12-1-2021

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

Yash Sondhi
Florida International University

Emily A. Ellis
Florida Museum of Natural History

Seth M. Bybee
Brigham Young University

Jamie C. Theobald
Florida International University

Akito Y. Kawahara
Florida Museum of Natural History

Follow this and additional works at: https://digitalcommons.fiu.edu/all_faculty

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

Yash Sondhi[1][✉], Emily A. Ellis [2], Seth M. Bybee[3], Jamie C. Theobald [1] & Akito Y. Kawahara [2]

Opsins, combined with a chromophore, are the primary light-sensing molecules in animals and are crucial for color vision. Throughout animal evolution, duplications and losses of opsin proteins are common, but it is unclear what is driving these gains and losses. Light availability is implicated, and dim environments are often associated with low opsin diversity and loss. Correlations between high opsin diversity and bright environments, however, are tenuous. To test if increased light availability is associated with opsin diversification, we examined diel niche and identified opsins using transcriptomes and genomes of 175 butterflies and moths (Lepidoptera). We found 14 independent opsin duplications associated with bright environments. Estimating their rates of evolution revealed that opsins from diurnal taxa evolve faster —at least 13 amino acids were identified with higher dN/dS rates, with a subset close enough to the chromophore to tune the opsin. These results demonstrate that high light availability increases opsin diversity and evolution rate in Lepidoptera.

[1] Department of Biology, Florida International University, Miami, FL, USA. [2] McGuire Center for Lepidoptera and Biodiversity, Florida Museum of Natural History, University of Florida, Gainesville, FL, USA. [3] Department of Biology, Brigham Young University, Provo, UT, USA. [✉]email: ysond001@fiu.edu

The evolution of light detection has allowed animals to monitor ambient brightness for circadian rhythms, determine the intensity of light from different directions for phototaxis, and form images for proper spatial vision[1]. With images, animals track and correct their own motion, target or avoid objects, and sometimes infer more elusive properties of a scene, such as object nearness or time until a collision. Although some animals achieve spatial vision without dedicated structures, like the diffuse photoreception of sea urchins[2], overwhelmingly, acuity is captured by eyes, the diverse and convergently evolved organs that arrange screening pigments, optical lenses, and photoreceptors, to focus and capture images[3]. When light is sufficient, some estimate of its wavelength composition can improve object discrimination even under diverse or uneven lighting[4]. This perception of color requires photoreceptors with different wavelength sensitivities whose signals are then compared by underlying opponent processes[5].

Wavelength sensitivities result largely (but not wholly) from a photoreceptor's visual pigments, which are formed by joining a light-sensitive, retinal-based chromophore to an opsin, a seven-transmembrane-domain G protein-coupled receptor protein. When a chromophore responds to a photon absorption that changes its conformation, the opsin transduces this into a biologically meaningful signal by activating an internal G protein signaling cascade. Large-scale phylogenomic analyses have found many duplications and losses in the opsin protein family across invertebrates[6,7]. The water flea, *Daphnia pulex* possesses 46 opsins[8], dragonflies have 15–33 opsins[9–11] and mantis shrimp have 12–33 opsins[12]. However, any more than four spectral channels offer diminishing returns in extracting information from natural scenes[13,14], so why do some animals have so many opsins?

Since they regulate visual light transduction, opsins are subject to strong adaptive evolution, but they can also evolve through non-adaptive mechanisms. 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 retention and diversification[15], such as with mate choice in guppies and butterflies[16,17], flower foraging in bees and wasps[18] and changing light intensity environment with many nocturnal animals[19,20].

Sensory modalities such as smell, electromagnetic reception, and touch can be more reliable than vision in dim environments[19]. Resource allocation trade-offs can cause a loss of stabilizing selection on genes of inefficient sensory systems, resulting in their downregulation or them becoming pseudogenes. This has been seen in nocturnal mammals[21], cave-dwelling crayfish[22], and deep-sea organisms[23–25]. If diminished light availability causes reduced opsin expression and loss, abundant light, conversely, may cause higher opsin expression, or prevent the loss of duplicated opsins and eventually lead to functional divergence. A comparison of visual genes between diurnal and nocturnal Lepidoptera revealed elevated opsin expression in the diurnal species[26]. Similarly, a study of opsin evolution across fireflies found higher amino acid transition rates in diurnal fireflies, across four independent diel switches[27].

Lepidoptera opsin diversity has been studied in a handful of model taxa—mostly diurnal butterflies[26,28,29]—but has yet to be studied comprehensively across the entire order and multiple diel niches. Opsins are characterized by the wavelength to which their response is maximum (λ_{max}). The λ_{max} is sufficient to approximate the response curve of most opsins[30]. Lepidoptera opsins are usually classified as UV/RH4, Blue/RH5, and LW/RH6 opsins with corresponding maximal responses (λ_{max}) in the ultraviolet (UV) (300–400 nm), blue (400–550 nm), and green/red (450–620

nm) wavelengths (LW). RH4–6 are implicated in color vision, but Lepidoptera also possess the non-visual RH7, which is associated with light sensing needed to maintain circadian rhythm[31].

Earlier studies on Lepidoptera opsin evolution include work by Briscoe[28], who analyzed the visual genes of eight Lepidoptera species including two moths, Xu et al.[32], who analyzed 30 species including 12 moths, and Feuda et al.[31] who analyzed 10 species including four moths. However, due to small sample sizes, these studies had limited statistical power. They used gene trees instead of species trees for selection analyses, which, if different from the species tree topology, can bias results[33]. The few studies that examined opsin diversity and diel-niche association[31,32] compare butterflies and moths, effectively using only a single diel switch. But Lepidoptera have more than 100 recorded diel transitions[34], 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 test if bright environments drive opsin diversification, we mine genomes and transcriptomes of 175 Lepidoptera species for visual opsins, combine our annotations with natural history diel-niche from the literature and map these traits onto a well resolved tree[35] to examine their evolution.

Results

Lepidoptera opsin duplications associated with bright environments: mapping opsin diversity and diel-niche. We examined patterns of opsin diversity across Lepidoptera by mining assembled transcriptomes[35] and annotated genomes from Lepbase[36] and Ensembl Metazoa[37] (Supplementary Data 1). Transcriptome quality was assessed using BUSCO[39] and visual score, calculated as the percentage of genes recovered from a well characterized set of Lepidoptera visual genes[26] (Supplementary Data 1,2). A phylogenetically informed annotation approach (PIA)[40] was used for opsin annotation (Supplementary Data 1), which we reconfirmed by building nucleotide and amino acid gene trees (Supplementary Figs. 1, 2). We recovered at least one opsin from 114 of 175 Lepidoptera taxa (Supplementary Data 1) and were able to confirm diel-niche for 104 of these species using the literature, in consultation with experts.

We mapped opsin diversity—specifically UV/RH4, Blue/RH5, LW/RH6, and RH7 opsin recovery—and diel-niche onto a well-resolved Lepidoptera phylogeny[35] (Fig. 1). Species with duplications were counted (Table 1) and the number of duplication events were estimated using a tree reconciliation analysis (Supplementary Data 3, Supplementary Figs. 3–5). Chi-squared tests determined if the number of diurnal species with duplications was more than what could be expected by chance. Duplication events occurred more often in diurnal lineages ($\chi^2 = 6.025$, p-value = 0.014, Table 1) and increased duplication in diurnal species was evident even after we excluded species with ambiguity in diel-niche assignment ($\chi^2 = 8.478$. p-value = 0.0035, Table 1).

Transcriptome quality was variable (Supplementary Data 1, 2) and to ensure the trends were not an artifact of this variation, we investigated whether transcriptome quality was correlated to phylogeny and found no evidence for such a correlation (BUSCO, $K = 0.3143298$, p-value = 0.151; visual score, $K = 0.3846586$, p-value = 0.071). To test if transcriptome quality was correlated with diel-niche after accounting for the effect of phylogeny, a Phylogenetic ANOVA was performed, but it failed to reveal any correlations (BUSCO, $F = 0.375$, d.f. = 4, p-value = 0.8256; visual score, $F = 0.25$, d.f. = 4, p-value = 0.8213).

We performed an ancestral state reconstruction (ASR) of diel-niche, each opsin and total number of opsins. Opsin losses are notoriously difficult to confirm[31] and while both tree reconciliation
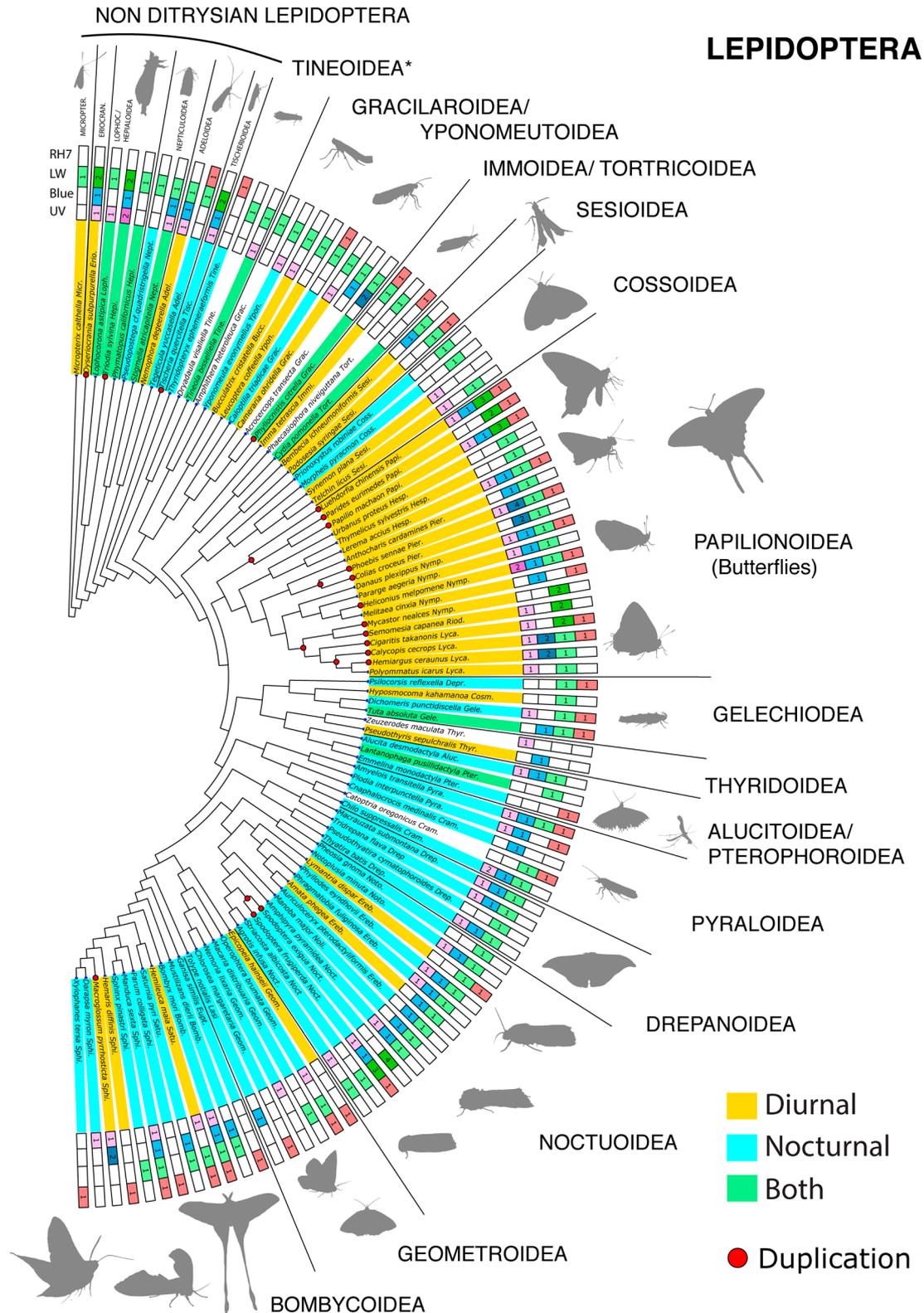
Fig. 1 Opsin annotation and diel activity are mapped onto a Lepidoptera species tree. Duplications are associated with taxa active in bright light. Taxa are color coded by diel-niche and RH4/UV, RH5/Blue, RH6/LW, and RH7 opsin recovery is marked. Red dots at nodes indicate duplication events identified by the tree reconciliation analysis for that lineage and darkened colors indicate a duplication in a particular opsin. Duplications (red dots) are more commonly associated with diurnal (yellow) and and partially diurnal or"both" taxa (green) which are more active in bright light environments. The tree is a pruned cladogram of tree of the most recent Lepidoptera phylogeny[35]. 98/104 taxa are included in the figure, taxa for which the identification were only till genus level are excluded, but all the species with duplications are shown. Asterisk (*) indicates that these superfamilies are not monophyletic. Family names and some superfamily names are abbreviated, see Supplementary Data 1 for expanded names and complete classification, raw data data used to create these figures can be found at ref. [38].

Table 1 Chi-square values for nocturnal vs. diurnal Lepidoptera species, including and excluding crepuscular and "both" species.

Dataset	Chi-square coefficient	p-value	No. of samples D:N	No. of duplications D:N
nocturnal vs. diurnal + crepuscular + both	10.11	0.0014	51:54	16:3
nocturnal vs. diurnal + crepuscular + both[#]	7.97	0.0047	51:46	16:3
diurnal vs. nocturnal + crepuscular + both	9.32	0.0022	42:63	14:5
diurnal vs. nocturnal + crepuscular + both[#]	7.41	0.0064	42:55	14:5
nocturnal vs. diurnal	10.67	0.001	42:54	14:3
nocturnal vs. diurnal[#]	8.478	0.0035	42:46	14:3
nocturnal vs. diurnal[**]	6.0259	0.014	42:46	14:5
nocturnal vs. diurnal[**] (yates-correction)	4.8171	0.02818	42:46	14:5

Rows 1–7 is count data of the species with identified opsins. Row 8–9 is the number of duplication events from the tree reconciliation analysis. 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 $p < 0.05$).
D diurnal, *N* nocturnal.
#excluding uncertain diel states.
**From tree reconciliation analyses.

and ASR results are included (Supplementary Figs. 3–10)., we refrain from interpreting the losses.

Ultraviolet (UV). 50% (57/114) of the taxa recovered UV/RH4 opsins (Fig. 1, Supplementary Figs. 3, 7). UV opsins had duplications in only 3 independent lineages; the diurnal *Heliconius melpomene*, in which UV opsin duplication has been recorded before[41] and crepuscular *Triodia sylvina* (Hepialidae), an ancient lineage of ghost moths, known for swarming at dusk[42]. The only nocturnal species with a UV duplication was *Chilo supressalis*, an important pest species[43].

Blue. 46.4% (53/114) of the taxa recovered Blue/RH5 opsins (Fig. 1, Supplementary Figs. 4, 8). Blue opsin duplications were present in 5 families, with 6 different duplication events, all occurring in diurnal species. The pierid species, *Colias croceus* and *Phoebis sennae*, and two lycaenid genera, *Calycopis* and *Hemiargus*, also have Blue opsin duplications. Behavioral and electrophysiological data have shown that these families, if not these individual species, have functional duplications[28,44,45]. *Macroglossum pyrrhosticta*, a diurnal hawkmoth (Sphingidae), also had a Blue duplication.

Long wavelength (LW). 83.3% (96/114) of taxa recovered LW/RH6 opsins, the highest capture rate of all three opsin families (Fig. 1, Supplementary Figs. 5, 9). LW opsins duplications were recorded in 8 families and 11 genera, with 10 duplication events. A Mann–Whitney U test indicated that there were more LW opsins in diurnal taxa than nocturnal taxa ($U = 662.0$, p-value = 0.033). Only two of these genera with duplications are nocturnal. We recovered previously reported LW duplications in one Lycaenidae, two Riodinidae, and three Papilionidae species[28]. Butterflies are among the few insects known to have sensitivity to the color red, through LW duplication or filtering pigments[28,46]. Thus, duplications may indicate true color expansion. Other diurnal or crepuscular species with LW duplication were *Dyseriocrania subpurpurella* (Eriocraniidae) and *Triodia sylvina* (Hepialidae).

The tiger moth *Callimorpha dominula* (Erebidae: Arctiinae) and the giant butterfly moth *Paysandisia archon* (Castniidae), two diurnal moths not in this dataset, also have LW duplications, and *Paysandisia* is known to have sensitivity to red wavelength light[31,44]. *Spodoptera* (Noctuidae), of which some species are invasive pests, is nocturnal but has a red sensitive LW duplication. Many *Spodoptera* species are migratory, and flying above the clouds of the night sky may free them from low light constraints[47,48]. *Tischeria quercitella* (Tischeriidae), a leaf mining

moth, is the only other nocturnal species found with a LW duplication, despite examining over 50 nocturnal species. In addition to the visual opsins, the non-visual RH7 opsin was also analyzed (Supplementary Fig. 10), but we recovered far fewer opsins with only 28% (33/114) of taxa recovering RH7 opsins. No duplications were found, consistent with other work[31].

Diurnal species opsins have higher selection rates than nocturnal species: Opsin selection in Lepidoptera. PAML estimated rates of selection (ω or dN/dS) and we tested if these rates differed between nocturnal and diurnal Lepidoptera. For datasets that showed significant differences, branch-site models in PAML and HyPhy were used to identify amino acids under selection (Supplementary Data 4). Sensitivity to different sample sizes and starting trees was tested (Fig. 2, Table 2, Supplementary Data 4). To ensure comparison of opsin rates across the same group of species, analyses were limited to species that recovered all three visual opsins. We were unable to use the entire dataset of recovered opsins for UV, Blue, and LW opsins because of low quality alignments and manually trimmed or removed sequences that reduced the alignment quality (see "Methods" for details).

We first ran PAML on a subset of the data ($n = 12$–14), similar in size to those used by Feuda et al.[31] and Xu et al.[32], to determine the effect of sample size, although taxon sampling across datasets was non-uniform. Of all three visual opsins, only UV opsin rates were a significantly better fit when partitioned into two rates, using diel niche, than one rate for all branches (p-value = <0.001, $\omega_{\text{diurnal}(d)} = 0.0447$, $\omega_{\text{nocturnal}(n)} = 0.0234$) (Fig. 2A, Table 2, Supplementary Data 4).

Next, using larger sample sizes and similar species across datasets ($n = 24$–27), we found significant differences between diel-niche for all three color opsins. The dN/dS rates were higher in diurnal species than nocturnal species. UV (p-value <0.001, $\omega_d = 0.04816$, $\omega_n = 0.01643$) and LW opsins (p-value <0.001, $\omega_d = 0.052$, $\omega_n = 0.0117$) had a greater magnitude of differences than Blue opsins (Fig. 2B, Table 2, Table S4). RH7 acted as a control because it is not involved in vision and had no significant differences for rates between diel-niches.

Since Blue and UV opsin recovery was sparse (~50% of the taxa) and gene tree topology might be biased, we ran the PAML analysis using a species tree instead of gene trees. The species tree models showed significant differences between diel-niches for all three visual opsins (p-value <0.001) (Fig. 2C, Table 2, Supplementary Data 4). RH7 was excluded from this analysis in order to compare across similar species trees, which was precluded by the poor overlap of species between RH7 and visual opsins.
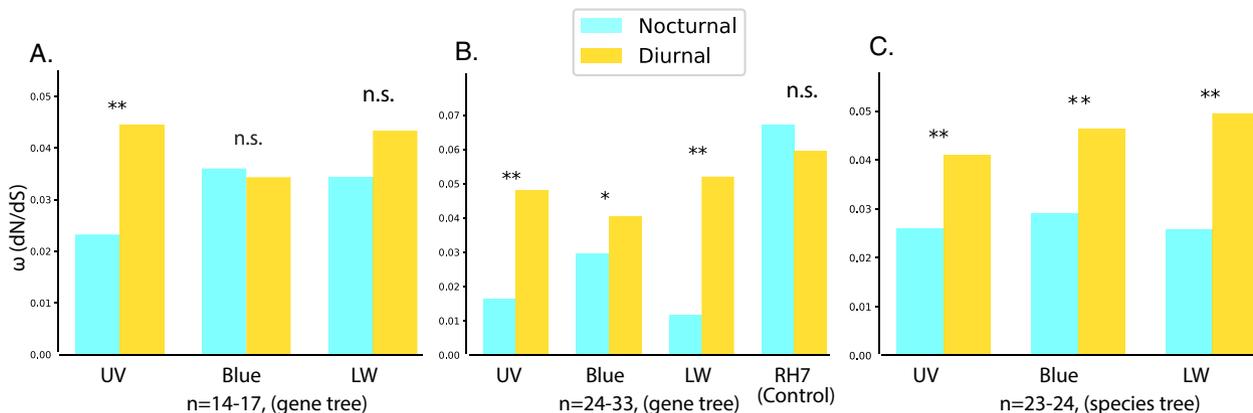
Fig. 2 Opsin selection (dN/dS) rates between nocturnal and diurnal Lepidoptera species. Different models show that rates for visual opsins are higher in diurnal taxa. A Blue, LW ($n = 17$ taxa) and UV ($n = 14$ taxa) opsin dataset were run using opsin gene trees. **B** Expanded dataset ($n = 24$ taxa) for UV, Blue, and LW opsins run using gene trees. RH7 was analyzed using more species ($n = 33$ taxa). **C** Expanded dataset run using a robust species trees (UV, LW $n = 24$ taxa, Blue = 23 taxa). *p-value < 0.05, **p-value < 0.01. The significance values indicate how well different models fit the data using likelihood ratio test (LRT). Model parameters used to run PAML and estimate the rates represented in the figures can be found at ref. [38]. Opsin selection rates modeled for various datasets with different sample sizes and different model parameters all show that RH4/UV, RH5/Blue, and RH6/LW visual opsins have higher dN/dS rates in diurnal species than nocturnal species. RH7, a non-visual opsin, shows no significant difference. Smaller sample sizes (**A**) often fail to detect differences, which were more apparent when using larger data sets (**B**). Using a robust species tree rather than the gene tree shows more consistent differences across the visual opsins (**C**).

Table 2 dN/dS (ω) rates for null (M_0) and branch models (M_2) for 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 (ω_1) and diurnal species (ω_2) for color opsins but not for RH7 opsin, which served as a control.

| Opsin	ω_0	ω_1	ω_2	n	LnL_0	LnL_2	d.f.	2LnL	p-value
M_2 vs. M_0 gene trees									
UV	0.0328	**0.0482**	**0.01643**	24	−13222.396	−13196.06	48	52.54	<0.001*
Blue	0.03251	0.0403	0.02945	23	−11310.806	−11308.083	46	5.449	0.0196
LW	0.03264	**0.0521**	**0.01173**	24	−11232.972	−11174.84	48	116.27	<0.001*
RH7	0.0653	0.0594	0.06719	33	−22578.67	−22577.68	66	1.978	0.15
UV	0.03058	**0.0445**	**0.02324**	14	−6892.501	−6887.289	28	9.6	<0.001*
Blue	0.0354	0.0343	0.03602	17	−9822.971	−9822.9126	34	0.115	0.735
LW	0.03893	0.0434	0.03445	17	−6690.525	−6689.4261	34	2.197	0.138
M_2 vs. M_0 species trees									
UV	0.03206	**0.0411**	**0.02609**	24	−13265.664	−13260.007	49	11.314	<0.001*
Blue	0.03361	**0.0465**	**0.02922**	23	−11333.887	−11328.244	47	11.287	<0.001*
LW	0.03309	**0.0495**	**0.02589**	24	−11247.362	−11233.584	49	28.779	<0.001*

Using species tree shows differences for all three opsins. ω_0: rates for null model, ω_1: rates for diurnal species, ω_2 for nocturnal species, n: number of taxa, $LnL_{0/2}$: log likelihood scores for M_0/M_2, d.f.: degrees of freedom for M_2. 2LnL: twice the LnL_2–LnL_0, * highly significant p-value (<0.001), with rates from corresponding models in bold.

To test if particular sites were under positive selection in diurnal species, we used site and branch site-models. The site models test if any sites have different dN/dS rates across all branches. The analyses found no significant sites under positive selection across all branches showing that if diel-niche is not considered, there is no underlying signal of elevated dN/dS rates. The branch-site models, which allow partition by diel-niche, did identify amino acid sites that were under a higher dN/dS in diurnal species in UV, Blue, and LW opsins, however it was not able to classify them as being under positive selection. This is expected in opsins because they have a low dN/dS rates overall[49], with $\omega < 0.5$, and the null models fixes dN/dS rates to $\omega = 1$, preventing it from classifying it as significant positive selection. Instead, we used more sensitive tests like the Contrast-FEL[50] and MEME models[51] from HyPhy[52] to verify the branch site models (Supplementary Data 4). These models look for differences in dN/dS rates in different sites among different branches, as opposed to looking for rates greater than one and thus are more

appropriate for opsins. They are more recent implementations and are designed to be more sensitive for detecting selection since they are tailored to detect specific kinds of evolutionary scenarios[50–52]. They showed similar results as the PAML analysis but revealed additional sites under higher dN/dS in diurnal opsins (Supplementary Data 4). There were some opsin specific patterns, UV and LW opsins had considerable overlap in recovered sites for the two different analyses of sites under selection, but Blue opsins had less overlap in the identified sites (Supplementary Data 4). Model sensitivity to input tree also had opsin specific trends. With Blue and LW opsins, sites under selection predicted by the models were unaffected by the choice of gene vs. species trees but it did affect the results of UV opsins (Table S4).

Xu et al.[32] reported elevated dN/dS rates in butterflies (diurnal) compared to moths (nocturnal), with LW, Blue, and UV opsins, showing a decreasing magnitude in differences. In contrast, our study finds UV opsins had the highest and most consistent dN/dS rate differences. Feuda et al.[31] used two independent diurnal
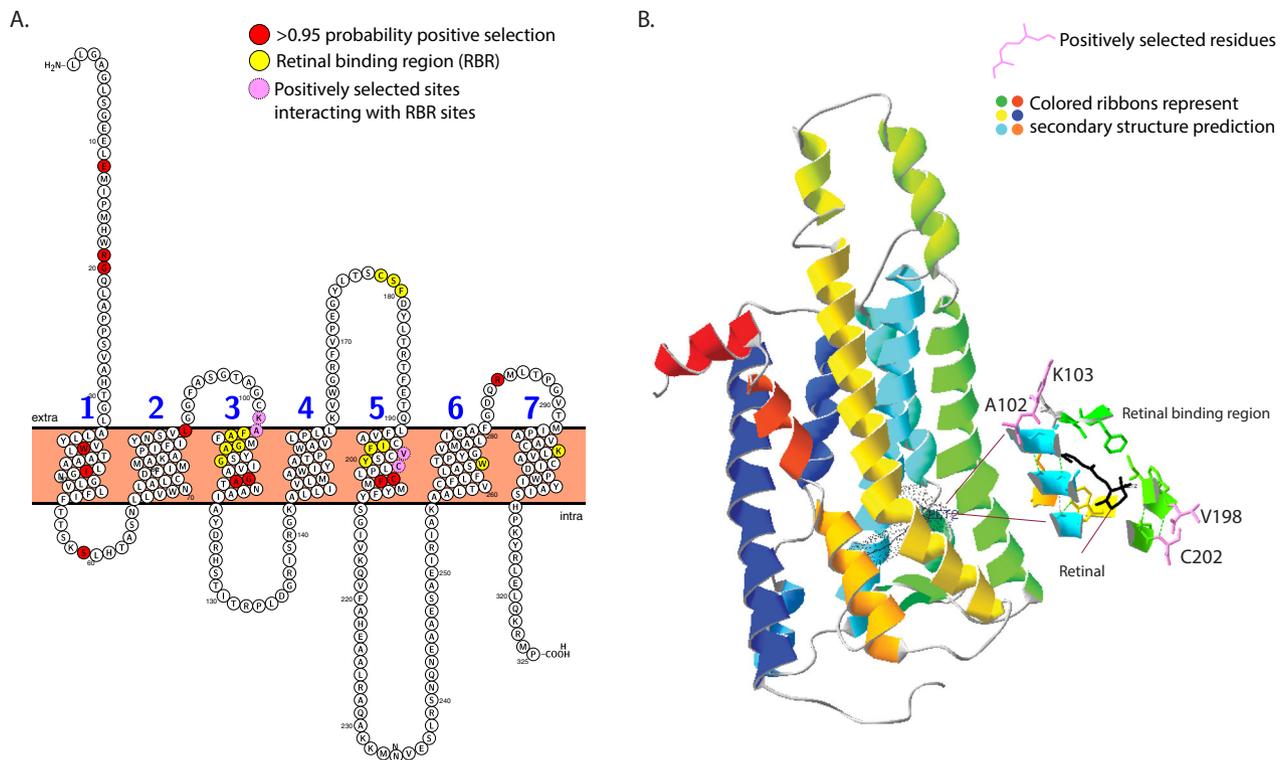
Fig. 3 Opsin structural prediction with retinal binding sites, sites with higher dN/dS (positively selection) and putative tuning sites, the subset that interacts with both. A UV-opsin PHOBIUS transmembrane prediction. Opsins sites with higher dN/dS rates and amino acids close to (<=4 Å) retinal, which form the retinal binding region (RBR) have been marked. **B** UV-opsin model made using squid rhodopsin as the template. Retinal is marked, secondary structure prediction of alpha helices are shown in different colors and the expanded view shows the positively selected sites that can interact (<=2 Å) with the retinal binding sites. Protein models used to get these metrics can be found at ref. [38]. Sites under positive selection are highlighted in red, sites in the retinal binding region in yellow and positively selected sites that are capable of interacting with the retinal binding sites in pink, identifying likely candidates for opsin tuning (**A**). Other positively selected are located within the transmembrane region and could be indirectly involved in opsin tuning (**B**).

transitions, with a total of 10 Lepidoptera taxa and have results similar to ours (Fig. 2B). UV and Blue and LW genes in nocturnal Lepidoptera underwent higher dN/dS rates in diurnal species but RH7 had almost similar levels of selection in both nocturnal and diurnal species.

Expansion of color vision through sequence tuning: mapping sites to predicted protein structure. We mapped the significant sites with higher selection rates (positive selected sites) onto the predicted protein structure (see "Methods" section) and the transmembrane helix predictions (Fig. 3A, B). Each class of opsin mapped sites to a unique pair of adjacent transmembrane helices (Supplementary Fig. 11). 13 amino acids were identified in the retinal binding region (Fig. 3A, B; Supplementary Data 4) and we mapped these retinal binding sites to transmembrane and 3D structures (Fig. 3A, B), to see if there were any interactions. The positively selected amino acids at sites 102, 103, 198, and 102 are bonded to the retinal binding sites (<2.0 Å) and hence are strong candidates for spectral tuning (Fig. 3, Supplementary Data 4). Eight other positively selected sites also map to transmembrane regions and may tune the opsin indirectly. In addition, UV 59 may be a convergent site of selection, as it has previously been identified in diel-transitions of fireflies[27].

Discussion
What does opsin duplication or loss mean for the visual system of an organism? One consequence of opsin duplication and divergence 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 and 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) due to Long Wavelength (LW) opsin duplication and divergence[29]. RH4–6 are implicated in color vision, but non-visual RH7 opsin is associated with circadian rhythm maintenance and has a phylogenetically scattered distribution among Lepidoptera.

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 33 color opsins[10,29,31,46], while beetles and scorpionflies show losses[53,54]. These studies have not examined potential links between diel activity and opsin diversity, or have failed to find consistent trends (reviewed in ref. [31]). 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[33,55].

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, one would sequence only eye tissue with muscle tissue from the body as a control, but complete

information about which tissues generated transcriptomes is often lacking. Further, researchers typically do not measure expression levels, which could identify non-functional opsins[56]. However, because multiple quality control metrics and annotation methods were used here, we believe we have obtained the best possible estimate of opsin diversity with these data. While some transcriptomes had lower quality, variation in quality was randomly distributed across the phylogeny (i.e., phylogenetic signal was low). Furthermore, after accounting for phylogeny, we were unable to find evidence for correlations of diel-niche with transcriptome quality, together indicating that the trends with diel-niche are unlikely to be an artifact of transcriptome quality. We are not confident that the losses we identify are true losses and refrain from making claims about trends for losses but include both the ancestral state reconstruction and tree reconciliation results (Supplementary Figs. 3, 4). Qualitatively, diurnal lineages and nodes appeared to have more opsins. Although we were not able to confirm this quantitively in the reconstructions, we did find the trend of higher diurnal opsin counts from the extant taxa for the better recovered LW opsins.

The selection models implemented in PAML are sensitive to a range of parameter conditions such as choice of tree topology, quality of alignment and sample size. We ran these models using a range of a parameters and presented the results. It is shown that branch length in the tree topology can affect these analysis[33] and therefore, we used the species tree rather than the gene trees models for estimating sites under positive selection. The species tree had better estimates of evolutionary branch length because it was constructed using phylogenomic data[35]. We also used several newer models, implemented in HyPhy, to detect sites under positive selection. These models are tailored to be more sensitive in the detection of specific kinds of evolutionary scenarios[50–52] and unlike the PAML these models do not assume dN/dS rates =>1, but just search for site classes with different rates and thus are better suited opsins, which almost never have any sites with dN/dS rates >=1.

Our analyses show that at least four positively selected sites are in the bio-physical range of direct interaction of the retinal binding pocket and hence could influence opsin spectral tuning in diurnal species by changing the chemical environment surrounding retinal. Feuda et al.[31] also map the positively selected amino acids to opsins[31], but they recover a large fraction of sites in terminal regions, not in the helices. This could be due to alignment methods (i.e., removing gap-filled regions). 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.

More than 75% of the opsin duplications events that we discovered across 10 independent lineages are in diurnal species, which additionally have higher selection (dN/dS) rates than nocturnal species. Differences in selection rate between the diel niches are visible across all three visual opsins, but are most robust in ultraviolet-sensitive opsins. Structural modeling shows that in diurnal species, most amino acids under positive selection are in the transmembrane helices and a subset interact 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. An alternative scenario, where a diverse repertoire of opsins allows the species to invade new light environments is also possible. However, ancestral state reconstruction indicates the shift to diurnality occurred prior to the duplication of opsins. However, because deep-time nodes are more difficult to reconstruct, the specific mechanism that gives rise to the results we document are difficult to resolve.

The ancestors of both diurnal and nocturnal lepidopterans were likely crepuscular with limited trichromatic vision[34], and if diel-niche shifts and opsin diversification are linked, we first need to know why their diel-niche switched. One hypothesis that has been proposed is that moths that were nocturnal switched to becoming diurnal to escape predatory bats[57]. Moving to diurnal behavior reduces light constraints and predation from bats, but may increase potential risk to visual predators such as birds, spiders, and parasitoids. Because the light environment during the day allows for visual aposematism, bright warning coloration is expected (and many diurnal species, both moths and butterflies, are more colorful than nocturnal moths[58]). Non aposematic *Heliconius* butterflies, for example, are slower than their counterparts because they are chemically defended[59], similarly there are many brightly colored, palatable mimics[60].

For species that switched to a nocturnal lifestyle from an ancestrally crepuscular state, light intensity was a limiting factor, and their eyes evolved to become superposition eyes. Superposition eyes effectively act as a large lens that increases the light available to each photoreceptor and these eyes have independently evolved in several nocturnal and crepuscular species[44,61]. The nature of visual pigments makes capturing color information harder as light intensity decreases. Stacking opsins for more sensitive receptors comes at the cost of getting a more broadband signal and a loss of color discrimination[62]. Color vision is slower than monochromatic vision and because the moth visual system also slows down to detect more light at night, seeing color is even more costly at night. The maintenance of trichromacy in nocturnal Lepidoptera is therefore puzzling: perhaps 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[63], or the strong innate attraction of flower foraging Lepidoptera to blue[64,65]. It is also possible multiple opsins may be maintained together for other reasons, for example, moths may overcome dim light constraints by pooling from spectrally distinct opsins to get a better representation of various visual stimuli. 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[66], suggesting that moths may have partitioned color and spatial vision in different regions.

Alternatively, selection could act on each opsin independently, regardless of their utility for color vision. UV and LW opsins models consistently showed differences across models. LW opsin recovery was almost complete, but UV and Blue opsin recovery was patchy. If this trend represents actual losses, though unlikely, 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[23,67,68]. UV contrast is a foraging cue for moths, for example nectar guides in many-night blooming flowers[69], and UV light is commonly used as an attractant to moth traps[70–72]. Because short wavelengths increase around twilight[73], UV light is a possible signal for pupil responses, (anecdotally more prevalent across nocturnal moths than butterflies), which could explain the lower selection rates in nocturnal species. LW sensitivity is useful for oviposition behavior in butterflies[74] and their high recovery and low selection rates in nocturnal species could be a signature of their role in oviposition. Butterflies display strong host plant specificity compared to moths[58,75,76], and butterfly LW opsins 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 stabilizing selection in nocturnal species due to integral functions.

In summary, we provide compelling evidence that patterns of opsin diversity and evolution across Lepidoptera are correlated with differences in diel-niche and opsin duplication and rates of evolution. Since light environment can drive gene evolution, there is a high chance that extreme increase in artificial light also plays a similar role. With the current rate of declines in insect populations it is imperative we understand the effects of light environment, especially artificial light on nocturnal animals, and prevent further declines. 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. In-vitro expression systems and CRISPR could also confirm that the amino acids identified result in spectral tuning of the photopigment. Examining opsin distribution patterns in eyes across lepidopterans could help to understand if there are sensitivity and color vision trade-offs 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 future research.

Methods

Lepidoptera opsin gene annotation. We annotated 162 Lepidoptera transcriptomes obtained from previously published studies (Supplementary Data 1). See supplementary information (Dataset S1, S3, S4) from Kawahara et al.[35] for details on transcriptome assemblies. Transcriptomes were annotated using Phylogenetically Informed Annotation (PIA)[40]. 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 for faster analysis on a high-performance cluster (https://github.com/xibalbanus/PIA2).

A set of well characterized metazoan visual opsin genes[40] 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 added opsin sequences from 13 Lepidoptera species obtained from BLAST searches of Lepbase[36], a repository of Lepidoptera genomes and transcriptomes, and Ensembl Metazoa[37]. We used *Manduca sexta* RH4, RH5, RH6, and, RH7 opsin sequences as queries for the BLAST search (e-value: 1.0e−10, num_alignments: 250).

The PIA analysis was conservative, it did not misidentify sequences, but it often picked up partial length opsins. If there was only one opsin in that class, we retained it, however for putative duplications, we used the local alignment option in Geneious v. 10.0.9 (https://www.geneious.com) to ensure they were not overlapping fragments, in most cases the duplications were easy to identify as duplications. In the cases where they were identical sequences, we retained the longer sequence and did not count that as a duplication. The Lepbase and Ensembl BLAST searches often mischaracterized which opsin family the genes belonged to and the opsin gene tree picked up on these errors (Supplementary Fig. 2).

Opsin gene tree reconstruction. The Lepidoptera cDNA sequences were collected and annotated using PIA as putative short wavelength (UV/RH4), medium wavelength (Blue/RH5), long wavelength (LW/RH6), and RH7 (Table S1). However, as a confirmation, we also constructed gene trees for all the sequences. Both trees were midpoint rooted using Archaeopteryx v0.9917beta (https://github.com/cmzmasek/archaeopteryx-js).We used FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) to order and color the nodes and convert the trees into cladograms for easier visualization (Supplementary Figs. 1, 2).

Gene tree estimation. MAFFT (v7.294b)[77] was used with default settings to align the sequences and IQ-TREE (multi-core v1.6.12)[78] to build a ML nucleotide gene tree (iqtree-s alignment_name.fasta -st DNA -bb 10000 -nt AUTO -alrt 1000) and an amino acid gene tree (iqtree -s alignment_name.fasta -st AA -bb 10000 -nt AUTO -alrt 1000). IQ-Trees uses ModelFinder[79] to identify the best model and ultrafast bootstrap[80] to calculate bootstrap scores. We used the gene trees to confirm opsin protein annotations taken from the genomes.

Diel niche assignment. We assigned diel-niche to compare opsin evolution, but only did so for species which 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) (Supplementary Data 1). The diel-niche was assigned using

published literature, natural history databases and in consultation with experts for more obscure species. 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[34]. 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[81]. We used different approaches for diel-niche assignment depending on analyses. For easy tree visualization, we used 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 with a "?" in the dataset. 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. In order to obtain a well-resolved Lepidoptera species tree, we pruned the species tree[82] from Kawahara et al.[35], using Python scripts using the package ETE v3[82], to include only species that had at least one identified opsin. These scripts then modified the pruned tree to show the number of opsins in various Lepidoptera (Fig. 1).

Opsin duplication. The python scipy stats package[83] 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). Duplication events resulting from the tree reconciliation analysis were also analyzed (Table 1) but duplications events at deep nodes were ignored for the purpose of the diel-niche associations (Supplementary Data 3) since there is uncertainty about ancestral diel-niches. Chi-squared tests were used to check if the proportion of duplications in each diel-niche was significantly different from a random distribution. Number of species with duplications and number of duplications events were both tested. Since there was no easy analog for number of lineages without duplications, we chose to the most conservative estimate and use the number of diurnal and nocturnal taxa. The Yate's correction for continuity was applied for the duplication event data implemented using a web tool (https://www.socscistatistics.com/tests/chisquare/). The Yate's correction is useful when the event numbers are low in any one category. We also compared the opsin counts for all opsins and for each opsin for extant diurnal and nocturnal taxa used the non-parametric Mann–Whitney U test implemented using scipy.stats package[83] after testing for normality using a D'Agostino and Pearson's test.

Selection analyses. The annotated opsin dataset from Lepidoptera was used for estimating rates of selection for 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, Blue, and LW) to make the analyses comparable between genes and datasets. Because RH7 was only used as a control, we used all taxa for which we recovered RH7. Geneious v. 10.0.9 (https://www.geneious.com) was used to sort and export the sequences. The sequences were manually cleaned in AliView v. 1.18.1[84] and Muscle v. 3.8.31[85] was used for aligning them. Sequences that were too long (<1200 base pairs) were trimmed and sequences shorter than 850 base pairs were removed. Further processing involved manual removal of highly divergent sequences with re-alignment till we obtained no major gaps in the alignment. After a mostly clean alignment was obtained, the sequences were trimmed from either ends to ensure the alignment contained no stop codons or incomplete codons. Final alignments were about 1000 nucleotide bases, roughly 330 amino acids, which falls within the length of characterized opsin sequences[86]. For RH7, the manual method resulted in a very short alignment, less than half the length of the other opsins, so Prank with TranslatorX[87] was used for RH7 even though it had more gaps. IQ-TREE (v. 1.6.12)[78] was used for building the gene trees from the alignments. PAML 4.9a[55] was used to generate various models of codon evolution and estimate site-wise synonymous (α) and non-synonymous (β) 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 from PAML to test for positive selection and Contrast-FEL[50] and MEME models[51] from HyPhy[52] using the Datamonkey[88] server to verify the branch site models.

Custom python scripts were used to filter the data by number of opsins annotated, compile sequences for the filtered species, and prune the species-tree for the analysis in PAML (Supplementary Information). We 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 using different alignment methods to test for sensitivity to these parameters (Supplementary Data 4).

Tree reconciliation. We used Notung (v-2.9.1.5)[89,90] to perform the tree reconciliation analyses· All trees were pruned using custom python scripts that used the package ETE v3[82]. The pruned species tree and gene trees from the mapping analysis were renamed using python scripts to fit the Notung input format requirements (Supplementary Figs. 3–5). Tree rearrangement minimized the duplication or lost cost to obtain an accurate estimate of duplications and losses in the tree[91].

Quality control. We ran two main checks on data quality in order to evaluate potential biases in sampling. First, we used the insect core-ortholog set BUSCO v4.0.6[39] to evaluate overall transcriptome quality, with the percentage of complete orthologs as a proxy. We then reasoned that while some transcriptomes may be high quality, it is possible that the transcriptomes excluded the head of the insect, or that the head may have been very small. In this case, the transcriptome may appear to be high quality (high BUSCO score) but would express few visual genes. To address this case, we used OrthoFinder[92] was used to create a visual gene quality score. Peptide files and a single peptide bait file with known *Manduca sexta* visual genes compiled by a previous study[26] were included in the OrthoFinder run. We removed all '?' characters in the bait file. The clustering algorithm was used to determine orthology to known visual genes. Then, for each sample, we marked presence (1+ gene copies present) or absence (no copies present) of each gene family. The percentage of visual gene families present in each transcriptome (out of a total of 47) serves as a visual gene quality assessment.

Phylogenetic signal and Phylogenetic ANOVA. We used the ape v. 5.3[93], phytools[94] and geiger[95] packages in R. Code was adapted from (https://lukejharmon.github.io/ilhabela/instruction/2015/06/02/ContinuousModels/). To test for phylogenetic signal, we used Blomberg's K from the phylosig module in phytools. We tested the phylogenetic ANOVA with the model "quality metric ~diel-niche", with the correlation set to Brownian.

Ancestral state reconstruction. We used the geiger[95] package in R with the SYM model and 10,000 repetitions, then used ape v. 5.3[93] to plot on the pruned tree. We generated 10,000 stochastic maps for each tree in SIMMAP[96], which is part of the R package phytools[94]. The methods were adapted from Kawahara et al.[34], which also mapped diel state, without prior information on opsin transition probability. SYM model was used 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[97]. 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 (Supplementary Fig. 6). We used UV, Blue, LW, total number of opsins, and diel-niche as the discrete characters for the ancestral state reconstruction.

Protein modeling and site mapping. We obtained predictions of transmembrane helix prediction for all three opsins using Phobius[98] implemented through Protter[99] using the first sequence from the branch-site alignments (Supplementary Fig. 11). Protter is a web-server based tool useful in annotating a protein sequence. It uses structural prediction tools, such as Phobius[98] to predict transmembrane domains, and allows the user to mark a custom set of amino acids. Only the UV opsin sequence recovered all 7 transmembrane domains, similar to the X-ray crystal structure of known invertebrate opsins and GPCR's, whereas both Blue and LW opsin only recovered six transmembrane domains. UV opsin also had consistently significant differences between diel-niche across the selection analyses and therefore, we used it for the 3D protein structure modeling using an online protein modeling tool, Swiss-model[100]. We chose the squid opsin X-ray structure as a template (2z73.A), because it had the highest identity-score and coverage (GMQE,0.73, identity = 33.6). We included retinal in the model from the squid structure and identified putative retinal binding sites, i.e., amino acids less than 4.0 Å from retinal (Fig. 3A, B). 4.0 Å is the length of weak hydrogen bonds (that longest bonds that opsin usually makes with retinal)[101,102].

Statistics and reproducibility. Chi-squared tests were used to assess differences in number of duplication events and total number of duplications for each diel-niche (Table 2) and the duplication events were corrected for Yate's continuity test. Nonparametric Mann–Whitney U tests were used to assess differences in opsin counts (Supplementary Data 1) after testing for normality using a D'Agostino and Pearson's test. ANOVA was used to test if transcriptome quality and visual gene recovery were correlated with diel-niche. Number of runs, support values and input parameters used for the phylogenetic reconstructions, such as the gene trees, ancestral state reconstruction and evolutionary rate estimation are mentioned in the methods and if possible the configuration files and data used at each step for each run are included in Supplementary Data[38]. The protein modeling and associated fits for each model are also included.

References

1. Nilsson, D. E. Eye evolution and its functional basis. *Vis. Neurosci.* **30**, 5–20 (2013).
2. Blevins, E. & Johnsen, S. Spatial vision in the echinoid genus Echinometra. *J. Exp. Biol.* **207**, 4249–4253 (2004).
3. Land, M. F. & Nilsson, D. E. *Animal Eyes*. (Oxford University Press, 2012).
4. Cuthill, I. C. et al. The biology of color. *Science* **357**, eaan0221 (2017).
5. Kelber, A. Colour in the eye of the beholder: receptor sensitivities and neural circuits underlying colour opponency and colour perception. *Curr. Opin. Neurobiol.* **41**, 106–112 (2016).
6. Porter, M. L. et al. Shedding new light on opsin evolution. *Proc. R. Soc. B Biol. Sci.* **279**, 3–14 (2012).
7. Henze, M. J. & Oakley, T. H. The dynamic evolutionary history of pancrustacean eyes and opsins. *Integr. Comp. Biol.* **55**, 830–842 (2015).
8. Colbourne, J. K. et al. The ecoresponsive genome of *Daphnia pulex*. *Science* **331**, 555–561 (2011).
9. Futahashi, R. Molecular mechanisms underlying color vision and color formation in dragonflies. *Divers. Evol. Butterfly Wing Patterns Integr. Approach* **17**, 303–320 (2017).
10. Bybee, S. M., Johnson, K. K., Gering, E. J., Whiting, M. F. & Crandall, K. A. All the better to see you with: a review of odonate color vision with transcriptomic insight into the odonate eye. *Org. Divers. Evol.* **12**, 241–250 (2012).
11. Futahashi, R. et al. Extraordinary diversity of visual opsin genes in dragonflies. *Proc. Natl Acad. Sci. USA* **112**, E1247–E1256 (2015).
12. Porter, M. L. et al. The evolution of complexity in the visual systems of stomatopods: Insights from transcriptomics. *Integr. Comp. Biol.* **53**, 39–49 (2013).
13. Marshall, J. & Arikawa, K. Unconventional colour vision. *Curr. Biol.* **24**, R1150–R1154 (2014).
14. Barlow, H. B. What causes trichromacy? A theoretical analysis using comb-filtered spectra. *Vis. Res.* **22**, 635–643 (1982).
15. Sharkey, C. R. et al. Overcoming the loss of blue sensitivity through opsin duplication in the largest animal group, beetles. *Sci. Rep.* **7**, 1–10 (2017).
16. Hoffmann, M. et al. Opsin gene duplication and diversification in the guppy, a model for sexual selection. *Proc. R. Soc. B Biol. Sci.* https://doi.org/10.1098/rspb.2006.3707. (2007).
17. Everett, A., Tong, X., Briscoe, A. D. & Monteiro, A. Phenotypic plasticity in opsin expression in a butterfly compound eye complements sex role reversal. *BMC Evol. Biol.* **12**, 232 (2012).
18. Dyer, A. G. Discrimination of flower colours in natural settings by the bumblebee species Bombus terrestris (Hymenoptera: Apidae). *Entomol. Gen.* **28**, 257–268 (2006).
19. Tierney, S. M. et al. Consequences of evolutionary transitions in changing photic environments. *Austral Entomol.* **56**, 23–46 (2017).
20. Warrant, E. J. & Johnsen, S. Vision and the light environment. *Curr. Biol.* **23**, R990–R994 (2013).
21. Zhao, H. et al. The evolution of color vision in nocturnal mammals. *Proc. Natl Acad. Sci. USA* **106**, 8980–8985 (2009).
22. Stern, D. B. & Crandall, K. A. Phototransduction gene expression and evolution in cave and surface Crayfishes. *Integr. Comp. Biol.* **58**, 398–410 (2018).
23. Musilova, Z. et al. Vision using multiple distinct rod opsins in deep-sea fishes. *Science* **364**, 588–592 (2019).
24. Tobler, M., Coleman, S. W., Perkins, B. D. & Rosenthal, G. G. Reduced opsin gene expression in a cave-dwelling fish. *Biol. Lett.* **6**, 98–101 (2010).
25. Schweikert, L. E., Fitak, R. R., Caves, E. M., Sutton, T. T. & Johnsen, S. Spectral sensitivity in ray-finned fishes: Diversity, ecology and shared descent. *J. Exp. Biol.* **221**, jeb189761 (2018).

26. Macias-Muñoz, A., Olguin, A. G. R. & Briscoe, A. D. Evolution of phototransduction genes in lepidoptera. *Genome Biol. Evol.* **11**, 2107–2124 (2019).

27. Sander, S. E. & Hall, D. W. Variation in opsin genes correlates with signalling ecology in North American fireflies. *Mol. Ecol.* **24**, 4679–4696 (2015).

28. Briscoe, A. D. Reconstructing the ancestral butterfly eye: focus on the opsins. *J. Exp. Biol.* **211**, 1805–1813 (2008).

29. Arikawa, K. The eyes and vision of butterflies. *J. Physiol.* **595**, 5457–5464 (2017).

30. Stavenga, D. G. On visual pigment templates and the spectral shape of invertebrate rhodopsins and metarhodopsins. *J. Comp. Physiol. A* **196**, 869–878 (2010).

31. Feuda, R., Marlétaz, F., Bentley, M. A. & Holland, P. W. H. Conservation, duplication, and divergence of five opsin genes in insect evolution. *Genome Biol. Evol.* **8**, 579–587 (2016).

32. Xu, P. et al. The evolution and expression of the moth visual opsin family. *PLoS ONE* **8**, e78140 (2013).

33. Diekmann, Y. & Pereira-Leal, J. B. Gene tree affects inference of sites under selection by the branch-site test of positive selection. *Evol. Bioinforma.* **11s2**, S30902 (2015).

34. Kawahara, A. Y. et al. Diel behavior in moths and butterflies: a synthesis of data illuminates the evolution of temporal activity. *Org. Divers. Evol.* **18**, 13–27 (2018).

35. Kawahara, A. Y. et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl Acad. Sci. USA* **116**, 22657–22663 (2019).

36. Challis, R. J., Kumar, S., Dasmahapatra, K. K. K., Jiggins, C. D. & Blaxter, M. Lepbase: the Lepidopteran genome database. Preprint at bioRxiv https://doi.org/10.1101/056994. (2016).

37. Kersey, P. J. et al. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **46**, D802–D808 (2018).

38. Sondhi, Y. et al. Data from: Light environment drives evolution of color vision genes in butterflies and moths, Dryad, Dataset, https://doi.org/10.5061/dryad.gmsbcc2kr (2021).

39. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. In (ed. Kollmar, M.) Gene Prediction. *Methods Mol. Biol.* **1962**, (Humana, New York, NY., 2019). https://doi.org/10.1007/978-1-4939-9173-0_14

40. Speiser, D. I. et al. Using phylogenetically-informed annotation (PIA) to search for light-interacting genes in transcriptomes from non-model organisms. *BMC Bioinforma.* **15**, 350 (2014).

41. Briscoe, A. D. et al. Positive selection of a duplicated UV-sensitive visual pigment coincides with wing pigment evolution in *Heliconius* butterflies. *Proc. Natl Acad. Sci. USA* **107**, 3628–3633 (2010).

42. Andersson, S., Rydell, J. & Svensson, M. G. E. Light, predation and the lekking behaviour of the ghost swift *Hepialus humuli* (L.) (Lepidoptera, Hepialidae). *Proc. R. Soc. Lond. Ser. B* **265**, 1345–1351 (1998).

43. Yin, C. et al. ChiloDB: a genomic and transcriptome database for an important rice insect pest *Chilo suppressalis*. *Database* **2014**, bau065–bau065 (2014).

44. Pirih, P. et al. The giant butterfly-moth *Paysandisia archon* has spectrally rich apposition eyes with unique light-dependent photoreceptor dynamics. *J. Comp. Physiol. A* **204**, 639–651 (2018).

45. Kinoshita, M. & Arikawa, K. Color and polarization vision in foraging *Papilio*. *J. Comp. Physiol. A* **200**, 513–526 (2014).

46. Chen, P. J., Awata, H., Matsushita, A., Yang, E.-C. & Arikawa, K. Extreme spectral richness in the eye of the common bluebottle butterfly, *Graphium sarpedon*. *Front. Ecol. Evol.* **4**, 18 (2016).

47. Langer, H., Hamann, B. & Meinecke, C. Tetrachromatic visual system in the Moth *Spodoptera exempta* (Lepidoptera). *J. Comp. Physiol. A* **129**, 235–239 (1979).

48. Fu, X., Feng, H., Liu, Z. & Wu, K. Trans-regional migration of the beet armyworm, *Spodoptera exigua* (Lepidoptera: Noctuidae), in North-East Asia. *PLoS ONE* **12**, e0183582 (2017).

49. Gutierrez, E. et al. The role of ecological factors in shaping bat cone opsin evolution. *Proc. R. Soc. B Biol. Sci.* **285**, 20172835 (2018).

50. Murrell, B. et al. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, e1002764 (2012).

51. Kosakovsky Pond, S. L. & Frost, S. D. W. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol.* **22**, 1208–1222 (2005).

52. Kosakovsky Pond, S. L. et al. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).

53. Böhm, A., Meusemann, K., Misof, B. & Pass, G. Hypothesis on monochromatic vision in scorpionflies questioned by new transcriptomic data. *Sci. Rep.* **8**, 1–8 (2018).

54. Lord, N. P. et al. A cure for the blues: Opsin duplication and subfunctionalization for short-wavelength sensitivity in jewel beetles (Coleoptera: Buprestidae). *BMC Evol. Biol.* **16**, 107 (2016).

55. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

56. Wagner, G. P., Kin, K. & Lynch, V. J. A model based criterion for gene expression calls using RNA-seq data. *Theory Biosci.* **132**, 159–164 (2013).

57. Yack, J. E. & Fullard, J. H. Ultrasonic hearing in nocturnal butterflies. *Nature* **403**, 265–266 (2000).

58. Scoble, M. *The Lepidoptera. Form, Function and Diversity*. (Oxford University Press, 1992).

59. Dudley, R. Biomechanics of flight in neotropical butterflies: morphometrics and kinematics. *J. Exp. Biol.* **150**, 37–53 (1990).

60. Kunte, K. Female-limited mimetic polymorphism: a review of theories and a critique of sexual selection as balancing selection. *Anim. Behav.* **78**, 1029–1036 (2009).

61. Nilsson, D. E., Land, M. F. & Howard, J. Optics of the butterfly eye. *J. Comp. Physiol. A* **162**, 341–366 (1988).

62. Cronin, T. W., Johnsen, S., Marshall, N. J. & Warrant, E. J. *Visual Ecology*. (Princeton University Press, 2014).

63. Henze, M. J., Lind, O., Mappes, J., Rojas, B. & Kelber, A. An aposematic colour-polymorphic moth seen through the eyes of conspecifics and predators – Sensitivity and colour discrimination in a tiger moth. *Funct. Ecol.* **32**, 1797–1809 (2018).

64. Kelber, A. Innate preferences for flower features in the hawkmoth *Macroglossum stellatarum*. *J. Exp. Biol.* **200**, 827–836 (1997).

65. Yurtsever, S., Okyar, Z. & Guler, N. What colour of flowers do Lepidoptera prefer for foraging? *Biology* **65**, 1049–1056 (2010).

66. White, R. H. The retina of *Manduca sexta*: rhodopsin expression, the mosaic of green-, blue- and UV-sensitive photoreceptors, and regional specialization. *J. Exp. Biol.* **206**, 3337–3348 (2003).

67. Jacobs, G. H., Fenwick, J. A. & Williams, G. A. Cone-based vision of rats for ultraviolet and visible lights. *J. Exp. Biol.* **204**, 2439–2446 (2001).

68. Höglund, J. et al. Owls lack UV-sensitive cone opsin and red oil droplets, but see UV light at night: retinal transcriptomes and ocular media transmittance. *Vis. Res.* **158**, 109–119 (2019).

69. Hirota, S. K., Miki, N., Yasumoto, A. A. & Yahara, T. UV bullseye contrast of *Hemerocallis* flowers attracts hawkmoths but not swallowtail butterflies. *Ecol. Evol.* **9**, 52–64 (2019).

70. Robinson, H. S. & Robinson, P. J. Some notes on the observed behaviour of Lepidoptera ln flight in the vicinity of light sources. *Entomol. Gaz.* **1**, 3–15 (1950).

71. Lamarre, G. P. A. et al. Stay out (almost) all night: contrasting responses in flight activity among tropical moth assemblages. *Neotrop. Entomol.* **44**, 109–115 (2015).

72. van Langevelde, F., Ettema, J. A., Donners, M., WallisDeVries, M. F. & Groenendijk, D. Effect of spectral composition of artificial light on the attraction of moths. *Biol. Conserv.* **144**, 2274–2281 (2011).

73. Johnsen, S. et al. Crepuscular and nocturnal illumination and its effects on color perception by the nocturnal hawkmoth *Deilephila elpenor*. *J. Exp. Biol.* **209**, 789–800 (2006).

74. Kelber, A. Ovipositing butterflies use a red receptor to see green. *J. Exp. Biol.* **202**, 2619–2630 (1999).

75. Ehrlich, P. R., Raven, P. H. & Butterflies and plants: a study in coevolution. *Evolution* **18**, 586–608 (1964).

76. Holm, S. et al. No Indication of High Host-Plant Specificity in Afrotropical Geometrid Moths. *J. Insect Sci.* **19**, 1 (2019).

77. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

78. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–W235 (2016).

79. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

80. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

81. Broadhead, G. T., Basu, T., von Arx, M. & Raguso, R. A. Diel rhythms and sex differences in the locomotor activity of hawkmoths. *J. Exp. Biol.* **220**, 1472–1480 (2017).

82. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

83. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* https://doi.org/10.1038/s41592-019-0686-2. (2020).

84. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).

85. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

86. White, R. H., Xu, H., Münch, T. A., Bennett, R. R. & Grable, E. A. The retina of *Manduca sexta*: rhodopsin expression, the mosaic of green-, blue- and UV-sensitive photoreceptors, and regional specialization. *J. Exp. Biol.* **206**, 3337–3348 (2003).

87. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* **38**, W7–W13 (2010).

88. Weaver, S. et al. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. *Mol. Biol. Evol.* **35**, 773–777 (2018).

89. Chen, K., Durand, D. & Farach-Colton, M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* **7**, 429–447 (2000).

90. Stolzer, M. et al. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics* **28**, i409–i415 (2012).

91. Durand, D., Halldórsson, B. V. & Vernot, B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* **13**, 320–335 (2006).

92. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

93. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

94. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

95. Pennell, M. W. et al. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinformatics* **30**, 2216–2218 (2014).

96. Bollback, J. P. SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC Bioinforma.* **7**, 88 (2006).

97. Revell, L. J. Two new graphical methods for mapping trait evolution on phylogenies. *Methods Ecol. Evol.* **4**, 754–759 (2013).

98. Kall, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432 (2007).

99. Omasits, U., Ahrens, C. H., Müller, S. & Wollscheid, B. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics* **30**, 884–886 (2014).

100. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).

101. Jeffrey, G. A. *An Introduction to Hydrogen Bonding*. (Oxford University Press, 1997).

102. Sekharan, S. & Morokuma, K. Why 11- cis -retinal? why not 7- cis -, 9- cis -, or 13- cis- retinal in the eye? *J. Am. Chem. Soc.* **133**, 19052–19055 (2011).

Author contributions

Y.S.: conceptualization, methodology, software, visualization, writing–original draft, data curation, formal analysis, project administration. E.E.: conceptualization, methodology, software, data curation, resources, writing–review and editing. S.B.: methodology, validation, writing–review and editing. J.T.: writing–review and editing, visualization, validation, conceptualization. A.K.: conceptualization, methodology, resources, writing–review and editing, visualization, supervision, funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-01688-z.

Correspondence and requests for materials should be addressed to Y.S.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.