

Genome assembly of a moth-like butterfly (*Macrosoma leucophasiata*) reveals convergent adaptation of visual genes to nocturnality

Corresponding Author: Professor Akito Kawahara

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I have read the manuscript by Singh and colleagues with great interest. The manuscript details the genome assembly of a species of Hedyliidae butterfly: *Macrosoma leucophasiata*, a member of the Lepidoptera family that reverts to a nocturnal lifestyle, similar to the majority of moths. The study primarily focuses on one specific aspect of convergent evolution: the evolution of phototransduction genes. Specifically, the authors conduct a gene expansion/contraction analysis on 31 genes and a positive selection scan on a subset of 7 opsins genes. While I strongly appreciate the biological interest of the study as a valuable system for studying convergent evolution, I see it as a missed opportunity to delve more deeply into patterns of convergence in various aspects of molecular biology. Additionally, I find the methodology implemented in the study somewhat superficial. I will provide feedback that I hope will enhance the study's quality, potentially making it more appealing to a broader audience interested in this unique group of butterflies.

The introduction is well-written, and I don't have much to add, except for two aspects. Firstly, I believe there is an unintended mistake in the description of the phylogeny of these butterflies. In lines 57-59, it is stated that the split of "Hedyliidae likely diverged from their sister group [...] less than 30 Mya." I find this inaccurate, as both the referenced papers and Figure 1 indicate that the split between Hedyliidae and Hesperidae occurred approximately 95 Mya, while the diversification of Hedyliidae themselves occurred around 30 Mya, as reported by Kawahara et al. in 2023, almost at the base of the butterfly's radiation estimated at ~100 Mya. This is an interesting aspect for describing and understanding how convergence occurred. This leads me to the second aspect of the introduction that, as a reader, I would have appreciated: a hypothesis on how convergence might have occurred. For example, the early split from other diurnal butterflies might indicate that the specialization to a diurnal lifestyle was not too strong. Additionally, a more detailed explanation of hypotheses testing on opsins and phototransduction genes, beyond just stating, "We focused our hypotheses on opsin genes, [...] we tested if opsin genes in hedyliid sharing the same diel niche show sequence similarity." What do they mean by "sequence similarity," and how are hedyliid opsins related to butterflies or moths? I suggest expanding on these aspects.

In my opinion, the most problematic aspect of the manuscript is the methods section. While I don't have many comments on the genome assembly, I am more concerned about its annotation. Because no RNAseq data was used with BRAKER, which is the standard configuration, and only homology information was adopted, I wonder how they ran BRAKER. There is not much description of how they aligned aa onto the genome, and why they didn't use more appropriate approaches such as GALBA, a branch of BRAKER specifically designed for gene prediction with only protein alignment. While I don't deny that BRAKER might do a good job, I didn't see any check on the quality of the prediction, such as a BUSCO run on the BRAKER output. I also have concerns about the list of species used for comparative genomics (supplementary table 1). Firstly, I think it should be listed in phylogenetic order to help readers understand the number of species for each clade and why that specific list was chosen. It was mentioned that only chromosome-level assemblies were used, but why only two genomes from the sister species of Hesperidae? I checked on NCBI, and there are 8 genomes assembled at the chromosome level due to the phylogenetic depth with the sister clade. Picking only 2 when more are available seems like a missed opportunity.

On the more technical aspect of the phylogeny, BUSCO genes were used, but no dataset, version, or method to identify them was specified. Also, apparently, single gene trees were generated directly from the alignment without any kind of filtering for paralogs or gaps. In my personal experience, this straightforward approach can lead to a lot of noise; I don't think

it's a very good approach. Many papers process the alignment before a phylogeny is carried out. I suggest doing the same, as it is the standard approach nowadays. Also, I didn't spot the resulting tree anywhere except in Figure 1, but bootstrap or other statistics (e.g.: concordance factor) regarding the robustness of the topology or branch length were not reported. This leads me to the following aspect, which I think was a bit superficial: the ultrametric tree reconstruction, which is key to inferring patterns of expansion/contraction and their rates. A simple penalized likelihood approach was applied without any indication of how the calibration was done. Again, I think it's a rather simplistic approach. I wonder why more complex approaches, such as Bayesian approaches, were not used. Works employing both Bayesian approaches and treePL showed contrasting results, with treePL leading to either older time estimates or younger node ages. These works also reported contrasting results regarding the precision of treePL time estimates (Costa et al., 2022). Considering this dataset with very few taxa, I can't see why more effort wasn't put into this aspect.

On the description of the methods used to annotate genes related to vision, I found it quite hard to follow. I wonder if they can streamline this section a bit better. I was very surprised that only the opsin genes were tested for positive diversifying selection. It's not clear if they used the whole tree from Figure 2 or used single trees, one per opsin. It's also not clear if and how they corrected the p-values for multiple tests. But what was more surprising was to see how they used not a single test to explicitly test for convergence or, for example, check whether there was a significant change in the selective regime between diurnal and nocturnal butterflies. There are many tests that could be run; I'll list a few. They could implement BUSTED-Phenotype to test the differential regime of selection, RELAX to test specifically for the relaxation or intensification of selection; both these are part of the toolkit of analyses in the HyPhy framework. For a more formal test of convergence, they could use RERconverge (DOI: 10.1093/bioinformatics/btz468) or CSUBST (DOI: 10.1038/s41559-022-01932-7).

The results section follows roughly the same pattern; there's nothing inherently wrong about it, but it feels a bit shallow. The CAFE analysis for gene turnover is very shallow, with no indication of the turnover rate, reflecting its method section, which is almost a single line in the method. They miss the opportunity to explore these results a bit more. Are the contractions they report the effect of the gene annotation? What are the expansions? Regardless of the copy number variation of the phototransduction genes, is there any meaningful result? For example, DAGL seems to be duplicated; any sign of selection? Is the other copy a pseudogene? Are the two copies in- or out-paralogs? Regarding the innexin gene loss, did they check if the synteny of the genomic location where they should be located is conserved? I found no trace of the meme analysis in the results. If they explain the results, I couldn't find a trace. I was expecting some consideration on sites under selection, maybe mapped onto the model of the protein structure.

Overall, I believe in the quality of the data, but I also strongly believe that a more accurate implementation of the methodologies with a more in-depth explanation of the resulting analyses would really benefit the quality of this very interesting study.

Reviewer #2

(Remarks to the Author)

The authors present a genome assembly of the nocturnal hedyliid butterfly *Macrosoma leucophasiata*. They use this genome to compare sequence variation of vision-related genes between *M. leucophasiata* to a diversity of butterfly and moth species. They find that several opsin genes of *M. leucophasiata* share sequence similarity with homologs of more distantly related moth species. They also investigate a subset of these opsin genes in detail and identify signatures of positive selection in the *M. leucophasiata*. The authors conclude that these patterns suggest convergent evolution in the vision genes of this species and moths.

The paper is well written and based on what is reported the assembly is well done, documented, and annotated. I do however have several questions about evidence for and interpretation of functional adaptation/evolution.

Major concerns:

1. I was disappointed that the genome assembly was not accessible. PRJNA1033655 on NCBI only has biosample and SRA records associated with it. Therefore, I am unable to independently judge the quality of the assembly and annotation.
2. The title suggests that "signatures of adaptation" will be revealed, but there is no analysis of the functional consequences of the sequence variation reported. The authors note that they "... focused our hypotheses on opsin genes, given their strong genotype-phenotype..." but they do use this opportunity. The adaptive arguments could be strengthened by mapping sites under strong selection to the function domains of the opsins as was done in Figure 3 of - <https://doi.org/10.1038/s42003-021-01688-z>. I would also be interested to know how the sites of strong selection for nocturnality in this study compare to those identified for bright-light adaptation in <https://doi.org/10.1038/s42003-021-01688-z>.
3. The possible loss of innexin genes is very interesting, but a more detailed analysis is needed. Are these arrayed on the same chromosome and have been lost through recombination? If so can you provide evidence of such structural variation at this locus? If they have not been lost through recombination, do you find evidence of pseudogenization?

Specific comments:

Title - It would be more accurate to say "new genome assembly" I think the genome has been around for a few million years :)

Lines 4-5 - I disagree with this statement. Adaptive changes in the visual system with light environment transitions are a classic topic in sensory neurobiology. These transitions are well understood to have shaped modern visual systems. For example, in the human lineage, the contraction and subsequent expansion of the opsin complement have been linked to the transition to an ancient transition to nocturnality and a more recent transition back to diurnality

Line 17 - What about the poles? These habitats are subject to continuous light or dark depending on the time of year

Line 27 - this process is typically described as “capture” rather than “intake”

Line 29 - Insects have “evolved” rather than “developed”

Line 30-32 - Can the authors provide a citation for this statement?

Line 35 - This wording is a bit strange. “Also” compare to what? “subtle”? The preceding paragraph describes profound differences in eye morphology associated with light environment.

Line 41 - Which patterns? I do not know of any way to infer temporal resolution or absolute sensitivity from the genome.

Line 51-52 - Can the authors provide a citation for this statement?

Line 59-61 - Can the authors explain the significance of this eye design? Are there genomics signatures we should expect to be associated with this design?

Line 66-68 - Please provide a citation for the conservation of repertoire size.

Line 82-84 - Interesting - could these be from a parasite infection or microbiota? Would be helpful to note in the accession meta-data of the raw reads so folks with an interest in parasites/microbiota can investigate in the future.

Line 95 - Is this level of repeat content typical in butterflies?

Lines 119-121 - Blastp? Were the gene sequences translated to the predicted AA sequence before searching? Please clarify.

Lines 139 - 143 - Sup Data 10 - Please annotate the trees to make this pattern more clear. Also, please provide a key for the species name abbreviations and annotate nocturnal vs. diurnal.

Line 144 - This is a question beyond the scope of the manuscript, but I am curious how these models of selection deal with the fact that, for diploid genomes, the reference genome sequences may be consensus mosaics of two different alleles.

Lines 151-152 - please define what the different rates classes mean.

Lines 171-172 - *M. leucophasiata* pteropsin is missing from figure 2. Why?

Line 199 - DAGL-Beta - please spell out the full name of this gene and briefly describe its putative function.

Line 214-216 - As mentioned above, this loss needs greater exploration. If you investigate the syntenic regions are there pseudogenes present? Is there evidence of recombination at the locus (loci) that could explain loss?

Line 227-228 - as mentioned above, it is difficult to understand the significance of this result without an introduction to the function of this gene.

Lines 242-244 - How?? What evidence do you have that the changes you observed are influencing sensitivity?

Line 262-265 - Although it may not be possible to precisely infer spectral tuning, it would be helpful to present the sites under selection. Where in the protein are they? What domains? e.g. Figure 3 from - <https://doi.org/10.1038/s42003-021-01688-z>

Line 438 - A search for PRJNA1033655 returns only biosample and SRA data. The assembled and annotated genome is not available

Line 440 - an all databases search of NCBI for OR825356 returns 0 results

Figure 1 - Why does this heat map have 5 or 6 levels of intensity? What are these data? The purported range is from 0 to 2, so there should only be three levels 0 copies, 1 copy, or 2 copies.

Figure 2. - Why is *M. leucophasiata* pteropsin missing from this figure, but included in later selection analyses?

(Remarks to the Author)

In this study, Singh and colleagues sequenced a novel moth genome, explored the evolutionary dynamics of opsins and phototransduction genes in response to changes in the photic niche and proposed a convergent evolution of opsin genes.

Overall, I find the manuscript intriguing, and the analyses are cutting-edge. However, I have a few concerns that need to be addressed:

1. The results section requires more background information in the results on the methods used. For instance, I understood the authors' use of GeneRax just at the end of the manuscript.
2. If the authors have identified a new opsin paralog (refer to Figure 2) is an important result. However, before reaching this conclusion, the results need to be compared with previous studies, i.e. <https://doi.org/10.1093/molbev/msad241>, by combining the two datasets together and evaluating the placement of the unknown opsins
3. The authors should calculate duplications and losses for phototransduction genes using GeneRax, as they did for opsins.
4. The multifunctionality of opsins, as indicated by several studies (e.g., <https://doi.org/10.1093/molbev/msac033>, [10.1146/annurev-cellbio-100616-060432](https://doi.org/10.1146/annurev-cellbio-100616-060432), and [0.1016/j.cub.2020.05.009](https://doi.org/10.1016/j.cub.2020.05.009)), should be discussed in the context of positive selection.
5. The observed cluster of hedylid opsins, which differs from the species tree, is an interesting observation. However, a simple UFB tree is insufficient, as the clade might be artefactual. The analysis of the opsins' genes should be performed using Transfer Bootstrap Expectation as implemented in IQTREE, which will allow us to account for possible unstable sequences. Additionally, it would be interesting to perform an approximately unbiased test and compare the two hypotheses.

Minor comments:

- Line 93: There seems to be an issue with this sentence. What was the DFAM database used for?
- Line 106: More information is needed here. Which model was used? What is well-supported?
- Line 111-112: OrthoFinder is applied to proteins, not transcripts. Please clarify.
- Please explain the methods used for annotating the HOGs. Did you use Egnog?
- How was the opsin phylogeny performed? What are the "Unclassified (UnRh)" opsins(see above)?
- Figure 1: Please use a different scale from white to cyan. It is challenging to discern between species where genes are absent and those with two copies.
- Figure 2: Please add support values for key nodes. If possible, avoid using circular tree output as it is difficult to read.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

I carefully evaluate all the responses to my questions/clarifications. I have to say that I'm delighted with all the modifications the authors made to the current version of the manuscript. I think all the suggestions made by all the referees substantially improved the manuscript's quality.

With the risk of being pedantic, I have few and minor comments.

I think there's something wrong with the current title for the manuscript. It reads:

"Genome reveals molecular signature of adaptation [...]". I saw this was the suggestion from another referee, I'd suggest something like: "The genome assembly of *Macrosoma leucophasiata* (Hedylidae) reveals molecular signature of convergent molecular adaptation to nocturnality in visual genes of moth-like butterflies"

I think it would be more appropriate. this is just a suggestion, I leave it to the authors to decide whether they want to change it.

I found a lot of methodological references in the result section. Something like: "A preliminary survey using Genomescope 2.0 estimated [...] / "filtering haplotigs via purging and non-target sequence removal using BlobTools" / "we modeled and soft-masked repeat regions using RepeatModeler2 and RepeatMasker [...]" / "transposable elements (TE) repeat families was generated corresponding to the Dfam database [...]" / "For gene annotation, we used the resulting soft-masked genome to run the BRAKER3 pipeline with protein evidence from the OrthoDB v1141 catalog [...]", just to list a few are more appropriate in the methods section.

In the Discussion, the paragraph about "Nvel genome assembly for Hedylidae", I don't think is particularly relevant. I think there are so many groups of invertebrates where we can say the same. Probably the scientific invertebrate communities, the lepidoptera's one has been blessed with so many few genomes being sequenced and published every month or so.

In the Discussion, the paragraph about the "Novel genome assembly for Hedyllidae" does not seem particularly relevant. Many invertebrate groups can make similar claims. In comparison to other scientific communities focused on invertebrates, the one for Lepidoptera has seen a significant number of genomes sequenced and published every month or so.

Reviewer #2

(Remarks to the Author)

With this revision, the authors have addressed my major concerns. However, there several places in the manuscript where questions remain or clarification is needed:

Line 14 - Why is "Blue" capitalized here?

Line 32-34 and 63-67 - Superposition eye structure is introduced as a major feature of nocturnal adaptation, but is not discussed in the rest of the paper. Are there specific molecular signatures that we should expect to see in animals with superposition eyes? Are there present in *M. leucophasiata*?

Line 76-77 - Throughout it is assumed that a shift to nocturnality will result in selection and adaptive change in the opsins. However, a shift to nocturnality could also reduce selective pressures if, for example, color vision becomes relatively less important in the animal's life. This might lead to neutral processes dominating the evolution of these genes.

Line 132 - Interesting observation. I realize this is beyond the scope of this paper, but might these changes in odorant receptor and P450 repertoire size be linked to shifts in larval hosts?

Line 137 - Provide a brief note of the thresholds used in the BLASTp search

Line 142 - I am a little unclear what 3,721 genes is referring to. Are there distinct genes or does this include homologs of the same gene in different species?

Line 154-155 and Figure 1 - Given the evidence that the missing innexins are most likely an artifact of the assembly it is not accurate to state they are missing in the text or represent them as missing in Figure 1. The figure 1 legend states "...losses of innexin in *M. leucophasiata*"

Line 159-161 - I am confused by this statement, how can the contig from *M. leucophasiata* be linked to a *T. sylvestris* chromosome? Do you mean the contig is mapping to that chromosome?

Line 167 - What is meant by cluster together? What metric? AA, nucleotide identity, or some other feature of the gene?

Line 211 - "but LWRh recovered 5 and 6 respectively" - what does respectively refer to? Are there two copies of this gene?

Line 213 - What is the relevance of squid here? From the rebuttal letter, I understand that this template was used in an early study, but it does not seem necessary to discuss here.

Line 244-245 - Could the rarity of gains and losses be attributable to the distinct photic niche of the larva? Are they diurnal?

Line 253 - subterranean?? Do you mean aquatic or amphibious?

Line 273 - "nosed into the assembling process" - what does this mean?

Line 288 - 289 - Can you speculate on DAGLB's role in adaptation? In vertebrates, it is involved in endocannabinoid signaling, not sure how that relates to dimlight vision.

Line 306 - What other factors might result in grouping with nocturnal species? Is introgression a possibility? Could this possibility be excluded with an ABBA BABA test?

Line 319-320 - What previous study?

Reviewer #3

(Remarks to the Author)

I am satisfied with almost all the authors' replies; however, I have two final comments:

The term "unclassified opsin" is confusing. Macias-Muñoz et al., 2019 is a paper on Hydra opsin, and Kanost et al., 2016 is a genome paper not focused on opsin (opsin is just one of the supplementary figures). Insect opsins are divided into C, R, RGR/Go, and probably arthropsin (<https://doi.org/10.1093/gbe/evw015>, <https://doi.org/10.1073/pnas.1424670112>). Please name the clades accordingly.

As I suggested in previous comments, please add the BS (UFB and TBE) for key nodes in Figure 2.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Responses to reviewers' comments

Reviewer #1 (Remarks to the Author):

I have read the manuscript by Singh and colleagues with great interest. The manuscript details the genome assembly of a species of Hedyliidae butterfly: *Macrosoma leucophasiata*, a member of the Lepidoptera family that reverts to a nocturnal lifestyle, similar to the majority of moths. The study primarily focuses on one specific aspect of convergent evolution: the evolution of phototransduction genes. Specifically, the authors conduct a gene expansion/contraction analysis on 31 genes and a positive selection scan on a subset of 7 opsins genes. While I strongly appreciate the biological interest of the study as a valuable system for studying convergent evolution, I see it as a missed opportunity to delve more deeply into patterns of convergence in various aspects of molecular biology. Additionally, I find the methodology implemented in the study somewhat superficial. I will provide feedback that I hope will enhance the study's quality, potentially making it more appealing to a broader audience interested in this unique group of butterflies.

Response: Thank you for your positive and constructive feedback on the manuscript. We have tried to revise the manuscript as per your suggestions and list out the individual responses below.

The introduction is well-written, and I don't have much to add, except for two aspects. Firstly, I believe there is an unintended mistake in the description of the phylogeny of these butterflies. In lines 57-59, it is stated that the split of "Hedyliidae likely diverged from their sister group [...] less than 30 Mya." I find this inaccurate, as both the referenced papers and Figure 1 indicate that the split between Hedyliidae and Hesperidae occurred approximately 95 Mya, while the diversification of Hedyliidae themselves occurred around 30 Mya, as reported by Kawahara et al. in 2023, almost at the base of the butterfly's radiation estimated at ~100 Mya. This is an interesting aspect for describing and understanding how convergence occurred. This leads me to the second aspect of the introduction that, as a reader, I would have appreciated: a hypothesis on how convergence might have occurred. For example, the early split from other diurnal butterflies might indicate that the specialization to a diurnal lifestyle was not too strong. Additionally, a more detailed explanation of hypotheses testing on opsins and phototransduction genes, beyond just stating, "We focused our hypotheses on opsin genes, [...] we tested if opsin genes in species sharing the same diel niche show sequence similarity." What do they mean by "sequence similarity," and how are hedylid opsins related to butterflies or moths? I suggest expanding on these aspects.

Response: Yes, the description of the phylogeny had an error. Thank you for the correction and the suggestion to consider this aspect of the early split in the introduction. We have now made the correction and edited the sentences at [line 60-63](#), to further expand on it. We have also revised the text to expand on the points raised above regarding the hypotheses and hedylid opsins from [line 75-86](#).

In my opinion, the most problematic aspect of the manuscript is the methods section. While I don't have many comments on the genome assembly, I am more concerned about its annotation.

Because no RNAseq data was used with BRAKER, which is the standard configuration, and only homology information was adopted, I wonder how they ran BRAKER. There is not much description of how they aligned aa onto the genome, and why they didn't use more appropriate approaches such as GALBA, a branch of BRAKER specifically designed for gene prediction with only protein alignment. While I don't deny that BRAKER might do a good job, I didn't see any check on the quality of the prediction, such as a BUSCO run on the BRAKER output. I also have concerns about the list of species used for comparative genomics (supplementary table 1). Firstly, I think it should be listed in phylogenetic order to help readers understand the number of species for each clade and why that specific list was chosen. It was mentioned that only chromosome-level assemblies were used, but why only two genomes from the sister species of Hesperidae? I checked on NCBI, and there are 8 genomes assembled at the chromosome level due to the phylogenetic depth with the sister clade. Picking only 2 when more are available seems like a missed opportunity.

Response: Thank you for the comments. We chose BRAKER3 instead of GALBA for two reasons. First, GALBA uses protein sequences from closely related species, which does not apply to *Macrosoma* as it's the first genome in the family. On the other hand, the GeneMark-EP+ protein training pipeline in BRAKER3 can take protein sequences from further distant species with accuracy dropping only very little. Second, we considered BRAKER with the suggestions of the developers as they noted: "BRAKER pipeline will very likely produce more accurate results than GALBA in small and medium size genomes". For the gene model quality checking, we used BUSCO on the gene model, please see the method at line 395-396. The result was reported in Supplementary Table 4 but now we have added result information in the result section in line 110-113. For the supplementary table, thank you for the suggestion, we have reordered the table to match the phylogeny. And we chose the two skippers (Hesperidae) for two reasons. First, we would like to make the phylogeny (both species tree and gene trees) concise and balanced, as we would like to compare the genome and phototransduction genes to many other butterfly species. Second, we chose species based on the clear notation/reference about their diel niche and we think the diel niche of at least some of the skippers with a genome are not quite clear. To avoid the potential uncertainty in determining diel niche evolution, we decided to stay with the species with diel niche well documented. We hope these explanations are satisfactory.

On the more technical aspect of the phylogeny, BUSCO genes were used, but no dataset, version, or method to identify them was specified. Also, apparently, single gene trees were generated directly from the alignment without any kind of filtering for paralogs or gaps. In my personal experience, this straightforward approach can lead to a lot of noise; I don't think it's a very good approach. Many papers process the alignment before a phylogeny is carried out. I suggest doing the same, as it is the standard approach nowadays. Also, I didn't spot the resulting tree anywhere except in Figure 1, but bootstrap or other statistics (e.g.: concordance factor) regarding the robustness of the topology or branch length were not reported. This leads me to the following aspect, which I think was a bit superficial: the ultrametric tree reconstruction, which is key to inferring patterns of expansion/contraction and their rates. A simple penalized likelihood approach was applied without any indication of how the calibration was done. Again, I think it's a rather simplistic approach. I wonder why more complex approaches, such as Bayesian approaches,

were not used. Works employing both Bayesian approaches and treePL showed contrasting results, with treePL leading to either older time estimates or younger node ages. These works also reported contrasting results regarding the precision of treePL time estimates (Costa et al., 2022). Considering this dataset with very few taxa, I can't see why more effort wasn't put into this aspect.

Response: The database we used to run BUSCO is lepidoptera_odb10. This information was only in the gene annotation section (materials and methods) so we added it to genome assembly and decontamination section, too. Please see revised text at line 411. For single gene trees, at least for opsin genes, we did eye-examine all the gene alignments to ensure the quality of the alignments. We did not remove any sequence including duplicated genes as they all look normal to us and these orthogroups have passed the false duplication control. To clarify that the eye-examination was done, we revised the text at line 508-518 to include this information. For the species tree, we describe the result at line 110-113 and we noticed that the figure description does not include the explanation of the black dots on the nodes which are the fully supported branches by SH-aLRT and ultrafast bootstrap. The description has been added to the figure 1 caption at line 870-871. For ultrametric tree reconstruction, we agree that some details are missing in the method section. We have added more detailed descriptions to line 440-443. Thank you for your suggestions. For Bayesian approaches, we did not apply RelTime or MCMCtree because we believe that the relative branch lengths will be insensitive to the calibration points we applied to the tree. And since the divergence time estimation (the absolute branch lengths) is not our main focus, we think it is safe enough to use the relative branch length inferred from treePL to perform expansion/contraction analysis. Again, we hope this explanation about not using the Bayesian approach is satisfactory.

On the description of the methods used to annotate genes related to vision, I found it quite hard to follow. I wonder if they can streamline this section a bit better. I was very surprised that only the opsin genes were tested for positive diversifying selection. It's not clear if they used the whole tree from Figure 2 or used single trees, one per opsin. It's also not clear if and how they corrected the p-values for multiple tests. But what was more surprising was to see how they used not a single test to explicitly test for convergence or, for example, check whether there was a significant change in the selective regime between diurnal and nocturnal butterflies. There are many tests that could be run; I'll list a few. They could implement BUSTED-Phenotype to test the differential regime of selection, RELAX to test specifically for the relaxation or intensification of selection; both these are part of the toolkit of analyses in the HyPhy framework. For a more formal test of convergence, they could use RERconverge (DOI: 10.1093/bioinformatics/btz468) or CSUBST (DOI: 10.1038/s41559-022-01932-7).

Response: Thank you for this point; we agree that the procedure is complicated. We have added a visualized workflow to Supplementary Figure 8 and we hope this figure can help readers to understand the method easier. Please see the changes in Supplementary Information file and main manuscript at line 458. For why we focused only on opsin genes among those phototransduction genes, we examined the gene trees for other phototransduction genes as well,

but did not see strong support for the discordant tree topology where *M. leucophasiata* clusters with the nocturnal moths. Hence, we only focused the tests on opsins.

Single gene trees, per opsin, and their respective alignment files were used as input for the Hyphy analyses. We have now specified it in the results (Line 180-181) and the methods (Line 522) as it was not clearly stated earlier. The p-values reported in Supp. Table 2 with aBSREL results are obtained after Bonferroni-Holm correction. This is implemented in aBSREL when using the 'foreground' method to control for family-wise error rates and the output JSON file shows the corrected p-values (Ref- Smith et al., 2015). For formally testing convergence, we actually tried CSUBST but have faced difficulty from the tree topology. Briefly, the opsin gene trees that have *M. leucophasiata* clustered with nocturnal moths cannot be used to run CSUBST because in CSUBST, molecular convergence cannot be detected between immediate sisters. Here we attached the issue we addressed for more details of the conversation: <https://github.com/kfuku52/csubst/issues/44>.

The results section follows roughly the same pattern; there's nothing inherently wrong about it, but it feels a bit shallow. The CAFE analysis for gene turnover is very shallow, with no indication of the turnover rate, reflecting its method section, which is almost a single line in the method. They miss the opportunity to explore these results a bit more. Are the contractions they report the effect of the gene annotation? What are the expansions? Regardless of the copy number variation of the phototransduction genes, is there any meaningful result? For example, DAGL seems to be duplicated; any sign of selection? Is the other copy a pseudogene? Are the two copies in- or out-paralogs? Regarding the innexin gene loss, did they check if the synteny of the genomic location where they should be located is conserved? I found no trace of the meme analysis in the results. If they explain the results, I couldn't find a trace. I was expecting some consideration on sites under selection, maybe mapped onto the model of the protein structure.

Response: Thank you for your suggestions. We agree that the information of CAFE analysis we provided in both method and result sections are insufficient. For the method section, we added the number of gamma categories (k) we used in the analysis which is based on several test runs with different numbers of k, and we let CAFE to estimate the gene family evolutionary rate (lambda) (Line 449-451). Here, we have listed our answers to each of the questions:

- *Are the contractions they report the effect of the gene annotation?* **R:** We don't think the gene annotations have a strong impact on the repertoire size change analysis because the BUSCO completeness scores of the gene models are generally high (>92.9% except for one species, please see Supplementary Table 4), and the species with slightly lower gene model BUSCO score do not show more gene contraction than other species.

- *What are the expansions?* **R:** We interpret this question as the unclear definition of expansion in the result section. We have added additional definition to the term at line 130-131.

- *Regardless of the copy number variation of the phototransduction genes, is there any meaningful result?* **R:** We didn't see rapidly evolving genes from those candidate vision-related HOGs. And for other rapidly evolving genes, most of them are annotated with retrotransposon-related genes which have been identified to have frequent and drastic rapid repertoire size changes but the underlying mechanism and functions are possibly unclear. Since we are focusing on the

phototransduction-related genes, we constrained the interpretation on those rapidly evolving genes.

- *DAGL seems to be duplicated; any sign of selection? Is the other copy a pseudogene? Are the two copies in- or out-paralogs?* **R:** We think the duplicated DAGL-beta gene is unlikely to be pseudogene because 1) BRAKER is designed primarily to predict active protein-coding genes and 2) both copies have start and stop codons with similar length to the sequences of other species. And we are unable to tell if this duplication is in- or out-paralogs as there are many other *Macrosoma* species remaining unsampled so the time when the duplication occurred is undetermined. What we can say now is that, based on the gene tree topology (shown in Supplementary Data 10), it is likely that the gene duplication occurred after the emergence of *Macrosoma*, but not sure if this is specific to this species. Plus the function of DAGL-beta in moths and butterflies is still unclear, we have taken additional caution not overinterpreting the evolutionary consequence of this duplication. We add an additional statement about when the duplication is likely to occur at line 286-288. A similar idea applies to the losses of innexin genes; we believe that we will have a much clearer picture about the meaning of these gene loss and gain once we have more genomes from other *Macrosoma* species.

- *Regarding the innexin gene loss, did they check if the synteny of the genomic location where they should be located is conserved?* **R:** Thank you for the suggestion, this is a great idea. We have looked at the synteny position by comparing the orthologs from the other two skipper species and found that the four missing innexin genes are located in the same region which is absent in our *M. leucophasiata* genome. Because there should have been about 33 busco genes in that missing region (33 busco genes in the corresponding region of a skipper species and missing in our *M. leucophasiata* genome), we further checked our Hifiasm raw contig and unitig assemblies for busco genes. We found those missing busco genes can be recovered only in the raw unitig assembly, suggesting that during the assembling process, the region was not assembled for some unclear reasons. We have revised our manuscript with the new findings, please see the changes at line 155-165, 140-153, and 492-501.

- *I found no trace of the meme analysis in the results. If they explain the results, I couldn't find a trace. I was expecting some consideration on sites under selection, maybe mapped onto the model of the protein structure.* **R:** We have expanded on the meme results and mapped the sites under positive selection to opsin protein models. We primarily report the results for UV and Blue opsin since the sequences were complete enough to recover 7 transmembrane helices which makes the 3d models more likely to be accurate. The two long wavelength opsins had 5 and 6 predicted transmembrane domains which would make the 3d models less accurate (see additional supplementary figure). We modeled the 3d structure using swiss-pdb with the jumping spider crystal structure as a template and mapped the sites under positive selection and compared their proximity to the retinal binding region (RBR) in 3D space (4 Angstrom from retinal). The position of the retinal was determined by structural alignment of the jumping spider opsin with the retinal ligand and the predicted Hedyliidae opsin sequence structure. We find that UV opsin did not have many sites surrounding or close to the RBR, but the Blue opsin did have 4 sites that were within 3-6 Angstrom from this retinal interacting amino acids. These results are significant and have been included in a new main figure (Figure 3) and the results for UV and Blue have

been included in the supplementary figures. Additional results (line 209-221) and methods (line 535-546) have also been added further addressing these points.

Overall, I believe in the quality of the data, but I also strongly believe that a more accurate implementation of the methodologies with a more in-depth explanation of the resulting analyses would really benefit the quality of this very interesting study.

Reviewer #2 (Remarks to the Author):

The authors present a genome assembly of the nocturnal hedyliid butterfly *Macrosoma leucophasiata*. They use this genome to compare sequence variation of vision-related genes between *M. leucophasiata* to a diversity of butterfly and moth species. They find that several opsin genes of *M. leucophasiata* share sequence similarity with homologs of more distantly related moth species. They also investigate a subset of these opsin genes in detail and identify signatures of positive selection in the *M. leucophasiata*. The authors conclude that these patterns suggest convergent evolution in the vision genes of this species and moths.

The paper is well written and based on what is reported the assembly is well done, documented, and annotated. I do however have several questions about evidence for and interpretation of functional adaptation/evolution.

Major concerns:

1. I was disappointed that the genome assembly was not accessible. PRJNA1033655 on NCBI only has biosample and SRA records associated with it. Therefore, I am unable to independently judge the quality of the assembly and annotation.

Response: We apologize that the data was not available at the moment when the MS was reviewed. We have double-checked it on NCBI and the assembly is now publicly available on NCBI with the project accession number PRJNA1033655.

2. The title suggests that “signatures of adaptation” will be revealed, but there is no analysis of the functional consequences of the sequence variation reported. The authors note that they “... focused our hypotheses on opsin genes, given their strong genotype-phenotype...” but they do use this opportunity. The adaptive arguments could be strengthened by mapping sites under strong selection to the function domains of the opsins as was done in Figure 3 of - <https://doi.org/10.1038/s42003-021-01688-z>. I would also be interested to know how the sites of strong selection for nocturnality in this study compare to those identified for bright-light adaptation in <https://doi.org/10.1038/s42003-021-01688-z>.

Response: We repeated the analyses as done in <https://doi.org/10.1038/s42003-021-01688-z> but used PyMol and the crystal structure of the spider opsin assuming it would be more closely related than the squid opsin. We find less support for the functional pockets for the UV opsin, but surprisingly, two of the amino acids (88S and 149G), while they do not directly appear to map to

retinal, were also recovered as sites under positive selection from the previous study (59S and 116G. They are both 60 (61 vs. 57) amino acids apart and appear on the same regions in the 2d structure. We don't see any such pattern for Blue opsin, with the hedylid opsin showing many more sites under positive selection. We think this is added support for the importance of these amino acids in tuning, but being cautious about over-interpreting our results, add a few points about this in the discussion ([See Lines 313-325](#)).

3. The possible loss of innexin genes is very interesting, but a more detailed analysis is needed. Are these arrayed on the same chromosome and have been lost through recombination? If so can you provide evidence of such structural variation at this locus? If they have not been lost through recombination, do you find evidence of pseudogenization?

Response: Thank you for the great advice, we agree that we should look at the possible explanations for the missing innexin genes. As was also suggested by the other reviewer, we compared the genome with other two skipper species and found that the four missing innexin genes are located in a DNA segment which is absent in our *M. leucophasiata* genome. And the 33 busco genes in the orthologous region of a skipper species are also missing in our *M. leucophasiata* genome. We found those missing busco genes can be recovered only in the raw unitig assembly, suggesting that during the assembling process, the region was not assembled for some unclear reasons. We have revised our manuscript with the new findings, please see the changes at line [155-165](#), [140-153](#), and [492-501](#).

Specific comments:

Title - It would be more accurate to say “new genome assembly” I think the genome has been around for a few million years :)

Response: We have edited the title of the manuscript accordingly.

Lines 4-5 - I disagree with this statement. Adaptive changes in the visual system with light environment transitions are a classic topic in sensory neurobiology. These transitions are well understood to have shaped modern visual systems. For example, in the human lineage, the contraction and subsequent expansion of the opsin complement have been linked to the transition to an ancient transition to nocturnality and a more recent transition back to diurnality

Response: That is a fair argument and we agree with it. However, considering the point that non-model species such as Hedylidae and other insects still have a lot to be explored, we have edited the sentence as follows: *While the genetic and neural basis of visual adaptation is well-studied in some model systems, our understanding of these adaptations across diverse lineages remains incomplete.*

Line 17 - What about the poles? These habitats are subject to continuous light or dark depending on the time of year

Response: Yes, it applies to poles as well. Phenomenon such as the midnight sun causes arrhythmic behavioural patterns in animals.

Line 27 - this process is typically described as “capture” rather than “intake”

Response: The sentence has been edited to incorporate the change in the revision, as suggested.

Line 29 - Insects have “evolved” rather than “developed”

Response: The sentence has been edited to incorporate the change in the revision, as suggested.

Line 30-32 - Can the authors provide a citation for this statement?

Response: A citation (Greiner, A. 2006) has been added for the statement.

Line 35 - This wording is a bit strange. “Also” compare to what? “subtle”? The preceding paragraph describes profound differences in eye morphology associated with light environment.

Response: We have edited the sentence in **Line 35**, as follows: “Adapting to dimly lit environments entails adjustments of the visual and nervous system in different ways.”

Line 41 - Which patterns? I do not know of any way to infer temporal resolution or absolute sensitivity from the genome.

Response: The sentence has now been edited in **Line 39-42** to address the limitations as well as the usefulness of genomics in studying visual systems of animals.

Line 51-52 - Can the authors provide a citation for this statement?

Response: We have added a citation (Kinoshita & Arikawa, 2023) for this statement.

Line 59-61 - Can the authors explain the significance of this eye design? Are there genomics signatures we should expect to be associated with this design?

Response: The significance of superposition eyes in Hesperidae and Hedylidae is explained in the revised text in **Line 62-67**.

There are differences in genomic signatures between the two eye designs which are not exclusively limited to the anatomical differences. Liénard et al. (2022) discuss examples of how functions of vision genes can be explored from a genomic and functional aspect.

Line 66-68 - Please provide a citation for the conservation of repertoire size.

Response: The sentence has been modified now and the citation for Macias-Muñoz et al., 2019 has been added.

Line 82-84 - Interesting - could these be from a parasite infection or microbiota? Would be helpful to note in the accession meta-data of the raw reads so folks with an interest in parasites/microbiota can investigate in the future.

Response: We have searched for relevant literature but found little information about parasitism of Euglenozoa in moths and butterflies (mostly in Bees and other insects). We agree that the sequence data could be helpful for other studies so we have made the sequence available in our figshare repository (<https://doi.org/10.6084/m9.figshare.24807792.v2>).

Line 95 - Is this level of repeat content typical in butterflies?

Response: We cannot comment on typical repeat content, as it seems to be quite variable across the families and is often positively correlated with genome size. For Hesperidae and Nymphalidae members, 37-50% repeat content is commonly seen in publicly available genome assemblies.

Lines 119-121 - Blastp? Were the gene sequences translated to the predicted AA sequence before searching? Please clarify.

Response: We apologize for the confusion. We used protein sequences predicted from the primary transcript to blast against the published phototransduction-related genes. The description has been revised accordingly at [Line 135](#).

Lines 139 - 143 - Sup Data 10 - Please annotate the trees to make this pattern more clear. Also, please provide a key for the species name abbreviations and annotate nocturnal vs. diurnal.

Response: Thank you for the suggestion. We found many gene trees are very big and difficult to visualize as an image in a page, we thus provide the tree file in newick format to make it more reusable. The trees can be viewed in general tools such as TreeViewer or Figtree. A key for the species name abbreviations is given in the Supplementary information word document.

Line 144 - This is a question beyond the scope of the manuscript, but I am curious how these models of selection deal with the fact that, for diploid genomes, the reference genome sequences may be consensus mosaics of two different alleles.

Response: Based on our understanding, these dN/dS based selection analyses assume that the genetic variations in these protein coding genes are conserved within species, compared to interspecific variations. Thus, we believe that the intraspecific variation will not be considered. And since the genome was assembled with good-coverage long read sequences, we believe that the haplotypes are well phased and the genes used in the selection analyses are the primary assembly that has more uninterrupted contigs.

Lines 151-152 - please define what the different rates classes mean.

Response: An explanation of the ω rate classes has been added now in [Line 184-187](#).

Lines 171-172 - *M. leucophasiata* pteropsin is missing from figure 2. Why?

Response: Thank you for pointing it out. The *pteropsin* gene was initially detected in the *M. leucophasiata* annotation but the gene got removed from the final gene model upon filtering based on the criteria described in [Line 458-472](#). Hence, *pteropsin* got excluded from the gene tree and selection analyses for *M. leucophasiata*. We have now corrected the sentence in the results as well as the figure caption of Supplementary figure 4.

Line 199 - DAGL-Beta - please spell out the full name of this gene and briefly describe its putative function.

Response: Thank you. We have revised our manuscript accordingly at line [248](#), and the brief function was described at line [285-286](#).

Line 214-216 - As mentioned above, this loss needs greater exploration. If you investigate the syntenic regions are there pseudogenes present? Is there evidence of recombination at the locus (loci) that could explain loss?

Response: We agree that more exploration is needed. Please see our response for the comment at major concern 3.

Line 227-228 - as mentioned above, it is difficult to understand the significance of this result without an introduction to the function of this gene.

Response: For the function of this gene, there is no clear evidence to what function of the gene in insects. But it was proposed to be related to phototransduction cascade and neural development, based on the gene expression data from different tissues. We briefly address this at line 285-287.

Lines 242-244 - How?? What evidence do you have that the changes you observed are influencing sensitivity?

Response: We agree that there was not enough evidence to state that our findings support the premise, hence we have removed the sentence in line 304. Moreover, given the new analyses with the protein models, we have further revised the discussion.

Line 262-265 - Although it may not be possible to precisely infer spectral tuning, It would be helpful to present the sites under selection. Where in the protein are they? What domains? e.g. Figure 3 from - <https://doi.org/10.1038/s42003-021-01688-z>

Response: We repeated the analyses as done in <https://doi.org/10.1038/s42003-021-01688-z> but used PyMol and the crystal structure of the spider opsin assuming it would be more closely related than the squid opsin. We find less support for the functional pockets for the UV opsin, but surprisingly, two of the amino acids (88S and 149G), while they do not directly appear to map to retinal, were also recovered as sites under positive selection from the previous study (59S and 116G. They are both 60 (61 vs. 57) amino acids apart and appear on the same regions in the 2d structure. We don't see any such pattern for Blue opsin, with the hedyliid opsin showing many more sites under positive selection. We think this is added support for the importance of these amino acids in tuning, but being cautious about over-interpreting our results, add a few points about this in the discussion (See line 313-325). The methods (line 535-546) and results section (line 209-221) have also been revised accordingly.

Line 438 - A search for PRJNA1033655 returns only biosample and SRA data. The assembled and annotated genome is not available

Response: We apologize that the data was not available at the moment when the MS was being reviewed. We have double-checked it on NCBI and the assembly is now publicly available on NCBI.

Line 440 - an all databases search of NCBI for OR825356 returns 0 results

Response: Thank you. Please see our response above (Line 438).

Figure 1 - Why does this heat map have 5 or 6 levels of intensity? What are these data? The purported range is from 0 to 2, so there should only be three levels 0 copies, 1 copy, or 2 copies.

Response: The heatmap data are the average of the gene copy number in the gene family. We followed the definition of phototransduction-related gene family in Macias-Muñoz et al, 2019, and in our orthology analysis, many HOGs (orthogroups) are under a gene family. For example, the opsin family has BRh (HOG0004874), UVRh (N0.HOG0007637), pteropsin (HOG0004386) etc, where each HOG could have different gene copy number in different species; and each species could have different copy number in different HOGs. As such, the “mean gene copy number” of a gene family of a species is the average of gene copy number from the HOGs in the gene family. This is why we could have float points in the data matrix.

Figure 2. - Why is *M. leucophasiata* pteropsin is missing from this figure, but included in later selection analyses?

This was an error on our part. The *pteropsin* gene was initially detected in the *M. leucophasiata* annotation but the gene got removed from the final gene model upon filtering based on the criteria described in **Line 458-472**. Hence, *pteropsin* got excluded from the gene tree and selection analyses for *M. leucophasiata*. Fig. 2 still shows data for *pteropsin* but is missing for the hedylid. We have corrected the results as mentioned above in the previous comment. However, *pteropsin* is unlikely to play a role in insect vision and hence is not as important here.

Reviewer #3 (Remarks to the Author):

In this study, Singh and colleagues sequenced a novel moth genome, explored the evolutionary dynamics of opsins and phototransduction genes in response to changes in the photic niche and proposed a convergent evolution of opsin genes.

Overall, I find the manuscript intriguing, and the analyses are cutting-edge. However, I have a few concerns that need to be addressed:

1. The results section requires more background information in the results on the methods used. For instance, I understood the authors' use of GeneRax just at the end of the manuscript.

Response: We have now added a line in the subheading “Selection analyses” (Results) to refer to the gene tree reconciliation performed prior to performing the selection analysis **Line 182-183**.

2. If the authors have identified a new opsin paralog (refer to Figure 2) is an important result. However, before reaching this conclusion, the results need to be compared with previous studies, i.e. <https://doi.org/10.1093/molbev/msad241>, by combining the two datasets together and evaluating the placement of the unknown opsins

Response: Thank you for the suggestion. However, we do not report a new opsin paralog in the manuscript or in Figure 2. The Unclassified (*UnRh*) opsin is a candidate retinochrome which is expressed in the crystalline cone cells, pigment cells surrounding the ommatidium and the head region. We do not have enough data to comment on the *UnRh* opsins reported in the species from our dataset as expression data would help understand it better. Please refer to our response to the minor comment below regarding *UnRh*.

3. The authors should calculate duplications and losses for phototransduction genes using GeneRax, as they did for opsins.

Response: In the current manuscript, we do not discuss the duplications and losses among the opsins of the lepidoptera species (for eg., as in Sondhi et al., 2019), hence these were not reported in detail for other phototransduction genes as well. GeneRax does show this data and uses it for reconciliation but is not reliable as a standalone tool to comment on the duplications and losses, without further validation/examination by statistical tests and other tools. For this reason, we only used GeneRax for the opsin gene trees before performing the selection analyses.

4. The multifunctionality of opsins, as indicated by several studies (e.g., <https://doi.org/10.1093/molbev/msac033>, [10.1146/annurev-cellbio-100616-060432](https://doi.org/10.1146/annurev-cellbio-100616-060432), and [0.1016/j.cub.2020.05.009](https://doi.org/10.1016/j.cub.2020.05.009)), should be discussed in the context of positive selection.

Response: This is a relevant point and we have revised the last part of the discussion to incorporate these aspects into the manuscript in [line 340-342](#).

5. The observed cluster of hedyliid opsins, which differs from the species tree, is an interesting observation. However, a simple UFB tree is insufficient, as the clade might be artefactual. The analysis of the opsins' genes should be performed using Transfer Bootstrap Expectation as implemented in IQTREE, which will allow us to account for possible unstable sequences. Additionally, it would be interesting to perform an approximately unbiased test and compare the two hypotheses.

Response: Thank you for the constructive suggestion. We agree that the clade could be artefactual, so we took the suggestion and performed additional IQtree runs using Transfer Bootstrap Expectation (TBE) on all the opsin genes (please see the trees in Supplementary Data 11). The results are very similar to the original runs with some branch swaps at the nodes with low branch support. For an approximately unbiased test, we constrained the monophyly of genes from butterfly species as suggested by the species tree, and tested if the constrained tree could be rejected by the AU test. We focused on the opsin genes that have *M. leucophasiata* clustering with other nocturnal moths (BRh and UVRh) but excluded LWRh because there are too many duplications in the gene and possibly the out-paralogs has violated the assumption of butterfly LWRh gene being monophyly. The results show that both BRh and UVRh have their constrained tree rejected, supporting the gene tree topology instead of species tree topology. Please see the revised manuscript at line [172-175](#) and [511-518](#).

Minor comments (Reviewer 3):

- Line 93: There seems to be an issue with this sentence. What was the DFAM database used for?

Response: The Dfam database was not directly used for repeat masking by us. However, we meant that RepeatModeler2 uses the Dfam database to classify and refine the repeat families that are then used by RepeatMasker for masking. We have edited the sentence in [Line 109](#) to state that the repeat families were identified based on the Dfam database.

- Line 106: More information is needed here. Which model was used? What is well-supported?

Response: The sentence has now been modified slightly at [line 122](#). We meant that the inferred phylogeny is well-supported, and we mention the same in the figure caption for the Figure 1 as well.

- Line 111-112: OrthoFinder is applied to proteins, not transcripts. Please clarify.

Response: Thank you for the correction, we have clarified it at [Line 127](#).

- Please explain the methods used for annotating the HOGs. Did you use Eggnog?

Response: Yes we used the consensus function from all the genes in the HOG to represent its function. We have added this information to the methods at line [453-454](#).

- How was the opsin phylogeny performed? What are the "Unclassified (UnRh)" opsins(see above)?

Response: The opsin phylogeny was performed following the steps described in the section "Vision gene evolution", [Line 453-473](#).

We have labeled the Unclassified (UnRh) opsins as described in Macias-Muñoz et al, 2019 and first reported in Kanost et al., 2016 (Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, *Manduca sexta*. Insect Biochem Mol Biol. 76:118–147). Vision genes from each species were identified using the three benchmark species, which also includes *M. sexta*, and hence the annotation of *UnRh* opsins was done in this manner. To get a clearer understanding of *UnRh*, gene expression data would be helpful, as discussed thoroughly in Macias-Muñoz et al., 2019.

- Figure 1: Please use a different scale from white to cyan. It is challenging to discern between species where genes are absent and those with two copies.

Response: We attempted to use different color palettes for this figure and settled on using this gradient as it showed the pattern most clearly. Additionally, we hope the discussion of the gene family analyses and the results on innexin and DAGLB gene families in the main text would also help the readers discern the species where two copies are present and where an intermediate gene copy number is present.

- Figure 2: Please add support values for key nodes. If possible, avoid using circular tree output as it is difficult to read.

Response: Thank you for the suggestion about the tree shape. However, we believe that the circular tree outputs are commonly used to present gene trees for emphasizing specific topology as we have in our opsin gene tree here. After considering the suggestion, and trying the rectangular output we decided to keep the original tree with branch supports annotated as it appeared better visually after annotating the different opsins with the colors and positively selected nodes. We hope our explanation is satisfactory.

Response letter

We thank all three reviewers for their efforts and time in providing constructive criticisms and insightful comments that helped improve the quality of the manuscript. In this letter we have addressed the comments by the reviewers for the second revision and have added our point-by-point replies below. New changes in the main text and figures are also highlighted with tracked changes in another version of the file.

Reviewer #1 (Remarks to the Author):

I carefully evaluate all the responses to my questions/clarifications. I have to say that I'm delighted with all the modifications the authors made to the current version of the manuscript. I think all the suggestions made by all the referees substantially improved the manuscript's quality.

With the risk of being pedantic, I have few and minor comments.

I think there's something wrong with the current title for the manuscript. It reads: "Genome reveals molecular signature of adaptation [...]". I saw this was the suggestion from another referee, I'd suggest something like: "The genome assembly of *Macrosoma leucophasiata* (Hedylidae) reveals molecular signature of convergent molecular adaptation to nocturnality in visual genes of moth-like butterflies" I think it would be more appropriate. this is just a suggestion, I leave it to the authors to decide whether they want to change it.

Response: We agree that the title was not the most appropriate option and in light of the previous comment from the referee, we have changed the title in this revision to: "*Genome assembly of a moth-like butterfly reveals convergent adaptation of visual genes to nocturnality*"

I found a lot of methodological references in the result section. Something like: "A preliminary survey using Genomescope 2.0 estimated [...]" / "filtering haplotigs via purging and non-target sequence removal using BlobTools" / "we modeled and soft-masked repeat regions using RepeatModeler2 and RepeatMasker [...]" / "transposable elements (TE) repeat families was generated corresponding to the Dfam database [...]" / "For gene annotation, we used the resulting soft-masked genome to run the BRAKER3 pipeline with protein evidence from the OrthoDB v1141 catalog [...]", just to list a few are more appropriate in the methods section.

Response: Thank you for your suggestion, we left the brief descriptions about what analysis was used in each step because the result section comes before the methods. We have carefully reexamined the paragraph and believe that those

descriptions provide the minimum information about methods and not overlap with the method sections.

In the Discussion, the paragraph about the "Novel genome assembly for Hedylidae" does not seem particularly relevant. Many invertebrate groups can make similar claims. In comparison to other scientific communities focused on invertebrates, the one for Lepidoptera has seen a significant number of genomes sequenced and published every month or so.

Response: Thank you for your suggestions. We agree that this section is less relevant to our main finding. However, we do wish to address the problem that there seems to be a strong bias in what taxa to be sequenced in Lepidoptera. And we think it would be very beneficial if more unpopular taxa were sequenced. We thus hope that we can keep this section without distracting the reading flow.

Reviewer #2 (Remarks to the Author):

With this revision, the authors have addressed my major concerns. However, there several places in the manuscript where questions remain or clarification is needed:

Line 14 - Why is "Blue" capitalized here?

Response: This was a typographical error and has been changed to "blue".

Line 32-34 and 63-67 - Superposition eye structure is introduced as a major feature of nocturnal adaptation, but is not discussed in the rest of the paper. Are there specific molecular signatures that we should expect to see in animals with superposition eyes? Are there present in *M. leucophasiata*?

Response: We agree that the introduction of superposition eyes is not discussed further and it is difficult to comment on with the current data. Hence, we have removed the lines 63-67 and focus the introduction on how this transition is a useful place to test visual gene evolution.

Line 76-77 - Throughout it is assumed that a shift to nocturnality will result in selection and adaptive change in the opsins. However, a shift to nocturnality could also reduce selective pressures if, for example, color vision becomes relatively less important in the animal's life. This might lead to neutral processes dominating the evolution of these genes.

Response: This is a valid hypotheses, in part resolved by testing the nature of selection acting on these genes as well as by examining the 3D protein model of the opsin, a response dominated by neutral selection would lead to a different pattern than we see, but we include this as a possible outcome and a reason for inferring the nature of selection acting on the opsins.

Line 132 - Interesting observation. I realize this is beyond the scope of this paper, but might these changes in odorant receptor and P450 repertoire size be linked to shifts in larval hosts?

Response: We agree that this is an interesting and unexpected result. We believe it's worth further investigation. However, we can not expand our interpretation of this result because the phylogenetic scale here is large, where we are comparing the genomes of *Macrosoma* to other butterflies and the families with rapid size changes could happen in Hedyliidae, *Macrosoma*, or this species. We simply report these results and more genomes from Hedyliidae could help solve this problem.

Line 137 - Provide a brief note of the thresholds used in the BLASTp search

Response: E-value threshold of 1E-50 is stated in the methods at line 463. We would like to refrain from adding a note about the methods in the results, in light of the comment from Rev. 1 about mentions of methods in the results section.

Line 142 - I am a little unclear what 3,721 genes is referring to. Are there distinct genes or does this include homologs of the same gene in different species?

Response: Thank you for pointing out the issue. We added 218 genes from the 23 additional HOGs using keyword searches in the EggNOG annotations. To clarify these numbers, we added additional information to the sentence at line 141-142.

Line 154-155 and Figure 1 - Given the evidence that the missing innexins are most likely an artifact of the assembly It is not accurate to state they are missing in the text or represent them as missing in Figure 1. The figure 1 legend states "...losses of innexin in *M. leucophasiata*"

Response: The revised figure caption reads as follows: "Figure 1. Analysis of visual genes shows high conservation within day and night flying lepidoptera with gene duplication in the *DAGL* in *M. leucophasiata*."

Line 159-161 - I am confused by this statement, how can the contig from *M. leucophasiata* be linked to a *T. sylvestris* chromosome? Do you mean the contig is mapping to that chromosome?

Response: We apologize for the confusion. We have changed the wording from linked to mapped.

Line 167 - What is meant by cluster together? What metric? AA, nucleotide identity, or some other feature of the gene?

Response: Thank you for catching the unclear statement. The clusters are defined by the ML gene tree topology. We have added this information to the sentence at line 167.

Line 211 - "but LWRh recovered 5 and 6 respectively" - what does respectively refer to? Are there two copies of this gene?

Response: Sorry for the unclear statement. Yes, there are two copies of LWRh and we just revised the sentence at line 210-211 to clarify that the two copies of LWRh do not recover seven predicted transmembrane helices.

Line 213 - What is the relevance of squid here? From the rebuttal letter, I understand that this template was used in an early study, but it does not seem necessary to discuss here.

Response: We agree with the reviewer, we have removed the mention of squid. We used it since it was the other model template that had high matching scores with the swiss-model, but the spider sequence is a better match and the comparison makes less sense in the context of the current study.

Line 244-245 - Could the rarity of gains and losses be attributable to the distinct photic niche of the larva? Are they diurnal?

Response: Thank you for bringing up the interesting idea. Usually caterpillars (larvae of moths and butterflies) don't use vision signals much with small and low resolution eyes, especially compared to their lifestyle as adults which heavily relies on vision for flying orientation and food searching with high resolution eyes and color vision. There is also very limited literature about the photic niche of most caterpillars in general with the exception of pest species, so it is hard to test this hypothesis. However, the more dramatic transition in eye shape seems to be in adult compound eyes of these nocturnal butterflies (superposition vs. apposition) (which is not the case for larva with simple ocelli), we believe that the vision evolution could have taken place in adults and very unlikely in caterpillars.

Line 253 - subterranean?? Do you mean aquatic or amphibious?

Response: We used the same terminology as used in the cited paper (Langille et al., 2022). They studied species that primarily live underground and also included species from surface pools.

Line 273 - "nosed into the assembling process" - what does this mean?

Response: We removed this phrase and replaced it with *"interfered with the assembly outcome"*.

Line 288 - 289 - Can you speculate on DAGLB's role in adaptation? In vertebrates, it is involved in endocannabinoid signaling, not sure how that relates to dimlight vision.

Response: We did not speculate the function in vision of butterflies because its function has not been determined in any other insects. Based on Macias-Muñoz et al., 2019, DAGLB putatively plays a role in phototransduction, but it is also found expressing in other tissue rather than just the head of moth. Thus we decided to carefully address its possibility of being associated with phototransduction and not to extend our interpretation.

Line 306 - What other factors might result in grouping with nocturnal species? Is introgression a possibility? Could this possibility be excluded with an ABBA BABA test?

Response: Thank you for the suggestion. We did not consider introgression because Hydelidae emerged around 20 million years after the first butterfly which was likely to be diurnal (judged by the diurnality of Papilionidae, the basal clade of butterfly). In addition, the genetic distance to other nocturnal moths from any Hedyliidae are phylogenetically very distant, so it is very unlikely that the convergence is the result of introgression in this case. For ABBA/BABA analysis, we did not consider it as we don't have good outgroup taxa if considering a nocturnal moth being the P3 ingroup donating the genes.

Line 319-320 - What previous study?

Response: We have edited the sentence and now write "*previous study (Sondhi et al., 2021) on lepidoptera opsins (59S and 116G)*"

Reviewer #3 (Remarks to the Author):

I am satisfied with almost all the authors' replies; however, I have two final comments:

The term "unclassified opsin" is confusing. Macias-Muñoz et al., 2019 is a paper on Hydra opsin, and Kanost et al., 2016 is a genome paper not focused on opsin (opsin is just one of the supplementary figures). Insect opsins are divided into C, R, RGR/Go, and probably arthropsin (<https://doi.org/10.1093/gbe/evw015>, <https://doi.org/10.1073/pnas.1424670112>). Please name the clades accordingly.

Response: Sorry for the confusion. We understand that the term "unclassified" is confusing as we are not the first one finding this gene. However, we believe that the "unclassified opsin (UnRh)" has been formally defined in Macias-Muñoz et al., 2019 (this author has two papers in 2019, one on moth opsin and one on Hydra opsin, and the UnRh was defined in the first paper: <https://doi.org/10.1093/gbe/evz150>) with some functional tests, and thus should be a valid name. Since there was no further examination on this gene published since and thus no new naming convention on this gene to follow as we know, we decided to use this term to make it trackable.

As I suggested in previous comments, please add the BS (UFB and TBE) for key nodes in Figure 2.

Response: Thank you for the suggestion. We have marked circles on the tree at figure 2 to show the support values (both SH-aLRT and ultrafast bootstrap) at the major nodes including those close to the root and include nocturnal branches.