Manuscript type: Research Article

Title: Light environment drives evolution of color vision genes in butterflies and moths.

Authors: Yash Sondhi¹(ORCID: 0000-0002-7704-3944), Emily A. Ellis²(ORCID:0000-0003-3343-0923), Jamie C. Theobald¹, Akito Y. Kawahara² (ORCID: 0000-0002-3724-4610).

Author Affiliation:
¹Department of Biology, Florida International University, Miami FL 33174 USA
²McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL 32611 USA

Corresponding author: Yash Sondhi, ysond001@fiu.edu

Keywords: color vision; diel-niche; insects; Lepidoptera; transcriptomes; vision

Abstract

Opsins are the primary light-sensing molecules in animals. Opsins have peak sensitivities to specific wavelengths which allows for color discrimination. The opsin protein family has undergone duplications and losses, dynamically expanding and contracting the number of opsins, throughout invertebrate evolution, but it is unclear what drives this diversity. Light availability, however, appears to play a significant role. Dim environments are associated with low opsin diversity in deep-sea fishes and cave-dwelling animals. Correlations between high opsin diversity and bright environments, however, are tenuous. Insects are a good system to test whether opsin expansion is associated with greater light availability because they are enormously diverse and consequently display large variation in diel activity. To test this, we used 200 insect transcriptomes and examined the patterns of opsin diversity associated with diel-niche. We focused on the butterflies and moths (Lepidoptera) because this group has significant variation in diel-niche, substantial opsin recovery (n=100), and particularly well-curated transcriptomes. We identified opsin duplications using ancestral state reconstruction and examined rates of opsin evolution, and compared them across diel-niches. We find Lepidoptera species active in high light environments have independently expanded their opsins at least 10 times. Opsins from diurnal taxa also evolve faster; 13 amino acids were identified across different opsins that were under diversifying selection. Structural models reveal that four of these amino acids overlap with opsin color-tuning regions. By parsing nocturnal and diurnal switches, we show that light environment can influence gene diversity, selection, and protein structure of opsins in Lepidoptera.
Introduction

Opsins are the primary light sensing molecule in animals. Lepidoptera has over 150,000 species with large variation in the activity time and light environment (diel-niche). Opsin diversity has yet to be studied comprehensively across the entire order and multiple diel-niches. Here we provide a novel link between visual genotype and light environment phenotype. We analyze a large dataset of 200 insect transcriptomes to study opsin diversity and its association with diel-niche. By comparing opsin diversity across 100 nocturnal and diurnal lepidopterans, we find that opsin diversity increases in bright light environments. Bright light environments are also associated with higher rates of opsin evolution, and some of these rapidly evolving sites are implicated in tuning vision.

Vision is a fast and reliable sensory modality, useful to detect shape, color and distance of signals. Eyes are diverse, convergently evolved structures that detect the direction and intensity of light in an animal’s environment. Although eyes are often considered inseparable from vision, light sensitivity likely predates the origin of eyes. Bacteria, fungal spores and dinoflagellates lack identifiable eyes, but can still detect and respond to light.

The visual pigment rhodopsin is the primary light sensing molecule in all animals ranging from insects to primates. Rhodopsin, or opsin, transduces light into a biologically meaningful signal by activating a G-Protein and initiating the phototransduction signaling cascade. Large-scale phylogenomic analyses have found many duplications and losses — which we refer to as expansion and contraction — in the opsin protein family across invertebrates. The water flea, Daphnia pulex, possesses 46 opsins, dragonflies have 12-15 and mantis shrimp have 12-33. However, any more than four spectral channels offer diminishing returns in extracting information from natural scenes, so why do some animals have so many opsins?

Opsins may expand and contract for non-adaptive reasons. Non-adaptive forces usually cause random sequence evolution, and unless the duplicated opsins are co-opted for visual use, they usually become pseudogenes. Adaptive evolutionary forces are more likely to cause consistent, repeated, and persistent patterns of opsin diversification, such as with mate choice in guppies and butterflies, flower foraging in bees and wasps, and light intensity environment with many nocturnal animals.

Sensory modalities such as smell, electromagnetic reception and touch are more reliable than vision in dim environments. Resource allocation trade-offs can cause loss of stabilizing selection on inefficient sensory systems and can cause a downregulation of opsins, leading to them becoming pseudogenes, a pattern seen in nocturnal mammals, cave-dwelling crayfish, and deep-sea
organisms\textsuperscript{20–22}. If diminished light availability causes reduced opsin expression and loss, abundant light, conversely, may cause higher opsin expression, duplications, and diversification. A comparison of visual genes between diurnal and nocturnal Lepidoptera revealed elevated opsin expression in the diurnal species\textsuperscript{23}. Similarly, a study of opsin evolution across fireflies revealed higher amino acid transition rates in diurnal fireflies, across 4 independent diel switches\textsuperscript{24}.

Lepidoptera opsin diversity has been studied in handful of model taxa — mostly diurnal butterflies\textsuperscript{23,25,26} — but opsin diversity has yet to be studied comprehensively across the entire order and multiple diel niches. Briscoe (2008) analysed the visual genes of 8 Lepidoptera species including two moths\textsuperscript{25}, Xu et al. (2013) analysed 30 Lepidoptera including 12 moth species\textsuperscript{25}, and Feuda et al. (2016) analysed 10 Lepidoptera species including four moths\textsuperscript{28}. These studies annotate genomic opsins, and used gene trees instead of species trees for selection analyses, both of which can bias results. The few studies that examined opsin diversity and diel-niche association\textsuperscript{27,28} compare butterflies and moths, effectively using only a single diel switch. Lepidoptera have more than 100 recorded diel transitions\textsuperscript{29}, and only by examining multiple independent diel-niche switches can we understand how light environment and diel-niche drive the evolution of their visual systems. To this end, we mine 176 Lepidoptera species transcriptomes for visual opsins and combine our annotations with publicly available data to present the most comprehensive survey of opsins across Lepidoptera. In parallel, we compiled diel-activity from the literature, natural history databases, and in consultation with experts.

Results

Opsin diversity patterns across insects

We examined patterns of opsin diversity across insects using assembled transcriptomes from InsectBase\textsuperscript{30} along with opsin annotations from Ensembl Metazoa\textsuperscript{31} and previous studies\textsuperscript{28}. Diel-niche assignment came from literature, and transcriptomes were annotated using a phylogenetically informed annotation approach (PIA)\textsuperscript{32}. We recovered at least one opsin from more than half (28/50) of the insect transcriptomes, with a total of 79 opsin sequences. The final dataset included 45 insects (Table S2).

We reconfirmed the annotations using nucleotide and amino acid gene trees (Fig. S1, S2, supplementary information), and summarized opsin presence-absence and diel niche on a pruned insect tree (Fig. 1A). Our dataset of insect transcriptomes serves as a positive control for the annotation pipeline since the results agree with the broad patterns of opsin diversity published before\textsuperscript{28}. However, the insect opsin dataset recovered opsins from too few nocturnal taxa to
statistically compare opsin evolution between diel-niches. Therefore, we used similar methods on dataset of Lepidoptera with sufficient diel variation and better phylogenetic coverage to address this question.

**Lepidoptera opsin diversity and diel-niche association**

We recovered at least one opsin from 114 of 175 Lepidoptera taxa, identified 265 opsin sequences (Table S1), and confirmed the opsin annotations by building gene trees (Fig. S1).

We mapped opsin diversity—specifically UV/RH4, Blue/RH5, LW/RH6 and RH7 opsin presence and absence—and diel-niche onto a well-resolved Lepidoptera phylogeny (Fig. 1B). We find that duplications occur more often in diurnal species than expected by chance (Table 1). Increased duplication in diurnal species was evident even after we excluded species with ambiguity in diel-niche assignment (Table 1).

We compared ancestral state reconstructions (ASR) of diel-niche and total number of opsins. We find that diurnal lineages and nodes have ancestrally higher opsin numbers (Fig. S3A). Opsin losses are notoriously difficult to confirm, since inferred losses could be due to poor opsin recovery or low-quality transcriptomes. To distinguish these from true losses, we performed an ASR for all four opsins (Fig. S3A-D). If closely related lineages display a loss, ancestral nodes will show a high probability of loss (>50%); whereas, if losses are random, perhaps due to low quality data, then ancestral nodes will have 50% chance of loss.

**Ultraviolet (UV):** 50% (56/114) of the taxa recovered UV/RH4 opsins (Fig. 1, Fig. S3B). UV opsins had duplications in only 2 independent lineages; the diurnal *Heliconius melpomene*, in which UV opsin duplication has been recorded before and crepuscular *Triodia sylvi*na (Hepialidae), an ancient lineage of ghost moths, known for swarming at dusk. We did not find multiple lineages with a high probability of loss, but deeper nodes had a 50% chance of loss, likely indicating incomplete data, as opposed to true losses.

**Blue:** 46.4% (53/114) of the taxa recovered Blue/RH5 opsins (Fig. 1, Fig. S1-3). Blue opsin duplications were present in 5 families and 7 genera, all diurnal species. The pierid *Phoebis sennae* and two lycaenid genera *Hemiargus* and *Polymnatus* also have blue opsin duplications. Behavioral and electrophysiological data have shown that these families, if not in these individual species, have functional duplications. *Macrogolossum pyrrhosticta*, a diurnal sphingid hawkmoth, also had a blue duplication. Our ASR suggests true losses in four lineages in the Tortricidae, Crambidae,
Noctuinae and Macroglossinae. For all other nodes, the chance of a loss was about 50%, likely indicating issues in opsin recovery and annotation.

**Long Wavelength (LW):** 83.3% (95/114) of taxa recovered LW/RH6 opsins, the highest of all three opsin families (Fig. 1, Fig. S3). LW opsins duplications were recorded in 8 families and 11 genera. Only two of these genera are nocturnal. We recovered previously reported LW duplications in one Lycaenidae, two Riodinidae, and three Papilionidae species. Butterflies are among the few insects that can detect red, through LW duplication or filtering pigments. Thus, any duplications might indicate true color expansion. Other diurnal or crepuscular species that had LW duplication were *Dyseriocrania subpurpurella* (Eriocraimiidae) and *Triodia sylvina* (Hepialidae).

The tiger moth *Callimorpha dominula* and the castniid, *Paysandisia archon* – two diurnal moths not in this dataset – also have LW duplications, and *Paysandisia* at least can detect red wavelengths. *Spodoptera*—of which some species are invasive pests—is nocturnal but has a red sensitive LW duplication. Many *Spodoptera* species are migratory, flying above the clouds in night skies may free them from low light constraints. *Tischeria quercitella*, a leaf mining moth, is the only other nocturnal species we found with a LW duplication, despite examining over 50 nocturnal species. Losses were identified in three distinct lineages, Geometrinae, Macroglossinae and, Smerinthinae; their presence was easier to confirm in LW than UV or Blue because of more complete LW recovery (Fig. S3D). Some Macroglossinae hawkmoths with a LW loss have a duplicated blue opsin, and electrophysiological investigation has shown that *Macroglossum stellatarum*, from the same clade, has three distinct sensitivity peaks.

**RH7:** 28% (33/114) of taxa recovered RH7 opsins with no duplications, consistent with other work (Feuda et al. 2016). Since RH7 is involved in circadian rhythm light sensing, it is possible that the time of sample collection affected the expression of opsins and might explain the lower recovery rate across our dataset.

**Opsin selection in Lepidoptera**

We used PAML to estimate rates of selection (ω) and tested if these rates differed between nocturnal and diurnal Lepidoptera. For datasets that showed significant differences (ω), we used branch-site models in PAML to identify amino acids under selection. We tested if the analysis was sensitive to different sample sizes and starting trees (Fig. 2, Table 2, Table S4). To ensure that we compared opsin rates across the same group of species, we limited our analyses to species that recovered all three visual opsins. We were unable to use the entire dataset of recovered opsins for
UV, Blue and Green opsins because the alignment included gaps, which could reduce the accuracy of the model. Since PAML requires stop codon free and unambiguous alignments, we removed sequences that created ambiguity in the alignment and trimmed the ends — not the middle, which confounds structural modelling — to ensure a clean alignment.

First, we ran PAML on a subset of the data (n=12–14) of similar size as Feuda et al. (2016) and Xu et al. (2013) to determine sample size effect, although taxon sampling across datasets was non-uniform. Of all three opsins, only UV opsin rates were significantly different between diel-niche for this dataset (Fig. 2A).

We next used a larger sample size and similar species across datasets (n=24–27). We found highly significant differences between diel-niche for UV and LW opsins (p-value <0.001) and moderately significant differences for Blue opsins (p-value <0.05). Each show relaxed selection in diurnal species. We used RH7 as a control because it is not involved in vision and found no significant differences for RH7 rates between diel-niches (Fig. 2B).

Lastly, we ran the PAML analysis using a species tree instead of gene trees, since Blue and UV opsin recovery was sparse (~50% of the taxa) and gene trees may be biased. The species tree models showed highly significant differences between diel-niches for UV, Blue and LW opsins (Fig. 2C). We did not include RH7 in this analysis in order to compare across similar species trees, which was precluded by the poor overlap of species between RH7 and visual opsins.

We tested if particular sites were under positive selection in diurnal species using branch and branch site-models. There was no significant signature of positive selection across the opsin when comparing with the null model for branch models, however this is expected for highly conserved proteins like opsins. The branch-site models, however, identified amino acid sites that were under positive selection/relaxed purifying selection in diurnal species in UV/RH4, Blue/RH5 and LW/RH6 opsins (Table S4).

Xu et al. (2013) reported elevated dN/dS (ω) rates in butterflies (diurnal) compared to moths (nocturnal)27, with LW, Blue and UV opsins, showing a decreasing magnitude in differences. In contrast, we find UV opsins had the highest and most consistent ω rate differences. Feuda et al. (2016) used two independent diurnal transitions, with a total of 10 Lepidoptera taxa and have results similar to ours (Fig. 2B). We found that UV/RH4 and Blue/RH5 and LW/RH6 genes in nocturnal Lepidoptera underwent strong purifying selection and more diversifying selection in diurnal species. RH7 had almost similar levels of selection in both nocturnal species and diurnal
species. Only UV opsin showed consistent differences through the range of analysis parameters, therefore, we used it for 3D protein modelling.

**Mapping amino acid sites under selection to predicted protein structure**

The site-selection models failed to return any significant sites under selection, but when we took diel-niche into consideration and used branch-site models, they returned several sites under positive selection (Table S4). We mapped the significant sites recovered from the branch site model onto the predicted protein structure, obtained using transmembrane helix prediction. We found that each class of opsin mapped to a unique pair of adjacent transmembrane helices, likely responsible for tuning rhodopsin (Fig. S4). We found that the for Blue and LW opsins, sites under selection, predicted by the models, were unaffected by the choice of gene vs. species trees. The choice of tree, however, affected the results of the UV opsin selection models. Since species tree models predicted more sites in the helical regions—which can affect and tune the opsin—we infer that the species tree models are more accurate. Additionally, UV 59 may be a convergent site of selection, as it has previously been identified in diel-transitions of fireflies\(^{24}\).

**Expansion of color vision through sequence tuning**

Since the UV opsin recovered all 7 transmembrane domains, similar to the x-ray crystal structure of squid rhodopsin, we modelled it using an online protein modelling tool, Swiss-model\(^{42}\), and identified putative retinal binding sites, i.e. amino acids less than a minimum distance from the retinal molecule. We found that majority of the amino acid sites under positive selection are close to amino acids of the retinal binding region (Fig. 3). Feuda et al. (2016) also map the positively selected amino acids to opsins\(^{28}\), but they recover a large fraction of sites in terminal regions, not in the helices. This could be because their alignment method removes gap-filled regions, reducing the accuracy of structural prediction. We refrained from removing gap filled regions from the alignment, instead, we performed end trimming and limited the analysis to species that resulted in a gapless alignment. Our analyses show that these sites likely alter the opsin spectral tuning in diurnal species by changing the chemical environment surrounding retinal. Now that genetic manipulation in Lepidoptera is feasible using CRISPR-Cas9, one can determine the relative impact of these site-specific mutations on Lepidoptera opsin sensitivity and overall behavior.

**Discussion**
What does opsin expansion or contraction mean for the visual system of an organism? One consequence of opsin expansion is the potential for improved color discrimination. Color vision usually requires at least two opsins with a partial overlap in spectral sensitivity and even small changes to the amino acid sequence can modify the tune the opsin, shifting the perceived color space of an organism. Alternatively, more dramatic shifts in detectable color space can occur through loss or gain of opsins. For example many butterflies can see in the red (620-780 nm) through Long Wavelength (LW) opsin duplication and divergence. RH1–6 are implicated in color vision, but non-visual RH7 is also phylogenetically scattered across insects, and associated with circadian rhythm maintenance.

Examining opsin diversity across multiple lineages that have switched diel-niche can help determine how light environment affects opsin evolution. Dragonflies, mosquitoes and butterflies show multiple opsin duplications with as many as 15 color opsins, while beetles and scorpionflies show losses identifying them as potential monochromats. These studies have not examined potential links between diel activity and opsin diversity, or have failed to find consistent trends. Analyses are confounded by uncertainty in diel-niche assignment and a lack of multiple independent diel-switches. Systematic error — including shallow sequencing and poorly resolved species trees — could also obscure any trends.

The greatest limitation of large-scale gene mining approaches is the reduced power to detect absences. Low coverage transcriptomes from older studies, mixed tissue sources, as well as varied assembly and sequencing methods, all increase heterogeneity in opsin recovery. Ideally, we would sequence only eye tissue with muscle tissue from the body as a control, but we frequently lack compete information about which tissues generated the transcriptomes. Further, we do not measure expression levels, which could identify non-functional opsins. However, as we verified our pipeline with known data and used multiple annotation methods and ancestral state reconstruction to identify false positives, we believe we have obtained the best estimate of opsin diversity possible from this dataset.

More than 80% of the opsin duplications we discovered across 10 independent lineages are in diurnal species, which additionally have more variation in opsin tuning sites than nocturnal species. Differences in selection rate between the diel niches are visible across all three visual opsins, but most robust and consistent in ultraviolet-sensitive opsins. Structural modelling shows that in diurnal species, amino acids opsin under positive selection overlap with the retinal binding region. Each of these observations supports our prediction that transitions from dim to bright-light environments
drive opsin diversification in Lepidoptera. Only an adaptive selective force would likely show consistent patterns across multiple independent duplication events. The ancestors of both diurnal and nocturnal lepidopterans were likely crepuscular with limited trichromatic vision, so what is likely to have driven this selection?  

Moving to diurnal behavior reduces light constraints but increases the risk of predators, which may in turn drive selection of either chemical defenses or faster escape flights. Chemical defenses are often precursors to bright warning coloration which can then become targets for sexual selection, which then requires accurate color perception. Aposematic Heliconius butterflies, for example, use wing-coloration as mate choice signals. Lepidopterans are fast in the air, with speeds up to 3.5 m s\(^{-1}\) in free flying Danaus sp. and 5.3 m s\(^{-1}\) in Manduca sexta. Aposematism and fast flight appear to be alternative strategies, since most aposematic butterflies are slower than their counterparts. Butterflies and diurnal hawkmoths are both predominantly nectar feeders and the energetic requirements for fast flight may select for efficient nectar foraging; for which an expanded color space would be beneficial. In sum, the transition to a diurnal light environment may set into motion a diverse trajectory of evolutionary scenarios, each of which may increase opsin diversity.

For species that switched to a more nocturnal lifestyle, light intensity was a limiting factor, and their eyes underwent transitions to become superposition eyes; transitions that happened multiple times across Lepidoptera. The nature of visual pigments makes capturing color information harder as light intensity drops; stacking more rhodopsin to make more sensitive receptors comes at the cost of getting a more broadband signal and losing color discrimination. Superposition eyes effectively act as a large lens in order to increase the light available at each photoreceptor. The maintenance of trichromacy in nocturnal Lepidoptera is therefore puzzling: either color vision has some critical function in nocturnal Lepidoptera, or opsins are maintained for functions independent of color vision.

As a critical function, color can serve as a short-range cue useful for mating and foraging, such as tiger moths that can distinguish conspecifics using color markings that are unrecognizable to birds, or the strong innate attraction of flower foraging Lepidoptera to blue. Moths may overcome dim light constraints by pooling from different opsins. The most detailed study of opsin distribution across a moth eye shows that there are very different dorsal-ventral patterns in Manduca compared to butterflies, suggesting that moths may have partitioned color and spatial vision in different regions.

Alternatively, selection could act on opsins independently, regardless of their utility for color vision. UV and LW opsins models consistently showed differences across models, but Blue opsin
rates had a smaller differences in all but one model. LW opsin recovery was almost complete, but UV and Blue opsin recovery was patchy. If this trend represents actual losses, it supports the idea that each opsin is maintained independently and LW might be more critical. UV light sensitivity is prevalent in nocturnal animals, even dichromats such as rodents, owls and deep-sea fishes\textsuperscript{20,60,61}. UV contrast is a foraging cure for moths, as nectar guides in many-night blooming flowers\textsuperscript{62}, and an attractant\textsuperscript{63–65}. Because short wavelengths increase around twilight\textsuperscript{66}, UV light is a possible signal for pupil responses, (anecdotally more prevalent across nocturnal moths than butterflies), which could explain the purifying selection in nocturnal species. LW sensitivity is useful for oviposition behavior in butterflies\textsuperscript{67} and their high recovery and strong purifying selection in nocturnal species could be a signature of their role in oviposition. Butterflies display strong host plant specificity compared to moths,\textsuperscript{68–70} and butterfly LW may have diversified for finer oviposition site discrimination. Irrespective of their role in color vision, LW and UV opsins may diversify in diurnal species but be held under purifying selection in nocturnal species due to integral functions.

In summary, patterns of opsin diversity and evolution across Lepidoptera and insects show that differences in diel-niche correlate with opsin duplication and rates of evolution. Future studies on fine-scale expression patterns in closely related diel-transitions can examine opsin expression levels and immunohistochemistry to determine if the duplications are functional, and in-vitro expression systems and CRISPR could confirm that the amino acids identified tune color vision. A more thorough sampling across insect transcriptomes might identify deeper evolutionary patterns. And examining opsin distribution patterns across lepidopterans could help to understand if sensitivity and color vision trade-off in different eye regions. This study provides a library of opsin sequences, useful for these sorts of studies, as well as opsin diversity patterns across different species, which can help inform which species to probe.

**Methods**

**Transcriptome annotations and database mining**

**Insect opsin gene annotation**

We annotated 30 assembled insect transcriptomes from InsectBase\textsuperscript{30} and Ensembl Metazoa\textsuperscript{71}, limiting our analyses to head or whole-body tissue to enhance recovery of visual genes (Table S3). Transcriptomes were annotated using Phylogenetically Informed Annotation (PIA)\textsuperscript{32}. PIA is a bioinformatic pipeline that queries transcriptomes using pre-existing reference gene-sets and places them on a supplied amino acid gene tree. It identifies reading frame and creates a comprehensive
gene tree. A modified version of this pipeline was used to allow for faster analysis on a high-performance cluster (https://github.com/xibalbanus/PIA2, Supplementary Information).

A set of well characterized metazoan visual rhodopsin genes were used as the reference gene set for PIA, and parameters were set to ensure high fidelity of hits while allowing for partial length matches (minimum amino acid sequence length -30, gene search type -single, gene set -r_opsin, e-value threshold -e^-19, maximum of blast hits retained -100).

We further obtained protein sequences from published genome and opsin annotations of 20 additional taxa (Table S2). These sequences were included in the opsin annotation tree (Fig. 1), but were not run through the PIA pipeline– PIA can only use nucleotide transcript data. An amino acid opsin gene tree was built to confirm the annotations (Fig. S2).

**Lepidoptera opsin gene annotation**

We annotated 162 Lepidoptera transcriptomes obtained from previously published studies (Table S1). See Kawahara et al. (2019) Supplementary Information for details on transcriptomes. PIA was used with the same parameters as described in the previous section for insects.

Further manual curation was done using the reconstructed opsin tree (Fig. S1). Putative duplications were aligned to ensure they were not fragments of the same gene. We added opsin sequences from 14 Lepidoptera species obtained from blast searches of Lepbase, a repository of lepidoptera genomes and transcriptomes, (Challis et al. 2016) and Ensembl Metazoa. We used *Manduca sexta* RH4, RH5, RH6 and, RH7 opsin sequences as queries for the BLAST search (-evalue 1.0e^-10 -num_alignments 250).

**Opsin gene tree reconstruction**

The Lepidoptera and insect cDNA sequences were collected and annotated using PIA as putative short wavelength (SW/UV/RH4), medium wavelength (MW/Blue/RH5) and long wavelength (LW/RH6), RH7 like (Table S1). However, as a confirmation, we also constructed gene trees for all the sequences. These trees also allowed us to identify unknown opsins, measure divergence of duplications and, catch incorrect annotations.

**Nucleotide gene tree**

MAFFT (v7.294b) was used to align the sequences (- auto was used and MAFFT detected the sequences as nucleotide), and IQ-TREE (multi-core v1.6.12) to build a ML tree (iqtrees alignment_name.fasta -st DNA -bb 10000 -nt AUTO -alrt 1000). The tree was color-coded based on opsin clades, and taxon names were color-coded based on the PIA annotation, to find discrepancies between the two methods. The tree was midpoint rooted using Archaeopteryx (v0.9917
beta)(https://github.com/cmzmasek/archaeopteryx-js). Using melanopsins as an outgroup did not affect the opsin annotations, so we chose to exclude them from the gene trees (Fig. S1A–D, Supplementary Information).

**Amino acid gene tree**

MAFFT (- auto was used and mafft detected the sequences as amino acid) and IQ-TREE (iqtree -s alignment_name.fasta -st AA -bb 10000 -nt AUTO -alrt 1000) were used to build an amino acid gene tree. We used it to confirm opsin protein annotations taken from annotated genomes or other studies (Feuda et al. 2016).

**Diel niche assignment**

We assigned diel-niche to compare opsin evolution, but only did so for species for which we recovered at least one opsin, limiting further analysis to these taxa. Species were classified as “diurnal”, “nocturnal”, “crepuscular” (species active at dawn or dusk) or “both” (species with some activity during the day and night) (Table S1, S2). The diel-niche was assigned using published literature, natural history databases and in consultation with experts for more obscure species (Table S1, S2).

Diel-niche is often assigned based on whether an insect is attracted to light, and can be a reliable indicator of species that are strictly nocturnal and diurnal\(^29\). But it often fails for crepuscular species, and species that fly during the night and the day (“both”). For example, even though *Manduca sexta* is often considered crepuscular, one study has found it is almost entirely nocturnal when compared to *Hyles lineata*, which is active both during the night and day\(^75\).

We used different approaches for diel-niche assignment depending on analyses. For easy tree visualization, we used only three diel states, grouping “crepuscular” and “both” into a single category (Fig. 1). For statistical analysis, however, we tried all possible grouping combinations (Table 1). Because some diel-niche assignments are uncertain, we note this ambiguity in the dataset (Table S1).

For selection analysis, models only allow two groups, and we therefore categorized species into strictly “nocturnal” and “diurnal” by assigning “crepuscular” and “both” to the “diurnal” category.

**Species-tree reconstruction**

We required a well-resolved species tree for Lepidoptera and insects to perform ancestral state reconstruction, opsin rate analysis, and visualization of opsin diversity. The insect species tree (Fig. 1) was built by grafting an NCBI common taxonomy onto the Misof et al. (2014) tree, pruned using
Archaeopteryx. Because only a few taxa, spread across different families, were used in this tree, relationships at the family level and lower are only representative—they do not reflect current taxonomy. For the Lepidoptera species tree we pruned a well-resolved species tree from Kawahara et al. (2019) to include only species that had at least one identified opsin. We utilized the python package ETE v3 \(^{77}\) to prune and annotate the trees. The pruned tree was modified using python scripts (Supplementary Information) to show the number of opsins in various Lepidoptera taxa (Fig. 1).

**Statistical analysis**

The python scipy stats package \(^{78}\) was used to analyze the opsin duplications and their association with diel-niche across Lepidoptera. Custom scripts (Supplementary Information) were used to filter data—including only species for which we recovered at least one opsin—and redo the tests after excluding taxa with uncertainty or varying placement of “crepuscular” and “both” diel-niches as either “nocturnal” or “diurnal” (Table 1).

**Selection Analyses**

The annotated opsin dataset from Lepidoptera was used for analyzing patterns of selection acting on the opsins and comparing them across diel-niche. Each opsin family was analyzed separately, but we limited the analysis to taxa which had recovered at least one copy of all three visual opsin sequences (UV/RH4, Blue/RH5 and LW/RH6) to make the analyses comparable between genes and datasets. Because RH7 was only used as a control, we used all species from which we recovered RH7.

We used Geneious v. 10.0.9 (https://www.geneious.com) to sort and export the sequences. The sequences were manually cleaned by trimming longer sequences and removing sequences that were too short. AliView v. 1.18.1 \(^{79}\) with Muscle v. 3.8.31 \(^{80}\) was used for trimming and aligning. We did not remove gap-filled regions since this can bias structural prediction. Instead, we dropped sequences or trimmed edges to reduce gaps and stop codons while optimizing the length of alignment.

For RH7, the manual method resulted in a very short alignment less than half the length of the other opsins, so Prank with TranslatorX \(^{81}\) was used for RH7 even though it had more gaps. IQ-TREE (v. 1.6.12) \(^{74}\) was used for building the gene trees from the alignments. PAML 4.9a was used to generate various models of codon evolution and estimate site-wise synonymous (\(\alpha\)) and non-synonymous (\(\beta\)) rates. The likelihood ratio test was used to determine if a site has a significantly
deviant β/α (w) from the neutral/null model. We used branch, site, and branch-site models to test for positive selection and relaxation in purifying selection.

Custom python scripts were used to filtered the data by number of opsins annotated, compile sequences for the filtered species, and prune the Kawahara et al. (2019) tree to create the species-tree used in PAML (Supplementary Information). We then tested whether diel-niche (“nocturnal” or “diurnal”) has influenced opsin rate evolution for the various selection models. For selection analyses, we categorized the background branches as strictly “nocturnal” and everything else as “diurnal” and as a foreground branch. We ran these analyses under different conditions, varying the number of taxa, choice of gene tree versus species trees and with different alignment methods to test for sensitivity to these parameters (Table S4; Supplementary Information).

**Ancestral state reconstruction**

We used the geiger 82 package in R with the SYM model and 10,000 repetitions, then used ape to plot on the pruned tree. We generated 10,000 stochastic maps for each tree in SIMMAP 83, which is part of the R package phytools 84. The methods were adapted from Kawahara et al. (2018), which also mapped diel state, without prior information on opsin transition probability. We used the SYM model for each opsin class, since we did not want to assume anything about differences in rates for losses versus gains.

Stochastic character mapping is a Bayesian approach, supposedly better and more robust than other parsimony or likelihood methods because it allows changes along branches, not just tips, and makes use of data along the nodes to make predictions. It also permits the assessment of uncertainty in character history due to topology and branch lengths 85. SIMMAP does not allow for missing or unknown data. Therefore, all tips were coded with a discrete, unordered character state, and the taxa with missing traits were pruned from the dataset, causing some discrepancies when comparing reconstructions of opsin number and diel state (Fig. S3A). We used UV/RH4, Blue/RH5, LW/RH6, total number of opsins, and diel-niche as the discrete characters for the ancestral state reconstruction.

**Protein modeling and site mapping**

We mapped amino acid sites under relaxed purifying selection, from the branch-site models, to the protein structure using Protter (Omasits et al. 2014, http://wlab.ethz.ch/protter/). Protter is a web-server based tool useful in annotating a protein sequence. It uses structural prediction tools, such as
Phobius\textsuperscript{87} to predict transmembrane domains, and allows the user to mark a custom set of amino acids.

To obtain the retinal binding sites, we chose UV opsin because it recovered seven transmembrane helices in Protter, similar to other known invertebrate opsins. We modeled it using the online swiss model workspace\cite{Waterhouse2018}. We used a squid rhodopsin z-ray structure as a template (2z73.1), as it had a much higher identity-score than other rohdopsin structures. We used Swisspdb viewer\textsuperscript{88} to fit the template and the modeled protein using magic fit, and then selected sites within 4 Å of the retinal molecule in the modeled UV-opsin.

Acknowledgments

We thank Jessica Liberles and Joseph Ahrens for guidance with bioinformatic analyses. We thank Heather Bracken-Grissom and Jorge Perez-Moreno for discussions, advice and assistance with bioinformatics pipelines, Ravindran Palavalli-Nettimi, Jesse W. Breinholt for feedback on manuscript drafts, Scott Cinel, Harlan Gough, David Plotkin, Ryan St. Laurent, Caroline Storer, and other members of the Kawahara lab for their assistance throughout the project. We thank John P. Currea, Nick Palmero, and Carlos Ruiz, along with other members of the Theobald lab for comments and feedback on the manuscript. The authors acknowledge University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication (http://researchcomputing.ufl.edu)

Funding support for the project was through NSF DEB-1557007 and NSF IOS-1750833. The Florida International University Presidential fellowship supported YS.

References


73. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7:


87. Kall, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane


**Figure Legends**

**A.**

**B.**

Figure 1: Opsin annotation and diel activity for Insecta and Lepidoptera.

Taxa are color coded by diel-niche. Red dots at terminal nodes indicate duplications within a lineage for that species and darkened colors indicate a duplication in a particular opsin family. In the Lepidoptera tree, duplications (red dots) are more commonly associated with yellow and green.
colored taxa, which are at least partially active in bright light. In insects, data is sparse with too few nocturnal species (<25%) to establish such trends.

A. Opsin annotation for 45 available insect transcriptomes and previously published literature. An NCBI taxonomy tree was grafted onto an accurate insect phylogeny obtained from Misof et al. 2014. Relationships at the level of order are based on the NCBI taxonomy tree and are only meant to be indicative of familial relationships and do not reflect current phylogenetic relationships between families. Insect silhouettes were taken from PhyloPic. B. Opsin annotation of 100 Lepidoptera species on a Lepidoptera phylogeny. The tree is a pruned cladogram of tree of the most recent Lepidoptera phylogeny. (Kawahara et al. 2019).

Figure 2: Opsin selection rates between nocturnal and diurnal Lepidoptera species for RH4/UV, RH5/Blue, RH6/LW and RH7 opsins. Diurnal and nocturnal species show differences in colour opsins for a few opsins, but on running the models with more sampling or species trees, these differences become more apparent, however RH7, a non-visual opsin, does not follow this trend.

The test was two rates vs. 1 rate for the branch model vs. the null model with a Chi-square test. A. 14-17 taxon dataset run using opsin gene trees. B. 24 taxon dataset run using opsin gene trees. RH7 was analyzed using a larger dataset of 33 taxa. C. 24 taxon dataset run using species trees. * p-value <0.05, ** p-value <0.01
Figure 3: Opsin structural prediction with retinal binding sites and sites under diversifying selection. Note the overlap between a single site identified by the two models, other sites are in close proximity to each other, signifying that sites associated with rapid evolution in high light environments could be involved in opsin tuning by changing the chemical environment around retinal.

A. UV-opsin PHOBIUS transmembrane prediction. Opsins sites under relaxed purifying selection and amino acids around (<=4 Å) from the retinal binding region have been marked.

B. UV-opsin model made using squid rhodopsin as the template. Retinal is marked and secondary structure prediction of alpha helices are shown.
### Tables

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Chi-square coefficient</th>
<th>p-value</th>
<th>No. of samples D:N</th>
<th>No. of duplications D:N</th>
</tr>
</thead>
<tbody>
<tr>
<td>nocturnal vs. diurnal + crepuscular + both</td>
<td>10.11</td>
<td>0.014</td>
<td>51:54</td>
<td>16:3</td>
</tr>
<tr>
<td>nocturnal vs. diurnal + crepuscular + both#</td>
<td>7.97</td>
<td>0.0047</td>
<td>51:46</td>
<td>16:3</td>
</tr>
<tr>
<td>diurnal vs. nocturnal + crepuscular + both</td>
<td>9.32</td>
<td>0.0022</td>
<td>42:63</td>
<td>14:5</td>
</tr>
<tr>
<td>diurnal vs. nocturnal + crepuscular + both#</td>
<td>7.41</td>
<td>0.0064</td>
<td>42:55</td>
<td>14:5</td>
</tr>
<tr>
<td>nocturnal vs. diurnal</td>
<td>10.67</td>
<td>0.001</td>
<td>42:54</td>
<td>14:3</td>
</tr>
<tr>
<td>nocturnal vs. diurnal*</td>
<td>8.478</td>
<td>0.0035</td>
<td>42:46</td>
<td>14:3</td>
</tr>
</tbody>
</table>

Table 1: Chi-square values for nocturnal vs. diurnal Lepidoptera species, including and excluding crepuscular and “both” species. Duplications occur at a higher frequency in Diurnal species than expected by chance, even after accounting for various kinds of uncertainty in diel niche assignment. (All results are significant)

D: diurnal, N: nocturnal #=(excluding uncertain diel states).

<table>
<thead>
<tr>
<th>Opsin</th>
<th>$\omega_0$</th>
<th>$\omega_1$</th>
<th>$\omega_2$</th>
<th>n</th>
<th>LnL_0</th>
<th>LnL_2</th>
<th>df</th>
<th>2LnL</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M_2 vs. M_6 gene trees</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UV</td>
<td>0.0328</td>
<td>0.0482</td>
<td>0.01643</td>
<td>24</td>
<td>-13222.396</td>
<td>-13196.06</td>
<td>48</td>
<td>52.67</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Blue</td>
<td>0.03251</td>
<td>0.0403</td>
<td>0.02945</td>
<td>23</td>
<td>-11310.806</td>
<td>-11308.083</td>
<td>46</td>
<td>5.44</td>
<td>0.0196</td>
</tr>
<tr>
<td>LW</td>
<td>0.03264</td>
<td>0.0521</td>
<td>0.01173</td>
<td>24</td>
<td>-11232.972</td>
<td>-11174.84</td>
<td>48</td>
<td>116.27</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>RH7</td>
<td>0.0653</td>
<td>0.0594</td>
<td>0.06719</td>
<td>33</td>
<td>-22578.67</td>
<td>-22577.68</td>
<td>66</td>
<td>1.978</td>
<td>0.15</td>
</tr>
<tr>
<td>UV</td>
<td>0.03058</td>
<td>0.0445</td>
<td>0.02324</td>
<td>14</td>
<td>-6892.501</td>
<td>-6887.289</td>
<td>28</td>
<td>9.6</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Blue</td>
<td>0.0354</td>
<td>0.0343</td>
<td>0.03602</td>
<td>17</td>
<td>-9822.971</td>
<td>-9822.9126</td>
<td>34</td>
<td>0.115</td>
<td>0.735</td>
</tr>
<tr>
<td>LW</td>
<td>0.03893</td>
<td>0.0434</td>
<td>0.03445</td>
<td>17</td>
<td>-6690.525</td>
<td>-6689.4261</td>
<td>34</td>
<td>2.197</td>
<td>0.138</td>
</tr>
<tr>
<td><strong>M_2 vs. M_6 species trees</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UV</td>
<td>0.03206</td>
<td>0.0411</td>
<td>0.02609</td>
<td>24</td>
<td>-13265.664</td>
<td>-13260.007</td>
<td>49</td>
<td>11.314</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>Blue</td>
<td>0.03361</td>
<td>0.0465</td>
<td>0.02922</td>
<td>23</td>
<td>-11333.887</td>
<td>-11328.244</td>
<td>47</td>
<td>11.287</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>LW</td>
<td>0.03309</td>
<td>0.0495</td>
<td>0.02589</td>
<td>24</td>
<td>-11247.362</td>
<td>-11233.584</td>
<td>49</td>
<td>28.779</td>
<td>&lt;0.001*</td>
</tr>
</tbody>
</table>
Table 2: $\omega$ (dN/dS) rates for null ($M_0$) and branch models ($M_2$) for Ultraviolet (UV), Blue, Long wavelength (LW) and RH7 opsins using species and gene trees and different number of taxa.

Increasing the number of taxa shows significant differences between rates across nocturnal ($\omega_1$) and diurnal species ($\omega_2$) for color opsins but not for RH7 opsin, which served as a control. Using species tree shows differences for all three opsins.

$\omega_0$: rates for null model, $\omega_1$: rates for diurnal species, $\omega_2$ for nocturnal species, n: number of taxa, LnL$_{0/2}$: log likelihood scores for $M_0/M_2$, df: degrees of freedom for $M_2$, 2LnL: twice the LnL$_2$-LnL$_0$, * highly significant p-value (<0.01)