vs RD9 (without chromatin state information; **AUC-ROC = 0.634; AU-PRC = 0.603**) and RD4C (extreme redundancy gene pairs with chromatin state information; **AUC-ROC = 0.807; AU-PRC = 0.795**) vs RD4 (without chromatin state information; **AUC-ROC = 0.842; AU-PRC = 0.825**), respectively.

# The following modifications were applied to the section underlined:

**[ Results, Page 7-8, Lines 219-247]**

In contrast, [the models developed by Cusack et al., 2021 (**fig. 5F-I**) categorized redundancy into different definitions, covering a plethora of features with distinct transformations. Consequently, we opted to incorporate all CS metrics to model redundancy for each definition, resulting in four different sets:] RD4 (extreme redundancy, [where single-mutants have no abnormal phenotype, and the double-mutant is lethal;] without CS information), RD4C (with CS information), RD9 (inclusive redundancy, a general definition [that also included] RD4 gene pairs; without CS information) and RD9C (with CS information). [Analysis of models without CS information (RD4 and RD9) revealed that the number of variables and the relative importance of the six feature categories largely corroborated the discoveries in the reference] (**fig. 5F**). [In summary], the ranking from best to worst, based on median importance ranks in those categories for RD4/RD9-based models [(without CS information)], was functional annotation (37/16) > network properties (57.5/64.5) > evolutionary properties (76/110) > gene expression (104/105) > protein properties (145/88) > epigenetic modifications (121/127), [with] gene expression [being] the category with the highest number of variables in both cases. These findings validated the reproducibility of the models and [ensured rigorous interpretation of subsequent results. Considering RD4C/RD9C-based models (with CS information),] the CS feature category was sixth/second in importance rankings and [emerged as the first] in terms of the number of variables for both cases. [This suggests that CS information is more valuable when prediciting general (RD9 definition gene pairs) rather than extreme redundancy (RD4 definition gene pairs).] This [notion] was further verified [when comparing] SVM models (see **Methods**) with different sets using AUC-ROC and AU-PRC values (**fig. 5G-H**). While CS data [notably] improved predictions for general redundancy [(RD9C vs RD9, AUC-ROC = 0.665 vs 0.634, AU-PRC = 0.651 vs 0.603),] it also reduced the values for the extreme definition [(RD4C vs RD4, AUC-ROC = 0.807 vs 0.842, AU-PRC = 0.795 vs 0.825).] Finally, we [observed] that the intersection with the highest number of features was common to all sets, [suggesting] that the core predicting power remained constant for all the models, [thereby ensuring accurate comparisons between all mentioned models] (**fig. 5I**).

Figure 5 (like others) is very dense. I would recommend either breaking it up into multiple figures or moving some panels to the supplement.

# We agree with the reviewer’s observation. In the reviewed manuscript, we have split Figure 5 into one main and one supplementary figure.

***Reviewer: 2***

*Comments to the Author:*

The authors measured the conservation of chromatin states between three plant species and showed that this can help predict functional divergence between paralogs. This is a compelling result of broad value to the molecular biology and evolution community.

# We thank the reviewer for the time and effort taken in reviewing our manuscript. We really appreciate that the reviewer recognized the broad value of PlantFUNCO to the valuable molecular biology and evolution community.

*Points:*

I have a concern about the framing in the introduction. The authors mention transgenerational epigenetics. While no doubt interesting, this is generally regarded as a phenomenon relating to the inheritance of cytosine methylation (CG methylation which is stable through cell divisions). I am skeptical about the similarity in histone modifications between species being due to shared inheritance but rather conserved function of cis and trans epigenome regulators that result in the genes/regions showing similar chromatin states between species. Thus, I think the framing in the intro could be improved and made more clear.

# The reviewer is absolutely right. While we are keen on retaining the mention of transgenerational epigenetics in the introduction of the manuscript to support the utility of plant species in this type of analysis, we apologize for any inconvenience caused. We have made adjustments to the following lines to avoid confusion with the inheritance of histone modifications, as suggested in this comment:

**[ Introduction, Page 2, Lines 56-61]**

This brings us to evolutionary epigenomics, and more generally, evolutionary functional genomics [– which are emerging fields evaluating how alterations in the conservation of epigenome regulators and cytosine methylation over multiple generations represent a crucial form of plasticity and epigenetic adaptation. Regulatory elements states have begun to be regarded as major targets of evolution, given that their diversity plays a critical role in phenotypic variance across all organisms, enabling them to adapt to various environmental niches.]

**[ Introduction, Page 3, Lines 69]**

instance, [cytosine methylation] are more easily transgenerationally transmitted due to soft

Along these lines, it is a bit unclear why phenotypic plasticity is mentioned many times in the manuscript.

# We thank the reviewer once again for these remarks. We included the mention of phenotypic plasticity because it is regarded as an adaptative trait subject to natural selection, predominantly influenced by epigenomics and functional genomics. Thus, it establishes a theoretical foundation for the development of comparative and evolutionary-related functional genomics analyses and databases. To prevent confusion, we have replaced the term “phenotypic plasticity” in the following sentences:

**[ Abstract, Page 1, Lines 28-29]**

Taking all the above into account, PlantFUNCO aim to leverage data diversity and extrapolate molecular mechanisms findings from different model organisms to determine the extent of functional conservation, thus, deepening our understanding of how [plants epigenomics and functional genomics have evolved.]

**[ Discussion, Page 10, Lines 352]**

We introduced PlantFUNCO, a database that allows for further inspection of the crosstalk between evolution [and epigenomics/functional-genomics.]

**[ Discussion, Page 13, Lines 447-448]**

Overall, PlantFUNCO aims to leverage data diversity and extrapolate findings from different models to determine the extent of molecular conservation, thus deepening our understanding of how [plants epigenomics and functional genomics have fascinatingly evolved.]

This paper goes beyond showing a number of interesting results and also is presented as a database for community use (PlantFUNCO). It would be nice if there was some more documentation for their website. For example, the Downloads tab has a bunch of files, and it was hard (if not impossible) to easily find out what these files all are. For this to be of wide use, it would help to have some clear guides about what these files are and how users can use them. Similarly, the “Overlap Enrichment” tool on the website was hard to understand what its actually doing.

# The reviewer is right. “Download” tab now incorporates documentation for the available files (https://rocesv.github.io/PlantFUNCO/Download.html). “Overlap Enrichment” tab now provides a summary of the main purpose of the application (https://rocesv.github.io/PlantFUNCO/ToolsOverlapEnrichment.html).

Minor but general comment: the figures are very dense and I am concerned the text size will be too small in print.

# We agree with the reviewer’s observation. In the resubmitted manuscript, we have split Figure 2 into two new main figures, Figure 3 into one main and one supplementary figure, and Figure 5 into one main and one supplementary figure.