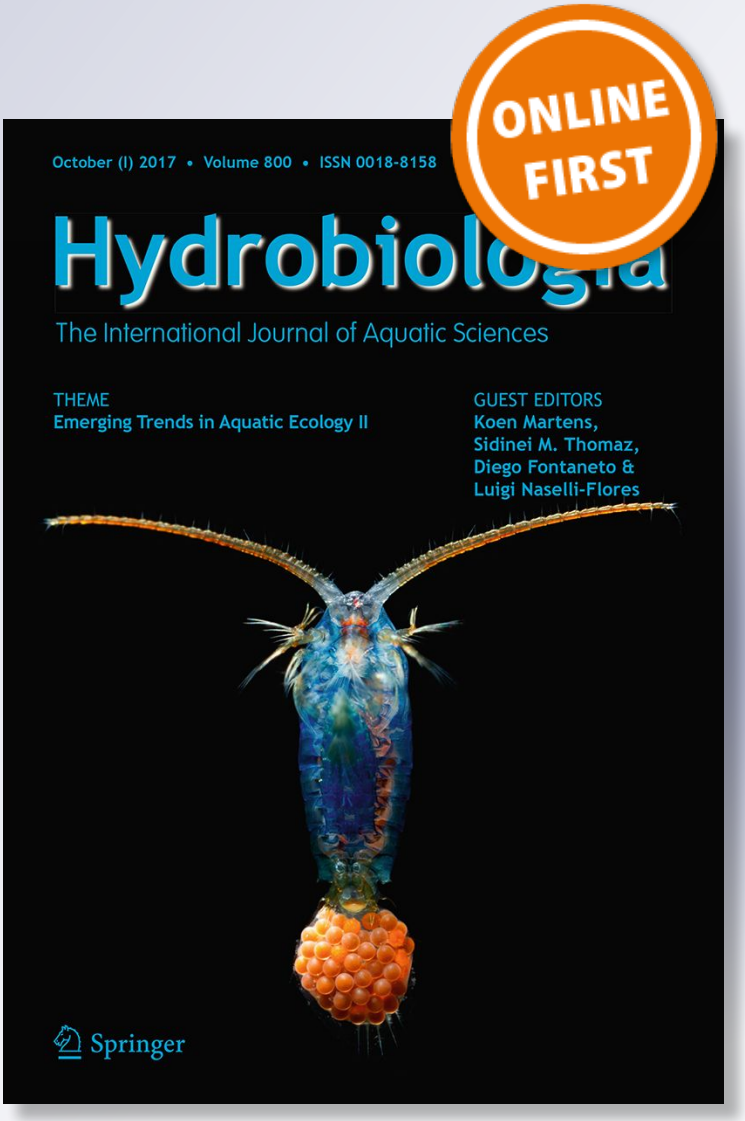
Phylogenetic annotation and genomic architecture of opsin genes in Crustacea

Jorge L. Pérez-Moreno, Danielle M. DeLeo, Ferran Palero & Heather D. Bracken-Grissom

Hydrobiologia
The International Journal of Aquatic Sciences

ISSN 0018-8158

Hydrobiologia
DOI 10.1007/s10750-018-3678-9



Springer

Hydrobiologia
https://doi.org/10.1007/s10750-018-3678-9

CRUSTACEAN GENOMICS

Phylogenetic annotation and genomic architecture of opsin genes in Crustacea

Jorge L. Pérez-Moreno · Danielle M. DeLeo · Ferran Palero · Heather D. Bracken-Grissom

Received: 15 October 2017 / Revised: 25 May 2018 / Accepted: 2 June 2018

Abstract A major goal of evolutionary biology is to understand the role of adaptive processes on sensory systems. Visual capabilities are strongly influenced by environmental and ecological conditions, and the evolutionary advantages of vision are manifest by its complexity and ubiquity throughout Metazoa. Crustaceans occupy a vast array of habitats and ecological niches, and are thus ideal taxa to investigate the evolution of visual systems. A comparative approach is taken here for efficient identification and classification of opsin genes, photoreceptive pigment proteins involved in color vision, focusing on two crustacean model organisms: *Hyalella azteca* and *Daphnia pulex*. Transcriptomes of both species were assembled de novo to elucidate the diversity and function of expressed opsins within a robust phylogenetic context. For this purpose, we developed a modified version of the Phylogenetically Informed Annotation tool's pipeline to filter and identify visual genes from transcriptomes in a scalable and efficient manner. In addition, reference genomes of these species were used to validate our pipeline while characterizing the genomic architecture of the opsin genes. Next-generation sequencing and phylogenetics provide future venues for the study of sensory systems, adaptation, and evolution in model and nonmodel organisms.

Keywords Evolution · Phototransduction · Protein · RNAseq · Transcriptomics · Vision

Guest editors: Guiomar Rotllant, Ferran Palero, Peter Mather, Heather Bracken-Grissom & Begoña Santos / Crustacean Genomics

Jorge L. Pérez-Moreno and Danielle M. DeLeo have contributed equally to this work.

J. L. Pérez-Moreno (✉) · D. M. DeLeo · H. D. Bracken-Grissom
Department of Biological Sciences, Florida International University – Biscayne Bay Campus, North Miami, FL 33181, USA
e-mail: jorge.perezmoreno@fiu.edu

F. Palero
Centre d'Estudis Avançats de Blanes (CEAB-CSIC), Carrer d'Accés a la Cala Sant Francesc 14, 17300 Blanes, Spain

Introduction

Opsins are photoreceptor molecules that play a crucial role in animal vision and can be found across metazoans (Terakita, 2005). As membrane-associated, G-protein-coupled receptors (GPCRs), opsins can function in both visual and nonvisual phototransduction, and in some instances as photoisomerases (Shichida & Matsuyama, 2009). Previous studies have classified opsins in three primary categories according to the type of G-protein to which they couple namely, "ciliary" (c-opsins), "rhabdomeric" (r-opsins), and RGR/Go opsins (Terakita, 2005; Feuda et al., 2014,

Published online: 14 June 2018

2016). Ciliary and rhabdomeric opsins diversified prior to the protostome-deuterostome split and are found in both invertebrates and vertebrates, which suggests that they co-occurred in a common ancestor (Kojima et al., 1997; Shichida & Matsuyama, 2009; Hering & Mayer, 2014; Ramirez et al., 2016). Opsin categories can additionally be further subdivided into subfamilies based on molecular phylogenetics and functional classifications (Terakita, 2005). These subfamilies share less than 20 percent amino acid identity (Fryxel & Meyerowitz, 1991) and comprise c-opsins (visual and nonvisual); tmt/encephalopsins; r-opsins; melanopsins; and photoisomerases/neuropsins. Photoreceptive opsins can be of either the ciliary-type, found largely in vertebrates (for exceptions see Arendt et al., 2004; Passamaneck et al., 2011; Bok et al., 2017; Tsukamoto et al., 2017), or the rhabdomeric-type found in mollusks, annelids, and the compound eyes of arthropods (Arendt et al., 2002; Shichida & Matsuyama, 2009; Gühmann et al., 2015), with the last being the focus of the present study.

Opsins form visual pigments capable of absorbing photons when bound to a chromophore, generally a vitamin A1 derivative (11-cis retinal). These visual pigments trigger conformational changes that activate G-proteins (Nathans, 1987) and elicit phototransduction signaling cascades. Key biological processes such as the regulation of circadian clocks, phototaxis, and vision have been shown to be linked to the phototransduction cascade (e.g., Arendt et al., 2004; review Shichida & Matsuyama, 2009). The set of amino acid residues that interact with the chromophore produce an environment suitable for the absorption of light with distinct wavelengths (Imai et al., 1997; Kuwayama et al., 2002) and thus influence spectral tuning (e.g., Porter et al., 2007; Katti et al., 2010). As the absorption spectrum of the photopigment is influenced by the amino acid composition of the opsin protein, slight variations can alter its physical and chemical properties and lead to visual pigments maximally sensitive to different wavelengths of light. This in turn would allow organisms to perceive and distinguish between lights of particular wavelengths. The direct association between amino acid composition of photoreceptive opsins and their spectral sensitivity make them amenable to functional classification by sequence analysis (Mirzadegan et al., 2003; Matsumoto & Ishibashi, 2016).

Three main approaches have been employed to characterize opsins from transcriptomic data: (I) Sequence similarity searches via pairwise alignments (i.e., BLAST); (II) Protein structure prediction through Hidden Markov Model (HMM) profile alignments; and (III) Phylogenetic inference. Functional annotation by means of sequence similarity is typically based on heuristic algorithms that search for matching nucleotide and/or amino acid sequences in curated databases (e.g., BLAST; Altschul et al., 1990). Sequences are locally or globally aligned and subsequently annotated based on inferred homology with statistically significant matches (Pearson, 2013). These comparisons, however, can rapidly become computationally expensive as the number of query and/or reference sequences increases (Suzuki et al., 2012). Although similarity searches via pairwise alignments are capable of identifying homologous sequences, their shortcomings are notorious when the queries consist of protein families with low sequence similarities, as is the case for opsins and other GPCRs (Pearson, 2013). Hidden Markov Model (HMM) methods offer an enticing alternative to pairwise alignments at similar computational costs (Eddy, 2011; Pearson, 2013). HMM profiles can also contain relevant information regarding protein structure, which translates to more accurate identification, classification, and annotation of proteins even when overall sequence similarity is low (Krogh et al., 1994; Yoon, 2009; Pearson, 2013). However, the efficacy of HMMs is intrinsically dependent on the quality of the training data, which is a nontrivial process in the case of understudied taxa or protein families (Rasmussen & Krink, 2003; Pearson, 2013). Therefore, the use of HMMs for annotation of GPCRs is hindered when independently verified sequences are not readily available. The robustness and suitability of phylogenetic approaches for functional annotation of opsins (and other proteins) is unparalleled, as it can readily overcome many of the deficiencies of other homology-based methodologies (Engelhardt et al., 2009; Gaudet et al., 2011; Speiser et al., 2014). The placement of proteins on a phylogenetic tree not only enables a rapid assessment of homology and efficient discrimination of false positives, but also allows for the inference of putative functions and roles within an evolutionary context (Engelhardt et al., 2009). This approach has been successful in classifying novel opsins (and other GPCRs) despite their characteristic low sequence

similarities and, in the case of nonmodel organisms, scarce genomic resources (Porter et al., 2007, 2012; Speiser et al., 2014). The main drawbacks of phylogenetic reconstruction as an efficient functional annotation method are possible difficulties aligning distantly related sequences, its propensity to be time-consuming (obtaining adequate references, computation of trees, etc.) and the steep learning curve to master these analyses, which might result in subjectivity and misinterpretations (Crisp & Cook, 2005).

Efforts to characterize opsins from high-throughput sequencing data in nonmodel Crustacea have primarily focused on transcriptomes, but without genomic validation (Porter et al., 2013; Wong et al., 2015; Biscontin et al., 2016). When available, genomes can provide valuable information regarding opsin gene duplication in an organism, as well as the relative locations of those genes. Gene locations allow for intra- and interspecific comparisons (Nordström et al., 2004) and to make inferences about the evolutionary history of opsin diversification (review Shichida & Matsuyama, 2009).

In this study, we modified the Phylogenetically Informed Annotation (PIA) tool's pipeline (Speiser et al., 2014) to conduct a robust and scalable phylogenetic annotation of visual opsins from transcriptomes of two crustacean model organisms, *Hyalella azteca* (Saussure, 1858) and *Daphnia pulex* Leydig, 1860. *Hyalella azteca* is a freshwater epibenthic amphipod, commonly used as a bioindicator species, which has one pair of pigmented compound eyes (Gonzalez & Watling, 2002). *Daphnia* pulex is a freshwater cladoceran that has a single but movable cyclopean, compound, and pigmented eye. Specifically, we made modifications for PIA to run on the command-line rather than on Galaxy's GUI and wrote wrapper scripts to facilitate the analyses. This resulted in a scalable and automated platform to annotate visual genes and pathways, while minimizing possible biases and subjectivity from manual curation. As genomes are available for both species, they were used to validate the annotations and make inferences about the genomic architecture and the opsin intron–exon gene structure within these species.

Methods

Data, quality control, and transcriptome assembly

Raw RNA sequencing data of the freshwater amphipod *H. azteca* and the model branchiopod *D. pulex* were downloaded from the NCBI's Sequence Read Archive (SRA). In order to facilitate de novo transcriptome assembly and accurate detection of complete opsin isoforms, the read files were trimmed taking into consideration factors such as length and quality of the sequencing reads, sequencing depth, and tissue type (Table 1).

Prior to the assembly process, quality of the raw sequencing reads was evaluated via FastQC (Andrews, 2010). The FastQC output was subsequently used to inform stringent quality and adaptor trimming with Trimmomatic 0.36 (parameters: "*ILLUMINACLIP:-TruSeq 3-PE.fa:2:30:10 CROP:140 HEADCROP:20 LEADING:15 TRAILING:15 SLIDINGWINDOW:4:20 MINLEN:36*"; Bolger et al., 2014). Clean sequencing reads were then assembled into a de novo transcriptome with the Trinity pipeline (version 2.5.0; Grabherr et al., 2011; Haas et al., 2013) using default parameters, a minimum contig length of 200 bp, and a *kmer* size of 23. Assembly summary statistics were calculated using Transrate 1.0.3 (Smith-Unna et al., 2016). BUSCO 3.0.2 (Benchmarking Universal Single-Copy Orthologs; Simão et al., 2015) was employed to assess the quality and completeness of the resulting transcriptomes. The latter method provides an accurate evaluation of transcriptomes in an evolutionary informed context by assessing the presence and completeness of universal single-copy orthologs (Simão et al., 2015). BUSCO analyses were conducted with the Arthropoda database of orthologous groups (n = 1066) sourced from OrthoDB (Waterhouse et al., 2013).

Identification and annotation of crustacean opsins

Identification and functional classification of putative opsin transcripts was achieved through the use of our modified version of the existing PIA tool (Speiser et al., 2014). While phylogenetic confirmation of BLAST similarity hits is becoming routine in model systems, PIA allows for the identification of proteins involved in visual pathways for nonmodel organisms in a computationally efficient manner (Speiser et al.,

Table 1 Raw data chosen for de novo transcriptome assembly and annotation of opsin proteins in *Hyalella azteca* and *Daphnia pulex*

Species	Megabytes	Megabases	Read lengths	Sequencing platform	Tissue type	SRA BioProject
Hyalella azteca	15,543	33,160	2 × 150 bp	Illumina HiSeq	Whole organism	PRJNA312414
Daphnia pulex	16,134	39,280				PRJNA380400

2014). This informative tool places putative visual gene transcripts (e.g., opsins), previously identified via BLAST searches against a custom database, in precomputed phylogenies of such genes. The resulting phylogenies can then be used to discriminate BLAST false positives and/or paralogous sequences from the transcripts of interest. While PIA has been used in previous studies to annotate genes in a phylogenetic context, it was originally designed as a workflow for the Galaxy Project (Afgan et al., 2016) and as such is dependent on a Graphical User Interface (Speiser et al., 2014). This workflow can become inefficient when conducting concurrent analyses of numerous transcriptomes. Further, curation of the phylogenetic gene trees produced by PIA for each input transcriptome is typically undertaken manually, which inevitably makes it sensitive to potential biases. Tree branch length cutoff values for a given gene (i.e., opsins) can, however, be determined empirically through a series of manual tree curation comparisons. The pipeline presented here is a modification to PIA's pipeline in which the authors wrote a wrapper script to enable its use as a command-line/automated workflow, which effectively increases its scalability allowing for the identification of visual opsins from multiple transcriptomic datasets through simple scripting. Although the pipeline was designed for analyses of visual pathways, it is possible to create custom databases and phylogenies for other genes/pathways. We refer the reader to the original publication of PIA for additional information regarding included genes and pathways (Speiser et al., 2014). The modified Phylogenetically Informed Annotation pipeline employed in this study, along with usage examples, will be made available at: https://github.com/xibalbanus/PIA2.

Once the transcriptome assembly was completed, our de novo assemblies were scanned with Biopython's *get_orfs_cds.py* script (Cock et al., 2009) to translate each transcript into its corresponding amino acid sequence. Open Reading Frames (ORFs) were then extracted via the same script to facilitate the PIA annotation process. After conclusion of PIA's main component (BLAST, MAFFT alignment, and phylogenetic placement via RAxML; Altschul et al., 1990; Stamatakis, 2014; Yamada et al., 2016), a script adapted from the Osiris Phylogenetics toolkit (*long_branch_finder.py*; Oakley et al., 2014) was used to identify transcripts that exceeded 4 × the Mean Absolute Deviation of the tree's branch lengths. This simple threshold proved effective at removing spurious BLAST hits in an unbiased manner. Subsequently, the previously identified false positives were pruned from our phytab-formatted hit-list (part of PIA's output) with the *prune_phytab_using_list.py* script, also adapted from Osiris (Oakley et al., 2014). The resulting list of putative opsins was then converted to FASTA format, and sequence redundancy was reduced by removing identical protein sequences with UCLUST (Edgar, 2010). The multiple sequence aligner MAFFT (Yamada et al., 2016) was then invoked to align our filtered putative opsins to a large opsin dataset (n = 910) compiled by the Porter Lab (University of Hawaii at Manoa), which includes representatives of the main opsin subfamilies. MAFFT alignment parameters were chosen to prioritize accuracy over speed and to allow for large unalignable regions that can be pervasive with divergent GPCRs ("*–ep 0 –genafpair –maxiterate 1000*"). Following the alignment procedure, a final phylogenetic reconstruction was undertaken with IQ-tree (Nguyen et al., 2015) for characterization and annotation of our PIA-identified putative opsins. IQ-tree compares favorably to alternatives (e.g., RAxML, FastTree, etc.) in recent benchmarks (Zhou et al., 2017), while also providing a more extensive choice of evolutionary models for phylogenetic inference. After proper consideration, IQ-tree was selected given that evolutionary model choice is important, and its choice would be limited in alternative software. Choosing an appropriate model is

especially relevant when inferring phylogenetic relationships in protein families with both highly conserved domains and hypervariable regions (e.g., opsins). The IQ-tree analysis was run with a LG general amino acid replacement matrix under a FreeRate model with 10 rate categories and empirical base frequencies (LG + R10 + F; Le & Gascuel, 2008; Soubrier et al., 2012) as suggested by ModelFinder (Kalyaanamoorthy et al., 2017). Branch support was assessed in tripartite by Ultra-fast bootstrap approximation (UFBoot; 10,000 replicates), a Shimodaira–Hasegawa–like approximate likelihood ratio test (SH-aLRT; 10,000 replicates), and an approximate Bayes test (Guindon et al., 2010; Anisimova et al., 2011; Minh et al., 2013).

Finally, the tool HHBlits 'HMM-HMM–based lightning-fast iterative sequence search' (Remmert et al., 2012) was used to confirm opsin identity based on profile HMMs using Uniclust30 (Mirdita et al., 2017) as the reference database. HHBlits was chosen as it incorporates highly sensitive sequence search methods (HMMs) in a fast, and more accurate manner compared to other sequence search tools like PSI-BLAST (Remmert et al., 2012).

Genomic Architecture of Annotated Opsins

Proteins encoded in the transcriptomes analyzed may not have corresponding annotations in public databases. Therefore, to validate our pipeline, the exon–intron architecture of the opsin genes obtained from transcriptomic data (see above) was annotated de novo using the recently assembled genomes of *H. azteca* (GCA_000764305.2; accession date: 20-07-2017) and *D. pulex* (GCA_000187875.1; accessed on 20-07-2017). The Benchmarking set of Universal Single-Copy Orthologs (BUSCO version 3; Simão et al., 2015) was used to ensure an adequate completeness of the genomes used for transcriptome/genome comparison. BUSCO provides quantitative measures for the assessment of genome assembly based on evolutionarily informed expectations of gene content from near-universal single-copy orthologs selected from OrthoDB v9. The tblastn algorithm v2.2.29 + was then used with default parameters in order to discriminate between exonic and intronic regions along the genomic scaffolds. When a significant blast hit was found (similarity > 80%; *e* value < 10^{-8}), the corresponding genomic region was annotated as exonic, or protein coding/expressed region. DNA regions located between two consecutive exons in the same genomic scaffold (chromosome) but with no corresponding counterpart in the expressed RNA were considered as introns. In addition, the nucleotide coding sequence of each putative opsin was mapped to their respective genomes using the spliced aligner HISAT2 (Kim et al., 2010). The mapping was done without penalties for noncanonical splicing using the following command and arguments: "*hisat2 -f -x index input.cds.fasta –score-min L,0,-4 –pen-noncansplice 3 -S output.sam*". Plots of the gene architecture and the exon length distribution were subsequently completed using the Integrated Genome Browser (Freese et al., 2016) and the software package Mathematica v.11.1 (Wolfram Inc., USA).

Results

Hyalella azteca's transcriptome assembly recovered 243,398 contigs with a mean sequence length of 1033.04 base pairs (Table 2). Of these, 61,401 sequences contained Open Reading Frames (ORFs) designating them as putative protein-coding genes. Similarly, our de novo transcriptome for *D. pulex* was comprised of 187,310 contigs with a mean sequence length of 848.76, and 38,157 sequences with ORFs. Additional metrics for our de novo transcriptomes, as well as for the reference genomes, are given in Table 2.

Completeness assessment of our de novo transcriptomes by Benchmarking Universal Single-Copy Orthologs (BUSCO) was favorable for both species. In *H. azteca*, we were able to find 990 (92.48%) complete sequences of the 1,066 arthropod genes used for benchmarking. An additional 49 (4.6%) were also present as fragmented sequences, and only 27 (2.6%) were not found. Similarly, *D. pulex*'s transcriptome was found to be nearly complete with 1,048 (98.4%) full-length BUSCO genes, 16 (1.5%) fragmented, and a marginal 2 (0.1%) missing. The reference genomes selected for validation were rather complete as well, with over 90% of the BUSCO genes being found complete. Interestingly, the proportion of missing BUSCOs was slightly higher for the genomic data compared to the transcriptomic data (Table 3).

Table 2 Summary statistics for the de novo transcriptome assemblies produced as part of this study and the corresponding genome assemblies

	Hyalella azteca		*Daphnia pulex*	
Metric	Genome	Transcriptome	Genome	Transcriptome
Number of sequences/contigs	23,426	243,398	18,989	187,310
Longest sequence/contig (bp)	2,207,822	16,780	528,830	27,096
Number of bases	550,886,000	251,440,760	197,206,000	158,981,525
Mean transcript/contig length (bp)	23,404	1,033.04	8,352	848.76
Number of transcripts/contigs > 1000 bp long	14,563	73,869	16,743	42,717
Number of transcripts/contigs > 10000 bp long	7,614	157	2,854	179
Number of transcripts with ORFs		61,401		38,157
Mean ORF percent		45.73		50.22
N50	114,415	1,929	49,250	1,404
N30		3,213		2,588
N10		5,560		5,122
GC content	0.38	0.42	0.40	0.39

Our custom version of the PIA tool's pipeline outputs a single FASTA file of amino acid sequences per transcriptome. This file contains the transcripts that remain after the removal of spurious BLAST hits and the merging/removal of duplicated and fragmented sequences, and should only contain those that are closely related to the gene of interest (i.e., opsins). This output can then be piped to a final step for functional annotation by phylogenetic inference. In our case, putative opsin sequences for both species were aligned to a curated dataset of different opsin types. This final step resulted in a large phylogeny (Fig. 1) where opsins are classified based on their phylogenetic position.

Following Trinity's definition of assembled genes/isoforms, our pipeline identified 1 SWS/UV opsin (2 isoforms), 3 LWS opsins (3 isoforms), and 1 opsin-like GPCR as an outgroup (Fig. 2; Table 4) in *H. azteca*'s transcriptome. On the other hand, *D. pulex's* transcriptome contained 2 different SWS opsins (4 isoforms), 6 LWS opsins (35 isoforms), 2 melanopsins (4 isoforms), and 1 opsin-like transcript that was placed within the outgroup clade (Fig. 2; Table 4).

The identity results of the HHBlits search using iterative pairwise alignments and profile HHMs are summarized in Table 5, along with the inferred classification of each putative opsin transcript based on their respective placement in the phylogeny (Figs. 1, 2). In addition, each sequence entry was annotated as visual or nonvisual based on the sequence homology inferred by both methods (represented by a black box when both methods are in agreement; and a

Table 3 Results of transcriptome completeness assessment by Benchmarking Universal Single-Copy Orthologs (BUSCO) using OrthoDB's Arthropoda database of orthologous genes

Species	Dataset	Complete BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total BUSCOs Searched
Hyalella azteca	Genome	970 (91.0%)	29 (2.7%)	67 (6.3%)	1,066
	Transcriptome	990 (92.8%)	49 (4.6%)	27 (2.6%)	
Daphnia pulex	Genome	1,038 (97.3%)	9 (0.8%)	19 (1.9%)	
	Transcriptome	1,048 (98.4%)	16 (1.5%)	2 (0.1%)	

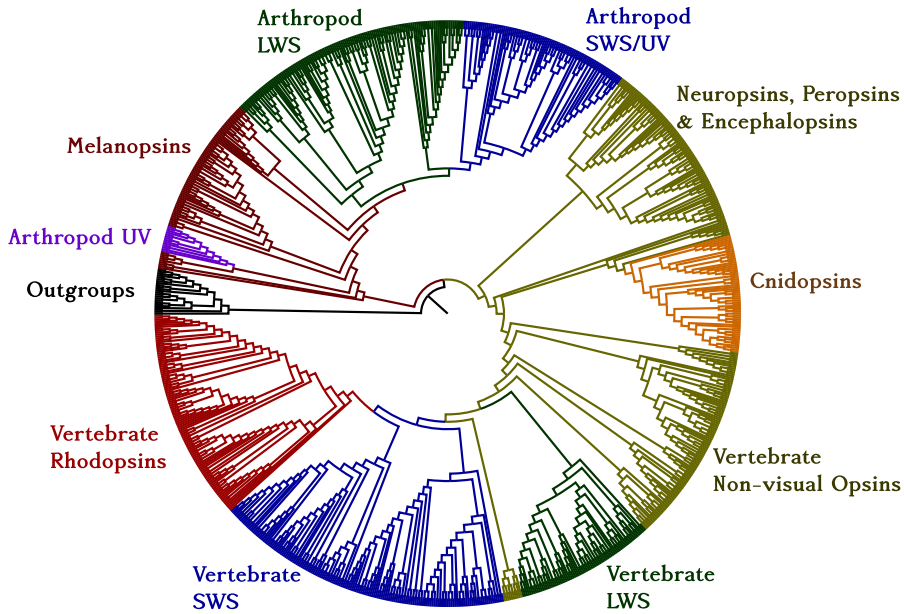


Fig. 1 Maximum-Likelihood phylogeny of opsins estimated using putative opsin proteins identified by our annotation pipeline from the *de novo* transcriptome assemblies of *Hyalella azteca* and *Daphnia pulex,* along with a dataset of reference opsin sequences. Clades are annotated with opsin types contained therein and, in the case of visual opsins, with their inferred spectral sensitivities

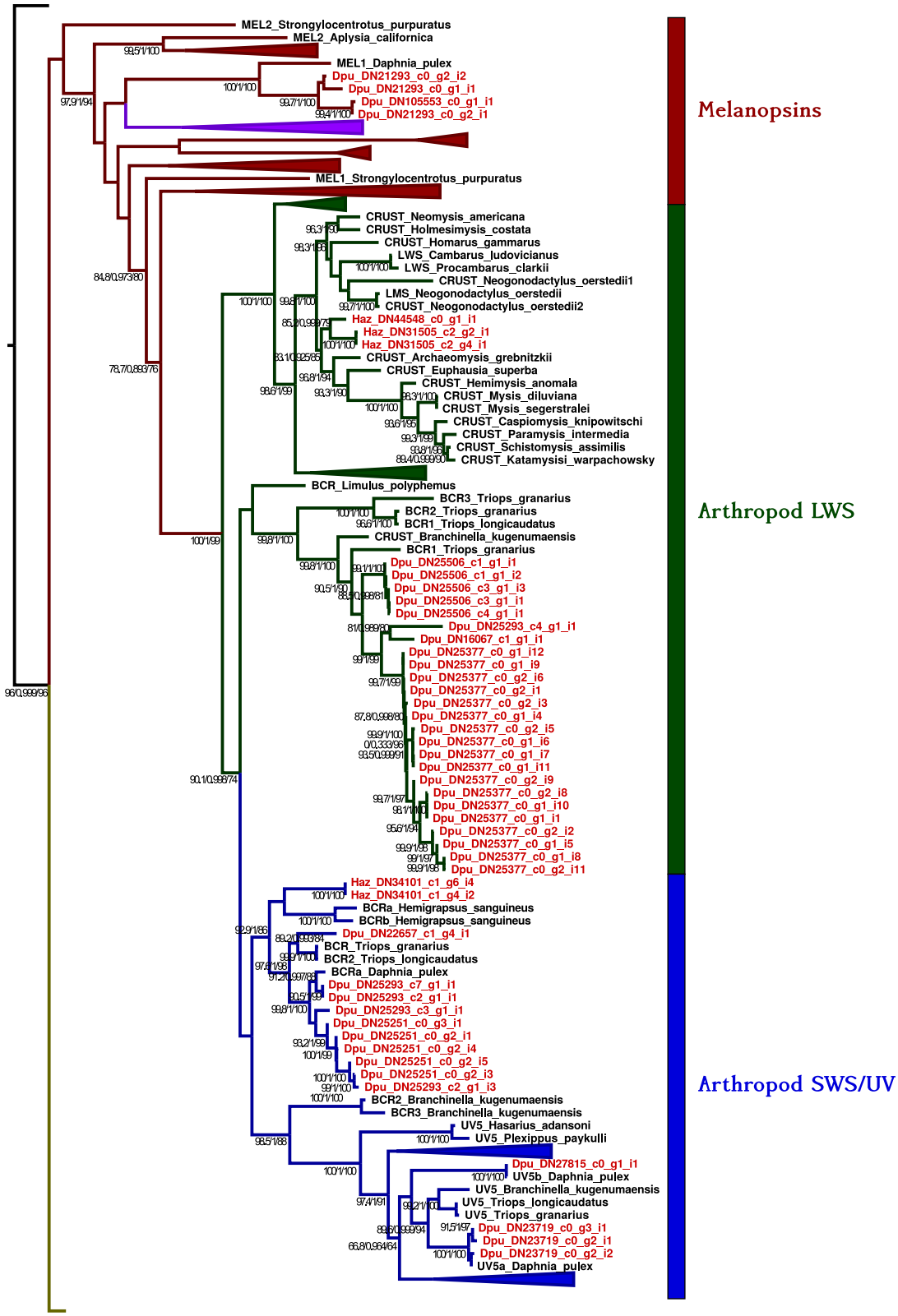
gray box when HMMs fail to identify them as a visual opsin; Table 5).

Genomic structure of annotated opsins

Every protein sequence predicted using the modified PIA pipeline gave at least one significant TBLASTN hit both in the *D. pulex* and *H. azteca* reference genomes. The observed distribution of introns within opsin genes appears to be variable both within and between genomes. To illustrate this variation, the Intron–Exon gene structure patterns of representative opsins were further characterized. The genomic region encoding for the SWS/UV opsin gene spanned about 4 kb in *Hyalella* and presented an extremely partitioned structure formed by at least seven different exons (Fig. 3). Interestingly, some LWS opsins within the *H. azteca* genome were located on the antisense strand and appear to be duplicated retrogenes (Fig. 3; see Discussion). Both SWS/UV and LWS opsins were also arranged following disparate gene architectures in the *D. pulex* genome (Fig. 4). LWS opsins presented slightly shorter introns on average than SWS/UV opsins, but the presence of gene duplications and genes with numerous introns were identified in most cases. Exon size distribution had similar shapes in both *D. pulex* and *H. azteca*, being multimodal for both genomes (Fig. 5). Nevertheless, *H. azteca* had a larger average exon size (Mean 410 bp; Median 235 bp) than *D. pulex* (Mean 225 bp; Median 164). Mapping results in SAM format are available for download from the following repository: https://github.com/xibalbanus/PIA2.

Discussion

Our results demonstrate the power of incorporating phylogenetic annotation toward the characterization and interpretation of large transcriptomic datasets of nonmodel organisms. Annotations via simple sequence similarity based methods like BLAST alone can result in false positives including, but not limited to, functional diversification following gene duplication events, domain shuffling, or even existing database errors (review Sjölander, 2004). Using the modified version of PIA allowed for the rapid and automated identification of false positives among the putative visual opsins curated for the two species of crustaceans, *H. azteca* and *D. pulex*. The modified PIA pipeline was able to successfully identify and filter opsins from the de novo transcriptomes in a fully automated manner with minimal manual curation. This automation is made possible mainly by the modifications and wrapper scripts that converted PIA from a Galaxy workflow to a command-line one, which effectively increases the scalability of the pipeline allowing for the identification of opsins (and



◀**Fig. 2** Expanded view of the melanopsin, Arthropod LWS, and Arthropod SWS/UV clades. Large noncrustacean clades have been collapsed for readability. Support values correspond to SH-aLRT/aBayes/UFBoot, and are not shown when UFBoot support < 75. Splits are considered highly supported when SH-aLRT >= 80%, aBayes = 0.95, and UFboot = 95%

other genes) from multiple transcriptomic datasets through simple scripting. Theoretically, this would allow for the annotation of dozens, if not hundreds, of transcriptomes at a time without the need of the excessive time-costs that a graphical user interface and manual curation of hundreds of phylogenetic trees would imply. Further improvements are certainly possible, particularly in terms of parallelization for its use in High Performance Computing environments for even greater computing speeds. Nevertheless, the current pipeline is dependent on its individual components and would thus require those to be made compatible with parallelization beforehand.

The initial hits recovered by a BLAST search using the original PIA opsin dataset recovered 11 putative opsin isoforms for *H. azteca* and 51 for *D. pulex*. Our pipeline removed 54.5% and 23.5% of those as spurious hits (Table 4) based on the chosen branch length thresholds. These thresholds can easily be adjusted for increased/decreased conservativeness if deemed necessary, which should be assessed on a gene-to-gene basis. The final phylogenetic inference took this a step further by classifying these opsins in statistically supported functional clades (Figs. 1, 2), which allowed for the determination of their putative photoreceptive roles. Both *H. azteca* and *D. pulex* transcriptomes were generated from whole organism RNA extractions. As opsins are known to function in various cells and tissues of arthropods, as well as the retina (e.g., Lampel et al., 2005), it is likely that the opsin groups identified here are expressed across several tissue types. Nonvisual opsins can be readily identified via phylogenetic inference provided that appropriate reference sequences are included in the multiple sequence alignments. HMM alignments were also used as a secondary source of evidence to confirm protein identities as well as to compare with the results of the phylogenetic annotation. HHMs were able to pair most putative visual opsins to the lateral compound eye opsins of arthropods for both species and, in the case of *D. pulex,* specifically to *Daphnia* class A rhodopsins. While there were a few discrepancies among annotation methods with regard to visual opsins (r-opsins) and melanopsins, this could be explained by their common origin (Porter et al., 2012). Melanopsins are very similar to the r-opsins found in invertebrates (Provencio et al., 1998, 2000) and can couple to similar signaling cascades (Isoldi et al., 2005; Panda et al., 2005; Qiu et al., 2005). In fact, the similarities between these opsin types are evident in our phylogenetic trees (Figs. 1, 2), showing a well-supported clade of arthropod UV opsins nested within the melanopsin clade. Partial sequences or existing database errors could also be a contributing factor to which BLAST and HMM approaches are more sensitive. Even though the HMM searches were not able to determine the functional classification of the opsins in terms of spectral sensitivity, they were confirmed as visual opsins (Table 5). Our results further support the notion that integrated annotation methods are advantageous and recommended to confirm the robustness of findings and annotations.

Opsin repertoire and spectral sensitivities

There are several subgroups of rhabdomeric visual opsins responsible for vision in crustaceans, each with characteristic absorption spectra when bound to a

Table 4 Number of genes and respective isoforms, as defined by Trinity, recovered for each type of opsin in *Hyalella azteca* and *Daphnia pulex*

Species	Short-wavelength sensitive/UV		Long-wavelength-sensitive		Peropsins/Neuropsins/ Encephalopsins		Melanopsins		Opsin-like transcripts (Outgroups)	
	Genes	Isoforms	Genes	Isoforms	Genes	Isoforms	Genes	Isoforms	Genes	Isoforms
Hyalella azteca	1	2	3	3	0	0	0	0	1	1
Daphnia pulex	5	15	3	24	0	0	2	4	1	1

Table 5 Identification and classification of *Hyalella azteca* and *Daphnia pulex* opsins based on phylogenetic inference and HMM profile alignments

				HHBlits w/ Uniclust30				
Species	**Sequence ID**	**Phylogeny Clade**	**Visual**	**Prob.**	**E-value**	**P-value**	**Score**	**Hit ID**
Hyalella azteca	Haz_DN255_c1_g1_i1	OUTGROUP		100.0	1.2E-66	4.3E-72	483.4	A0A0P5Y4K7_9CRUS Putative Tachykinin peptides receptor 99D (Fragment) OS=Daphnia magna PE=3 SV=1
	Haz_DN34101_c1_g4_i2	SWS		100.0	1.2E-77	4.3E-83	540.3	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Haz_DN34101_c1_g6_i4	SWS		100.0	1.9E-76	6.9E-82	541.7	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Haz_DN31505_c2_g2_i1	LWS		100.0	8.8E-80	3.1E-85	550.9	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Haz_DN31505_c2_g4_i1	LWS		100.0	3.8E-80	1.4E-85	555.0	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Haz_DN44548_c0_g1_i1	LWS		100.0	6.9E-72	2.5E-77	497.4	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
Daphnia pulex	Dpu_DN105553_c0_g1_i1	MELANOPSIN		100.0	8E-52	2.8E-57	336.2	H0UT82_CAVPO Uncharacterized protein OS=Cavia porcellus GN=GALR1 PE=3 SV=1
	Dpu_DN16067_c1_g1_i1	LWS		100.0	6.3E-74	2.2E-79	524.1	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN16472_c1_g1_i1	OUTGROUP		100.0	1.8E-73	6.2E-79	520.2	A0A0P5Y4K7_9CRUS Putative Tachykinin peptides receptor 99D (Fragment) OS=Daphnia magna PE=3 SV=1
	Dpu_DN21293_c0_g1_i1	MELANOPSIN		100.0	3.4E-40	1.3E-45	291.9	A0A0P5EI05_9CRUS Class a rhodopsin g-protein coupled receptor gprop2 OS=Daphnia magna PE=4 SV=1
	Dpu_DN21293_c0_g2_i1	MELANOPSIN		100.0	3.6E-62	1.3E-67	446.6	A0A1A6GZZ2_NEOLE Uncharacterized protein OS=Neotoma lepida GN=A6R68_00284 PE=4 SV=1
	Dpu_DN21293_c0_g2_i2	MELANOPSIN		100.0	3.9E-38	1.4E-43	282.8	A0A0F8BMY2_LARCR Melanopsin-B OS=Larimichthys crocea GN=EH28_08950 PE=3 SV=1
	Dpu_DN22657_c1_g4_i1	SWS/UV		100.0	1.7E-79	6.1E-85	557.5	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN23719_c0_g2_i1	SWS/UV		100.0	1.9E-78	6.8E-84	549.4	A0A0P5RD30_9CRUS Class a rhodopsin g-protein coupled receptor gprop2 OS=Daphnia magna PE=3 SV=1
	Dpu_DN23719_c0_g2_i2	SWS/UV		100.0	4.5E-78	1.6E-83	544.5	A0A0P5RD30_9CRUS Class a rhodopsin g-protein coupled receptor gprop2 OS=Daphnia magna PE=3 SV=1
	Dpu_DN23719_c0_g3_i1	SWS/UV		100.0	2E-78	7.4E-84	548.4	A0A0P5RD30_9CRUS Class a rhodopsin g-protein coupled receptor gprop2 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25251_c0_g2_i1	SWS/UV		100.0	2.6E-77	9.3E-83	532.3	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25251_c0_g2_i3	SWS/UV		100.0	3.4E-78	1.2E-83	547.0	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25251_c0_g2_i4	SWS/UV		100.0	1.9E-77	6.8E-83	533.3	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25251_c0_g2_i5	SWS/UV		100.0	8.2E-77	2.9E-82	529.8	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25251_c0_g3_i1	SWS/UV		100.0	2.6E-77	9.1E-83	537.5	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25293_c2_g1_i1	SWS/UV		100.0	5.7E-48	2.2E-53	330.3	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25293_c2_g1_i3	SWS/UV		100.0	2E-78	7.1E-84	548.9	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25293_c3_g1_i1	SWS/UV		100.0	5.7E-35	2E-40	227.3	D0E2W5_CHICK Uncharacterized protein OS=Gallus gallus GN=OPNVA PE=2 SV=1
	Dpu_DN25293_c4_g1_i1	SWS/UV		100.0	1.1E-75	3.9E-81	538.2	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25293_c7_g1_i1	SWS/UV		100.0	1.4E-33	5.1E-39	219.0	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN25377_c0_g1_i1	LWS		100.0	6.3E-75	2.2E-80	517.7	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g1_i10	LWS		100.0	3.1E-62	1.1E-67	425.5	A0A0P5KGY5_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=4 SV=1
	Dpu_DN25377_c0_g1_i11	LWS		100.0	4.1E-76	1.4E-81	528.5	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g1_i12	LWS		100.0	1.4E-75	4.9E-81	523.3	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g1_i4	LWS		100.0	1E-56	3.6E-62	386.9	A0A0P5KGY5_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=4 SV=1
	Dpu_DN25377_c0_g1_i5	LWS		100.0	5.3E-82	1.8E-87	558.9	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g1_i6	LWS		100.0	8E-77	2.8E-82	534.7	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g1_i7	LWS		100.0	2E-76	6.8E-82	529.2	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g1_i8	LWS		100.0	1.1E-80	3.7E-86	552.3	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g1_i9	LWS		100.0	4E-77	1.4E-82	533.7	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g2_i1	LWS		100.0	4.5E-76	1.6E-81	528.0	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g2_i11	LWS		99.9	8.6E-29	3.3E-34	200.2	E9FX22_DAPPU Octopamine receptor beta-2-like protein OS=Daphnia pulex GN=DAPPUDRAFT_347041 PE=3 SV=1
	Dpu_DN25377_c0_g2_i2	LWS		100.0	1.7E-75	6E-81	524.9	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g2_i3	LWS		100.0	1.2E-75	4.2E-81	525.7	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g2_i5	LWS		100.0	4.4E-76	1.5E-81	529.3	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g2_i6	LWS		100.0	4.2E-77	1.4E-82	535.1	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g2_i8	LWS		100.0	9.3E-76	3.2E-81	523.9	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25377_c0_g2_i9	LWS		100.0	4.3E-76	1.5E-81	528.5	A0A0P5USN7_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25506_c1_g1_i1	LWS		100.0	8.9E-35	3.2E-40	222.0	E6ZHB1_DICLA Beta-1 adrenergic receptor OS=Dicentrarchus labrax GN=ADRB1 PE=3 SV=1
	Dpu_DN25506_c1_g1_i2	LWS		100.0	1.1E-34	3.9E-40	221.4	E6ZHB1_DICLA Beta-1 adrenergic receptor OS=Dicentrarchus labrax GN=ADRB1 PE=3 SV=1
	Dpu_DN25506_c3_g1_i1	LWS		100.0	4.3E-80	1.4E-85	521.3	A0A0P4XHY0_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=3 SV=1
	Dpu_DN25506_c3_g1_i3	LWS		100.0	1.1E-60	4E-66	401.9	A0A0P5KGY5_9CRUS Class a rhodopsin g-protein coupled receptor gprop1 OS=Daphnia magna PE=4 SV=1
	Dpu_DN25506_c4_g1_i1	LWS		100.0	2.3E-48	8.2E-54	312.8	OPSL_LIMPO Lateral eye opsin OS=Limulus polyphemus PE=1 SV=1
	Dpu_DN27815_c0_g1_i1	SWS/UV		100.0	9.5E-71	3.4E-76	496.1	A0A0P5RD30_9CRUS Class a rhodopsin g-protein coupled receptor gprop2 OS=Daphnia magna PE=3 SV=1

Putative visual opsins are marked in black, under the "Visual" column, when both methodologies are in agreement and in gray when they differ

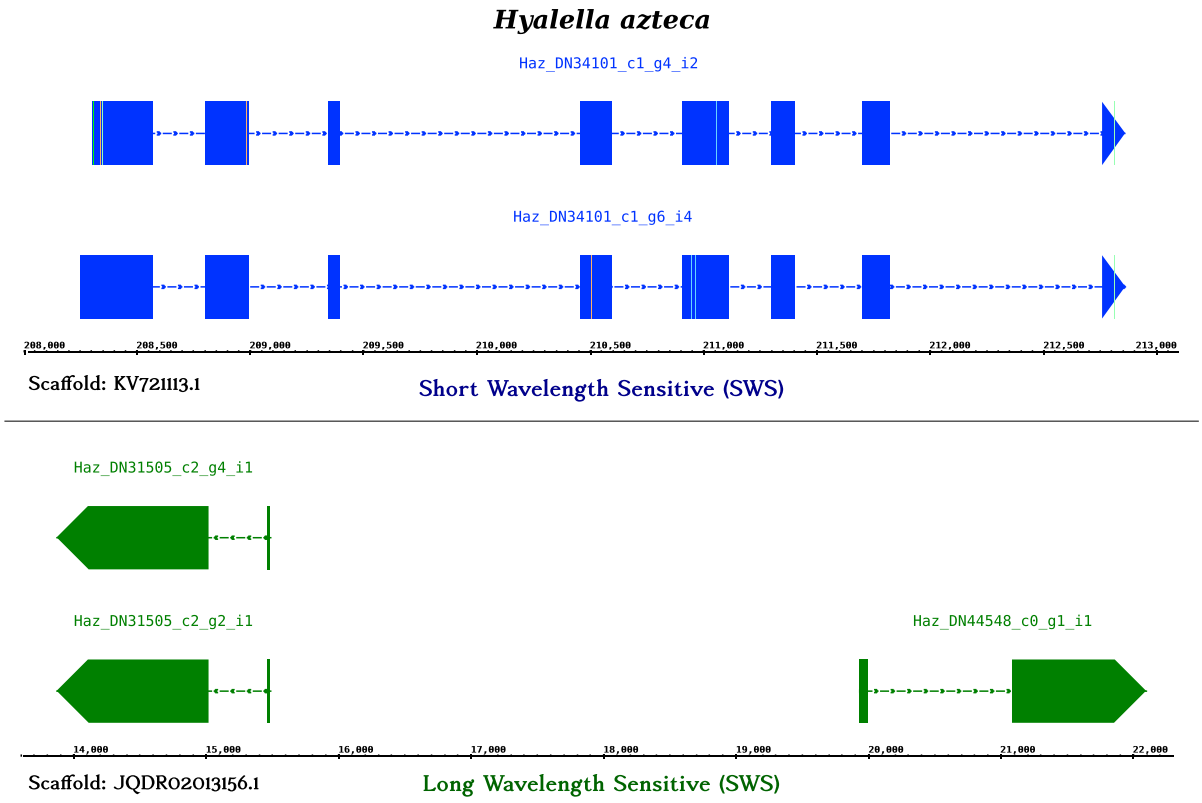


Fig. 3 Intron–Exon gene structure patterns of representative *Hyalella azteca* SWS/UV and LWS opsins. A SAM alignment file with all of the mapped transcripts is provided in the GitHub respository: https://github.com/xibalbanus/PIA2

chromophore (Kashiyama et al., 2009; Henze & Oakley, 2015). The number and type of opsins found throughout Crustacea can range greatly, partially owing to differences in methodologies—with no homologs found in freshwater *Bathynellacea* (Kim et al., 2017), one or two SWS visual opsins found in species of deep-sea shrimp (Wong et al., 2015) and brachyuran crabs (Sakamoto et al., 1996), and as many as 33 identified in stomatopods (Porter et al., 2009, 2013). The number of opsins and corresponding spectral sensitivity of an organism appear to correlate with its life-history, habitat, and the ecological niche it may occupy (Marshall et al., 2015; Stieb et al., 2017). This study represents the first transcriptomic exploration of *H. azteca's* opsin repertoire, which revealed several putative visual opsins (Fig. 2; Table 4). *Hyalella azteca* is a freshwater epibenthic amphipod commonly used as a bioindicator species. Though further evidence is required to make inferences regarding the expression and functionality of these putative opsins, the ability to differentiate between short and long wavelengths would allow *H. azteca* to discern between direct and reflected light from the benthos. Direct sunlight (or moonlight) tends to be abundant in short-wavelengths (< 450 nm) whereas reflected light from sources like leaves and sediment tends to be shifted toward longer (> 450 nm) wavelengths (Menzel, 1979). Our analyses revealed four distinct opsin genes (one SWS/UV and three LWS) expressed in its transcriptome, and suggests that *H. azteca* may be capable of discriminating between the aforementioned light sources. The authors hypothesize that if *H. azteca* does possess functional SWS and LWS visual opsins, this distinction could serve as an important environmental cue influencing their response to a variety of abiotic and biotic factors (e.g., refugia, vegetation, predators, prey). However, the authors note that additional studies incorporating electroretinographic analyses are needed to confirm if *H. azteca* can indeed discriminate between different wavelengths of light as the transcriptomic data suggests. Fewer opsins were found to be expressed in *H.*

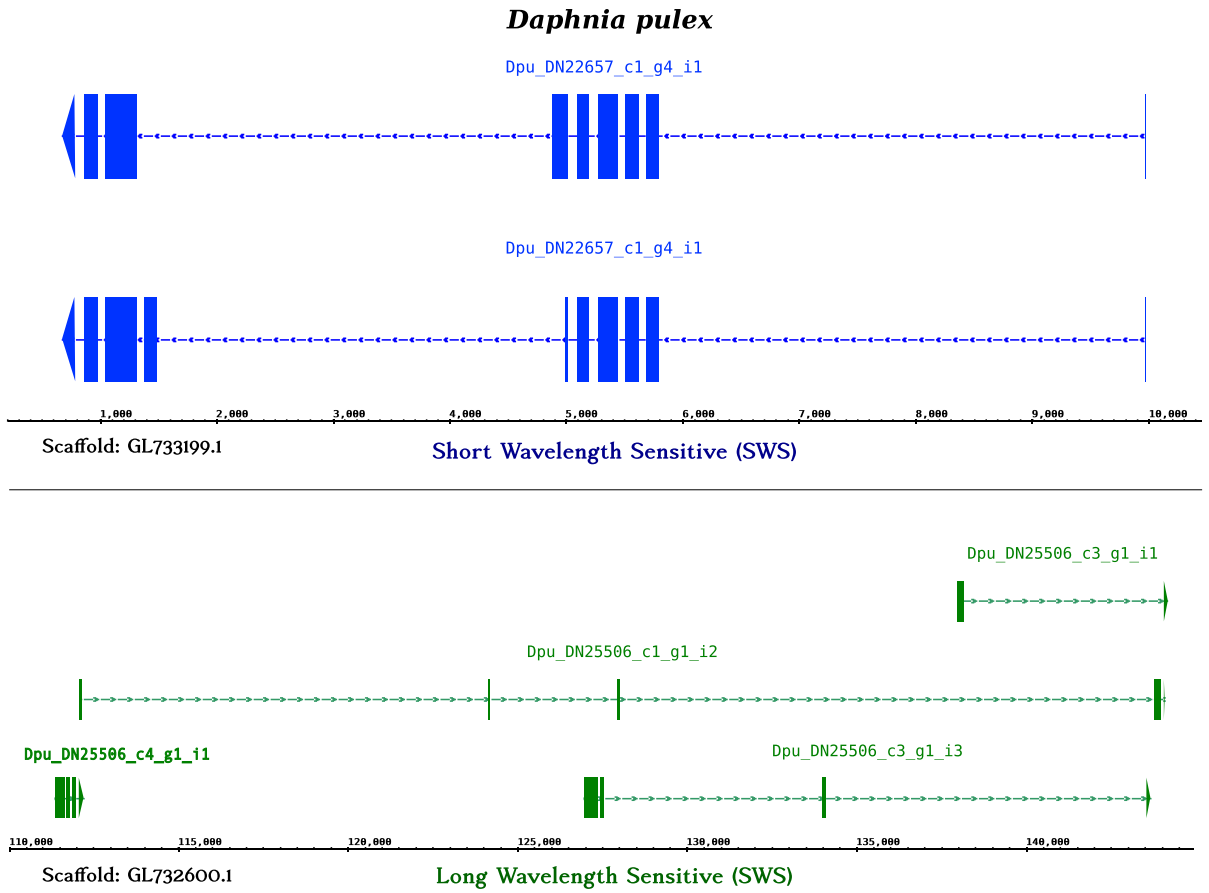


Fig. 4 Intron–Exon gene structure patterns of representative *Daphnia pulex* opsins SWS/UV and LWS opsins, which are arranged in the genome in distinct patterns according to opsin type. A SAM alignment file with all of the mapped transcripts is provided in the GitHub respository: https://github.com/xibalbanus/PIA2

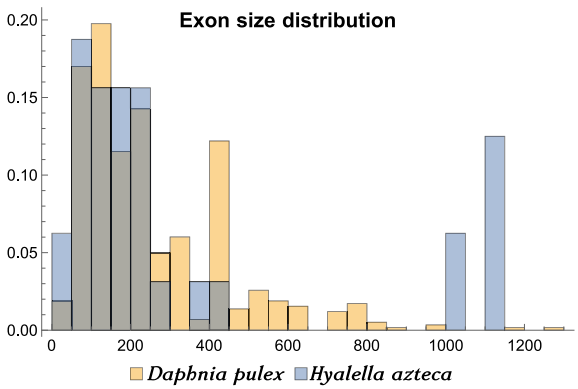


Fig. 5 Exon size distribution for both *Daphnia pulex* and *Hyalella azteca*. *X*-axis is in base pair units (bp), while *Y*-axis represents proportion of transcripts in said range

azteca compared to *D. pulex,* which is not surprising given the expansive opsin repertoire previously described for *Daphnia* (Colbourne et al., 2011; Brandon et al., 2017). *Daphnia* has both simple and compound eyes, which may contribute to the relatively large number of opsin isoforms expressed. Differences have been found in the number and type of opsin genes expressed among eye forms within the ostracod *Skogsberia lerneri* (Oakley & Huber, 2004) and similarly hypothesized for *Daphnia* (Brandon et al., 2017).

The subset of identified rhabdomeric opsins expressed in the *D. pulex* transcriptome allows for comparisons to prior studies characterizing the range of opsin types found in the *D. pulex* genome (Colbourne et al., 2011; Brandon et al. 2017). Colbourne et al. (2011) reported 25 medium- (MWS) and long-wavelength-sensitive (LWS) opsin genes as present in the *D. pulex* genome, but only 3 LWS opsin genes (and 24 isoforms) were identified in our

analyses. While it is possible that the additional opsin classes identified in previous genomic investigations were not expressed in the current *D. pulex* dataset, it is also possible these discrepancies are due to differences in classification schemes across Arthropoda, with 'blue-green' wavelengths currently grouped under SWS. An alternative explanation is that separate genes are being considered isoforms of each other by Trinity during the de novo assembly process. Considering the large number of "isoforms" and low number of "genes" identified in the *D. pulex* transcriptome, in contrast with previous genomic investigations (e.g., Brandon et al., 2017), this is likely a contributing factor to this observed discrepancy.

Genomic architecture and opsin gene duplications

Gene duplications play a fundamental role in genome evolution (Ohno, 1970; Kondrashov et al., 2002), with replicates occasionally evolving new biological functions (Zhang, 2003; Pegueroles et al., 2013). Some of these genome duplications may result in pseudogenes, loci whose nucleotide sequences are similar to a normal gene but that do not produce a functional product when translated. The "unprocessed" pseudogenes, can have all the normal parts of a protein-coding gene, but generally are nonfunctional due to coding errors (Lynch & Force, 1999). Occasionally, so-called "processed" pseudogenes lack the noncoding introns present in the original gene, and are thought to arise from mRNA reinserted into the genome by reverse transcription (Betrán & Long, 2002). Some of these "retrogenes" have been found to be actively transcribed, and the RNA product can be further processed to give two different molecules of RNA of smaller size that form elaborate secondary structures. These RNA regulatory molecules can control a variety of key genes involved in the regulation of the cell cycle and in cell growth (Tutar, 2012; Wen et al., 2012). Opsin genes with few or no introns, such as the LWS opsins our analyses identified in *H. azteca* (Fig. 3), have evolved in various metazoans (including crustaceans) and are thought to be functional photoproteins (Morris et al., 1993; Fitzgibbon et al., 1995; Porter et al., 2007; Liegertová et al., 2015), although it has been postulated that the expression of retrotransposed opsins is a form of transcriptional noise and a byproduct of transcriptional activity in the new genomic region (Xu et al., 2016). Opsin diversification and photopigment evolution seems to have been driven by duplicated opsin genes (e.g., Frentiu et al., 2007; Briscoe et al., 2010), as is the case of both ocular and extraocular cnidarian photoreceptors (Liegertová et al., 2015). Likewise, a functional LWS retrogene was recently found in the arthropod *Helicoverpa armigera,* although expression was believed to be under temporal compartmentalization and primarily expressed in larval stages (Xu et al., 2016). Our results provide further evidence supporting the importance of retrogenes in the evolution of the opsin gene family.

Concluding remarks

Our results support the use of integrative phylogenetic annotation in place of exclusively similarity-based approaches. This is an often overlooked but especially important consideration for the study of protein families (e.g., GPCRs) known for having large numbers of isoforms, multiple duplication events, low sequence similarities, and various combinations of highly conserved domains with hypervariable regions. Phylogenetic approaches are not only able to robustly evaluate homology in an evolutionary context, but they can also provide valuable functional information based on recovered branching patterns. In the case of opsins, this functional information can be insightful from a variety of perspectives, and aid in the formulation and testing of organismal, ecological, and evolutionary hypotheses. Many of these will be put to the test in the present and forthcoming genomic era, for which efficient and scalable methodologies and pipelines will be paramount.

Acknowledgements The authors would like to thank Megan Porter for access to her compilation of reference opsin data, Daniel Speiser and Todd Oakley for allowing us to modify the original PIA tool, and Katherine Dougan for advice during the preparation of this manuscript. JPM was supported by the Philip M. Smith Graduate Research Grant for Cave and Karst Research from the Cave Research Foundation, The Crustacean Society Scholarship in Graduate Studies, and Florida International University's Dissertation Year Fellowship. This work was partially funded by two grants awarded from the National Science Foundation: Doctoral Dissertation Improvement Grant (#1701835) awarded to JPM and HBG, and the Division of Environmental Biology Bioluminescence and Vision grant (DEB-1556059) awarded to HBG. FP acknowledges project CHALLENGEN (CTM2013-48163) of the Spanish Government and a postdoctoral contract funded by the Beatriu de Pinos Programme of the Generalitat de Catalunya (2014-

BPB-00038). The authors would like to thank the Instructional & Research Computing Center (IRCC) at Florida International University for providing High-Performance Computing resources that have contributed to the research results reported within this article. This is contribution #92 of the Marine Education and Research Center of the Institute for the Water and the Environment at the Florida International University.

References

Afgan, E., D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. Von Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko & J. Goecks, 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research 44: W3–W10.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers & D. J. Lipman, 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Andrews, S., 2010. FastQC A Quality Control tool for High Throughput Sequence Data. [available on internet at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/]

Anisimova, M., M. Gil, J.-F. Dufayard, C. Dessimoz & O. Gascuel, 2011. Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. Systematic Biology 60: 685–699.

Arendt, D., K. Tessmar, M.-I. M. de Campos-Baptista, A. Dorresteijn & J. Wittbrodt, 2002. Development of pigment-cup eyes in the polychaete Platynereis dumerilii and evolutionary conservation of larval eyes in Bilateria. Development 129: 1143–1154.

Arendt, D., K. Tessmar-Raible, H. Snyman, A. W. Dorresteijn & J. Wittbrodt, 2004. Ciliary photoreceptors with a vertebrate-type opsin in an invertebrate brain. Science 306: 869–871.

Betrán, E. & M. Long, 2002. Expansion of genome coding regions by acquisition of new genes. Genetica 115: 65–80.

Biscontin, A., E. Frigato, G. Sales, G. M. Mazzotta, M. Teschke, C. De Pittà, S. Jarman, B. Meyer, R. Costa & C. Bertolucci, 2016. The opsin repertoire of the Antarctic krill *Euphausia superba*. Marine Genomics 29: 61–68.

Bok, M. J., M. L. Porter & D.-E. Nilsson, 2017. Phototransduction in fan worm radiolar eyes. Current Biology 27: R681–R701.

Bolger, A. M., M. Lohse & B. Usadel, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Brandon, C. S., M. J. Greenwold & J. L. Dudycha, 2017. Ancient and recent duplications support functional diversity of *Daphnia* opsins. Journal of Molecular Evolution 84: 12–28.

Briscoe, A. D., S. M. Bybee, G. D. Bernard, F. Yuan, M. P. Sison-Mangus, R. D. Reed, A. D. Warren, J. Llorente-Bousquets & C.-C. Chiao, 2010. Positive selection of a duplicated UV-sensitive visual pigment coincides with wing pigment evolution in *Heliconius* butterflies. Proceedings of the National Academy of Sciences 107: 3628–3633.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski & M. J. L. de Hoon, 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423.

Colbourne, J. K., M. E. Pfrender, D. Gilbert, W. K. Thomas, A. Tucker, T. H. Oakley, S. Tokishita, A. Aerts, G. J. Arnold, M. K. Basu, D. J. Bauer, C. E. Caceres, L. Carmel, C. Casola, J.-H. Choi, J. C. Detter, Q. Dong, S. Dusheyko, B. D. Eads, T. Frohlich, K. A. Geiler-Samerotte, D. Gerlach, P. Hatcher, S. Jogdeo, J. Krijgsveld, E. V. Kriventseva, D. Kultz, C. Laforsch, E. Lindquist, J. Lopez, J. R. Manak, J. Muller, J. Pangilinan, R. P. Patwardhan, S. Pitluck, E. J. Pritham, A. Rechtsteiner, M. Rho, I. B. Rogozin, O. Sakarya, A. Salamov, S. Schaack, H. Shapiro, Y. Shiga, C. Skalitzky, Z. Smith, A. Souvorov, W. Sung, Z. Tang, D. Tsuchiya, H. Tu, H. Vos, M. Wang, Y. I. Wolf, H. Yamagata, T. Yamada, Y. Ye, J. R. Shaw, J. Andrews, T. J. Crease, H. Tang, S. M. Lucas, H. M. Robertson, P. Bork, E. V. Koonin, E. M. Zdobnov, I. V. Grigoriev, M. Lynch & J. L. Boore, 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331: 555–561.

Crisp, M. & L. Cook, 2005. Do early branching lineages signify ancestral traits? Trends in Ecology & Evolution 20: 122–128.

Eddy, S. R., 2011. Accelerated Profile HMM Searches. PLoS computational biology 7: e1002195–e1002195.

Edgar, R. C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461.

Engelhardt, B. E., M. I. Jordan, S. T. Repo & S. E. Brenner, 2009. Phylogenetic molecular function annotation. Journal of Physics: Conference Series 180: 012024.

Feuda, R., O. Rota-Stabelli, T. H. Oakley & D. Pisani, 2014. The Comb Jelly Opsins and the Origins of Animal Phototransduction. Genome Biology and Evolution 6: 1964–1971.Feuda, R., F. Marlétaz, M. A. Bentley, P. W.H. Holland, 2016. Conservation, Duplication, and Divergence of Five Opsin Genes in Insect Evolution. Genome Biology and Evolution 8: 579–587.

Fitzgibbon, J., A. Hope, S. J. Slobodyanyuk, J. Bellingham, J. K. Bowmaker & D. M. Hunt, 1995. The rhodopsin-encoding gene of bony fish lacks introns. Gene 164: 273–277.

Freese, N. H., D. C. Norris & A. E. Loraine, 2016. Integrated genome browser: visual analytics platform for genomics. Bioinformatics 32: 2089–2095.

Frentiu, F. D., G. D. Bernard, M. P. Sison-Mangus, A. Van Zandt Brower & A. D. Briscoe, 2007. Gene duplication is an evolutionary mechanism for expanding spectral diversity in the long-wavelength photopigments of butterflies. Molecular Biology and Evolution 24: 2016–2028.

Fryxel, K. J. & E. M. Meyerowitz, 1991. The evolution of rhodopsins and neurotransmitter receptors. Journal of Molecular Evolution 33: 367–378.

Gaudet, P., M. S. Livstone, S. E. Lewis & P. D. Thomas, 2011. Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. Briefings in Bioinformatics 12: 449–462.

Gonzalez, E. R. & L. Watling, 2002. Redescription of Hyalella azteca from Its type locality, Vera Cruz, Mexico

(Amphipoda:Hyalellidae). Journal of Crustacean Biology 22: 173–183.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. a Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, & A. Regev, 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29: 644–652.

Gühmann, M., H. Jia, N. Randel, C. Verasztó, L. A. Bezares-Calderón, N. K. Michiels, S. Yokoyama & G. Jékely, 2015. Spectral Tuning of Phototaxis by a Go-Opsin in the Rhabdomeric Eyes of *Platynereis*. Current Biology 25: 2265–2271.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk & O. Gascuel, 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59: 307–321.

Haas, B. J., A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. Leduc, N. Friedman & A. Regev, 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8: 1494–1512.

Henze, M. J. & T. H. Oakley, 2015. The Dynamic Evolutionary History of Pancrustacean Eyes and Opsins. Integrative and Comparative Biology 55: 830–842.

Hering, L. & G. Mayer, 2014. Analysis of the opsin repertoire in the tardigrade *hypsibius dujardini* provides insights into the evolution of opsin genes in panarthropoda. Genome Biology and Evolution 6: 2380–2391.

Imai, H., D. Kojima, T. Oura, S. Tachibanaki, A. Terakita & Y. Shichida, 1997. Single amino acid residue as a functional determinant of rod and cone visual pigments. Proceedings of the National Academy of Sciences 94: 2322–2326.

Isoldi, M. C., M. D. Rollag, A. M. de Lauro Castrucci & I. Provencio, 2005. Rhabdomeric phototransduction initiated by the vertebrate photopigment melanopsin. Proceedings of the National Academy of Sciences 102: 1217–1221.

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

Kashiyama, K., T. Seki, H. Numata & S. G. Goto, 2009. Molecular characterization of visual pigments in branchiopoda and the evolution of opsins in arthropoda. Molecular Biology and Evolution 26: 299–311.

Katti, C., K. Kempler, M. L. Porter, A. Legg, R. Gonzalez, E. Garcia-Rivera, D. Dugger & B.-A. Battelle, 2010. Opsin co-expression in *Limulus* photoreceptors: differential regulation by light and a circadian clock. Journal of Experimental Biology 213: 2589–2601.

Kim, D., B. Langmead & S. L. Salzberg, 2010. HISAT: a fast spliced aligner with low memory requirements. Nature Methods 12: 357–360.

Kim, B.-M., S. Kang, D.-H. Ahn, J.-H. Kim, I. Ahn, C.-W. Lee, J.-L. Cho, G.-S. Min & H. Park, 2017. First insights into the subterranean crustacean *Bathynellacea transcriptome*: transcriptionally reduced opsin repertoire and evidence of conserved homeostasis regulatory mechanisms. PloS One 12: e0170424.

Kojima, D., A. Terakita, T. Ishikawa, Y. Tsukahara, A. Maeda & Y. Shichida, 1997. A novel Go-mediated phototransduction cascade in scallop visual cells. Journal of Biological Chemistry 272: 22979–22982.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf, & E. V. Koonin, 2002. Selection in the evolution of gene duplications. Genome Biology 3: research0008–1.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander & D. Haussler, 1994. Hidden Markov Models in Computational Biology. Molecular Biology 235: 1501–1531.

Kuwayama, S., H. Imai, T. Hirano, A. Terakita & Y. Shichida, 2002. Conserved Proline Residue at Position 189 in Cone Visual Pigments as a Determinant of Molecular Properties Different from Rhodopsins. Biochemistry 41: 15245–15252.

Lampel, J., A. D. Briscoe & L. T. Wasserthal, 2005. Expression of UV-, blue-, long-wavelength-sensitive opsins and melatonin in extraretinal photoreceptors of the optic lobes of hawkmoths. Cell and Tissue Research 321: 443–458.

Le, S. Q. & O. Gascuel, 2008. An Improved General Amino Acid Replacement Matrix. Molecular Biology and Evolution 25: 1307–1320.

Liegertová, M., J. Pergner, I. Kozmiková, P. Fabian, A. R. Pombinho, H. Strnad, J. Pačes, Č. Vlček, P. Bartůněk & Z. Kozmik, 2015. Cubozoan genome illuminates functional diversification of opsins and photoreceptor evolution. Scientific Reports 5: 11885.

Lynch, M. & A. Force, 1999. The Probability of Duplicate Gene Preservation by Subfunctionalization. Genetics 154: 459–473.

Marshall, J., K. L. Carleton & T. Cronin, 2015. Colour vision in marine organisms. Current Opinion in Neurobiology 34: 86–94.

Matsumoto, T. & Y. Ishibashi, 2016. Sequence analysis and expression patterns of opsin genes in the longtooth grouper *Epinephelus bruneus*. Fisheries Science 82: 17–27.

Menzel, R., 1979. Spectral Sensitivity and Color Vision in Invertebrates In Autrum, H. (ed), Comparative Physiology and Evolution of Vision in Invertebrates. Springer Berlin Heidelberg, Berlin, Heidelberg: 503–580.

Minh, B. Q., M. A. T. Nguyen & A. von Haeseler, 2013. Ultrafast Approximation for Phylogenetic Bootstrap. Molecular Biology and Evolution 30: 1188–1195.

Mirdita, M., L. von den Driesch, C. Galiez, M. J. Martin, J. Söding & M. Steinegger, 2017. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Research 45: D170–D176.

Mirzadegan, T., G. Benkö, S. Filipek & K. Palczewski, 2003. Sequence Analyses of G-Protein-Coupled Receptors: Similarities to Rhodopsin. Biochemistry 42: 2759–2767.

Morris, A., J. K. Bowmaker & D. M. Hunt, 1993. The molecular basis of a spectral shift in the rhodopsins of two species of squid from different photic environments. Proceedings of the Royal Society B: Biological Sciences 254: 233–240.

Nathans, J., 1987. Molecular biology of visual pigments. Annual review of neuroscience 10: 163–194.

Nguyen, L.-T., H. A. Schmidt, A. von Haeseler & B. Q. Minh, 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32: 268–274.

Nordström, K., T. A. Larsson & D. Larhammar, 2004. Extensive duplications of phototransduction genes in early vertebrate evolution correlate with block (chromosome) duplications. Genomics 83: 852–872.

Oakley, T. H. & D. R. Huber, 2004. Differential Expression of Duplicated Opsin Genes in Two EyeTypes of Ostracod Crustaceans. Journal of Molecular Evolution 59: 239–249.

Oakley, T. H., M. A. Alexandrou, R. Ngo, M. S. Pankey, C. K. Churchill, W. Chen & K. B. Lopker, 2014. Osiris: accessible and reproducible phylogenetic and phylogenomic analyses within the Galaxy workflow management system. BMC Bioinformatics 15: 230.

Ohno, S., 1970. Evolution by gene duplication. George Allen and Unwin, London.

Panda, S., S. K. Nayak, B. Campo, J. R. Walker, J. B. Hogenesch & T. Jegla, 2005. Illumination of the Melanopsin Signaling Pathway. Science 307: 600–604.

Passamaneck, Y. J., N. Furchheim, A. Hejnol, M. Q. Martindale & C. Lüter, 2011. Ciliary photoreceptors in the cerebral eyes of a protostome larva. EvoDevo 2: 6.

Pearson, W. R., 2013. An Introduction to Sequence Similarity ("Homology") Searching In Baxevanis, A. D., G. A. Petsko, L. D. Stein, & G. D. Stormo (eds), Current Protocols in Bioinformatics. Wiley, Hoboken.

Pegueroles, C., S. Laurie & M. M. Albà, 2013. Accelerated Evolution after Gene Duplication: A Time-Dependent Process Affecting Just One Copy. Molecular Biology and Evolution 30: 1830–1842.

Porter, M. L., T. W. Cronin, D. A. McClellan & K. A. Crandall, 2007. Molecular Characterization of Crustacean Visual Pigments and the Evolution of Pancrustacean Opsins. Molecular Biology and Evolution 24: 253–268.

Porter, M. L., M. J. Bok, P. R. Robinson & T. W. Cronin, 2009. Molecular diversity of visual pigments in Stomatopoda (Crustacea). Visual Neuroscience 26: 255–265.

Porter, M. L., J. R. Blasic, M. J. Bok, E. G. Cameron, T. Pringle, T. W. Cronin & P. R. Robinson, 2012. Shedding new light on opsin evolution. Proceedings of the Royal Society B: Biological Sciences 279: 3–14.

Porter, M. L., D. I. Speiser, A. K. Zaharoff, R. L. Caldwell, T. W. Cronin & T. H. Oakley, 2013. The Evolution of Complexity in the Visual Systems of Stomatopods: Insights from Transcriptomics. Integrative and Comparative Biology 53: 39–49.

Provencio, I., G. Jiang, W. J. De Grip, W. PÄR HAYES, & M. D. Rollag, 1998. Melanopsin: An opsin in melanophores, brain, and eye. Proceedings of the National Academy of Sciences of the United States of America 95: 340–345.

Provencio, I., I. R. Rodriguez, G. Jiang, W. P. Hayes, E. F. Moreira & M. D. Rollag, 2000. A novel human opsin in the inner retina. Journal of Neuroscience 20: 600–605.

Qiu, X., T. Kumbalasiri, S. M. Carlson, K. Y. Wong, V. Krishna, I. Provencio & D. M. Berson, 2005. Induction of photosensitivity by heterologous expression of melanopsin. Nature 433: 745–749.

Ramirez, M. D., A. N. Pairett, M. S. Pankey, J. M. Serb, D. I. Speiser, A. J. Swafford & T. H. Oakley, 2016. The Last Common Ancestor of Most Bilaterian Animals Possessed at Least Nine Opsins. Genome Biology and Evolution 8: 3640–3652.

Rasmussen, T. K. & T. Krink, 2003. Improved Hidden Markov Model training for multiple sequence alignment by a particle swarm optimization—evolutionary algorithm hybrid. Biosystems 72: 5–17.

Remmert, M., A. Biegert, A. Hauser & J. Söding, 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods 9: 173–175.

Sakamoto, K., O. Hisatomi, F. Tokunaga & E. Eguchi, 1996. Two opsins from the compound eye of the crab Hemigrapsus sanguineus. Journal of Experimental Biology 199: 441–450.

Shichida, Y. & T. Matsuyama, 2009. Evolution of opsins and phototransduction. Philosophical Transactions of the Royal Society B: Biological Sciences 364: 2881–2895.

Simão, F. A., R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva & E. M. Zdobnov, 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Sjölander, K., 2004. Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 20: 170–179.

Smith-Unna, R., C. Boursnell, R. Patro, J. M. Hibberd & S. Kelly, 2016. TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Research 26: 1134–1144.

Soubrier, J., M. Steel, M. S. Y. Lee, C. Der Sarkissian, S. Guindon, S. Y. W. Ho & A. Cooper, 2012. The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates. Molecular Biology and Evolution 29: 3345–3358.

Speiser, D. I., M. Pankey, A. K. Zaharoff, B. a Battelle, H. D. Bracken-Grissom, J. W. Breinholt, S. M. Bybee, T. W. Cronin, A. Garm, A. R. Lindgren, N. H. Patel, M. L. Porter, M. E. Protas, A. S. Rivera, J. M. Serb, K. S. Zigler, K. a Crandall, & T. H. Oakley, 2014. Using phylogenetically-informed annotation (PIA) to search for light-interacting genes in transcriptomes from non-model organisms. BMC Bioinformatics 15: 350–350.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Stieb, S. M., F. Cortesi, L. Sueess, K. L. Carleton, W. Salzburger & N. J. Marshall, 2017. Why UV vision and red vision are important for damselfish (Pomacentridae): structural and expression variation in opsin genes. Molecular Ecology 26: 1323–1342.

Suzuki, S., T. Ishida, K. Kurokawa & Y. Akiyama, 2012. GHOSTM: A GPU-Accelerated Homology Search Tool for Metagenomics. PLoS ONE 7: e36060.

Terakita, A., 2005. The opsins. Genome biology 6: 213.

Tong, D., N. S. Rozas, T. H. Oakley, J. Mitchell, N. J. Colley & M. J. McFall-Ngai, 2009. Evidence for light perception in a bioluminescent organ. Proceedings of the National Academy of Sciences 106: 9836–9841.

Tsukamoto, H., I.-S. Chen, Y. Kubo & Y. Furutani, 2017. A ciliary opsin in the brain of a marine annelid zooplankton is ultraviolet-sensitive, and the sensitivity is tuned by a single

amino acid residue. Journal of Biological Chemistry 292: 12971–12980.

Tutar, Y., 2012. Pseudogenes. Comparative and Functional Genomics 2012: 1–4.

Waterhouse, R. M., F. Tegenfeldt, J. Li, E. M. Zdobnov & E. V. Kriventseva, 2013. OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research 41: D358–D365.

Wen, Y.-Z., L.-L. Zheng, L.-H. Qu, F. J. Ayala & Z.-R. Lun, 2012. Pseudogenes are not pseudo any more. RNA Biology 9: 27–32.

Wong, J. M., J. L. Pérez-Moreno, T.-Y. Chan, T. M. Frank & H. D. Bracken-Grissom, 2015. Phylogenetic and transcriptomic analyses reveal the evolution of bioluminescence and light detection in marine deep-sea shrimps of the family Oplophoridae (Crustacea: Decapoda). Molecular Phylogenetics and Evolution 83: 278–292.

Xu, P., R. Feuda, B. Lu, H. Xiao, R. I. Graham & K. Wu, 2016. Functional opsin retrogene in nocturnal moth. Mobile DNA 7: 18.

Yamada, K. D., K. Tomii & K. Katoh, 2016. Application of the MAFFT sequence alignment program to large data—re-examination of the usefulness of chained guide trees. Bioinformatics 32: 3246–3251.

Yoon, B.-J., 2009. Hidden Markov models and their applications in biological sequence analysis. Current genomics 10: 402–415.

Zhang, J., 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18: 292–298.

Zhou, X., X.-X. Shen, C. T. Hittinger, & A. Rokas, 2017. Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets. bioRxiv 142323.