University of New Hampshire
University of New Hampshire Scholars' Repository

Master's Theses and Capstones | Student Scholarship

Winter 2018

Phylogenetic Focusing Reveals the Evolution of Eumetazoan Opsins

Curtis Provencher
University of New Hampshire, Durham

Follow this and additional works at: https://scholars.unh.edu/thesis

Recommended Citation
Provencher, Curtis, "Phylogenetic Focusing Reveals the Evolution of Eumetazoan Opsins" (2018). *Master's Theses and Capstones*. 1248.
https://scholars.unh.edu/thesis/1248

This Thesis is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Master's Theses and Capstones by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.

PHYLOGENETIC FOCUSING REVEALS THE EVOLUTION OF EUMETAZOAN OPSINS

BY

CURTIS PROVENCHER

Baccalaureate Degree (BS), University of Vermont, 2016

Submitted to the University of New Hampshire

In Partial Fulfillment of

The Requirements for the Degree of

Master of Science

in

Genetics

December, 2018

This thesis/dissertation was examined and approved in partial fulfillment of the requirements for the degree of Master of Science in Genetics by:

David Plachetzki, Assistant Professor in Department of Molecular, Cellular, and Biomedical Sciences

Matthew MacManes, Assistant Professor in Department of Molecular, Cellular, and Biomedical Sciences

Kelley Thomas, Professor in Department of Molecular, Cellular, and Biomedical Sciences

On September 20th, 2018

Approval signatures are on file with the University of New Hampshire Graduate School.

TABLE OF CONTENTS

LIST OF TABLES…… iv

LIST OF FIGURES…… v

ABSTRACT…….. vi

CHAPTER PAGE

I. INTRODUCTION…… 1

II. METHODS…….. 7

III. RESULTS…… 19

IV. DISCUSSION…… 37

V. LIST OF REFERENCES…… 48

LIST OF TABLES PAGE

Table 1: Bait and Anchor detail…… 9

Table 2: Counts for sequences retained…… 20

Table 3: List of source information for published data used…… 31

Table 4: Programs used and models tested…… 41

LIST OF FIGURES

Figure 1: Phylogenetic Focusing Pipeline…….. 4

Figure 2: Species Tree…… 8

Figure 3: Taxon specific sampling example…… 11

Figure 4: “Total Tree” phylogeny…… 13

Figure 5: “Opsin and Outgroup” phylogeny…… 16

Figure 6: “Opsin must_have_K” phylogeny…… 28

Figure 7: Opsin phylogeny with additional data…….. 36

Figure 8: Opsin phylogeny from PhyloBayes MPI…….. 45

ABSTRACT

PHYLOGENETIC FOCUSING REVEALS THE EVOLUTION OF EUMETAZOAN OPSINS

By

Curtis Provencher

University of New Hampshire

Phylogenetic analyses of gene trees commonly begin by searching large molecular datasets from the taxa of interest using some known query sequence. Resulting sequences that exceed some threshold are then concatenated, aligned, and analyzed phylogenetically. This approach has revealed much about the evolutionary history of gene families, but several problems are apparent. Here we apply a new approach that we call Phylogenetic Focusing that circumvents some issues related to global search strategies. Our approach first circumscribes the largest possible orthogroup containing the gene family of interest and then proceeds to focus in on the gene family of interest based on iterative rounds of phylogenetic analyses. We demonstrate this approach by using the phylogeny of eumetazoan rhodopsin class GPCRs to focus in on a clade containing melatonin receptors, opsins, and other genes. Our results clarify the evolutionary history of eumetazoan rhodopsin class GPCRs, the subclade containing opsins, and provide new hypotheses on the functional significance of these genes in cnidarians.

INTRODUCTION

Light detection in most animals is mediated by the visual pigment protein, opsin (Ovchinnikov 1982; Shichida and Imai 1998; Hardie and Raghu 2001; Arendt 2003). Opsins are a member of the rhodopsin class G-protein protein coupled receptor (GPCR) superfamily and are characterized by having seven transmembrane helices and a lysine residue at position 296 in reference to the bovine rhodopsin sequence (Nathans and Hogness 1983; Hargrave and McDowell 1992; Yokoyama 2000). Lysine 296 serves as the binding site for a light sensitive chromophore which, when bound, forms a Schiff-base linkage triggering a phototransduction cascade (Land and Nilsson 2002; Terakita 2005). Opsins play a key role in the ability to sense light, so understanding the evolutionary history of these proteins is vital to our understanding of the evolution of photoreception and vision in animals.

Opsins have been classified into three major groups: ciliary (c-opsin, used mostly in vertebrate eyes), rhabdomeric (r-opsin, used in the eyes of arthropods, cephalopods and other protostomes), and Go-coupled/RGR/RRH (photoisomerases and related proteins) (Zucker et al. 1985; Arendt et al. 2004; Shichida and Matsuyama 2009). However, studies investigating opsins outside of model organisms have identified new subfamiles such as cnidopsin, pteropsin, chaopsin, and xenopsin, making opsin classification and evolution difficult to elucidate (Plachetzki et al. 2007; Verlarde et al. 2005; Picciani et al. 2018). Studies have shown that the last common ancestor of Bilateria most likely possessed opsins from all three of the major groups (Porter et al. 2011; Ramirez et al. 2016). Yet we know that animals such as cnidarians and other groups that predate Bilateria also display photosensitive behaviors controlled through the usage of

opsins (Singer et al. 1963; Plachetzki et al. 2012; Schnitzler et al. 2012). In 2011, Sweeney et al. found that spectral changes in the water caused by lunar phases is correlated to the mass spawning events seen in coral reefs. Plachetzki et al. in 2012 described how the hydrozoan, *Hydra magnipapillata*, uses opsin-based phototransduction to regulate the firing of the cnidarian specific cnidocyte cells. Cnidarian opsins have been a topic of debate since their discovery (Plachetzki et al. 2007; Suga et al. 2008). Cnidarians are the evolutionary sister to bilaterians, so their opsin complement has a direct bearing on our understanding of the evolution of opsins, and phototransduction in animals (Plachetzki et al. 2007). However, comprehensive studies that address what types of opsins are present in cnidarians have often resulted in poorly supported results based on only a few cnidarian sequences derived from a poor sample of extant taxonomic diversity. Thus, taxon sampling has been a critical impediment in understanding the opsin complement of cnidarians (Dunn et al. 2008; Pick et al. 2010). In order to fill this gap, genome scale datasets from a comprehensive sample of cnidarian taxa is required.

Recent studies have proposed multiple hypotheses regarding opsin evolution (Suga et al. 2008; Feuda et al. 2012; Hering and Mayer 2014; Ramirez et al. 2016; Picciani et al. 2018). While most studies have employed canonical phylogenetic methods based on maximum likelihood and Bayesian approaches, little agreement on the structure of the animal opsin phylogeny has resulted. This confusion can be linked to several critical aspects of previous analyses that often differ. First, many studies lack a large enough sample of cnidarian and early branching metazoan taxa to draw generalizable conclusions, potentially missing important aspects of early opsin evolution. For instance, the first study of cnidarian opsin phylogeny was based on only two genome sequences

(Plachetzki et al. 2007). Therefore, there is a clear need for studies to increase the number of genome scale datasets from early branching taxa, including cnidarians, to address opsin evolution and the origins of metazoan phototransduction. Additionally, the production of a growing number of genome scale datasets inevitably leads to the description of new opsin sequences and clades. However the classification of new data as opsins can often be misleading due to biases in the way the data are handled and confusion on the existence of consistently supported subclades of metazoan opsins (Ramirez et al. 2016; Vöcking et al. 2017).

Lastly, hand-curated opsin datasets may be useful for data exploration, but they are not exhaustive and fail to capture the totality of opsin loci present in genome scale datasets. To improve our understanding of opsin evolution we cannot rely on phylogenies built from a few opsin sequences that have been screened for certain diagnostic features. Lysine 296 (K296) is the classic diagnostic feature used to determine whether a newly found rhodopsin class GPCR is actually an opsin (Tsukamoto and Terakita 2010; Oakley and Speiser 2015). As common as it is to rely on K296 for opsin identification, this practice discards potential in-group opsins that may lack the K296 residue, but are still part of the opsin lineage. Such loci, if present, are part of the story of opsin evolution but are generally not included in analyses.

To circumvent these issues, we have created a methodology termed phylogenetic focusing, in which we circumscribe the largest possible orthogroup of a gene family of interest and, through exhaustive rounds of phylogenetic analyses, focus in on the clade of interest (See Figure 1 for pipeline). First we employ a global search strategy to identify rhodopsin class GPCRs, instead of using preexisting opsin datasets. Curated opsin data

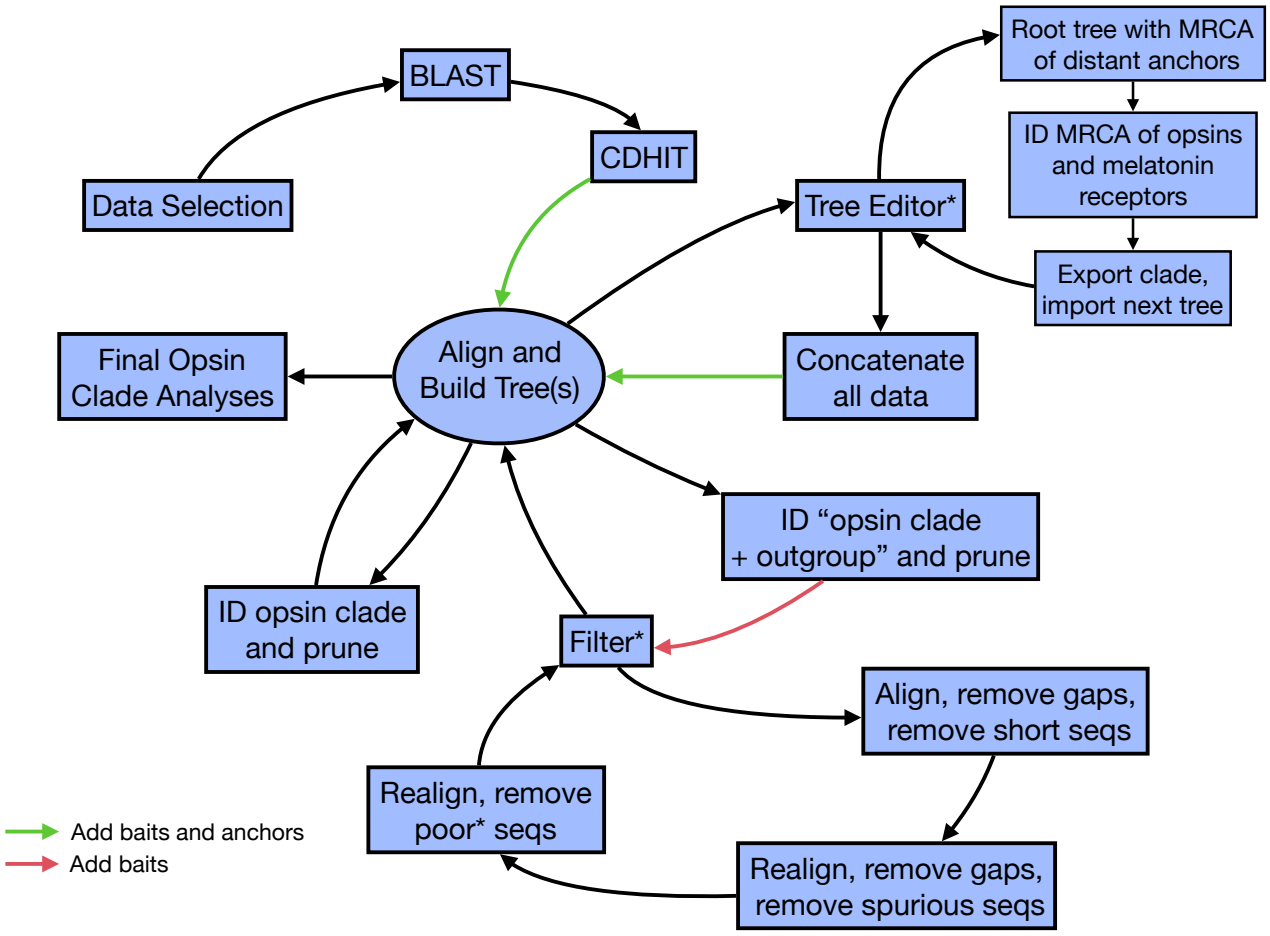

Figure 1. Phylogenetic focusing pipeline. The diagram depicts how the phylogenetic focusing process works starting from "Data Selection" and ending with "Final Opsin Clade Analyses". Green arrows denote when bait and anchor sequences are added into the dataset and the red arrow denotes when bait sequences only are added into the dataset. Each phase is discussed in detail in the methods section.

sets are often incomplete and regularly apply stringent filtering methods to ease the computational burden. Conversely, we start with the largest possible set of GPCR blast hits from our taxa and focus in on the monophyletic opsin clade through iterative rounds of phylogenetic focusing. This allows for opsin sequences to be identified in taxa that have and have not been screened before, for which we have no preconceived notions.

To assist with identification of the opsin clade from the GPCR family, we have gathered a set of well-characterized human and invertebrate GPCRs we refer to as "anchor" sequences. Our clade of interest is the alpha class of rhodopsin-like GPCRs, so our anchors fall into the gamma, beta, and delta classes. The anchors are added to certain datasets just prior to alignment and tree building, which allows us to extract sequences with similar motifs but may be distantly related phylogenetically. Unlike previous opsin studies the addition of anchors allows for us to truly take the largest possible orthogroup from the rhodopsin class GPCR family.

Sequence alignment is a vital part of the phylogenetic process and can be difficult with an abundance of data. We implement two different alignment approaches depending on the amount of data present in the current dataset. In the early phases of phylogenetic focusing, MAFFT v7.305b (Katoh et al. 2013) is used to align sequences due to its accuracy and computational efficiency with a high volume of data. PASTA (Mirarab et al. 2015) is used to align downstream datasets once the clade of interest has been identified via phylogenetic focusing. PASTA is a stepwise alignment program that is a highly accurate progressive extension of MAFFT, but is far too computationally expensive to work on alignments with more than ~1,000 sequences.

Here we use a large set of cnidarian data, along with data from bilaterians, ctenophores, sponges, and placozoans, to clarify the evolution of eumetazoan alpha-rhodopsin class GPCRs. By utilizing our new methodology with genomic and transcriptomic data from 95 taxa, we were able to uncover the largest and least biased representation of metazoan opsin evolution to date. We found strong support for a monophyletic cnidopsin clade, or cnidarian specific opsins, whose existence was previously a topic of debate. We find that the previously described cnidarian opsin class, cnidopsin, (Plachetzki 2007; 2010; 2012) is present in every major class of cnidarians, absent only from the parasitic myxozoan subclade. Additionally, we uncovered two clades of anthozoan specific opsins that appear to be unstable in the phylogeny. One clade appears to be an r-type ortholog present in anthozoans and the other is a larger hexacoral-specific clade that switches topology with adjustments in model selection of the sequences present in the dataset. Furthermore, the newly erected clade xenopsin (Ramirez et al. 2016; Vöcking et al. 2017) present in cnidarians, mollusks and other lophotrochozoans was not recovered as a well-supported monophyletic clade. Instead, we found that these previously reported xenopsin sequences fall into almost every clade of our final opsin phylogeny, indicating the striking polyphyly of this proposed group and likely phylogenetic error in previous analyses. This finding highlights the importance of exhaustive phylogenetic approaches that provide a realistic reflection of extant taxonomic diversity when trying to classify groups of proteins from genome-scale datasets across animals. Our findings are applicable to the phylogenetic analysis of any gene family.

METHODS

1. Data Selection

Genomic or transcriptomic data was collected from 60 cnidarian species to include taxa from every major lineage in the phylum including: Hydrozoa, Scyphozoa, Cubozoa, Staurozoa, Hexacorallia, Octocorallia, and Endocnidozoa. This data set was curated from Kayal et al. (2017), representing the current largest set of cnidarian sequence data. The genomes or transcriptomes of four ctenophores, four sponges, three xenacoelomorphs, one placozoan, four deuterostomes, and nine protostomes were also included into the data set, bringing the total taxon count to 95 species (See Figure 2 for species tree). Our dataset is unique in that it is very well sampled from early branching metazoans, which allows us to better understand the genes that were present before the evolution of opsins.

Well-characterized ciliary, rhabdomeric, and Go-coupled/RGR opsin sequences from taxa with fully annotated genomes were chosen as query sequences for the BLASTp search. Also, we chose a set of more distantly related alpha, beta, gamma, and delta rhodopsin class GPCRs as sequences we termed “anchors”. These sequences were not used in the blast search, but were used later in the pipeline to root phylogenies and to help focus in on the alpha rhodopsin class GPCRs, which contains our focal opsin clade of interest. Accession numbers and additional information on the query and anchor sequences are in Table 1.

Data Preparation

Sequence data was converted into protein space and special stop codon characters were removed. Sequence headers were then modified to match the following format: “>Genus_#” such as “>Nematostella_1”, with the number corresponding to a specific

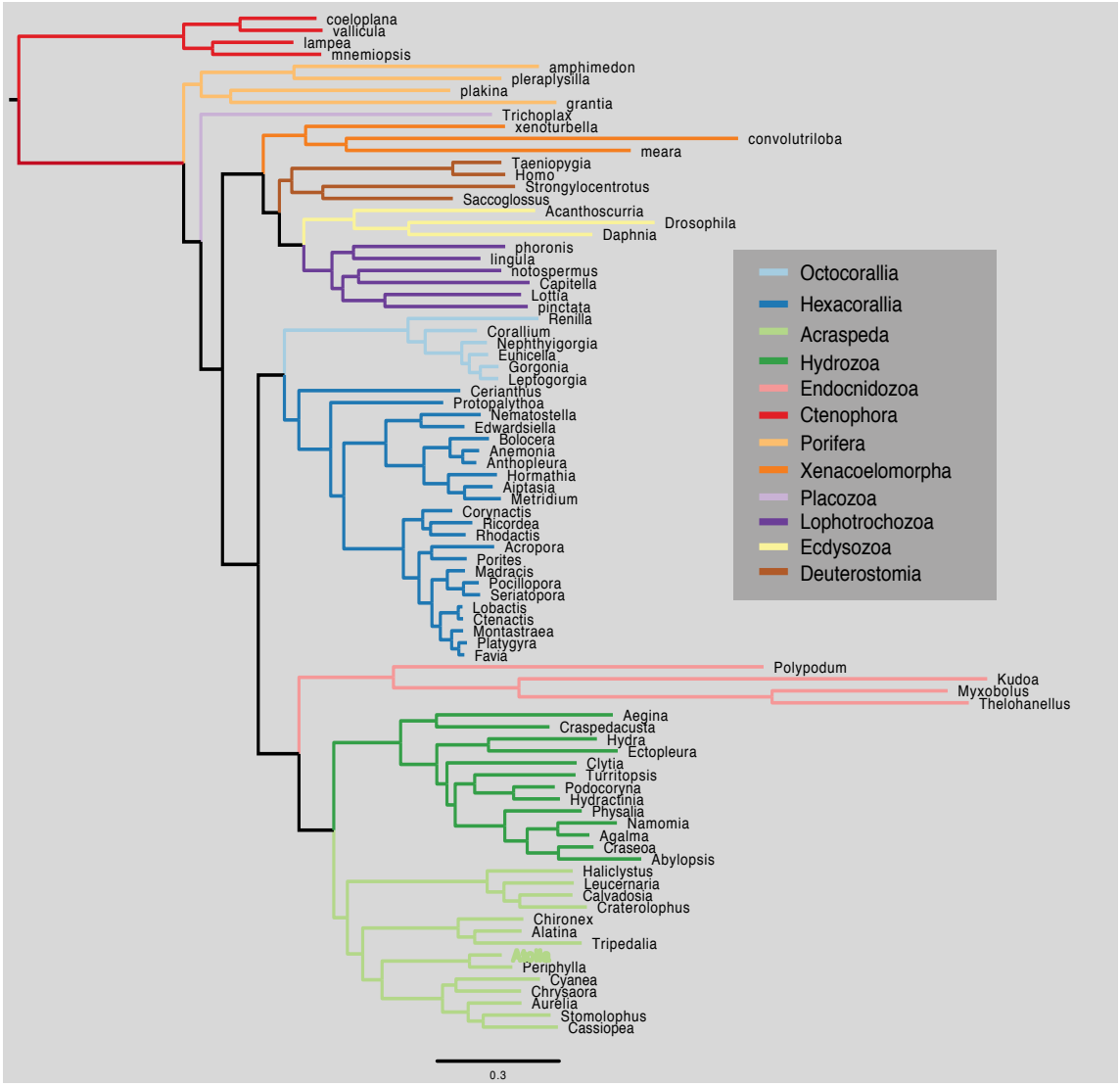

Figure 2. Species tree rooted with ctenophores depicting the relationship of all 95 taxa included in our analyses. Branch color denotes the taxonomic group. Genera are given at tips. For cases where two or more taxa from the same genus were included in the analyses (Hydra, Hydractinia, Myxobolus, Montastrea, Haliclystus, and Eunicella) only one branch was used to represent the genus.

Name Used	NCBI Name	NCBI ACCESSION
Homo_Encephalopsin	opsin-3 [Homo sapiens]	NP_055137.2
Homo_RGR_61744454	RPE-retinal G protein-coupled receptor isoform 2 [Homo sapiens]	NP_001012738.1
Homo_rhodopsin	rhodopsin [Homo sapien]	NP_000530.1
Homo_Long_wave_sensitive_opsin	RecName: Full=Long-wave-sensitive opsin 1; AltName: Full=Red cone photoreceptor pigment; AltName: Full=Red-sensitive opsin; Short=ROP	P04000.1
Homo_RRH	visual pigment-like receptor peropsin [Homo sapiens]	NP_006574.1
Homo_OPN4 (melanopsin)	Opsin 4 [Homo sapiens]	AAI13559.1
Drosophila_Rh1_opsin	neither inactivation nor afterpotential E [Drosophila melanogaster]	NP_524407.1
Gallus_melanopsin	melanopsin [Gallus gallus]	NP_001038118.1
Mus_rhodopsin	rhodopsin [Mus musculus]	AAA63392.1
Gallus_pinopsin	opsin [Gallus gallus]	AAB47565.1
Platyneries_Go_coupled_opsin2	Go coupled opsin 2 [Platynereis dumerilii]	AKS48307.1
Manduca_rhodopsin	RecName: Full=Opsin-3; Short=MANOP3; AltName: Full=Rhodopsin 3, short-wavelength; AltName: Full=Rhodopsin P450	O96107.1
Xenopus_rhodopsin	RecName: Full=Rhodopsin	P29403.1
Euprymna_rhodopsin	opsin [Euprymna scolopes]	ACB05673.1
Platynereis_ciliary	ciliary opsin [Platynereis dumerilii]	AAV63834.1
Helobdella_opsinB	opsin B [Helobdella robusta]	AID66634.1
Octopus_rhodopsin_P313562	RecName: Full=Rhodopsin	P31356.2
Homo_melatoninR	melatonin receptor type 1A [Homo sapiens]	NP_005949.1
Platynereis_melatoninR	melatonin receptor [Platynereis dumerilii]	AIT11923.1
Homo_GPR50	GPR50 protein [Homo Sapien]	AAI03697.1
Homo_Histamine_1_receptor	histamine H1 receptor [Homo sapiens]	NP_000852.1
Homo_GPR21	probable G-protein coupled receptor 21 [Homo sapiens]	NP_005285.1
Homo_GPR52	G-protein coupled receptor 52 [Homo sapiens]	NP_005675.3
Homo_dopamine_receptor	D(2) dopamine receptor isoform long [Homo sapiens]	NP_000786.1
Homo_orexin_receptor_1	orexin receptor type 1 [Homo sapiens]	NP_001516.2
Homo_RFamide_receptor	pyroglutamylated RFamide peptide receptor [Homo sapiens]	NP_937822.2
Homo_neurokinin_receptor	neurokinin A receptor [Homo sapiens]	AAC31760.1
Homo_neuropeptide_FF_receptor	neuropeptide FF receptor 2 isoform 1 [Homo sapiens]	NP_004876.2
Homo_galanin_receptor	galanin receptor type 3 [Homo sapiens]	NP_003605.1
Homo_mu_opioid_receptor_variant	mu opioid receptor variant MOR-1R [Homo sapiens]	AAK74189.1
Homo_somatostatin_receptor	somatostatin receptor [Homo sapiens]	AAA20828.1

Table 1. List of bait and anchor sequences used. NCBI names and accession numbers are included

protein from that species FASTA formatted file. In the case of two or more species from the same genus, the first letter or two from the species name will be added, as such: ">Hydra_m_1" or ">Hydra_vi_1".

3. Phylogenetic Focusing (Pipeline in Figure 1.)

To obtain opsin orthologs from each taxon, BLASTp searches were done using the chosen opsin query sequences as baits, an E-value cutoff of 1e-5, and keeping up to 50 target sequences. Hit sequences were written to new FASTA files for each taxon and put through CD-HIT v4.6 (Fu et al. 2012), removing sequences with 98% or higher redundancy to others in the set. This was done because much of our data were transcriptomes that were previously assembled with Trinity (Grabherr et al. 2011). Such datasets often include pseudoreplicates derived from the assembly process. Our use of CD-HIT v4.6 (Fu et al. 2012) in this way removes such pseudoreplicates. Query (opsins) and anchor sequences (distantly related alpha, beta, gamma, and delta rhodopsin class GPCRs) were then concatenated into the global FASTA file and aligned using MAFFT v7.305b (Katoh et al. 2013). Alignments were converted to phylip format and analyzed phylogenetically using RAxML v8.2.10 (Stamatakis 2014) with the PROTGAMMALGF setting, as LG+GAMMA has been shown as the best model for opsin gene trees when there is not enough data to inform a GTR model (Feuda et al. 2102; Ramirez et al. 2016). This procedure produced a rooted phylogenetic tree containing anchors, query sequences, and putative opsins for each of 95 taxa. The resulting 95 maximum likelihood (ML) trees were then put through a custom R script that works in three steps. The first tree is rooted with the clade containing the most recent common ancestor of the beta, gamma, and delta

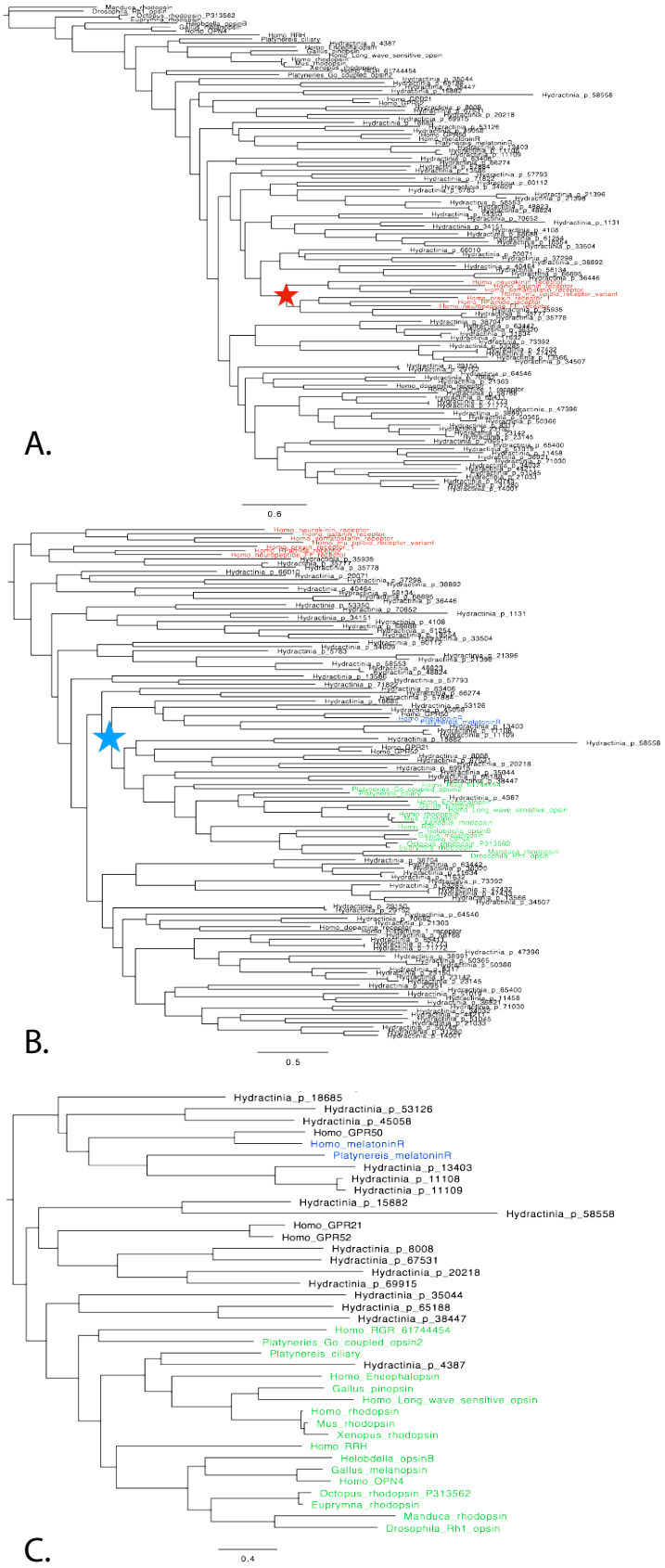

Figure 3. Example of what occurs during the Tree Editor phase of the phylogenetic focusing pipeline. This is done for each taxon, and this example is using *Hydractinia polyclina*. The red tips represent the beta, gamma, and delta anchor sequences, blue tips represent melatonin receptors, and green tips are the opsin bait sequences. **A.** The clade containing the most recent common ancestor (MRCA) of the beta, gamma, and delta anchor sequences is identified (red star) and used to root the tree. **B.** The clade containing the MRCA of melatonin receptors and opsin sequences is identified (blue star). **C.** The clade identified in step B is pruned off and used for downstream analyses.

rhodopsin class GPCR anchor sequences. Next, the clade containing the most recent common ancestor of opsins and melatonin receptors (an alpha anchor) is identified, pruned off, exported, and the next gene tree is imported. Melatonin receptors are commonly used as the outgroup to opsins (Feuda et al 2014; Hering and Mayer 2014). See figure 3 for a visual representation of what occurs in the tree editor script. In essence, this is phylogenetic focusing; progressively discarding sequences, as one gets closer to the focal gene family. Many of the sequences that were discarded may appear to be closely related through sequence similarity, but are phylogenetically quite distantly related. This allows us to zoom in on the clade and sequences of interest.

In total, 3,899 sequences made it through the R script from the 95 taxa and were concatenated together, forming what we termed the "total tree" data set. Query and anchor sequences were added into this data set. Alignment was then done using MAFFT v7.305b (Katoh et al. 2013) and gap sites were masked out using trimAl v1.4 (Capella-Gutiérrez et al. 2009) with the gap threshold set to 0.2. The alignment was then converted to phylip format and analyzed phylogenetically using RAxML v8.2.10 (Stamatakis 2014) with the PROTGAMMAGTR model. The GTR, or general time reversible, model has been shown to be the model that best estimates opsin evolution, but only when enough data is provided (Gatto et al. 2007). Without enough data to inform the model analyses can become over-parameterized, leading to erroneous phylogenetic signal.

Phylogenetic focusing continues by identifying the clade of interest (opsin and outgroup) from the first concatenated tree. The ML "total tree" (Figure 4) was rooted the same way as each taxon's gene tree, with the clade containing the most recent common ancestor of the beta, gamma, and delta rhodopsin class GPCR anchor sequences. Rooting

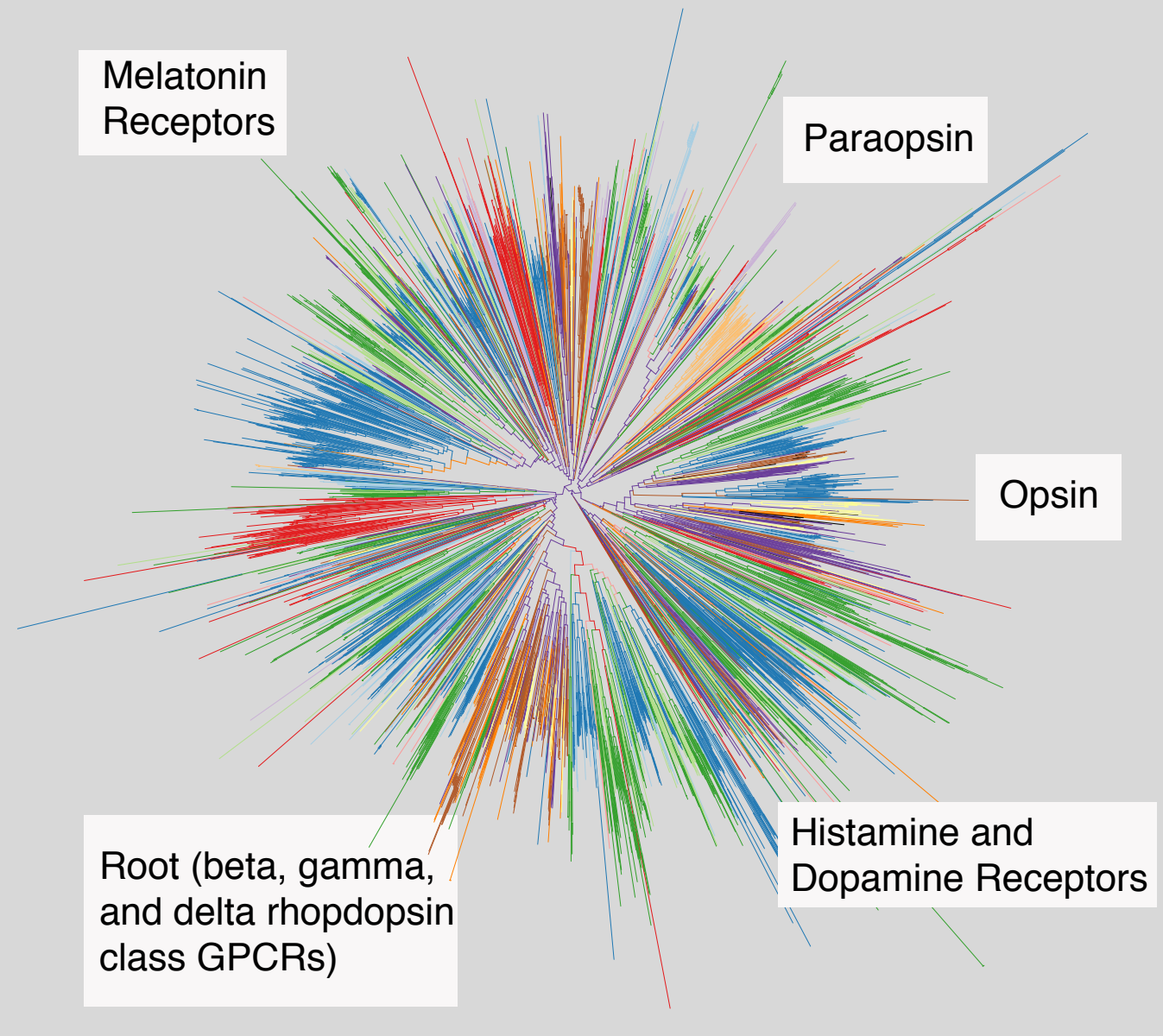

Figure 4. Maximum likelihood phylogeny formed in RAxML under the GTR model from the "Total Tree" dataset, which is the concatenation of all 95 species gene trees output from the R Tree Editor script. Branch color denotes the phylum or class of the taxa each gene was identified from. The tree is rooted with the clade containing our anchor sequences with the rest of the clades consisting of alpha rhodopsin class GPCRs. Melatonin receptors and Dopamine + Histamine receptors make up a monophyletic clade which is sister to the "Opsin and Outgroup" clade, which is pruned off for further analyses.

this way resulted in a monophyletic clade of alpha rhodopsin class GPCRs such as melatonin, histamine, and dopamine receptors as sister to a clade containing opsins. Canonical opsins fell out as sisters to a group of sequences we abbreviated as paraopsin. The monophyletic clade containing opsins plus paraopsin (1,049 sequences) was pruned off making the "opsin + outgroup" dataset. At this point filtering was necessary to remove sequences that were short, spurious, pseudogenized, or poorly assembled. Sequences within the first cut were pulled from the original FASTA files, concatenated with the opsin queries, aligned in MAFFT v7.305b (Katoh et al. 2013), and gap sites were masked using trimAl v1.4 (Capella-Gutiérrez et al. 2009). From here, sequences with less than 150 residues were removed for not providing enough informative information after trimming (182 sequences removed). The remaining 867 sequences were again pulled from the original FASTA files, concatenated with the opsin queries, aligned in MAFFT v7.305b (Katoh et al. 2013), and trimAl v1.4 (Capella-Gutiérrez et al. 2009) was used to mask out gap sites and then remove spurious sequences with the resoverlap and seqoverlap thresholds set at 0.55 and 55, respectively. These parameters were determined empirically. 844 sequences passed this threshold and were re-aligned in MAFFT v7.305b (Katoh et al. 2013). Lastly, we removed sequences with long insertions that disrupted the alignment. SEAveiw v4 (Gouy et al. 2010) was used to view the alignment and identify sequences with insertions greater than 25 amino acids to be removed. All but 7 of the 34 sequences removed in this step came from taxa with fully sequenced genomes leading us to believe most of these insertions were probably read-throughs from pseudogenes.

With the “opsin + outgroup” data set filtered, we began the next round of phylogenetic focusing, pruning off the monophyletic opsin clade. This FASTA file, including only those sequences in the monophyletic opsin clade plus its monophyletic sister clade, was aligned using PASTA (Mirarab et al. 2015) and gap sites were masked using PASTA’s run_seqtools package. The alignment was converted to phylip format and initial phylogenetic analyses were conducted using RAxML v8.2.10 (Stamatakis 2014) using the GTR model (Figure 5). The monophyletic paraopsin clade containing sequences from Porifera and Placozoa was used to root the tree allowing for the monophyletic opsin clade to be easily identified and pruned off. 368 sequences were retrieved including the opsin query sequences creating the initial “opsin clade” dataset and PASTA (Mirarab et al. 2015) was used for alignment. From here, the opsin data set underwent a filtering strategy commonly used for opsin identification. All sequences were checked for a lysine present at the retinal-binding site analogous to position 296 of bovine rhodopsin sequence (Nathans and Hogness 1983; Palczewski et al. 2000). Lacking a lysine means the chromophore will be unable to form a covalent bond to the opsin rendering this protein non-photoreceptive. Only 18 of the 368 sequences lacked K296 and were removed from the initial opsin dataset creating the “must_have_K” opsin data set.

The final opsin "must_have_K" data set consisted of 350 sequences with representatives from every group tested except Placozoa and Porifera, which we infer were lost from these taxa (Plachetzki et al. 2007; Feuda et al. 2012). Sequences were aligned using PASTA (Mirarab et al. 2015) and trees were made using the PROTGAMMAGTR and PROTGAMMAAUTO settings with 20 random start positions in RAxML v8.2.10 (Stamatakis 2014). Additionally, IQtree 1.6.0 (Wang et al. 2018) was

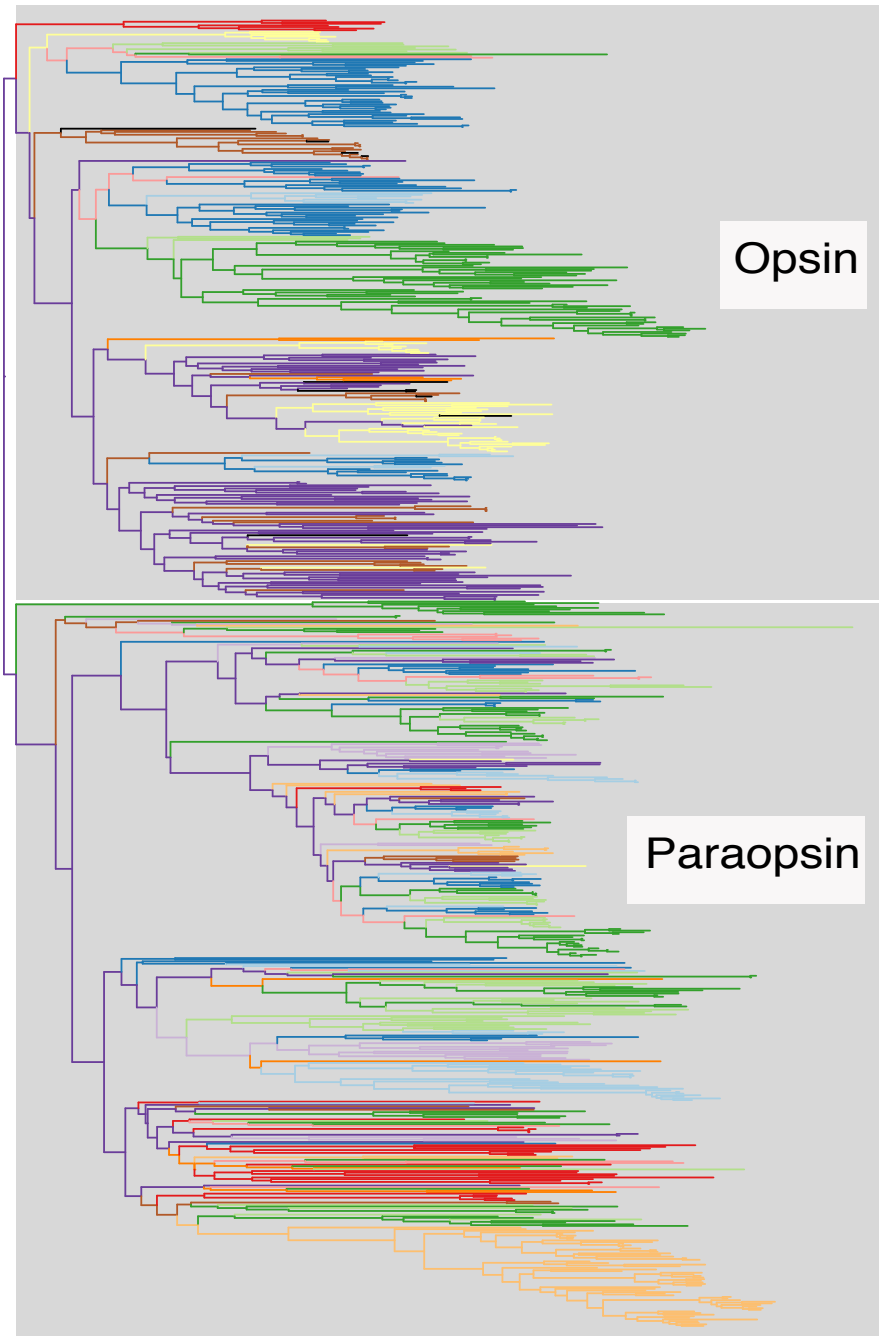

Figure 5. Maximum likelihood phylogeny formed using the GTR model in RAxML from the filtered "Opsin and Outgroup" dataset. Branch color denotes the phylum or class of the taxa each gene was identified from. STO (Sister To Opsins) is a monophyletic clade containing Placopsins (Feuda et al. 2012) that was used to root the tree. Placopsins are commonly used to root opsin phylogenies but due to out search procedure we have uncovered a large clade of sequences from Placozoa, Ctenophora, Porifera, and Cnidaria that for the most part, have not been described before. This clade lacks any human sequences and contains few from other bilaterians. The monophyletic opsin clade is pruned off for further analyses.

implemented because of its ability to create ML trees while considering site heterogeneity. IQtree 1.6.0 (Wang et al. 2018) was run using the GTR20, GTR20+C20, and GTR20+C60 models. The GTR20 model alone is a general time reversible model with 190 rate parameters. Adding +C20 and +C60 provides 20 and 60-profile mixture models, respectively, as variants of the CAT model for ML trees. These models deal with site-specific rate heterogeneity by allowing each position in the alignment to fall into 20 (+C20) or 60 (C60) categories of rate heterogeneity. All IQtree 1.6.0 (Wang et al. 2018) runs were done using –alrt 1000, which specifies 1000 replicates to perform SH-like approximate likelihood ratio test (SH-alrt), which is a single branch stability test (Wang et al. 2018).

General time reversible models, as empirical models, will always provide a strong model fit to the data (Feuda et al. 2012). However, these models fail when taxon sampling is low, causing model parameters to be incorrectly estimated. ModelFinder (Kalyaanamoorthy et al. 2017) was also used, as implemented in IQtree 1.6.0 (Wang et al. 2018) to find the best fitting fixed model according to the –Log likelihood, Akaike information criterion (AIC), the corrected AIC (AICc), and Bayesian information criterion (BIC). From the 546 fixed models tested, the LG+F+R8 model was chosen as the best fit. LG+F+R8 incorporates the LG model of amino acid substitution with a probability-distribution-free model of rate heterogeneity across sites. The benefit to this approach is that the distribution of rates-of-change across sites may take any shape, implying that estimates of rates and weights should be more accurate than those obtained under a gamma distribution. This model is more parameter rich than the gamma model potentially causing issues if not enough data is supplied. We estimated the phylogeny of

the "must_have_K" dataset under the LG+F+R8 model in IQtree 1.6.0 (Wang et al. 2018) with alrt support and bootstrapping (Hoang et al. 2018). We also estimated the phylogeny of the "must_have_K" dataset using Phylobayes MPI (Lartillot et al. 2013), which utilizes the GTR-CAT+ Γ.

4. Additional Data Sets

Cnidarians such as the cubozan *Tripedalia cystophora* and the hydrozoans *Cladonema radiatum* and *Podocoryna carnea* have been studied for possessing eyespots and genes involved in their development and photosensitivity have been identified (Suga et al. 2008; Koyanagi et al. 2008; Bielecki et al. 2014). We did not uncover any of the cubozoan ocular genes, most likely due to the poor quality of the transcriptomes used. However, in order to understand where these genes fall on the opsin phylogeny we made an additional data set using our opsin "must_have_K" set and including 36 published genes from the three taxa just mentioned. This dataset was called the “ocular” set and was aligned in PASTA (Mirarab et al. 2015) and a phylogeny was build using the GTR20+C20 model in IQTree 1.6.0 (Wang et al. 2018).

We failed to recover the xenopsin clade (Ramirez et al. 2016; Vöcking et al. 2017) in any of our analyses under any of our models. To explore this finding in greater depth we also build an additional dataset that concatenated 56 previously described xenopsin sequences from Vöcking et al. 2017 and Ramirez et al. 2016 to the "must_have_K" dataset. These sequences are derived from the lophotrochozoan *Lottia gigantea* and the anthozoan *Nematostella vectensis*. This “xenopsin” data set was treated the same way as the ocular with regards to alignment and tree building.

RESULTS

1. Initial Search and filtering

46,366 sequences were obtained from the initial blast search, averaging about 488 per taxon. CD-HIT v4.6 (Fu et al. 2012) removed roughly 85% of these sequences bringing the total count to 6,355. These sequences represent a non-redundant set of the alpha rhodopsin class GPCRs present in each taxon. Due to the repeating transmembrane domain motif and relatively short length, it is likely that some distantly related GPCRs were also identified as blast hits. It is important to remove as many of these distant GPCRs as possible before concatenating the data sets together for the best possible alignment. To achieve this, gene trees were made for each taxon and an R script was used to root each tree with a set of human sequences termed "anchors" which fall outside of the alpha class of rhodopsin-like GPCRs. The sequences that fell in between the anchors and melatonin receptors were not kept, as melatonin receptors are an accepted outgroup to opsins. An average of 65% of the sequences (3,868 in total) generated after the initial search and filtering steps were included in the "opsin + melatonin receptor" clade and were kept for further analysis. See Table 2 for further information on how many sequences were kept for each taxon throughout the analyses.

Total Tree

To remove any distantly related GPCR hits that managed to pass through filtering, the anchor sequences were included into the "total tree" dataset. Rooting with the anchors results in a topology that is similar to that of other GPCR evolution studies (Stevens et al. 2013). 89.8% of the remaining sequences (3502/3899) fell into a clade consisting of

Taxa	# of initial blast search hits	# seqs after cdhit	# seqs after R tree editor	# seqs in first cut	# seqs in opsin clade	# seqs in opsinclade musthavK
Abylopsis	208	44	37	12	0	0
Acanthoscurria	311	27	20	4	2	2
Acropora	857	130	95	14	13	13
Aegina	50	11	9	2	0	0
Agalma	797	97	42	27	12	10
Aiptasia	850	130	87	14	8	8
Alatina	744	79	41	10	0	0
Amphimedon	274	78	78	78	0	0
Anemonia	79	17	11	3	0	0
Anthopleura	847	102	15	14	11	11
Atolla	208	30	24	9	0	0
Aurelia	822	87	42	3	2	2
Bolocera	146	28	18	1	0	0
Brachionus	405	49	13	12	2	2
Calvadosia	420	46	46	13	1	1
Capitella	823	156	30	13	9	9
Cassiopea	375	52	48	9	0	0
Cerianthus	44	11	10	4	0	0
Chironex	647	51	34	8	0	0
Chrysaora	275	35	12	5	0	0
Clytia	497	49	43	1	1	1
Coeloplana	670	72	72	13	1	1
Convolutriloba	531	56	42	14	5	5
Corallium	823	87	29	9	5	5
Corynactis	850	87	75	12	11	11
Craseoa	484	87	72	16	3	1
Craspedacusta	795	93	57	21	10	10
Crassostrea	826	95	53	18	13	13
Craterolophus	126	14	11	4	0	0
Ctenactis	378	51	47	3	1	1
Cyanea	3	2	0	0	0	0

Daphnia	849	76	55	40	36	36
Drosophila	814	69	41	9	7	7
Ectopleura	661	77	60	16	0	0
Edwardsiella	738	103	9	2	2	2
Eunicella_c	622	89	63	11	0	0
Eunicella_v	339	49	32	2	0	0
Favia	435	59	39	5	2	2
Gorgonia	798	101	81	19	4	4
Grantia	17	1	1	1	0	0
Haliclystus_a	365	36	36	13	3	3
Haliclystus_s	305	42	28	8	1	1
Homo	828	91	53	11	11	11
Hormathia	102	19	11	1	0	0
Hydractinia_p	759	89	65	9	3	1
Hydractinia_s	496	77	5	0	0	0
Hydra_m	578	70	36	20	18	17
Hydra_o	130	21	19	7	1	1
Hydra_vi	125	24	18	2	0	0
Hydra_vu	740	117	41	24	11	8
Kudoa	1	1	1	1	0	0
Lampea	145	17	9	5	2	2
Leptogorgia	320	46	1	0	0	0
Leucernaria	789	77	25	15	3	2
Lingula	827	114	30	27	17	16
Lobactis	533	71	47	2	1	1
Lottia	823	115	62	25	14	14
Madracis	831	144	49	8	3	3
Meara	341	52	33	4	0	0
Metridium	109	17	17	1	0	0
Mnemiopsis	447	76	74	17	2	2
Montastraea_c	762	114	114	19	1	1
Montastraea_f	222	48	24	1	0	0
Myxobolus_c	18	3	3	3	0	0

Myxobolus_p	32	4	4	4	0	0
Namomia	258	39	35	16	4	2
Nematostella	858	173	173	28	21	20
Nephthyigorgia	34	6	5	0	0	0
Notospermus	826	90	27	12	6	5
Periphylla	83	18	9	3	0	0
Phoronis	807	82	55	17	7	7
Physalia	524	71	19	7	4	3
Pinctata	801	118	19	15	4	3
Plakina	119	14	14	10	0	0
Platygyra	377	69	26	9	1	1
Pleraplysilla	71	5	5	5	0	0
Pocillopora	572	103	18	3	1	1
Podocoryna	440	63	56	11	2	2
Polypodium	353	36	36	12	3	3
Porites	432	82	82	6	0	0
Protopalythoa	529	79	41	11	1	1
Renilla	775	102	102	23	2	2
Rhodactis	849	97	70	14	12	12
Ricordea	821	114	65	11	4	4
Saccoglossus	830	172	78	9	5	5
Seriatopora	347	61	46	7	1	1
Stomolophus	735	76	55	15	0	0
Strongylocentrotus	840	180	127	12	5	5
Taeniopygia	826	112	61	18	14	14
Trichoplax	831	141	94	31	0	0
Tripedalia	14	2	2	1	0	0
Turritopsis	404	52	34	8	0	0
Vallicula	637	84	83	18	2	2
Xenoturbella	307	52	32	2	0	0
TOTAL	46366	6355	3868	1032	351	333

Table 2. Counts for the number of sequences retained for each taxon during every round of phylogenetic focusing

opsins, melatonin receptors, and histamine + dopamine receptors. These are all alpha class rhodopsin-like GPCRs, which is a positive sign for our search and filtering strategy. Melatonin receptors plus histamine+dopamine receptors form a monophyletic clade that falls out as sister to a clade consisting of canonical opsins (Figure 4). As we are interested in opsin evolution, we did little to further investigate the sequences within the melatonin clade or histamine + dopamine clade. Further research into these sequences may shed light on the evolution of alpha rhodopsin class GPCRs.

First Cut (Opsin + Outgroup)

Focusing in on the opsin clade brings us to the "opsin + outgroup" dataset, consisting of an orthologous clade of opsins plus its evolutionary sister, an additional orthologous clade of opsins that has not been previously described. This group, which we call Paraopsins, is bounded by copious sequence representation from ctenophores, sponges, cnidarians, but very few from Bilateria (Figure 5). This clade that is the sister to opsins will be referred to as Paraopsins from here forward. This Paraopsin clade contains the placopsin sequences from *Trichoplax adhaerens* that were identified by Feuda et al. 2012, and a large group of sequences from the sponge *Amphimedon queenslandica* that were also described in the supplement by Srivastava et al. 2010. Additionally, this clade contains very few echinoderm and protostome sequences and only three chordate sequences from the zebra finch, *Taeniopygia guttata*. The single *Drosophila melanogaster* sequence present in the Paraopsin clade was identified as the Leucine-rich repeat-containing GPCR 1 (Lgr1) protein in FlyBase. Lgr1 is a known rhodopsin-like GPCR transmembrane receptor that binds glycoprotein hormones like follicle-stimulating

hormone, luteinizing hormone, and thyroid-stimulating hormone, but this assignment was based on blast similarity, not phylogenetic analysis (Rocco et al. 2016). Some of the Paraopsin sequences possess a lysine at position 296 in accordance with the bovine rhodopsin sequences (Nathans and Hogness 1983; Palczewski et al. 2000). This suggests that some of these Paraopsin proteins are indeed phototactic and exist outside of the monophyletic opsin clade, and that the lysine at this position has independently evolved on at least two occasions, multiple times. The vast majority of residues in position 296 of Paraopsins are not lysines. It is possible that selected Paraopsin sequences could represent independent origins of opsin-like photosensitivity.

The opsin and Paraopsin dataset contained 1,032 sequences excluding baits. Now that focused into the alpha class of rhodopsin-like GPCRs we no longer require the anchor sequences for filtering. Instead, filtering at this phase is done based on sequence and alignment quality. First, sequences that lack a sufficient span of informative data were removed. Sequences were aligned, gap sites masked out, and sequences with less than 150 amino acids worth of data were removed. The remaining 867 sequences were realigned and the alignment was then checked for spurious sequences using the –seqoverlap and –resoverlap settings set to 55 and 0.55, respectively, which removed 23 sequences. Finally, sequences with insertions greater than 25 amino acids in length that did not align to any other sequences were removed. This step removed 34 sequences, 27 of which came from taxa with fully sequenced genomes, leading us to believe these are likely read throughs from pseudogenes (see methods). These filtering steps removed 239 sequences in total and greatly improved the alignment quality. Traditional methods of alignment scores could not be used to compare pre and post filtering alignments because

sequence composition was not the same, so we used the amount of gap characters as a proxy for alignment quality. The initial "opsin + outgroup" alignment consisted of 40.8% gap characters (# of gap chars/ # of gap chars + # of AAs), but after filtering the amount of gap regions decreased significantly to 20.4%.

Once all filtering was complete, the final set of 810 sequences was aligned and ML trees were made using the PROTGAMMAGTR model in RAxML v8.2.10 (Stamatakis 2014). Rooting this tree with Paraopsins resulted in a monophyletic clade of 368 sequences including the opsin baits, which were pruned off for the next analysis.

Opsin Tree Analyses

To ensure the sequences that made it into the opsin data set were photoreceptive, the alignment was screened for a lysine at position 296 in accordance to the bovine rhodopsin sequence. Screening for lysine 296 is a common practice in opsin phylogenetics, as the lysine is needed for the chromophore to bind triggering a phototransduction cascade (Terakita 2005). Only 18 sequences lacked a lysine in this position and were removed from further analyses. The small number of sequences lacking lysine 296 is a positive sign for our search and filtering procedure showing that the vast majority of sequences that made it into the opsin clade are true photoreceptive opsins.

Choosing the best fitting model for amino acid evolution is a crucial and difficult step in the phylogenetic process. Modeling amino acid evolution correctly depends on the type of sequence data being analyzed, the amount of data provided, and how distantly related the sequences are in both a molecular and temporal sense. Multiple tree building

approaches were taken with the final dataset of 350 sequences to test the effects of model selection, rate heterogeneity, and compositional bias on the topology and support of the opsin phylogeny. RAxML v8.2.10 (Stamatakis 2014) was implemented using the PROTGAMMAAUTO and PROTGAMMAGTR models with 20 ML searches on 20 randomized stepwise addition parsimony trees and 1,000 bootstrap iterations. Both models recovered the same opsin topology and received low gamma based likelihood scores, indicating a good fit. Bootstrap support was low at certain internal nodes separating the opsin classes. Although bootstrap support for internal nodes was low, support for nodes with c-opsin, r-opsin, Go-coupled/RGR, and cnidopsin is relatively high. These bootstrapping results support the placement of sequences within their respective monophyletic clade, but do not provide strong support for the evolutionary relationships among opsin type.

IQTree 1.6.0 (Wang et al. 2018) was implemented in the tree building process for additional support and to help account for site-specific rate heterogeneity in a ML framework. Trees were built using the GTR20, GTR20+C20, and GTR20+C60 in IQTree 1.6.0 (Wang et al. 2018), each with 1,000 replicates of SH-like approximate likelihood ratio test (SH-alrt), which is a single branch stability test. The GTR20 model resulted in a slightly different topology from the other tree building methods, which is unusual, as the GTR20 model should be making similar estimations as the PROTGAMMAGTR model from RAxML v8.2.10 (Stamatakis 2014). However this model did receive the highest AIC, AICc, BIC, and –log likelihood scores out of all the analyses run, leading us to believe it is a poor fit. Adding +C20 and +C60 provides 20 and 60-profile mixture models, respectively, as variants of the CAT model for ML trees. This allows a GTR

model to be used that can account for rate heterogeneity to different capacities. Trees built from these models both resulted in the same topology (Figure 6) with very high alrt support for internal and external nodes. GTR20+C60 did have slightly better AIC, AICc, BIC, and –Log likelihood scores, but both were significantly better than GTR20 alone. 1,000 bootstrap iterations were done using IQTree 1.6.0 (Wang et al. 2018) for the GTR20+C20 tree and scores were much higher than the bootstrapping performed in RAxML v8.2.10 (Stamatakis 2014). Additionally, ModelFinder was implemented through IQTree 1.6.0 (Wang et al. 2018) to test over 500 different substitution model variations to find the one that was best fit for our opsin data set. Based on AIC, AICc, BIC, and –Log likelihood scores, LG+F+R8 was chosen as the best fitting model for having the lowest scores. LG+F+R8 incorporates the LG model of amino acid substitution with a probability-distribution-free model of rate heterogeneity across sites. The benefit to this approach is that the distribution of rates-of-change across sites may take any shape, implying that estimates of rates and weights should be more accurate than those obtained under a gamma distribution. The resulting tree produced the same topology as the RAxML v8.2.10 (Stamatakis 2014) runs and the GTR20+C20 and +C60 runs.

PhyloBayes MPI (Lartillot et al. 2013) was the last program used for tree building to provide a Bayesian Markov chain Monte Carlo (MCMC) approach that uses non-parametric methods for modeling among-site variation. Two PhyloBayes MPI (Lartillot et al. 2013) chains were run in tandem using 24 cores each for over 50 days to achieve the best possible convergence in tree space. Each chain generated over 60,000 trees. A burn-in of 1,000 trees and sub-sampling every 10 trees was done when comparing the

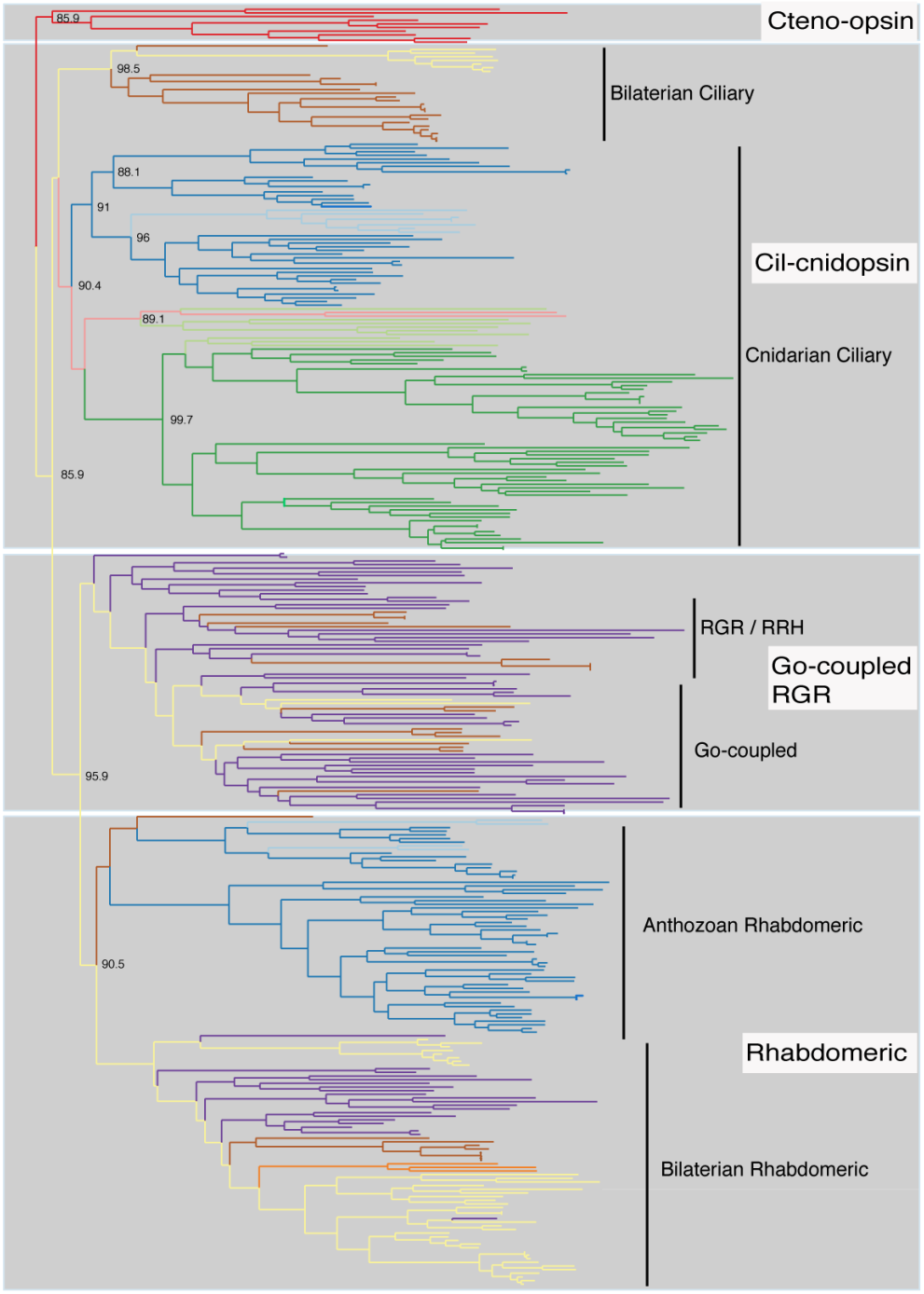

Figure 6. Phylogeny from the opsin clade data set consisting of 350 sequences all possessing K296, formed in IQTree with the GTR20+C20 model. The same topology resulted from using the following models: GTR20+C60, LG+F+R8 (best model identified through ModelTest), PROTGAMMAAUTO in RAxML, and PROTGAMMAGTR in RAxML. This is the topology referred to as "Topology 1" in the results section and is the most supported from this dataset. Branch color denotes the phylum or class of the taxa each gene was identified from.

discrepancies across all bipartitions. The maximum and mean differences across the 2,600 bipartitions were 0.1517 and 0.0074, respectively, and a consensus tree was obtained. A maximum difference less that 0.3 is considered acceptable and less than 0.1 is a good run (Lartillot et al. 2013). Taking this into consideration, our run was clearly in the acceptable range with the chains almost reaching convergence. The PhyloBayes MPI (Lartillot et al. 2013) consensus tree recovered strong support for the c-opsin, r-opsin, Go-coupled/RGR opsin, and cnidopsin clades, similar as all previous analyses. However, the anthozoan r-type opsin clade was split with a quarter of the sequences staying as anthozoan r-type and another group of sequences falling out with the ctenopsins and the xenacoelomorph opsins.

NoTung-2.9 (Stolzer et al. 2012) was used as a gene tree-species tree reconciliation method to gain additional support for rooting the opsin phylogeny with ctenophore opsins. By providing NoTung-2.9 (Stolzer et al. 2012) our opsin gene tree (Figure 6) and the species tree made up of the taxon datasets included in our analysis (Figure 1), the software identified the best location to root the tree based on the most parsimonious evolutionary route. Not surprisingly, NoTung-2.9 (Stolzer et al. 2012) identified ctenophores to be the outgroup for our opsin phylogeny providing the most parsimonious tree, which is a positive sign as the opsin clade in the "opsin and outgroup" tree was also rooted by the ctenophore opsins.

Additional Datasets

Studies investigating the biochemical function of ocular genes in cnidarians who possess eyes have uncovered the function of multiple genes involved in their

development (Suga et al 2008; Bielecki et al. 2014; Liegertová et al. 2015). These genes have been studied in cnidarians like the cubozoan, *Tripedalia cystophora*, and the hydrozoan, *Cladonema radiatum*. Transciptomes from these two species were included in our analyses, but due to poor quality of the cubozoan transcriptomes we did not uncover any of their opsins or ocular genes. However, to discover where these genes fall on our opsin phylogeny we have made an additional dataset consisting of the 350 sequences from our opsin phylogeny and 36 ocular genes with known biochemical function from the cubozoans *Carybdea rastonii* and *Tripedalia cystophora* and the hydrozoans *Podocoryna carnea* and *Cladonema radiatum* (Suga et al. 2008; Bielecki et al. 2014; Liegertová et al. 2015). See Table 3 for source information on sequences included in the additional datasets. Trees were formed using the PROTGAMMAGTR model in RAxML v8.2.10 (Stamatakis 2014) and the GTR20+C20 model in IQTree 1.6.0 (Wang et al. 2018) as previously described. The strategies resulted in slightly different topologies but the ocular genes included preformed the same way for both analyses, falling out into the cnidopsin clade with other taxa from Acraspeda.

Xenopsins were first documented by Ramirez et al. 2016 in a variety of lophotrochozoans and a few cnidarians as a monophyletic clade of opsins being sister to Go-coupled/RGR opsins. Since then additional researchers have continued to identify and classify new xenopsins in cnidarians and lophotrochozoans (Vöcking et al. 2017; Picciani et al. 2018). However the classification of xenopsins is purely phylogenetic and has not been based on any functional or biochemical criteria. We did not recover a monophyletic group that corresponded to xenopsin in any of our analyses of opsin phylogeny, under any model. In order to test the existence of xenopsins, we constructed an additional

Sequence included	Dataset	Accession number	Publication
Carybdea rastonii cubop mRNA for opsin, complete cds	Cnidarian Ocular	AB435549.1	Koyanagi et al. 2008
Podocoryna carnea PcopC mRNA for opsin, complete cds	Cnidarian Ocular	AB332435.1	Koyanagi et al. 2008
tr|A0A059UAP3|A0A059UAP3_TRICY Lens eye opsin (Fragment) OS=Tripedalia cystophora OX=6141 PE=2 SV=1	Cnidarian Ocular	A0A059UAP3	Bielecki et al. 2014
tr|A0A059NTG3|A0A059NTG3_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op4 PE=3 SV=1	Cnidarian Ocular	A0A059NTG3	Liegertová et al. 2015
tr|A0A059NTG8|A0A059NTG8_TRICY C-like opsin (Fragment) OS=Tripedalia cystophora OX=6141 GN=op8 PE=3 SV=1	Cnidarian Ocular	A0A059NTG8	Liegertová et al. 2015
tr|A0A059NTG7|A0A059NTG7_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op13 PE=3 SV=1	Cnidarian Ocular	A0A059NTG7	Liegertová et al. 2015
tr|A0A059NTD7|A0A059NTD7_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op5 PE=3 SV=1	Cnidarian Ocular	A0A059NTD7	Liegertová et al. 2015
tr|A0A059NTD1|A0A059NTD1_TRICY C-like opsin (Fragment) OS=Tripedalia cystophora OX=6141 GN=op17 PE=3	Cnidarian Ocular	A0A059NTD1	Liegertová et al. 2015
tr|A0A059NTG2|A0A059NTG2_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op9 PE=3 SV=1	Cnidarian Ocular	A0A059NTG2	Liegertová et al. 2015
tr|A0A059NTD5|A0A059NTD5_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op15 PE=3 SV=1	Cnidarian Ocular	A0A059NTD5	Liegertová et al. 2015
tr|A0A059NTC7|A0A059NTC7_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op6 PE=3 SV=1	Cnidarian Ocular	A0A059NTC7	Liegertová et al. 2015
tr|A0A059NTC8|A0A059NTC8_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op1 PE=3 SV=1	Cnidarian Ocular	A0A059NTC8	Liegertová et al. 2015
tr|A0A059NTC5|A0A059NTC5_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op16 PE=3 SV=1	Cnidarian Ocular	A0A059NTC5	Liegertová et al. 2015
tr|A0A059NTD2|A0A059NTD2_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op12 PE=3 SV=1	Cnidarian Ocular	A0A059NTD2	Liegertová et al. 2015
tr|A0A059NTG9|A0A059NTG9_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op3 PE=3 SV=1	Cnidarian Ocular	A0A059NTG9	Liegertová et al. 2015
tr|A0A059NTD6|A0A059NTD6_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op10 PE=3 SV=1	Cnidarian Ocular	A0A059NTD6	Liegertová et al. 2015
tr|A0A059NTD4|A0A059NTD4_TRICY C-like opsin OS=Tripedalia cystophora OX=6141 GN=op2 PE=3 SV=1	Cnidarian Ocular	A0A059NTD4	Liegertová et al. 2015
Cladonema radiatum CropB1 mRNA for opsin, complete cds	Cnidarian Ocular	AB332416.1	Suga et al. 2008
Cladonema radiatum CropB4 mRNA for opsin, complete cds	Cnidarian Ocular	AB332417.1	Suga et al. 2008
Cladonema radiatum CropM mRNA for opsin, complete cds	Cnidarian Ocular	AB332418.1	Suga et al. 2008
Cladonema radiatum CropO mRNA for opsin, complete cds	Cnidarian Ocular	AB332419.1	Suga et al. 2008
Cladonema radiatum CropC mRNA for opsin, complete cds	Cnidarian Ocular	AB332420.1	Suga et al. 2008
Cladonema radiatum CropE mRNA for opsin, complete cds	Cnidarian Ocular	AB332421.1	Suga et al. 2008
Cladonema radiatum CropD mRNA for opsin, complete cds	Cnidarian Ocular	AB332422.1	Suga et al. 2008
Cladonema radiatum CropH mRNA for opsin, complete cds	Cnidarian Ocular	AB332423.1	Suga et al. 2008

Cladonema radiatum CropI mRNA for opsin, complete cds	Cnidarian Ocular	AB332424.1	Suga et al. 2008
Cladonema radiatum CropL mRNA for opsin, complete cds	Cnidarian Ocular	AB332425.1	Suga et al. 2008
Cladonema radiatum CropF mRNA for opsin, complete cds	Cnidarian Ocular	AB332426.1	Suga et al. 2008
Cladonema radiatum CropG1 mRNA for opsin, complete cds	Cnidarian Ocular	AB332427.1	Suga et al. 2008
Cladonema radiatum CropG2 mRNA for opsin, complete cds	Cnidarian Ocular	AB332428.1	Suga et al. 2008
Cladonema radiatum CropN1 mRNA for opsin, complete cds	Cnidarian Ocular	AB332429.1	Suga et al. 2008
Cladonema radiatum CropN2 mRNA for opsin, complete cds	Cnidarian Ocular	AB332430.1	Suga et al. 2008
Cladonema radiatum CropK1 mRNA for opsin, complete cds	Cnidarian Ocular	AB332431.1	Suga et al. 2008
Cladonema radiatum CropK2 mRNA for opsin, complete cds	Cnidarian Ocular	AB332432.1	Suga et al. 2008
Cladonema radiatum CropJ mRNA for opsin, complete cds	Cnidarian Ocular	AB332433.1	Suga et al. 2008
Podocoryna carnea PcopB mRNA for opsin, complete cds	Cnidarian Ocular	AB332434.1	Suga et al. 2008
Podocoryna carnea PcopC mRNA for opsin, complete cds	Cnidarian Ocular	AB332435.1	Suga et al. 2008
Lottia gigantea, Uncharacterized protein	Xenopsin	V3Z0E3	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V3ZDT4	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V3ZSU7	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V3ZWI5	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4A259	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4A6Q4	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4AAH1	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4AS98	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4AUU9	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4B0S4	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4C2D5	Ramirez et al. 2016
Lottia gigantea, Uncharacterized protein	Xenopsin	V4CNF1	Ramirez et al. 2016
Lottia gigantea, hypothetical protein LOTGIDRAFT_72363	Xenopsin	XP_009051341.1	Ramirez et al. 2016
Nematostella vectensis, Predicted Protein	Xenopsin	A7RSR1	Ramirez et al. 2016

Nematostella vectensis, Predicted Protein	Xenopsin	A7SQJ5	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMW6	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMW7	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMX0	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMX1	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMX2	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMX3	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMX5	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMX6	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMX7	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMY1	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMY6	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMY7	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMY8	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMZ0	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMZ1	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMZ2	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMZ3	Ramirez et al. 2016
Nematostella vectensis, Opsin	Xenopsin	A9UMZ5	Ramirez et al. 2016
Nematostella vectensis A7RTL7	Xenopsin	A7RTL7	Vocking et al. 2017
Nematostella vectensis A7RVG8	Xenopsin	A7RVG8	Vocking et al. 2017
Nematostella vectensis A7RVG9	Xenopsin	A7RVG9	Vocking et al. 2017
Nematostella vectensis A7S8K8	Xenopsin	A7S8K8	Vocking et al. 2017
Nematostella vectensis A7SN09	Xenopsin	A7SN09	Vocking et al. 2017
Nematostella vectensis A7SN10	Xenopsin	A7SN10	Vocking et al. 2017
Nematostella vectensis A7SN12	Xenopsin	A7SN12	Vocking et al. 2017

Nematostella vectensis A9UMY0	Xenopsin	A9UMY0	Vocking et al. 2017
Nematostella vectensis A9UMY1	Xenopsin	A9UMY1	Vocking et al. 2017
Nematostella vectensis A9UMY2	Xenopsin	A9UMY2	Vocking et al. 2017
Nematostella vectensis A9UMY6	Xenopsin	A9UMY6	Vocking et al. 2017
Nematostella vectensis A9UMY7	Xenopsin	A9UMY7	Vocking et al. 2017
Nematostella vectensis A9UMY8	Xenopsin	A9UMY8	Vocking et al. 2017
Nematostella vectensis A9UMY9	Xenopsin	A9UMY9	Vocking et al. 2017
Nematostella vectensis A9UMZ0	Xenopsin	A9UMZ0	Vocking et al. 2017
Nematostella vectensis A9UMZ1	Xenopsin	A9UMZ1	Vocking et al. 2017
Nematostella vectensis A9UMZ2	Xenopsin	A9UMZ2	Vocking et al. 2017
Nematostella vectensis A9UMZ3	Xenopsin	A9UMZ3	Vocking et al. 2017
Nematostella vectensis A9UMZ4	Xenopsin	A9UMZ4	Vocking et al. 2017
Nematostella vectensis A9UMZ5	Xenopsin	A9UMZ5	Vocking et al. 2017
Nematostella vectensis A9UMZ6	Xenopsin	A9UMZ6	Vocking et al. 2017
Nematostella vectensis opsin A9UMX9	Xenopsin	A9UMX9	Vocking et al. 2017
Nematostella vectensis opsin A9UMY4	Xenopsin	A9UMY4	Vocking et al. 2017
Nematostella vectensis XP 001627311.1	Xenopsin	XP_1627311.1	Vocking et al. 2017
Nematostella vectensis XP 001631194.1	Xenopsin	XP_1631194.1	Vocking et al. 2017
Nematostella vectensis XP 001636803.1	Xenopsin	XP_1636803.1	Vocking et al. 2017

Table 3. Source information for the sequences included into the published xenopsin dataset and the cnidarian ocular gene data set.

dataset from our existing opsin sequences (Figure 6) and included 56 published xenopsins from Ramirez et al. 2016 and Vöcking et al. 2017 (xenopsins identified by Vöcking in *Nematostella vectensis* and xenopsins identified by Ramirez in *Lottia gigantea* and *Nematostella vectensis* were included). Trees were formed using the PROTGAMMAGTR model in RAxML v8.2.10 (Stamatakis 2014) and the GTR20+C20 model in IQTree 1.6.0 (Wang et al. 2018) as previously described. Both strategies resulted in the same topology. *Lottia gigantea* xenopsins from Ramirez fell out into the r-opsin and Go-coulped/RGR clade, and the *Nematostella vectensis* sequences fell into the cnidopsin and anthozoan specific opsin clades of our phylogeny. Sequences from Vöcking fell exclusively into the anthozoan specific opsin clades described previously, not into cnidopsin. See Figure 7 to see where all the published sequences from the additional data sets fell on our opsin phylogeny.

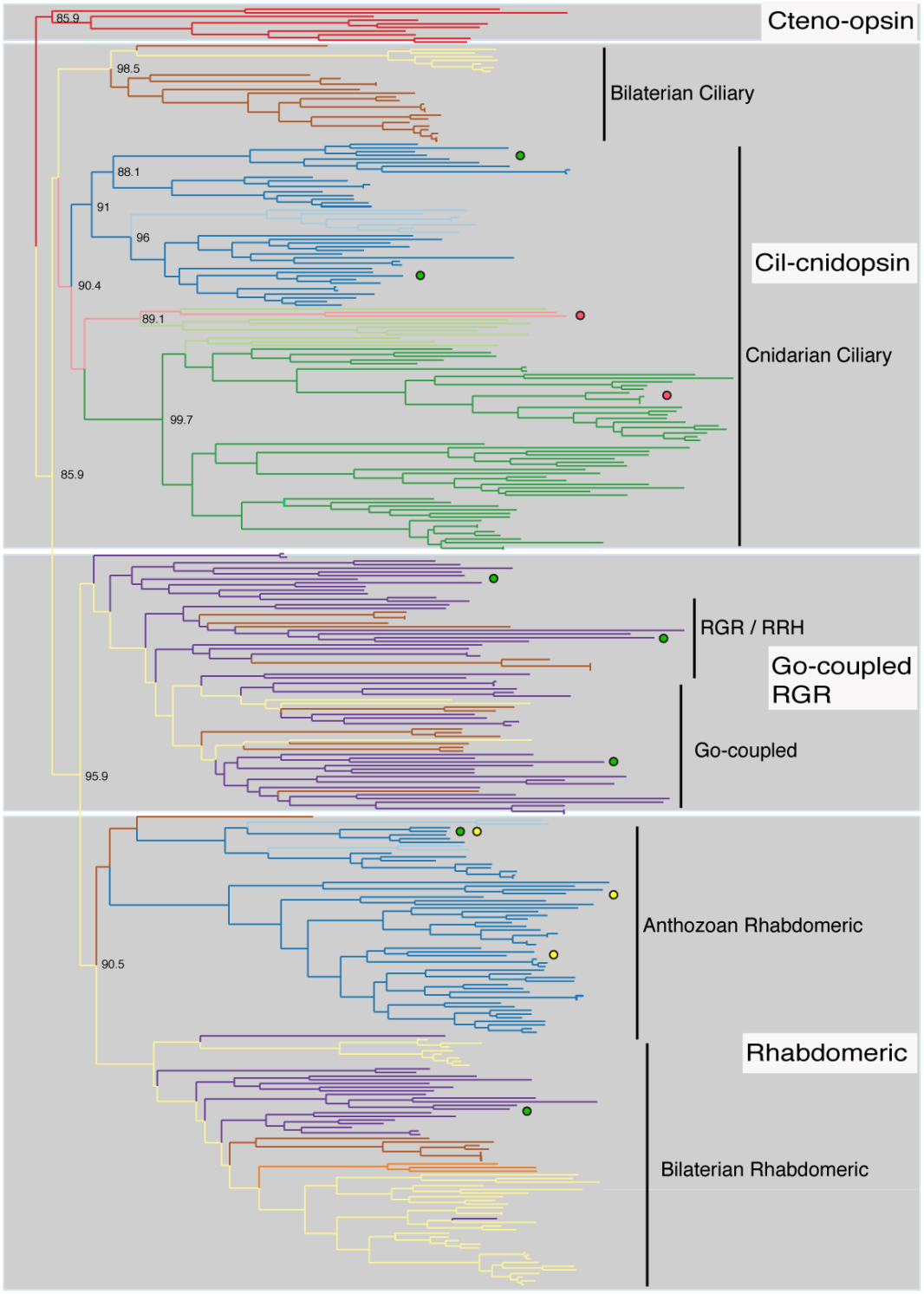

Figure 7. Opsin phylogeny with colored dots representing where the published cnidarian ocular genes (pink), and xenopsins from Ramirez et al. 2016 (green) and Vöcking et al. 2017 (yellow) fell on the phylogeny.

DISCUSSION

Taxon Specific Search Procedure

Our taxon specific filtering strategy is novel and should be considered for future phylogenomic studies. A BLAST search is a good start to the phylogenetic process for gathering data of interest, but depending on the quality of the genome or transcriptome being blasting against the initial results can vary. For example, if one is to do a BLAST search for the 20 top opsin hits against the human genome they are likely to identify all the opsin genes and a few other alpha rhodopsin class GPCRs due to the high quality of the genome. Yet if this same search is done against a poorly assembled transcriptome of a cnidarian they might be lucky to find 20 GPCRs of any type. By isolating the melatonin + opsin clade for each taxon we are making this search issue more manageable by removing unwanted, distantly related sequences before they become an issue in larger gene trees. The drawback to this strategy is when melatonin receptors fall out with or next to the anchor sequences (beta, gamma, and delta rhodopsin class GPCRS). This results in no sequences being filtered out, but can also be a sign of a large radiation of genes in between melatonin and opsins and a loss of other rhodopsin class GPCRs. This is the case for the sponges *Amphimedon queenslandica, Pleraplysilla spinifera,* and *Plakina jani*, the ctenophores *Grantia compressa*, *Coeloplana astericola*, and *Mnemiopsis leidyi,* the staurozoans *Haliclystus auricular* and *Calvadosia cruxmelitensis*, the anthozoans *Metridium senile*, *Montastraea cavemosa, Nematostella vectensis, Porites asteroidses,* and *Renilla reniformis,* and all the endocnidozoans tested.

First Cut (Opsin and Outgroup)

We have uncovered a large GPCR radiation in understudied organisms that is the true sister group to opsins (Paraopsins). Some of these proteins have been discussed before, such as the placopsins from Feuda et al. 2012, and LGR1 protein in *Drosophila melanogaster*. While little research has investigated the role of placopsins, the glycoprotein hormone ligand receptor LGR1 has been shown to play a key role fly development (Vandersmissen et al. 2014). While it is likely other genes within the Paraopsin clade are also glycoprotein receptors, the radiation of these proteins has occurred in entire phyla, such as Porifera and Ctenophora, and has never been discussed. Uncovering the radiation of this large group of GPCRs as the sister to opsins would likely have occurred decades ago if this radiation occurred in Bilateria. Sponges have been screened previously for opsins but all past studies have never found evidence for Poriferan opsins (Plachetzki et al. 2007; Suga et al. 2008; Feuda et al. 2012). However we know they possess phototactic abilities in larval form and have been shown to express other photopigments such as cryptochromes (Rivera et al. 2012). Paraopsin proteins may be photoreceptive, as some possess the diagnostic K296, and play a role in the way some animals detect light, but because these sequences are present in early branching animals that few spend time researching, they have gone undetected. Further research must be done into these Paraopsin proteins to determine their structure and function but until then little is known regarding these genes.

Opsin Tree Analyses

Opsin phylogenies often start with genes of interest and are screened for Lysine 296 to ensure the genes are able to bind a chromophore and are in fact an opsin. While this is informative, the data omitted are still part of the evolutionary history of opsins. Following the establishment of an orthologous clade of opsins, we employed this filtering tactic in our analysis, but this was only done after the group had already been established, in contrast to previous studies investigating opsin evolution (Plachetzki et al. 2007; 2010; 2012; Suga et al. 2008; Porter et al. 2011; Feuda et al. 2012; Hering and Mayer 2014; Ramirez et al. 2016; Vöcking et al. 2017). This approach allowed us to retrieve a set of opsin sequences using formal estimations of monophyly in an unbiased fashion. Once we reached the step of screening for K296 we only found 18 sequences out of 368 that lacked the lysine. Out of these 18 sequences all but 2 came from cnidarians, with most the representatives being hydrozoans. All these sequences fell out into in the cnidopsin clade, which brings up a few possible hypotheses. Hydrozoans have undergone gene duplication events exploring the opsin landscape more so than the other cnidarian classes, making these more recent innovations. Or these genes are artifacts from a gene duplication that occurred before the split of Hydrozoa from Acraspeda and they have become pseudogenized in Acraspeda. Although the genes lack K296, they are expressed in these taxa, as evidenced from the fact that they are derived from transcriptomes. Thus, these newly discovered K296-less opsins could be used for some function other than light detection.

The final opsin dataset underwent exhaustive analyses utilizing three different phylogenetic programs and was tested under multiple A.A. substitution models in each. We have taken the most exhaustive approach to date to identify the best fitting model and the true opsin topology (See Table 4) (Plachetzki et al. 2007; Suga et al. 2008; Porter et al. 2011; Hering and Mayer 2014; Ramirez et al. 2016; Vöcking et al. 2017). IQTree 1.6.0 (Wang et al. 2018) was found to be the program with the best AIC, BIC, log likelihood, and bootstrap scores resulting from the models tested. Particularly the GTR20+C60 and LG+F+R8 were the best, which previous studies have also found that the GTR and LG models often estimate opsin substitution rates the best (Feuda et al. 2012; Ramirez et al. 2016; Picciani et al 2018). The amount of data provided to a GTR model can be a concern if there is not enough to inform the 190 rates in the matrix. Recovering the same topology with a GTR model as a precomputed fixed substitution model is a positive sign that the GTR model is informed and the more complex model is reliable for our opsin dataset. While it appears most researchers are in agreement regarding the best fitting model for opsin evolution, the conflict with opsin phylogenetics appears to be more about taxon selection and opsin identification. Researchers must include a large enough sample of taxa from every eumetazoan group in order to capture the entire evolutionary history of this protein family. We note that this is computationally challenging but it can be done and will become easier in time.

Opsin Clade Topology

In our analyses of the final dataset (both ML and BI) we recovered the three major bilaterian subgroups such as c-opsin, r-opsin, and Go-coupled/RGRs. We also

Program:	RAxML	IQTree	PhyloBayes MPI
Models:	PROTGAMMAGTR	GTR20	CAT-GTR
	PROTGAMMAAUTO	GTR20+C20	
		GTR20+C60	
		LG+F+R8 (ModelFinder)	

Table 4. All the models variations tested with the opsin clade must have K data set under the program used. See Methods for further details on each model and program

recovered a cnidarian specific group of opsins with representatives from all major cnidarian classes. This clade has been documented previously as cnidopsin (or cnidarian xenopsins) (Plachetzki et al. 2007; Ramirez et al. 2016; Vöcking et al. 2017; Picciani et al. 2018) and its existence has been debated, but with increased cnidarian data we find strong support for a monophyletic cnidopsin clade. Cnidopsin contains 38 sequences from Hexacorallia and seven from Octocorallia (Anthozoa), two sequences from Endocnidozoa, 51 from Hydrozoa, seven from Staurozoa, two from Scyphozoa, but none from Cubozoa. Cubozoan opsins have been documented previously (Bielecki et al. 2014; Liegertová et al. 2015) from taxa we included in our analyses such as *Tripedalia cystophora*. However using our methodology, there is no way to account for poor genome/transcriptome quality. Cubozoan sequences were present in the initial opsin and outgroup dataset, but they were lost in the filtering steps for lacking enough informative data, due to poor quality input data.

A clade consisting of seven ctenophore sequences, one sequence from the endocnidozoan, *Polypodium hydriforme,* and two from the xenacoelomorph, *Convolutriloba macropyga,* was also recovered. Ctenophore opsins, or "ctenopsin" have been documented before and fell out as the sister clade to cnidopsin (Hering and Mayer 2014). We have also found support for ctenopsin being the sister to cnidopsin. However, this finding suggests that ctenophores and xenacoelomorphs share a type of ancient opsin that has only been retained in endocnidozoans, or that the xenacoelomorph and endocnidozoan sequences fell out with ctenopsin due to long branch effects, a common artifact of phylogenetic estimation. We also find support that the ctenophore opsins are the root of the opsin phylogeny through reconciled tree analysis in NoTung-2.9 (Stolzer

et al. 2012). Additionally, in the "opsin and outgroup" tree, ctenopsins were included in the opsin ingroup but also fell out as the root for opsins in that analysis. Rooting with ctenophore opsins results in cnidopin being sister to Bilaterian ciliary opsins, and Go-coupled/RGR opsins being sister to Bilaterian rhabdomeric + anthozoan specific opsins.

The anthozoan specific opsins that fall outside of the cnidopsin clade were the only unstable group in our analysis, as the only clade to move depending on the model. Anthozoan specific opsins have also been documented before (Plachetzki et al. 2007; Feuda et al. 2012; Ramirez et al. 2016) but normally only from *Nematostella vectensis* and usually as two separate clades. Anthozoan opsins 1 have been reported as the outgroup to all opsins and anthozoan opsins 2 as the sister to ciliary opsins, consistent with their membership in cnidopsin (Plachetzki, 2007; Hering and Mayer 2014; Vöcking et al. 2017). In all the analyses except PhyloBayes MPI (Lartillot et al. 2013), which was unresolved, we recovered a monophyletic clade of anthozoan specific opsins as the sister to rhabdomeric opsins. Additionally, this clade can be split into two groups, the first being specific to hexacorals containing sequences from *Nematostella, Aiptasia, Anthopleura, Edwardsiella, Protopalythoa, Acropora, Corynactis, Rhodactis, Ricordea, Seriatopora, Montastrea*, and *Platygyra*. The second group contains representatives from the octocorals *Corallium* and *Gorgonia,* and sequences from the hexacorals *Corynactis, Madracis, Nematostella, Aiptasia, Anthopleura, Rhodactis, Favia, Ctenactis,* and *Lobactis*. Usually anthozoan specific opsin clades only contain sequences data for a few hexacorals, but including abundant data for both hexacorals and octocorals allows for a monophyletic anthozoan opsin clade with many representatives to be identified as the sister to bilaterian r-opsins. The placement of this monophyletic clade is supported by

high alrt and bootstrap support, but the internal node splitting the two groups just discussed is low. However, in the PhyloBayes MPI (Lartillot et al. 2013) analyses, these two groups are split, leaving the octocoral+hexacoral group as the sister to bilaterian r-type opsins and moving the hexacoral only group as sister to cnidopsin (Figure 8). We are unable to determine with certainty where the hexacoral specific clade falls on the phylogeny, but their function should be investigated to uncover what role they play in anthozoan sensory perception.

Medusozoa Ocular Genes

The first additional dataset consisted of the 350 opsin genes identified through phylogenetic focusing plus 36 genes from the cubozoans *Carybdea rastonii* and *Tripedalia cystophora* and the hydrozoans *Podocoryna carnea* and *Cladonema radiatum* (Suga et al. 2008; Bielecki et al. 2014; Liegertová et al. 2015). We did not capture any cubozoan genes through our pipeline but were still curious to see where the previously described genes involved in cnidarian eye development fell on the phylogeny. All the cubozoan ocular genes fell out with the opsin genes from Acraspeda (Staurozoa, Cubozoa, and Scyphozoa) that were identified through our phylogenetic focusing pipeline. Similarly all the hydrozoan ocular genes fell out with the hydrozoan opsin genes. Both analyses show that the medusozoan ocular genes fall out with cnidopsin, providing strong support that the genes identified through phylogenetic focusing are also involved in cnidarian phototransduction, eye development, and potentially other sensory functions. Further investigation into these cnidopsin genes may provide insight into how other cnidarians use them for various sensory behaviors.

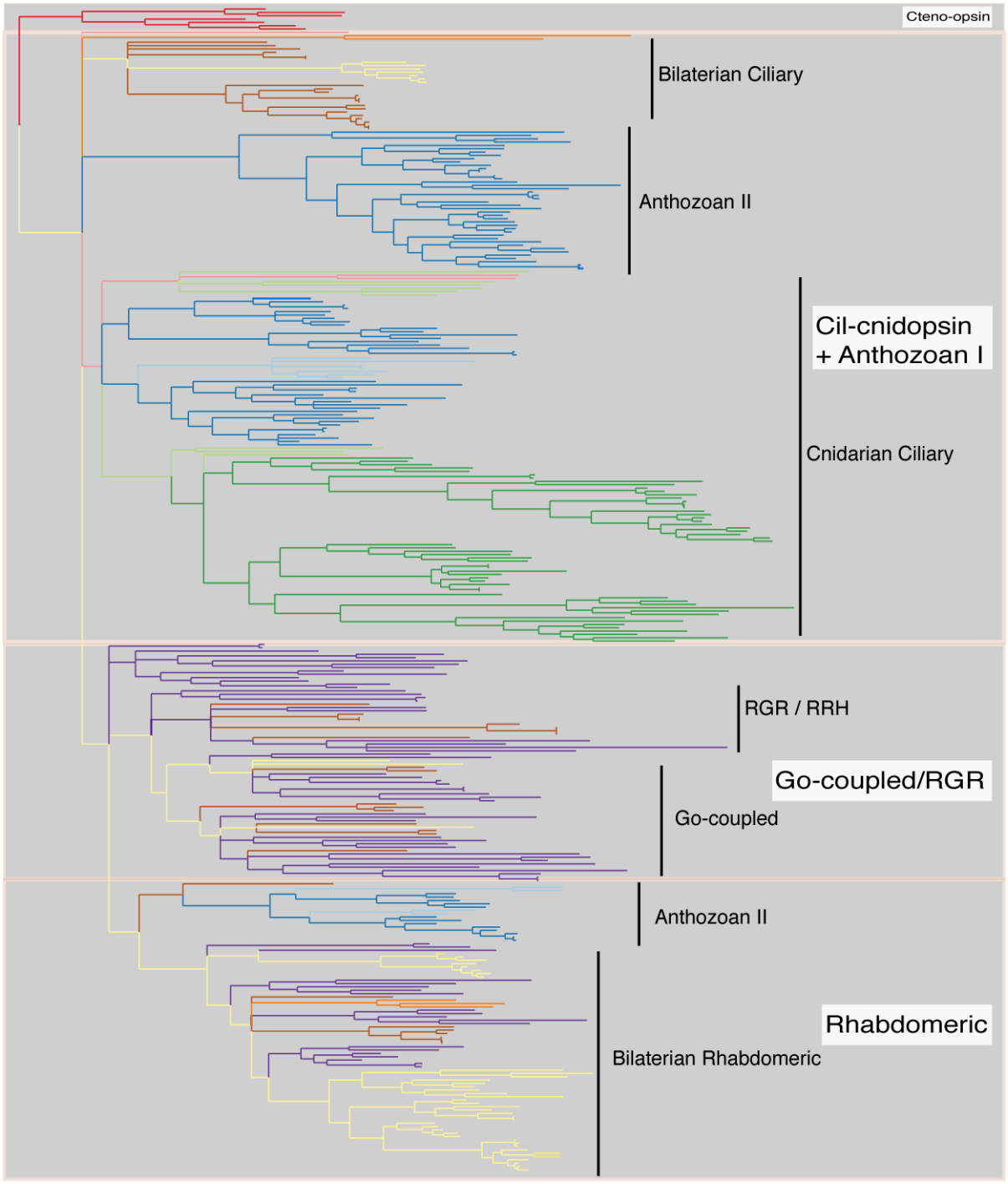

Figure 8. Opsin phylogeny from PhyloBayes MPI resulting in a different topology that splits the anthozoan specific opsins. Relationship of ciliary+cnidopsin+Anthozoan II is unresolved, but we do find support for the Go-coupled/RGR+rhabdomeric opsins.

No support for the Xenopsin clade

The last additional data set created included our opsin "must_have_K" data plus 56 xenopsins identified from Ramirez et al. (2016) and Vöcking et al. (2017). Adding xenopsins to our data and rooting with ctenophore opsins resulted in a different topology but the same clades were retained. Interestingly the xenopsins from Vöcking et al. (2017) were placed into nearly every clade. Xenopsins from *Nematostella vectensis* fell into the anthozoan specific clades, while the Ramirez et al. (2016) xenopsins from *Nematostella vectensis* fell out into cnidopsins. Lastly, the Ramirez et al. (2016) xenopsins from *Lottia gigantea* fell into the bilaterian r-opsin and Go-coupled/RGR opsin clades (Figure 7). These results highlight the confusion of opsin evolution and classification and the lack of support for the so-called xenopsins. With increasing amounts of sequence data new opsin sequences are being identified in a variety of different organisms and often given a name before thorough phylogenetic analysis is applied. This situation is compounded in gene families like opsin, which are short, highly diverse, and often under different selective regimes.

Opsin Phylogeny sensitivity

Our multiple analyses to form an accurate opsin phylogeny has shed light on how sensitive the topology is to change with the addition or removal of sequences. Once the opsin clade was isolated from the “opsin and outgroup” dataset we began an exhaustive approach to uncover the true topology, but it was soon noticed that different topologies would often result from the different data sets tested, such as the “opsin clade”, “must_have_K”, “opsin and ocular”, and “published xenopsin” data sets discussed

previously. While the topology of the tree changed between datasets, the sequences within each clade were retained allowing for a good sense of what opsins are present in different animal classes.

The anthozoan specific opsins seem to be the most unstable class of opsins, with the hexacoral specific clade being even more unpredictable than the hexacoral+octocoral clade, which consistently falls out as the sister to r-opsins. These results are similar to those of Feuda et al. (2012) where they recovered three cnidarian specific clades, one as sister to c-opins (most likely our cnidopsin clade), one as sister to r-opsins (most likely our Hexacoral+Octocoral clade), and one as sister to Go/RGR opsins (most likely our unstable Hexacoral specific clade). However these analyses fall short with regards to cnidarian taxon sampling, and only screened the genomes of the hydrozoan *Hydra magnipapillata* and hexacoral *Nematostella vectensis*. By including a significantly larger sample of cnidarians from all major classes we were able to uncover that the cnidopsin clade is a true class of cnidarian specific opsins. Hering and Mayer (2014) also reported cnidopsin as the sister to c-opsins, and our topology is further supported from the findings of Plachetzki et al. (2007) and Porter et al. (2011) with cnidopins being sister to c-opsins, and r-opsins being sister to Go/RGR opsins.

LITERATURE CITED

Ovchinnikov, Y.A. (1982). Rhodopsin and bacteriorhodopsin: structure—function relationships. FEBS Letters *148*, 179–191.

Hargrave, P.A., and McDowell, J.H. (1992). Rhodopsin and phototransduction: a model system for G protein-linked receptors. The FASEB Journal *6*, 2323–2331.

Shichida, Y., and Imai, H. (1998). Visual pigment: G-protein-coupled receptor for light signals. CMLS, Cell. Mol. Life Sci. *54*, 1299–1315.

Arendt, D. (2003). Evolution of eyes and photoreceptor cell types. Int. J. Dev. Biol. *47*, 563–571.

Yokoyama, S. (2000). Molecular evolution of vertebrate visual pigments. Progress in Retinal and Eye Research *19*, 385-419.

Blackshaw, S., and Snyder, S.H. (1999). Encephalopsin: a novel mammalian extraretinal opsin discretely localized in the brain. J. Neurosci. *19*, 3681–3690.

Zuker, C.S., Cowman, A.F. and Rubin, G.M. (1985). Isolation and structure of a rhodopsin gene from D. melanogaster. Cell *40*, 851–858.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Trong, I.L., Teller, D.C., Okada, T., Stenkamp, R.E., et al. (2000). Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. Science *289*, 739–745.

Terakita, A. (2005). The opsins. Genome Biology *6*, 213.

Terakita, A., and Nagata, T. (2014). Functional Properties of Opsins and their Contribution to Light-Sensing Physiology. Zoological Science *31*, 653–659.

Oakley, T.H., and Speiser, D.I. (2015). How Complexity Originates: The Evolution of Animal Eyes. Annu. Rev. Ecol. Evol. Syst. *46*, 237–260.

Shichida, Y., and Matsuyama, T. (2009). Evolution of opsins and phototransduction. Philos Trans R Soc Lond B Biol Sci *364*, 2881–2895.

Arendt, D., Tessmar-Raible, K., Snyman, H., Dorresteijn, A.W., and Wittbrodt, J. (2004). Ciliary Photoreceptors with a Vertebrate-Type Opsin in an Invertebrate Brain. Science *306*, 869–871.

Plachetzki, D.C., Degnan, B.M., and Oakley, T.H. (2007). The Origins of Novel Protein Interactions during Animal Opsin Evolution. PLOS ONE *2*, e1054.

Velarde, R.A., Sauer, C.D., O. Walden, K.K., Fahrbach, S.E., and Robertson, H.M. (2005). Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. Insect Biochemistry and Molecular Biology *35*, 1367–1377.

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

Picciani, N., Kerlin, J.R., Sierra, N., Swafford, A.J.M., Ramirez, M.D., Roberts, N.G., Cannon, J.T., Daly, M., and Oakley, T.H. (2018). Prolific Origination of Eyes in Cnidaria with Co-option of Non-visual Opsins. Current Biology *0*.

Porter, M.L., Blasic, J.R., Bok, M.J., Cameron, E.G., Pringle, T., Cronin, T.W., and Robinson, P.R. (2011). Shedding new light on opsin evolution. Proc. R. Soc. B rspb20111819.

Plachetzki, D.C., Fong, C.R., and Oakley, T.H. (2012). Cnidocyte discharge is regulated by light and opsin-mediated phototransduction. BMC Biology *10*, 17.

Schnitzler, C.E., Pang, K., Powers, M.L., Reitzel, A.M., Ryan, J.F., Simmons, D., Tada, T., Park, M., Gupta, J., Brooks, S.Y., et al. (2012). Genomic organization, evolution, and expression of photoprotein and opsin genes in Mnemiopsis leidyi: a new view of ctenophore photocytes. BMC Biology *10*, 107.

Sweeney, A.M., Boch, C.A., Johnsen, S., and Morse, D.E. (2011). Twilight spectral dynamics and the coral reef invertebrate spawning response. Journal of Experimental Biology *214*, 770–777.

Tsukamoto, H. and Terakita, A. (2010). Diversity and functional properties of bistable pigments. Photochem. Photobiol. Sci. *9*, 1435–1443.

Feuda, R., Hamilton, S.C., McInerney, J.O., and Pisani, D. (2012). Metazoan opsin evolution reveals a simple route to animal vision. PNAS *109*, 18868–18872.

Hering, L., and Mayer, G. (2014). Analysis of the Opsin Repertoire in the Tardigrade Hypsibius dujardini Provides Insights into the Evolution of Opsin Genes in Panarthropoda. Genome Biol Evol *6*, 2380–2391.

Suga, H., Schmid, V., and Gehring, W.J. (2008). Evolution and functional diversity of jellyfish opsins. Curr. Biol. *18*, 51–55.

Vöcking, O., Kourtesis, I., Tumu, S.C., and Hausen, H. (2017). Co-expression of xenopsin and rhabdomeric opsin in photoreceptors bearing microvilli and cilia. ELife *6*.

Rivera, A.S., Pankey, M.S., Plachetzki, D.C., Villacorta, C., Syme, A.E., Serb, J.M., Omilian, A.R., and Oakley, T.H. (2010). Gene duplication and the origins of

morphological complexity in pancrustacean eyes, a genomic approach. BMC Evolutionary Biology *10*, 123.

Wilden, U., and Kuhn, H. (1982). Light-dependent phosphorylation of rhodopsin: number of phosphorylation sites. Biochemistry *21*, 3014-3022.

Passano, L.M., and McCullough, C.B. (1962). The light response and the rhythmic potentials of hydra. Proc Natl Acad Sci USA *48*, 1376-1382.

Singer, R.H., Rushforth, N.B., and Burnett, A.L. (1963). The photodynamic action of light on hydra. J Exp Zool *154*, 169-173.

Hardie, R.C., and Raghu, P. (2001). Visual transduction in Drosophila. Nature *413*, 186-193.

Land, M., and Nilsson, D.E. (2002). Animal Eyes (Oxford Univ Press, Oxford, UK).

Nilsson, D.E., Gislén, L., Coates, M.M., Skogh, C., and Garm, A. (2005). Advanced optics in a jellyfish eye. Nature *435*, 201–205.

Nathans, J., and Hogness, D.S. (1983). Isolation, sequence analysis, and intron-exon arrangement of the gene encoding bovine rhodopsin. Cell *34*, 807–814.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E.A., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., et al. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. Nature *466*, 720–726.

Chapman, J.A., Kirkness, E.F., Simakov, O., Hampson, S.E., Mitros, T., Weinmaier, T., Rattei, T., Balasubramanian, P.G., Borman, J., Busam, D., et al. (2010). The dynamic genome of *Hydra*. Nature *464*, 592–596.

Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., et al. (2007). Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. Science *317*, 86–94.

Kayal, E., Bastian, B., Pankey, M.S., Ohdera, A., Medina, M., Plachetzki, D.C., Collins, A., and Ryan, J.F. (2017). Comprehensive phylogenomic analyses resolve cnidarian relationships and the origins of key organismal traits (PeerJ Inc.).

Gatto, L., Catanzaro, D., and Milinkovitch, M.C. (2007). Assessing the Applicability of the GTR Nucleotide Substitution Model Through Simulations. Evol Bioinform Online *2*, 145–155.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol *29*, 644–652.

Porath-Krause, A.J., Pairett, A.N., Faggionato, D., Birla, B.S., Sankar, K., and Serb, J.M. (2016). Structural differences and differential expression among rhabdomeric opsins reveal functional change after gene duplication in the bay scallop, Argopecten irradians (Pectinidae). BMC Evolutionary Biology *16*, 250.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol *30*, 772–780.

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

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. Mol Biol Evol *27*, 221–224.

Mirarab, S., Nguyen, N., Guo, S., Wang, L.-S., Kim, J., and Warnow, T. (2015). PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. J Comput Biol *22*, 377–386.

Wang, H.-C., Minh, B.Q., Susko, E., and Roger, A.J. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. Syst Biol *67*, 216–235.

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

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite *Mixtures* of Profiles in a Parallel Environment. Syst Biol *62*, 611–615.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E.A., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., et al. (2010). The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature *466*, 720–726.

Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol *35*, 518–522.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. (2012). Inferring Duplications, Losses, Transfers, and Incomplete Lineage Sorting with Non-Binary Species Trees. Bioinformatics *28*, 409-415.

Bielecki, J., Zaharoff, A.K., Leung, N.Y., Garm, A., and Oakley, T.H. (2014). Ocular and Extraocular Expression of Opsins in the Rhopalium of Tripedalia cystophora (Cnidaria: Cubozoa). PLoS One *9*.

Rivera, A.S., Ozturk, N., Fahey, B., Plachetzki, D.C., Degnan, B.M., Sancar, A., and Oakley, T.H. (2012). Blue-light-receptive cryptochrome is expressed in a sponge eye lacking neurons and opsin. Journal of Experimental Biology *215*, 1278–1286.

Koyanagi, M., Takano, K., Tsukamoto, H., Ohtsu, K., Tokunaga, F., and Terakita, A. (2008). Jellyfish vision starts with cAMP signaling mediated by opsin-Gs cascade. PNAS *105*, 15576–15580.

Stevens, R.C., Cherezov, V., Katritch, V., Abagyan, R., Kuhn, P., Rosen, H., and Wüthrich, K. (2013). The GPCR Network: a large-scale collaboration to determine human GPCR structure and function. Nat Rev Drug Discov *12*, 25–34.

Rocco, D.A., and Paluzzi, J.-P.V. (2016). Functional role of the heterodimeric glycoprotein hormone, GPA2/GPB5, and its receptor, LGR1: An invertebrate perspective. Gen. Comp. Endocrinol. *234*, 20–27.

Vandersmissen, H.P., Van Hiel, M.B., Van Loy, T., Vleugels, R., and Broeck, J. V. (2014). Silencing D. melanogaster lgr1 impairs transition from larval to pupal stage. Gen. Comp. Endocrinol. *209*, 135-147

Liegertová, M., Pergner, J., Kozmiková, I., Fabian, P., Pombinho, A.R., Strnad, H., Pačes, J., Vlček, Č., Bartůněk, P., and Kozmik, Z. (2015). Cubozoan genome illuminates functional diversification of opsins and photoreceptor evolution. Sci Rep *5*, 11885.

Technau, U., and Steele, R.E. (2011). Evolutionary crossroads in developmental biology: Cnidaria. Development *138*, 1447–1458.

Plachetzki, D.C., Fong, C.R., and Oakley, T.H. (2010). The evolution of phototransduction from an ancestral cyclic nucleotide gated pathway. Proceedings of the Royal Society of London B: Biological Sciences *277*, 1963–1969.

Borowiec, M.L., Lee, E.K., Chiu, J.C., and Plachetzki, D.C. (2015). Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. BMC Genomics *16*, 987.

Plachetzki, D.C., and Oakley, T.H. (2007). Key transitions during the evolution of animal phototransduction: novelty, "tree-thinking," co-option, and co-duplication. Integr. Comp. Biol. *47*, 759–769.

Pick, K.S., Philippe, H., Schreiber, F., Erpenbeck, D., Jackson, D.J., Wrede, P., Wiens, M., Alié, A., Morgenstern, B., Manuel, M., et al. (2010). Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol. Biol. Evol. *27*, 1983–1987.

Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., et al. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. Nature *452*, 745–749.

Terakita, A., and Nagata, T. (2014). Functional Properties of Opsins and Their Contribution to Light-Sensing Physiology. Zoological Science *30*, 653-659.

Kuhn, H. (1980). Light- and GTP-regulated interaction of GTPase and other proteins with bovine photoreceptor membranes. Nature *283*, 587–589

Koyanagi, M., Takano, K., Tsukamoto, H., Ohtsu, K., Tokunaga, F., and Terakita, A. (2008). Jellyfish vision starts with cAMP signaling mediated by opsin-G(s) cascade. Proceedings of the National Academy of Sciences of the United States of America *105*, 15576–15580.