Phylogenomics of Opsin Genes in Diptera Reveals Lineage-Specific Events and Contrasting Evolutionary Dynamics in *Anopheles* and *Drosophila*

Roberto Feuda [1,2,*], Matthew Goulty[1], Nicola Zadra[3,4], Tiziana Gasparetti[4], Ezio Rosato[1], Davide Pisani[5], Annapaola Rizzoli[3], Nicola Segata[4], Lino Ometto [6], and Omar Rota Stabelli[3,7,*]

[1]Department of Genetics and Genome Biology, University of Leicester, UK

[2]Department of Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn, Naples, Italy

[3]Research and Innovation Centre, Fondazione Edmund Mach (FEM), San Michele all'Adige, Italy

[4]Department CIBIO, University of Trento, Italy

[5]School of Earth Sciences, University of Bristol, UK

[6]Department of Biology and Biotechnology, University of Pavia, Italy

[7]Center Agriculture Food Environment (C3A), University of Trento, Italy

*Corresponding authors: E-mails: rf190@leicester.ac.uk; omar.rotastabelli@unitn.it.

Abstract

Diptera is one of the biggest insect orders and displays a large diversity of visual adaptations. Similarly to other animals, the dipteran visual process is mediated by opsin genes. Although the diversity and function of these genes are well studied in key model species, a comprehensive comparative genomic study across the dipteran phylogeny is missing. Here we mined the genomes of 61 dipteran species, reconstructed the evolutionary affinities of 528 opsin genes, and determined the selective pressure acting in different species. We found that opsins underwent several lineage-specific events, including an independent expansion of Long Wave Sensitive opsins in flies and mosquitoes, and numerous family-specific duplications and losses. Both the *Drosophila* and the *Anopheles* complement are derived in comparison with the ancestral dipteran state. Molecular evolutionary studies suggest that gene turnover rate, overall mutation rate, and site-specific selective pressure are higher in *Anopheles* than in *Drosophila*. Overall, our findings indicate an extremely variable pattern of opsin evolution in dipterans, showcasing how two similarly aged radiations, *Anopheles* and *Drosophila*, are characterized by contrasting dynamics in the evolution of this gene family. These results provide a foundation for future studies on the dipteran visual system.

Key words: Diptera, evolution, opsin, flies, mosquitoes.

Significance

Diptera is an insect order including flies, mosquitoes, and various other species of economic importance. Their vision is mediated by the opsin genes, which have been studied in a few key model species. However, a comprehensive comparative genomic analysis does not exist, impairing our understanding of the evolutionary history of these genes in this order. In this work, we perform the first genome-scale analysis of opsin gene evolution in Diptera. We investigate their pattern of duplication, selection, and expression in more than 60 species that belong to 10 different families. Our results clarify the evolution of the opsin genes in dipterans, in particular in fruit flies and mosquitoes, and represent the foundation for functional studies on their visual system.

Introduction

The ability to detect and respond to specific visual stimuli and light conditions is fundamental in defining animal biology and ecology, including mating, and predatory and foraging behavior (Tierney et al. 2012, 2015; Futahashi et al. 2015; Feuda et al. 2016; van der Kooi et al. 2021). In all animals, visual processing is mediated by opsins, a group of photosensitive G-protein coupled receptors, which originated in prebilateria metazoans by an ancient duplication from nonlight-sensitive receptors; subsequent duplications generated C-opsins, R-opsins, and Go-opsins (Feuda et al. 2012; Ramirez et al. 2016). Opsins are generally expressed in photoreceptor cells, where they mediate light sensing (Fain et al. 2010). The modification of opsin complement (such as gene duplication or loss) and/or specific functional amino acid mutations in opsin genes can confer the ability to adapt to new ecological niches, for example by providing the ability to respond to different wavelengths of light (Feuda et al. 2016; Tierney et al. 2012, 2015; Sondhi et al. 2020; see van der Kooi et al. 2021 for a recent review). However, increasing evidence indicates that, at least in the model organism *Drosophila melanogaster*, the function of the opsins is not restricted to photoreceptor cells but extends to different sensory modalities, such as mechanosensation (Zanini et al. 2018), taste (Leung et al. 2020), temperature sensation (Sokabe et al. 2016), and circadian clocks (Ni et al. 2017).

Diptera is an insect order containing more than 125,000 species (Skevington and Dang 2002), representing approximately 10% of animal diversity. This order comprises *Drosophila* and several species of economic importance, such as agricultural pests (e.g., fruit flies of genera *Bactrocera* and *Ceratitis*, as well as *Drosophila suzukii*) and vectors of infectious disease (e.g., *Glossina* tsetse flies and mosquitoes of the *Aedes*, *Anopheles*, and *Culex* genera; White and Elson-Harris 1992; Attardo et al. 2014, 2019; Neafsey et al. 2015; Rota-Stabelli et al. 2020; Zadra et al. 2021). Dipterans are characterized by a great variety of morphological, physiological, and ecological behaviors resulting from a rapid radiation (Wiegmann et al. 2011). Dipterans are also characterized by a large variation in sensitivity to light (van der Kooi et al. 2021). Even within the same genus, it is possible to observe diurnal, nocturnal, and crepuscular species (supplementary table S1, Supplementary Material online and references therein).

In Diptera, and insects in general, the visual process is mediated by R-opsins which are classified according to the wavelength at which they show maximum absorbance: Long-Wavelength Sensitive opsins (LWS, sometimes known as LW) can respond to green light, Short Wavelength Sensitive opsins (SWS, sometimes known as SW) to blue light, UV opsins to ultraviolet light, and *Rh7* opsins to a broad spectrum of light (Briscoe and Chittka 2001; Henze and Oakley 2015; Feuda et al. 2016; Ni et al. 2017; Sakai et al. 2017; Fleming et

al. 2018; van der Kooi et al. 2021). In *D. melanogaster*, the opsins are well characterized and seven genes/proteins have been identified: *Rh5* respond to blue light (SWS), *Rh1*, *Rh2*, and *Rh6* to green (LWS) light and *Rh3* and *Rh4* to UV light (Carulli et al. 1994; Bao and Friedrich 2009; Sakai et al. 2017). The absorbance of *Rh7* is particularly broad with a maximum in the UV light, but with a long tail encompassing the blue and cyan wavelengths (Sakai et al. 2017). *Rh7* is expressed in a limited number of neurons in the central brain, including some of those responsible for circadian activity (Ni et al. 2017; Ma et al. 2021). *Rh2* is expressed in the ocelli (Pollock and Benzer 1988) and possibly in the R7 photoreceptor cells. All other opsins are expressed mainly in retinal photoreceptor cells which in different combinations are used to define the visual competence of different photoreceptor subtypes (Courgeon and Desplan 2019). However, it is not known if the opsin repertoire is conserved throughout the genus *Drosophila*.

The opsin complement has been characterized in some other dipterans, such as in various *Glossina* species (Attardo et al. 2019), in *Lucilia cuprina* (Anstead et al. 2016), and in *Calliphora vicina* (Schmitt et al. 2005) where a similar opsin complement has been identified (*Rh1*, *Rh2*, *Rh3*, *Rh5*, and *Rh6*). A recent analysis performed by Giraldo-Calderón et al. (2017) on three species of Culicidae, i.e. *Aedes aegypti*, *Culex quinquefasciatus*, and *Anopheles gambiae* identified a series of duplication events affecting LWS-*Rh6* in this clade. However, whether these duplications are shared with other Culicidae remain unclear. Furthermore, other opsin genes such as arthropsins (belonging to the R-opsins), C opsins, and RGR/GO opsins have been identified in some insect groups (Futahashi et al. 2015; Fleming et al. 2018; Almudi et al. 2020). However, their function, presence, and potential distribution in dipterans are ambiguous (Velarde et al. 2005).

Despite the key role played by opsins in sensory biology, we lack a systematic understanding of their evolution along the dipteran phylogeny. How many opsins were present in the last dipteran common ancestor? Do the opsins in the different groups undergo similar evolutionary patterns? A rigorous comparison of opsin content in model genus *Drosophila* and *Anopheles* has also never been undertaken, leaving open such questions as whether the opsin repertoire is conserved throughout the genus and if selective forces are acting differently in different species. To address these questions, we investigated the evolution of opsin genes in 61 dipteran species sampled from ten different families and reconstructed their pattern of gene duplication and loss. We focused on the two iconic genera, *Drosophila* and *Anopheles*, and investigated the expression and occurrence of positive selection acting on the different opsin genes. Overall, our comparative genomics investigation provides an updated overview on the pattern of duplication and loss, as well as evidence for lineage-specific evolutionary histories of opsin genes in
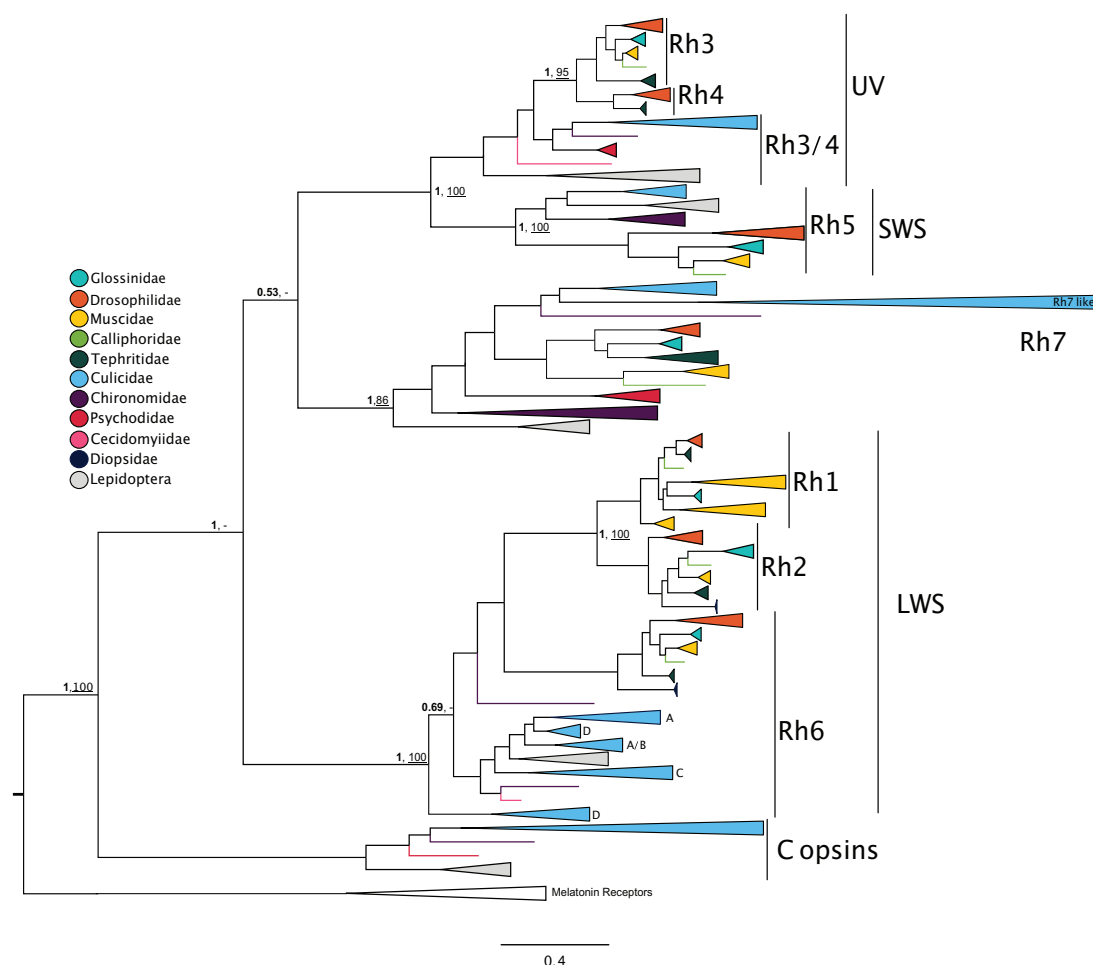
Fɪɢ. 1.—The phylogenomics of opsins in Diptera. This figure shows the phylogenetic relationships between 528 opsins from 61 dipteran species. The Bayesian PP (in bold) and the UltraFast bootstrap (underlined) support is shown for key nodes. The colors indicate the different Diptera subgroups analyzed in this work. Letters A, B, C, and D indicate the different paralogs of *Rh6* we identified in Culicidae. The full trees can be found in supplementary figures S1 and S2, Supplementary Material online.

Diptera, and provides a foundation for future functional studies on the dipteran visual system.

Results and Discussion

Dipterans Have At Least Eight Paralogous Opsin Groups

We based our analyses on the genomic data of 61 species belonging to ten different families of dipterans (supplementary table S1, Supplementary Material online). Overall, the proteome completeness values estimated by BUSCO ranged from 68.5% in *Phlebotomus papatasi* to 99.93% in *D. melanogaster* (supplementary table S1, Supplementary Material online). To identify the opsin genes, we used a combination of BLAST, motif search, and manual curation to minimize the possibility of false negatives (see Materials and Methods and Feuda et al. 2016).

We identified a total of 528 opsins across the 61 species (see supplementary table S1, Supplementary Material online

for the distribution of opsins in each species). We reconstructed their evolutionary affinities by Maximum Likelihood and Bayesian Inference (fig. 1, supplementary figs S1 and S2, Supplementary Material online) using the amino acid GTR-G model, which has previously been shown to fit opsin alignments better than other models (Feuda et al. 2012; Vöcking et al. 2017). Both approaches revealed that there are at least eight paralogous opsin groups in Diptera, which we named based on the *Drosophila* nomenclature as *Rh*1-7 and C. We also found that arthropsins and RGR/GO opsins have been lost in Diptera. Furthermore, both methods recover a similar topology, whereby LWS opsins (*Rh1*, *Rh2*, and *Rh6*) are monophyletic (Posterior Probability [PP = 1] and BS = 100) and *Rh5* is the sister group to *Rh3* plus *Rh4* (PP = 1 and BS = 100). The position of *Rh7* is uncertain, with Maximum Likelihood favoring their position as sister group to all the remaining R-opsins (BS = 92, supplementary fig. S1, Supplementary Material online). In summary, our phylogenetic analysis recovered the
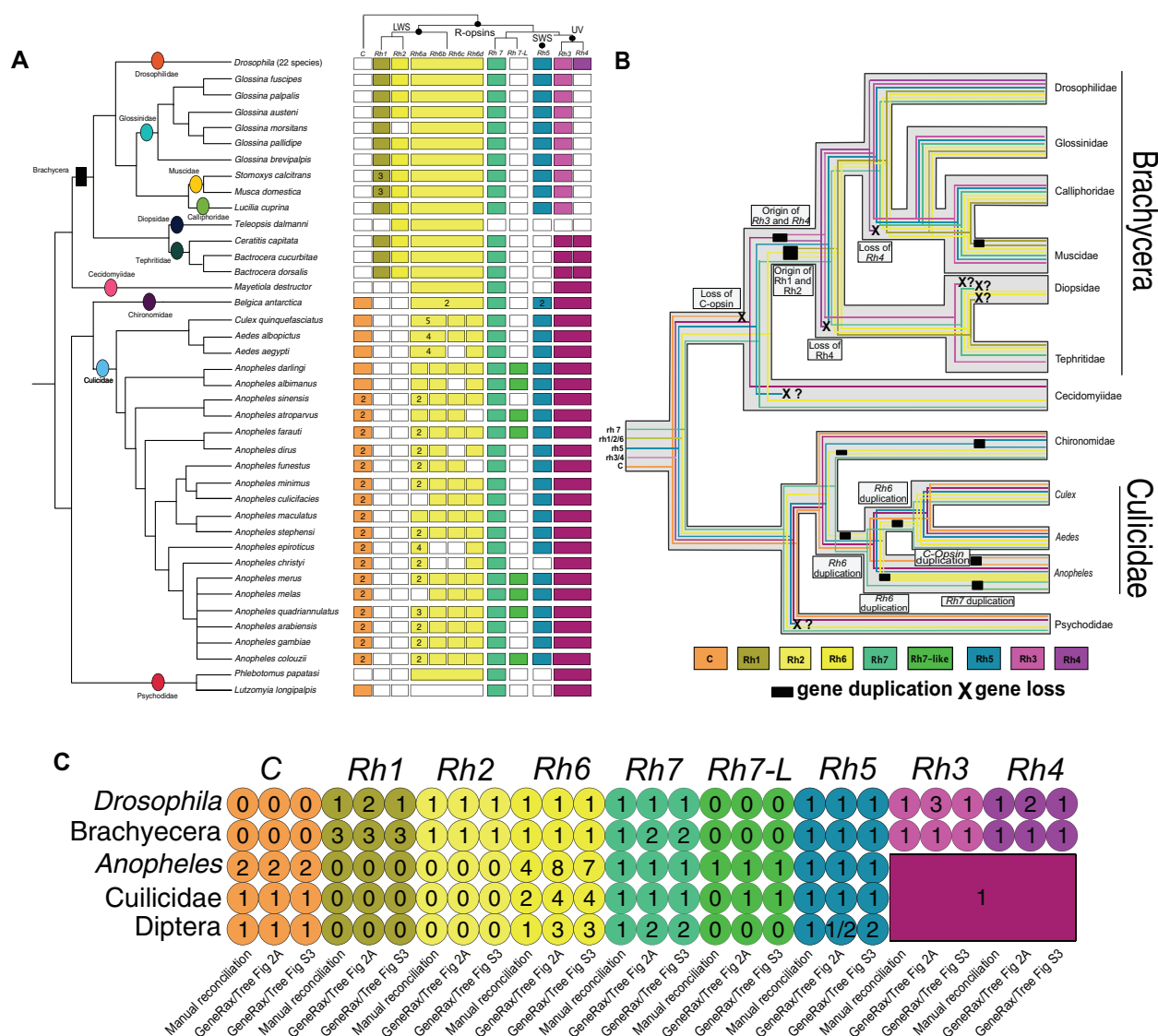
Fig. 2.—Opsins evolution in Diptera. (A) Opsin gene complements in Diptera. The phylogenetic tree was obtained from Wiegmann et al. (2011). Gene nomenclature has been obtained from *Drosophila melanogaster*. The numbers in the boxes indicate the copies of opsin genes identified; white boxes indicate that genes have not been found. (B) Synopsis of the patterns of opsin duplications and losses in Diptera subgroups. Lineage-specific events are marked with a question mark if they were inferred from one single representative genome. (C) Estimated number of ancestral Rh across five nodes. For each opsin paralog, we report the estimate using three different analytical procedures (manual reconciliation, GeneRax on tree of fig. 2A, GeneRax on tree of supplementary fig. S3, Supplementary Material online).

monophyly of all the main opsin groups (e.g., LWS, UV) with a high support value, which suggests the presence of eight opsin groups in dipterans.

Dipteran Opsins Have Undergone Lineage-Specific Diversification

To better understand the opsin distribution and evolutionary dynamics in the various dipteran groups, we mapped their presence/absence on the Diptera phylogeny (fig. 2A) and performed a manual as well as a statistical gene-tree/species-tree

reconciliation (fig. 2B and C). The results indicate that the opsin repertoire underwent significant rearrangements on the dipteran phylogeny in a lineage-specific manner.

In Brachycera (the clade comprising *Drosophila*), the opsins complement is derived in comparison to the ancestral dipteran condition. We confirm previous findings that c-opsin and RGR/Go have been lost in all Brachycera (Feuda et al. 2016) and provide evidence that four paralogs—*Rh1*, *Rh2*, *Rh3*, and *Rh4*—are present only in this group. The observation that at least one duplication from the ancestral *Rh1/2/6* and *Rh3/4* genes is shared between *Drosophila*, tephritid fruit flies,

Muscidae house flies, and *Glossina* tze-tze flies indicates that these duplications happened early in Brachycera evolution (fig. 2*B*). We further observe various lineage-specific events, such as the loss of *Rh4* in the common ancestor of Glossinidae, Muscidae, and Calliphoridae, duplications of *Rh1* in Muscidae, the loss of *Rh2* in the tsetse fly *Glossina morsitans* (Attardo et al. 2019), and the loss of all opsins except for *Rh2* and *Rh6* in Diopside stalk-eyed flies. Interestingly, when we map introns' presence/absence (supplementary table S2, Supplementary Material online) in the different opsins, the results indicate that *Rh3* genes in all *Drosophila* species are intronless, suggesting their possible origin as retrotransposons (Booth and Holland 2004; Xu et al. 2016).

In the family Culicidae (mosquitoes, e.g., *Culex*, *Anopheles*, and *Aedes*), opsins' repertoire is markedly different from that observed in the Brachycera clade (fig. 2). For example, eight out of the 19 *Anopheles* species have a divergent copy of the *Rh7* gene (figs 1 and 2*A*), whose phylogenetic distribution suggests that it was present in the ancestral *Anopheles* and secondarily lost in some species. The most remarkable difference we observed between Brachycera and Culicidae is the impressive series of duplications of the *Rh6* gene in the latter, where it ranges from three copies in *Anopheles melas* and *Anopheles christyi* to seven in *C. quinquefasciatus*. We identified four *Rh6* paralogs according to their relatedness (fig. 1) and distribution across the dipteran phylogeny (fig. 2*B*), which we named *Rh6a*, *b*, *c*, and *d*. These duplications have already been identified in three Culicidae species (Giraldo-Calderón et al. 2017), but our data indicate that this pattern is present in all the sampled Culicidae species. This pattern of presence/absence suggests that at least two concomitant duplications of *Rh6* happened in the Culicidae common ancestor, followed by additional lineage-specific duplications. To account for some *Anopheles* genomes characterized by low coverage genomes (supplementary table S1, Supplementary Material online), we further performed a manual search of all the missing genes to exclude the possibility of false negatives (see Materials and Methods). Despite this careful manual curation, we could not untangle the precise evolutionary relationships of these duplications within *Anopheles*, because some species lack well-assembled and high-quality genomes (see supplementary table S1, Supplementary Material online). However, we found that, similar to *Rh3* in *Drosophila*, multiple *Rh6* paralogs lack introns (supplementary table S2, Supplementary Material online), suggesting that these newly evolved genes may have originated from a retrotransposition event (Booth and Holland 2004; Xu et al. 2016).

Overall, our findings indicate that the opsin complement in the Brachycera and Culicidae is quite derived in comparison to ancestral dipteran, with the two groups having independently duplicated the LWS opsins. To further identify the ancestral opsin complements in key nodes (i.e. dipterans, Brachyecera,

Drosophila, Culicidae, and *Anopheles*), we performed a manual reconciliation as well as a gene tree-species tree reconciliation using the species tree obtained from the BUSCO single-copy orthologs and GeneRax (Morel et al. 2020) (see Materials and Methods for more details). The resulting ML tree recovers the traditional topology for dipterans, except for the position of Psychodidae (supplementary fig. S3, Supplementary Material online). This tree, and a modified topology matching traditional dipteran relationships (Wiegmann et al. 2011), was used to reconcile the opsin phylogeny. In general, the computational reconciliations identified a similar pattern of duplications compared with the manual reconciliation (fig. 2*C*, supplementary figs S4 and S5, Supplementary Material online). Most of the differences concern *Rh6* in *Anopheles*, where GeneRax identified a large number of *Rh6* copies. We think that this incongruence can be explained by the taxonomic levels used to perform the reconciliation (species vs. genus) and the limited performance of GeneRax in dealing with incomplete lineage sorting (Morel et al. 2020), a phenomenon that is known to have shaped the mosquitos' evolutionary history (Wen et al. 2016). We further tested the possible misleading effect of incomplete genomes by repeating the gene-tree species tree reconciliation after removal of all the species with a BUSCO Value <85%. The results (supplementary fig. S6, Supplementary Material online) suggest that for the key nodes of figure 2*C*, there are no differences between the two data sets.

The Evolution of Opsin Genes in *Drosophila*, *Aedes*, and *Anopheles* Species

Our reconciliation analyses indicate that starting from a repertoire of five (or nine according to GeneRax) opsin genes, the complement substantially diverged in the Brachycera clade compared with the Culicidae family, with several lineage-specific events (fig. 2*B*). The question arises as to whether these newly duplicated genes are expressed in photoreceptor cells and are associated with divergence and specialization of the visual system. In *D. melanogaster*, there is ample evidence that all opsin genes, including the newly duplicated intronless *Rh3*, are expressed and functional in photoreceptor cells and combinatorially define the different visual neural circuits (Courgeon and Desplan 2019).

However, the expression of opsin genes in other cells of the visual system remains poorly understood. We then investigated the pattern of opsin expression in other cell types of the *Drosophila*'s optic lobe by mining single-cell RNA-seq data previously obtained from Davis et al. (2020). Our data indicate that opsins expression is not restricted to the photoreceptor cells and that they contribute to different aspects of the visual neural circuits. For example, the *Rh7* mRNA is detected in the lamina neurons L1-2 (supplementary fig. S7*A*, Supplementary Material online) that regulate motion. Furthermore, the function of the newly duplicated opsin genes in *D. melanogaster* may not be restricted to the visual process: for instance, it has been recently
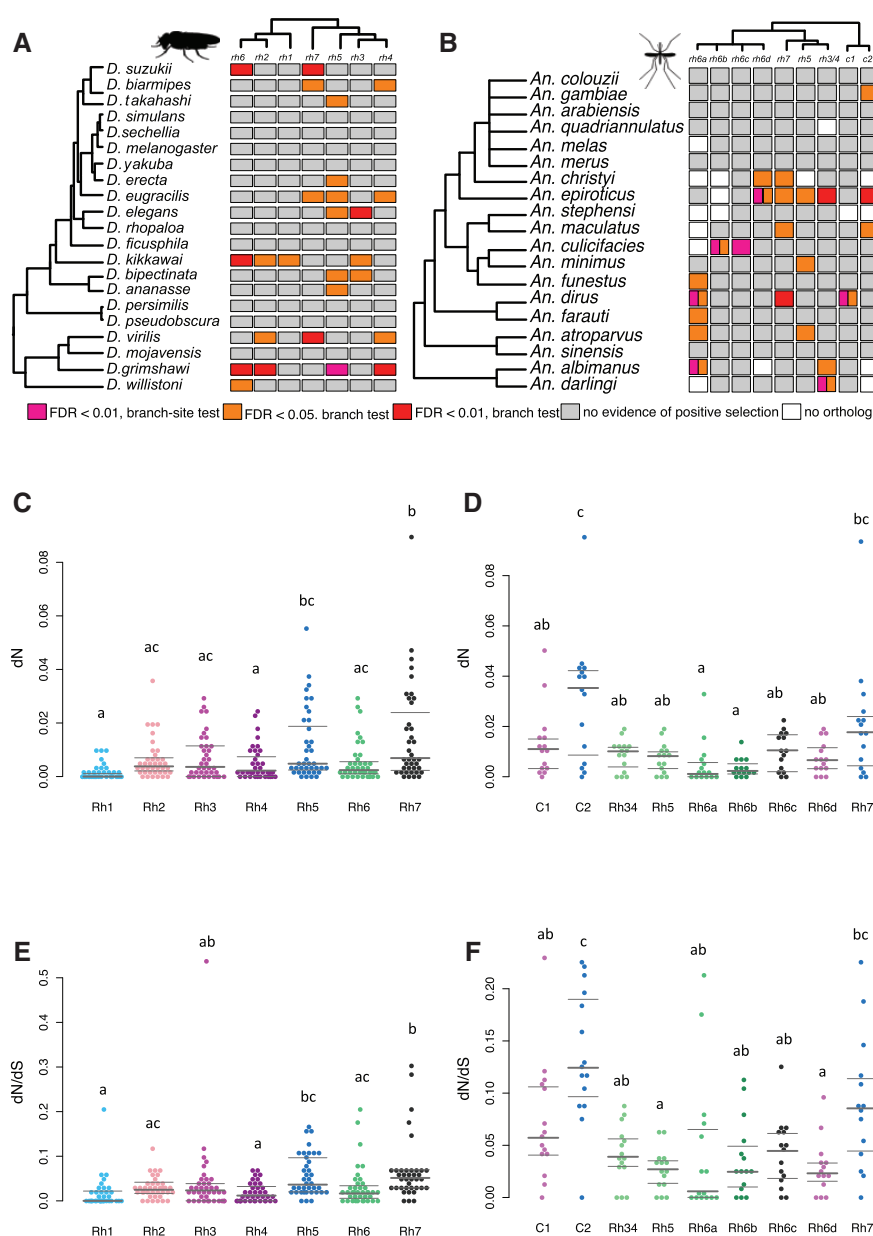
FIG. 3.—Pattern of positive selection and molecular evolution of the opsin genes in *Drosophila* and *Anopheles*. Genes under selection according to the branch or branch-site tests in one species of *Drosophila* and *Anopheles* are shown in (*A*) and (*B*), respectively. We also report the rate of protein evolution (d_N) and the level of selective pressure (d_N/d_S) across opsin phylogenies in *Drosophila* (*C* and *E*) and *Anopheles* (*D* and *F*). Genes are strong determinants of the variance in d_N and d_N/d_S values in both *Drosophila* (ANOVA, $F[6,266] = 7.43$, $P < 10^{-6}$; and $F[6,259] = 4.88$, $P = 0.0001$, respectively) and *Anopheles* (ANOVA, $F[8,126] = 6.37$, $P < 10^{-6}$; and $F[8,124] = 8.60$, $P < 10^{-8}$, respectively). Different letters identify significant statistical differences between genes at adjusted $P < 0.05$, according to a Tukey's HSD (honestly significant difference) multiple comparison test. Median and quantiles are shown as grey lines for each gene. These analyses were performed including parts of the alignments where one or more sequences contained a gap. FDR, false discovery rate. Detailed information for each gene is in supplementary table S4, Supplementary Material online.

been proposed that *Rh1*, *Rh4*, and *Rh7* are involved in taste (Leung and Montell 2017; Leung et al. 2020), suggesting a co-option of visual genes in different sensory pathways. In mosquitos, the information on opsin gene expression is scant. However, it is interesting to note that the R7 photoreceptor of *A. aegypti* may express, depending on their actual position

in the retina, the LWS (*Rh6a-AAop2* or *Rh7-Aaop10*), the SWS *Aaop9* (*Rh5*), and the UV-(*Rh3-Aaop8*) opsins (Hu et al. 2014; Rocha et al. 2015). We further investigated the expression of opsin genes in *A. gambiae* and *A. aegypti* by analyzing available microarray data sets (Baker et al. 2011; Leming et al. 2014). The results indicate that *Rh6a*, *Rh3/4*, and *Rh5* are statistically over-

expressed in the head of *A. gambiae* (supplementary table S3, Supplementary Material online), whereas all nine opsins we identified in *A. aegypti* are expressed in the head (see supplementary fig. S7B, Supplementary Material online). Although this expression data is not eye-specific, these results suggest that these opsins are potentially expressed and contribute to the mosquitos' visual system. We advocate that more specific gene expression studies (focused on the eye rather than on the whole head) are necessary to determine with confidence whether these genes are actually being expressed in the eye and if they have a role in color vision.

Opsins in *Drosophila* and *Anopheles* Underwent Substantial Divergent Molecular Evolution

Our results indicate that the opsin complement is quite divergent across the various dipteran families. We then asked if the pattern of opsin evolution also differs within different genera. To maximize the power of our analyses and inferences, we focused on two genera for which we had around 20 genomes each: *Drosophila* and *Anopheles*. Interestingly, whereas all *Drosophila* species have exactly the same opsin complement, indicating a frozen repertoire over circa 60 Myr, the similarly aged *Anopheles* genus is characterized by an extremely plastic opsin repertoire that includes lineage-specific duplications of *Rh7* (*Rh7*-like) and C opsin, and various instances of duplications and losses of *Rh6* (fig. 2 and supplementary table S1, Supplementary Material online).

To clarify the pattern of selection acting on the opsin genes in these two dipteran genera, we produced manually curated opsin alignments for each paralog group and estimated the selective pressure using PAML (see Materials and Methods for details). Importantly, the cross-group comparison is possible because both these two genera have a similar evolutionary history: both emerged in the Paleogene (between 100 and 30 Mya according to Neafsey et al. 2015; Obbard et al. 2012) and have a similar number of generations per year (up to 10). We found that the differences between *Drosophila* and *Anopheles* are not restricted to the opsin repertoire alone but extend to the pattern of molecular evolution of the opsin genes. While these two groups show an unusual signature of selection for a similar number of genes (26 and 23 respectively, in color in fig. 3A and B), in *Anopheles*, we observe more events of site-specific positive selection (1 in *Drosophila* and 7 in *Anopheles*, magenta squares in fig. 3A and B). A second difference concerns the rate of amino acid evolution. Opsin genes are subject to different molecular constraints in the two groups, as supported by a slightly lower selective pressure in *Anopheles* than in *Drosophila* (fig. 3E and F; overall $d_N/d_S = 0.0573$ and $d_N/d_S = 0.0374$, respectively). This is because while the two clades are characterized by a similar rate of synonymous nucleotide substitution (on average $d_S = 0.2012$ and $d_S = 0.1969$, respectively; data not shown), the two are characterized by different nonsynonymous rates (fig.

3C and D; on average $d_N = 0.0118$ in *Anopheles* and $d_N = 0.0073$ in *Drosophila*).

Our molecular evolution results indicate a much higher variability in selective pressure across opsin genes in *Anopheles* than in *Drosophila*. These different evolutionary patterns are independent of data treatment: when regions with gaps are removed from the alignments (supplementary figs S8–S10, Supplementary Material online), we observe lower substitution rates in *Anopheles* (because the orthologs in this genus are less constrained and accumulated more indels), but trends are consistent. Overall, our results indicate that in the genus *Anopheles* the opsin genes experienced a different evolutionary path and were subject to an accelerated rate of evolution compared with the *Drosophila* species. This is consistent and complementary with the more dynamic pattern of gene deletions/duplications we identified in *Anopheles*. Whereas almost all *Drosophila* species are diurnal, *Anopheles* can be both nocturnal and/or crepuscular (supplementary table S1, Supplementary Material online), suggesting that their extremely flexible opsin repertoire is playing an active role in their adaptation to different lifestyles. Importantly, our results do not allow us to determine whether the differences in the selective pressures are indicative of actual selective forces happening in the visual system (e.g., spectral tuning) or in the other sensory modalities.

Conclusion and Future Perspectives

Here we have characterized the evolutionary history of the opsin genes in ten dipteran families, focusing on the fine-scaled molecular evolution of model organisms *Drosophila* and *Anopheles*. Overall, we found that different dipterans underwent distinct patterns of deletions/duplications (figs 1 and 2) and positive selection (fig. 3). One of the key findings is the derived complement (*Rh1*, *Rh2*, *Rh3*, and *Rh4*) of the Brachycera species, including the model organism *Drosophila*. These genes' recent evolutionary origin suggests that the nonvisual opsin function in *Drosophila* (Leung and Montell 2017; Leung et al. 2020) probably represents a lineage-specific co-option event (Pisani et al. 2020) and implies that *Drosophila*'s opsins cannot be used to infer the ancestral function of these genes (Leung et al. 2020). Our data indicate that the opsin complement is even more dynamic in mosquitos, particularly concerning *Rh6* and *Rh7*. Moreover, our analyses revealed that the *Anopheles* lineages had experienced more instances of site-specific selective pressure and faster evolutionary rates than the *Drosophila* lineages (fig. 3).

In the absence of functional studies, it is impossible to assign an unequivocal role to the pattern of duplication and positive selection we have identified. However, our results allow us to formulate working hypotheses that can be experimentally tested in future studies. For example, the high heterogeneity in the selective pressure acting on *Drosophila* Rh7

(fig. 3E), coupled with its fast evolution (i.e., the high d_N, fig. 3C) and expression in the clock-neurons (Ma et al. 2021) may be associated with divergence in the circadian clock in species with different ecology and latitudinal distribution (Menegazzi et al. 2017). This might explain the findings in the agricultural pest *D. suzukii*, a species characterized by significant selective pressure affecting *Rh6* and *Rh7* (fig. 3). These changes are interesting from an applicative point of view, as it is possible that they are linked to the peculiar circadian activity (Hansen et al. 2019), color recognition pattern (Little et al. 2019), and even gustatory preferences (Crava et al. 2016; Leung et al. 2020) associated with this species' peculiar ecological lifestyle. In *Anopheles*, opsin function is less well understood than *Drosophila* (Montell and Zwiebel 2016). However, different mosquito species may be active during specific periods of the day or night, when light is characterized by a different wavelength composition (Downes 1969; Sawadogo et al. 2014). The opsins' unique capacity to tune their maximum absorbance to specific light conditions might therefore have had a role in the evolution of these ecological differences (Jenkins and Muskavitch 2015). Indeed, we hypothesize that the high variability in the selective pressure affecting *Rh7* and C-opsin in the *Anopheles* species (fig. 3F) may be linked to differences in their adaptation to light detection, including the possible function in the circadian clock.

Future works should concentrate on the physiological significance of the duplication/losses we have identified, as well as seeking to understand the functional role of the sites under positive selection. This requires, for example: 1) the validation of the candidate's selected sites using PCR; 2) a 3D reconstruction of the various opsins, which is complicated by the high divergence between orthologs and the absence of a validated 3D structure for most of the opsins; and 3) site-specific mutants to validate any possible function. Overall, our work serves as a comparative genomic overview of opsin evolution in dipterans and represents the foundation for future studies aimed at improving our understanding of dipteran visual biology and the management of economically relevant species such as mosquitoes (e.g., Zhan et al. 2020) and fruit-eating flies.

Materials and Methods

Opsin Identification

We downloaded 61 predicted proteomes from 10 Diptera families (Culicidae, Chironimidiae, Psychodidiae, Drosophilidae, Tephritidae, Glossinidae, Calliphoridae, Muscidae, Diopsidae, Ceciomyiidae; supplementary table S1, Supplementary Material online). We evaluated their quality by assessing their completeness with BUSCO (Simão et al. 2015), using the 1,367 single-copies orthologs of the insects lineage data set. To identify the opsin genes, we employed a combination of BLAST and motif search similarly to Feuda et

al. (2016). In brief, the sequences from Feuda et al. (2016, 2012) and Ramirez et al. (2016) were used to mine every genome. From this analysis, every gene with an e-value $<10^{-10}$ was retained as a putative opsin gene and was subject to a motive search using Prosite (Sigrist et al. 2013) and an annotation using BLASTP against the Uniprot90 Database. To be considered an opsin, either one of two conditions was sufficient: the sequence must contain a retinal binding domain or have an opsin as the first hit in the BLAST search. Using this approach and after a preliminary manual annotation, we identified 528 opsin genes (supplementary table S1, Supplementary Material online). Alignment and trees are available on Bitbucket (https://bitbucket.org/Feuda-lab/opsin_diptera/src/master/).

Manual Curation

The data set obtained was eventually manually curated. For example, we first checked for missing data. We selected sequences that lacked part of the opsin protein, and, where possible, we retrieved the missing data using BLAST (tblastn) on the assembled genomes. Second, we looked for putative false duplications in the tree, and in the case where we found a species-specific duplication in our subsequent analyses, we removed the incomplete sequence. Moreover, we looked for unexpected opsin losses to assess whether the missing genes were true losses or artifacts (false negatives). In some cases, we found the missing gene in the genome of interest and the sequence was added manually to the alignment.

For some mosquito species, we lacked well-assembled genomes and, therefore, accurate gene models, which may have caused misrepresentation in the exact number of *Rh6* copies in each *Anopheles* lineage and blurred the fine-scale duplications/losses pattern. We therefore carefully and manually validated the *Rh6* genes in the *Anopheles* species. Using such an approach allowed us to increase the length of many orthologs, most importantly, allowing us to detect instances of false positives: cases where putative duplicated contigs or allele variants from heterozygotes genomes could be mistaken for species-specific duplications.

We further manually inspected for possible pseudogenes. For the *Drosophila* and *Anopheles* species, we manually curated all the alignments in order to perform d_N/d_S studies (see below) to exclude pseudogenes, because we could not find signature of pseudogenes ($d_N/d_S = 1$), nor we detect internal stop codons. For all other species, we inspected the alignment by eye when the gene was characterized by extremely long branches.

Gene Structure Characterization

We investigated intron presence in the 61 dipteran species under study using Vector base (Giraldo-Calderón et al. 2015), Ensembl (Yates et al. 2020), and in some cases by manual curation. The full gene region of each of the seven opsins in

Drosophila was further inspected in the FlyBase genome browser for detailed intron length, which was mapped separately in supplementary table S3, Supplementary Material online. To assess significant events of intron length variation, we developed a method that assumed a normal distribution for the length of each intron and highlighted significant introns whose length was larger (or shorter) than the mean plus twice the standard deviation for that intron (estimated excluding from the target intron).

Phylogenetic Analysis

To identify the phylogenetic relationships between the opsin genes, we performed a phylogenetic analysis using Bayesian and Maximum Likelihood inferences. The opsins data set was first merged to a subsample of the insect data set of Feuda et al. (2016), and the sequences aligned using MAFFT v7.4 (Katoh et al. 2002) with default parameters. The maximum likelihood tree was performed using IQTree 2.0 (Nguyen et al. 2015) under the GTR-G4 model. The Bayesian tree was performed using Phylobayes-MPI (Lartillot et al. 2013) under the GTR-G4 model and node support was estimated using PP.

GeneRax Analysis

We used the 1,367 BUSCO single-copy orthologs (see above) to assemble a supermatrix composed of 505,000 amino acid positions. The species tree was estimated using the single-copy gene hits from the BUSCO analyses (see above). The sequences of each BUSCO gene were extracted and aligned using MAFFT v7.4 (Katoh et al. 2002) and trimmed using gblocks v0.91 (Talavera and Castresana 2007) (allowing half gaps, minimum block length 5, maximum contiguous non-conserved positions 5% and 75% of sequences present in flank positions), then all alignments were concatenated using FASconcat v1.11 (Kuck and Meusemann 2010). The concatenated alignment consisted of 73 species (61 Diptera, 4 Lepidoptera) with 504,666 nucleotide positions and was analyzed under LG+F+I+G4. Both selection and phylogenetic inference were performed in IQ-TREE2 (Nguyen et al. 2015). This species tree (supplementary fig. S3, Supplementary Material online), and a manually modified version matching that presented in figure 2 and Wiegmann et al. (2011), were used for species tree-gene tree reconciliation analysis using GeneRax (Morel et al. 2020) alongside the opsin gene tree resolved using GTR-G (supplementary fig. S2, Supplementary Material online). Reconciled trees are displayed in supplementary figures S4–S6, Supplementary Material online. Alignment and trees are available on bitbucket (https://bitbucket.org/Feuda-lab/opsin_diptera/src/master/).

Positive Selection

The coding sequence of each opsin subgroup (*Rh1*, *Rh2*, *Rh3*, *Rh4*, *Rh5*, *Rh6*, *Rh7*, and *C*) were aligned separately for the 21 *Drosophila* and 19 *Anopheles* species using the PRANK (Loytynoja and Goldman 2008) codon model, which produces fewer false positives in positive selection analysis (Markova-Raina and Petrov 2011). Each alignment was manually curated to avoid spurious divergence signals that may have biased the subsequent analyses, and we generated two sets of alignments, one using all sites and a second where all the regions containing gaps were removed. We inferred the level of selective pressure acting on each of the 7 Drosophila opsins using PAML 4.7 (Yang 2007). Rates of synonymous (d_S) and nonsynonymous substitution (d_N), as well as their ratio $\omega = d_N/d_S$ (which measures levels of selective pressure acting on a gene), were estimated by the "free-ratio" model using the unrooted species tree topology inferred above. In this analysis, alignments included only sequences from those species that were represented in all opsin alignments to allow cross opsin-gene comparisons in *Anopheles*. Heterogeneity in the selective pressure was inferred using a branch-test to compare the likelihood of a single ω model across branches (model = 0 and N_S sites = 0) versus one assuming two distinct ω, one for each terminal branch, one at a time (i.e., for each *Drosophila* and *Anopheles* species in their respective data sets), and another for rest of the tree. To further identify the occurrence of positive selection on specific sites, we employed the branch-site test (branch-site model A, test 2; model = 2 and NS sites = 2; null model has parameters fix_$\omega = 1$, $\omega = 1$; the positive selection model fix_$\omega = 0$, $\omega = 1$, with each species set as foreground species in separate analysis, see above). Both tests were estimated using either the whole alignment (clean = 0) or removing parts of the alignment where one or more sequences contained a gap (clean = 1). We tested twice the difference between the log-likelihood of the two models for both tests using a χ^2 test with 1 degree of freedom. To account for multiple testing, we estimated the false discovery rate of each test using the *q*-value approach (Storey 2002) implemented in R. All statistics are summarized in supplementary table S4, Supplementary Material online.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

Author Contributions

Data Availability

The data underlying this article are available on bitbucket (https://bitbucket.org/Feuda-lab/opsin_diptera/src/master/).

Literature Cited

Almudi I, et al. 2020. Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings. Nat Commun. 11(1):2631.

Anstead CA, et al. 2016. A blow to the fly—*Lucilia cuprina* draft genome and transcriptome to support advances in biology and biotechnology. Biotechnol Adv. 34(5):605–620.

Attardo GM, et al. 2019. Comparative genomic analysis of six Glossina genomes, vectors of African trypanosomes. Genome Biol. 20(1):187.

Attardo GM, et al.; International Glossina Genome Initiative. 2014. Genome sequence of the tsetse fly (*Glossina morsitans*): vector of African trypanosomiasis. Science 344(6182):380–386.

Baker DA, et al. 2011. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, *Anopheles gambiae*. BMC Genomics 12:296.

Bao R, Friedrich M. 2009. Molecular evolution of the Drosophila retinome: exceptional gene gain in the higher Diptera. Mol Biol Evol. 26(6):1273–1287.

Booth HAF, Holland PW. 2004. Eleven daughters of NANOG. Genomics 84(2):229–238.

Briscoe AD, Chittka L. 2001. The evolution of color vision in insects. Annu Rev Entomol. 46:471–510.

Carulli JP, Chen D-M, Stark WS, Hartl DL. 1994. Phylogeny and physiology of Drosophila opsins. J Mol Evol. 38(3):250–262.

Courgeon M, Desplan C. 2019. Coordination of neural patterning in the Drosophila visual system. Curr Opin Neurobiol. 56:153–159.

Crava CM, Ramasamy S, Ometto L, Anfora G, Rota-Stabelli O. 2016. Evolutionary insights into taste perception of the invasive pest *Drosophila suzukii*. G3 (Bethesda) 6(12):4185–4196.

Davis FP, et al. 2020. A genetic, genomic, and computational resource for exploring neural circuit function. eLife. 9:e50901.

Downes JA. 1969. The swarming and mating flight of Diptera. Annu Rev Entomol. 14(1):271–298.

Fain GL, Hardie R, Laughlin SB. 2010. Phototransduction and the evolution of photoreceptors. Curr Biol. 20(3):R114–R124.

Feuda R, Hamilton SC, McInerney JO, Pisani D. 2012. Metazoan opsin evolution reveals a simple route to animal vision. Proc Natl Acad Sci U S A. 109(46):18868–18872.

Feuda R, Marletaz F, Bentley MA, Holland PW. 2016. Conservation, duplication, and divergence of five opsin genes in insect evolution. Genome Biol Evol. 8(3):579–587.

Fleming JF, et al. 2018. Molecular palaeontology illuminates the evolution of ecdysozoan vision. Proc Biol Sci. 285(1892):20182180.

Futahashi R, et al. 2015. Extraordinary diversity of visual opsin genes in dragonflies. Proc Natl Acad Sci U S A. 112(11):E1247–E1256.

Giraldo-Calderón GI, et al.; VectorBase Consortium. 2015. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res. 43(Database issue):D707–D713.

Giraldo-Calderón GI, Zanis MJ, Hill CA. 2017. Retention of duplicated long-wavelength opsins in mosquito lineages by positive selection and differential expression. BMC Evol Biol. 17(1):84.

Hansen CN, et al. 2019. Locomotor behaviour and clock neurons organisation in the agricultural pest *Drosophila suzukii*. Front Physiol. 10:941.

Henze MJ, Oakley TH. 2015. The dynamic evolutionary history of pancrustacean eyes and opsins. Integr Comp Biol. 55(5):830–842.

Hu X, Leming MT, Whaley MA, O'Tousa JE. 2014. Rhodopsin coexpression in UV photoreceptors of *Aedes aegypti* and *Anopheles gambiae* mosquitoes. J Exp Biol. 217(Pt 6):1003–1008.

Jenkins AM, Muskavitch MAT. 2015. Crepuscular behavioral variation and profiling of opsin genes in *Anopheles gambiae* and *Anopheles stephensi* (Diptera: Culicidae). J Med Entomol. 52(3):296–307.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Kuck P, Meusemann K. 2010. FASconCAT: convenient handling of data matrices. Mol Phylogenet Evol. 56(3):1115–1118.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 62(4):611–615.

Leming MT, Rund SSC, Behura SK, Duffield GE, O'Tousa JE. 2014. A database of circadian and diel rhythmic gene expression in the yellow fever mosquito *Aedes aegypti*. BMC Genomics 15:1128.

Leung NY, et al. 2020. Functions of opsins in Drosophila taste. Curr Biol. 30(8):1367–1379.e6.

Leung NY, Montell C. 2017. Unconventional roles of opsins. Annu Rev Cell Dev Biol. 33:241–264.

Little CM, Rizzato AR, Charbonneau L, Chapman T, Hillier NK. 2019. Color preference of the spotted wing Drosophila. Sci Rep. 9(1):16051.

Loytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320(5883):1632–1635.

Ma D, et al. 2021. A transcriptomic taxonomy of Drosophila circadian neurons around the clock. Elife. 10. doi:10.7554/eLife.63056.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 21(6):863–874.

Menegazzi P, et al. 2017. Adaptation of circadian neuronal network to photoperiod in high-latitude European Drosophilids. Curr Biol. 27(6):833–839.

Montell C, Zwiebel LJ. 2016. Chapter ten—mosquito sensory systems. In: Raikhel AS, editor. Advances in Insect Physiology. Progress in Mosquito Research. Vol. 51. Academic Press. p. 293–328. doi:10.1016/bs.aiip.2016.04.007.

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. Mol Biol Evol. 37(9):2763–2774.

Neafsey DE, et al. 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science 347(6217):1258522.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Ni JD, Baik LS, Holmes TC, Montell C. 2017. A rhodopsin in the brain functions in circadian photoentrainment in Drosophila. Nature 545(7654):340–344.

Obbard DJ, et al. 2012. Estimating divergence dates and substitution rates in the Drosophila phylogeny. Mol Biol Evol. 29(11):3459–3473.

Pisani D, Rota-Stabelli O, Feuda R. 2020. Sensory neuroscience: a taste for light and the origin of animal vision. Curr Biol. 30(13):R773–R775.

Pollock JA, Benzer S. 1988. Transcript localization of four opsin genes in the three visual organs of Drosophila; RH2 is ocellus specific. Nature 333(6175):779–782.

Ramirez MD, et al. 2016. The last common ancestor of most Bilaterian animals possessed at least nine opsins. Genome Biol Evol. 8(12):3640–3652.

Rocha M, et al. 2015. Expression and light-triggered movement of rhodopsins in the larval visual system of mosquitoes. J Exp Biol. 218(Pt 9):1386–1392.

Rota-Stabelli O, et al. 2020. Distinct genotypes and phenotypes in European and American strains of Drosophila suzukii: implications for biology and management of an invasive organism. J Pest Sci. 93(1):77–89.

Sakai K, et al. 2017. Drosophila melanogaster rhodopsin Rh7 is a UV-to-visible light sensor with an extraordinarily broad absorption spectrum. Sci Rep. 7(1):7349.

Sawadogo PS, et al. 2014. Swarming behaviour in natural populations of Anopheles gambiae and An. coluzzii: review of 4 years survey in rural areas of sympatry, Burkina Faso (West Africa). Acta Trop. 132(Suppl):S42–S52.

Schmitt A, Vogt A, Friedmann K, Paulsen R, Huber A. 2005. Rhodopsin patterning in central photoreceptor cells of the blowfly Calliphora vicina: cloning and characterization of Calliphora rhodopsins Rh3, Rh5 and Rh6. J Exp Biol. 208(Pt 7):1247–1256.

Sigrist CJA, et al. 2013. New and continuing developments at PROSITE. Nucleic Acids Res. 41(Database issue):D344–D347.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.

Skevington JH, Dang PT. 2002. Exploring the diversity of flies (Diptera). Biodiversity 3(4):3–27.

Sokabe T, Chen H-C, Luo J, Montell C. 2016. A switch in thermal preference in Drosophila larvae depends on multiple rhodopsins. Cell Rep. 17(2):336–344.

Sondhi Y, Ellis EA, Theobald JC, Kawahara AY. 2020. Light environment drives evolution of color vision genes in butterflies and moths. bioRxiv doi:10.1101/2020.02.29.965335.

Storey JD. 2002. A direct approach to false discovery rates. J R Stat Soc B (Stat Methodol). 64(3):479–498.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56(4):564–577.

Tierney SM, et al. 2015. Opsin transcripts of predatory diving beetles: a comparison of surface and subterranean photic niches. R Soc Open Sci. 2(1):140386.

Tierney SM, et al. 2012. Photic niche invasions: phylogenetic history of the dim-light foraging augochlorine bees (Halictidae). Proc Biol Sci. 279(1729):794–803.

van der Kooi CJ, Stavenga DG, Arikawa K, Belušič G, Kelber A. 2021. Evolution of insect color vision: from spectral sensitivity to visual ecology. Annu Rev Entomol. 66:435–461.

Velarde RA, Sauer CD, Walden KK, Fahrbach SE, Robertson HM. 2005. Pteropsin: a vertebrate-like non-visual opsin expressed in the honey bee brain. Insect Biochem Mol Biol. 35(12):1367–1377.

Vöcking O, Kourtesis I, Tumu SC, Hausen H. 2017. Co-expression of xenopsin and rhabdomeric opsin in photoreceptors bearing microvilli and cilia. Elife. 6:e23435. doi:10.7554/eLife.23435.

Wen D, Yu Y, Hahn MW, Nakhleh L. 2016. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol Ecol. 25(11):2361–2372.

White IM, Elson-Harris MM. 1992. Fruit flies of economic significance: their identification and bionomics. [cited 2020 Apr 8]. Available from: https://www.cabdirect.org/cabdirect/abstract/19921161954

Wiegmann BM, et al. 2011. Episodic radiations in the fly tree of life. Proc Natl Acad Sci U S A. 108(14):5690–5695.

Xu P, et al. 2016. Functional opsin retrogene in nocturnal moth. Mob DNA. 7:18.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586–1591.

Yates AD, et al. 2020. Ensembl 2020. Nucleic Acids Res. 48(D1):D682–D688.

Zadra N, Rizzoli A, Rota-Stabelli O. 2021. Chronological incongruences between mitochondrial and nuclear phylogenies of Aedes mosquitoes. Life (Basel) 11(3):181.

Zanini D, et al. 2018. Proprioceptive opsin functions in Drosophila larval locomotion. Neuron 98(1):67–74.e4.

Zhan Y, Alberto DAS, Rusch C, Riffell JA, Montell C. 2020. Aedes aegypti vision-guided target recognition requires two redundant rhodopsins. bioRxiv doi:10.1101/2020.07.01.182899.

Associate editor: Andrea Betancourt