

An ocean of opsins

Giacinto De Vivo¹, Eric Pelettier^{2,3}, Roberto Feuda^{4,5,*}, Salvatore D'Aniello^{1,*}

¹ Biology and Evolution of Marine Organisms (BEOM), Stazione Zoologica Anton Dohrn
Napoli, Napoli (Italy)

² Metabolic Genomics, Genoscope, Institut de Biologie François-Jacob, CEA, CNRS,
Université Paris Saclay, Evry (France)

³ Research Federation for Global Ocean System Ecology & Evolution, CNRS, Paris
(France)

⁴ Neurogenetics Group, University of Leicester (UK)

⁵ Department of Genetics, Genomics and Cancer Sciences, University of Leicester (UK)

Correspondence: Roberto Feuda <rf190@leicester.ac.uk>

Salvatore D'Aniello <salvatore.daniello@szn.it>

Abstract

In this study, we explored the diversity and evolution of opsins using meta-omic data from the TARA Oceans and TARA Polar Circle expeditions, one of the largest marine datasets available. By using molecular clustering and phylogenetic analyses, we identified and characterized opsins across various taxonomic groups and gene subfamilies. Our results indicate that the majority of opsin sequences belong to arthropods and vertebrates. However, we also detected sequences from all major opsin subfamilies, including r-opsin, c-opsin, xenopsin, and Group-4 opsins. Despite the broad taxonomic scope, no novel opsin gene families were discovered, suggesting a comprehensive understanding of opsin diversity in marine ecosystems. Additionally, we present novel opsin sequences from less-studied taxa, which may serve as a valuable resource for future research into opsin function and diversity.

Key words: Phylogenetics; Metatranscriptomics; Metagenomics; Photoreception; Plankton; Marine Biodiversity.

Introduction

Plankton forms the foundation of the marine food web and harbors a vast portion of the biodiversity of the oceans. To better understand this diversity, the TARA Oceans expeditions embarked on a global journey, collecting and characterizing plankton from ecological, morphological, and molecular perspectives (Karsenti et al., 2011; Bork et al., 2015; Pesant et al., 2015; de Vargas et al., 2015; Carradec et al., 2018; Salazar et al., 2019; Sunagawa et al., 2020). With over 210 sampling locations worldwide and depths reaching up to 1000 meters, TARA have generated one of the most taxonomically rich genomic and transcriptomic datasets available. This remarkable resource provides an unprecedented opportunity to explore the evolution of gene families across a wide array of marine organisms.

Despite extensive research into plankton diversity, much of the marine biome remains unexplored at the molecular level, especially in terms of gene families crucial for ecological adaptations. Among these gene families, opsins hold a particularly important role for species interactions and ecology. They are key proteins required for animal vision, acting as light sensitive molecules that enable photoreception (Terakita, 2005; Porter et al., 2012). Traditionally opsins have been categorized in three main paralogue groups: ciliary opsin (c-opsin), rhabdomeric opsin (r-opsin), and Group-4 opsins (Go-opsins, neuropsin, RGR, and peropsin/retichronome) (Feuda et al., 2012; Fleming et al., 2020). However, recent studies have challenged this view, suggesting the existence of additional paralog groups. Indeed, new subfamilies such as xenopsin, non-canonical r-opsin, bathyopsin, chaopsin and pseudopsin have been proposed (D'Aniello et al., 2015; Ramirez et al., 2016; De Vivo et al., 2023). The relationships between these subfamilies are still subject to debate. Some phylogenies suggesting the monophyly of Group-4 opsins and c-opsins with r-opsins sister group to them (Davies et al., 2010; Feuda et al., 2012; Feuda et al., 2014; Yoshida et al., 2015; Vöcking et al., 2017; Rawlinson et al., 2019; Bonadè et al., 2020; Fleming et al., 2020; McCulloch et al., 2023), while others support the monophyly of Group-4 and r-opsins (Porter et al., 2012). In contrast, xenopsins have been variably positioned, sometimes closer to c-opsins (Yoshida et al., 2015; Vöcking et al., 2017; Vöcking et al., 2021; McCulloch et al., 2023) and at other times closer to Group-4 opsins (Ramirez et al., 2016; Rawlinson et al., 2019; Bonadè et al., 2020).

A key limitation in understanding opsin diversity and evolution stems from a focus on well-characterized species, whose genomes or transcriptomes have been sequenced. This narrow taxonomic scope may obscure the full range of opsin diversity and hamper a

comprehensive understanding of their evolutionary history. To address these gaps, this study investigates opsin diversity within the meta-transcriptomic (metaT) and meta-genomic (metaG) datasets of the TARA Oceans and TARA Arctic expeditions—the largest -omics dataset of marine plankton (Karsenti et al., 2011; Bork et al., 2015; Pesant et al., 2015; de Vargas et al., 2015; Carradec et al., 2018; Salazar et al., 2019; Sunagawa et al., 2020). This dataset offers an unprecedented chance to examine opsin diversity across a taxonomically and ecologically diverse range of marine organisms, providing new insights into the evolution and function of this critical gene family. We provide a quantitative description of opsin diversity, examining their taxonomic distribution and subfamilies composition within the TARA datasets, contributing to a better understanding of opsin evolution across the marine biome.

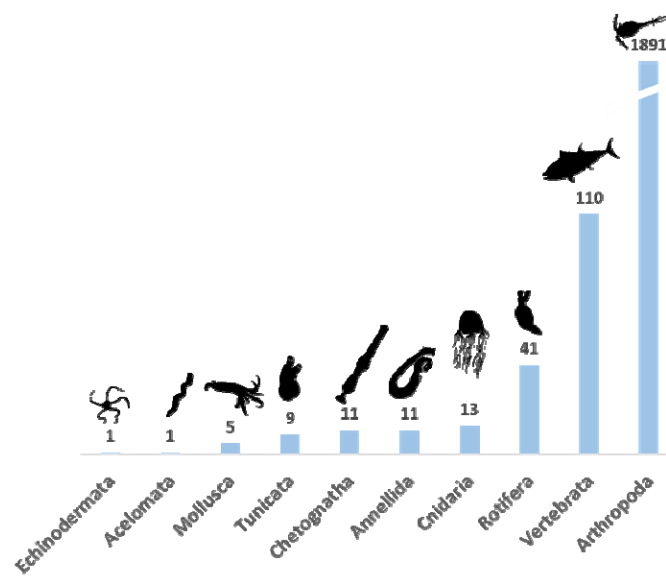
Results

We analysed over 158 million eukaryotic gene sequences obtained from plankton samples across more than 210 sampling stations globally. Using a combination of BLAST, clustering, and annotation methods (Figure S1), we identified 2093 opsin sequences. To ensure a robust taxonomic classification, we integrated these sequences with 567 well-annotated reference sequences (2093 TARA sequences + 567 reference sequences; File S1). A phylogenetic reconstruction was then performed, using Maximum Likelihood under the WAG+CAT model on FastTree (Price et al., 2010) allowing us to map the diversity of opsins across various phyla.

Our results indicate that the vast majority of opsin (~80%) were assigned to arthropods, which dominate zooplankton communities. Other taxa identified include cnidarians, acoelomates, molluscs, annelids, chaetognaths, rotifers, echinoderms, and vertebrates (Fig. 1A). The sequence diversity spanned across all r-opsin, c-opsin, xenopsin and Group-4 clades, with the exception for RGR opsins that have not been found (Fig. 1B). This absence of RGR opsins raises questions about potential lineage-specific losses or methodological biases, warranting further exploration. Among arthropods, we identified 1570 r-opsin, 189 c-opsin, 88 peropsin, 5 Go-opsin, and 36 neuropsins. Vertebrates, the second most represented group with a total of 110 opsin sequences, comprised 94 c-opsins and 16 peropsins. Rotifers were the third most abundant group, with 41 sequences in total: 5 r-opsins, 27 xenopsins, and 9 peropsins. Cnidarians, the fourth most represented group, had 13 sequences, all of which clustered with known cnidarian c-opsins. In chaetognaths, we found 2 xenopsins and 9 peropsins consistent

with known absence of canonical visual opsins (e.g., r- and c-opsin) in this taxon despite having eyes (Wollesen et al., 2023). All 11 annelid opsins, as well as the single opsin found in acoelomates, belonged to the r-opsin clade. Among the molluscs, we identified 5 sequences: 2 r-opsins, 2 Go-opsins, and 1 neuropsin. The only echinoderm sequence in our dataset clustered with bathyopsins. Additionally, 9 tunicate sequences clustered ambiguously with cnidarian c-opsins, likely due to their high sequence divergence. However, based on reciprocal BLAST results, we classified them as peropsins/retinochromes. Finally, three additional sequences could not be classified with confidence and were discarded from further analysis.

A



B

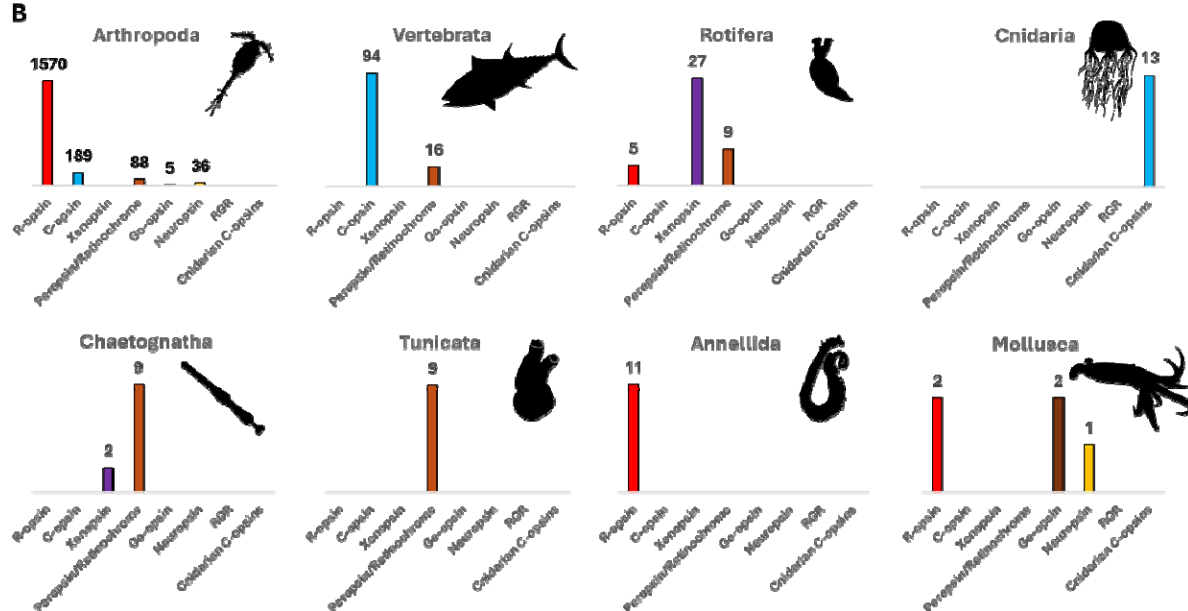


Figure 1. Opsin taxonomical distribution mined in TARA. (A) A plot showing the estimated number of opsin genes in TARA per taxonomical group. (B) Distribution of opsin subfamilies within different taxonomical groups. Animal silhouettes from Phylopic.

We then investigated the relationships among opsin subfamilies by performing a second phylogenetic analysis, aimed at reducing potential bias caused by the high number of arthropod sequences. To achieve this, we generated a second dataset in which 95% of arthropod sequences were removed (Dataset 2 = 298 TARA sequences + 567 reference sequences; File S2) and recompute the analysis using Maximum Likelihood under the WAG+CAT model on FastTree. Nodal support was estimated using the Shimodaira-Hasegawa test on the three alternate topologies (Price et al., 2010).

The resulting phylogenetic tree (Fig. 2 and Fig. S2) supports the existence of the four classical monophyletic groups: r-opsin with a bootstrap (BS) support of 0.99, c-opsin (BS=0.92), xenopsin (BS=0.97), Group-4 opsin (BS=0.94). Group-4 includes peropsin/retinochrome (BS=0.98), RGR (BS=0.99), Go-opsins (BS=0.85), and neuropsins (BS=0.97). The earliest emerging opsin group contains divergent sequences belonging to cnidarians, with r-opsin acting as the sister group to all other opsins. The second diverging group is c-opsin although this node is not well supported (BS=0.38) (Fig. 2; Fig. S3). Xenopsin, formed a clade with a group of cnidarian sequences that usually cluster with c-opsin (referred 'cnidarian c-opsin', BS=0.96) and all together with Group-4 they form a large clade (BS=0.80). A small group referred as bathyopsin (BS=0.86; Ramirez et al., 2016) is found as sister of the xenopsin/Group-4 clade (BS=0.92).

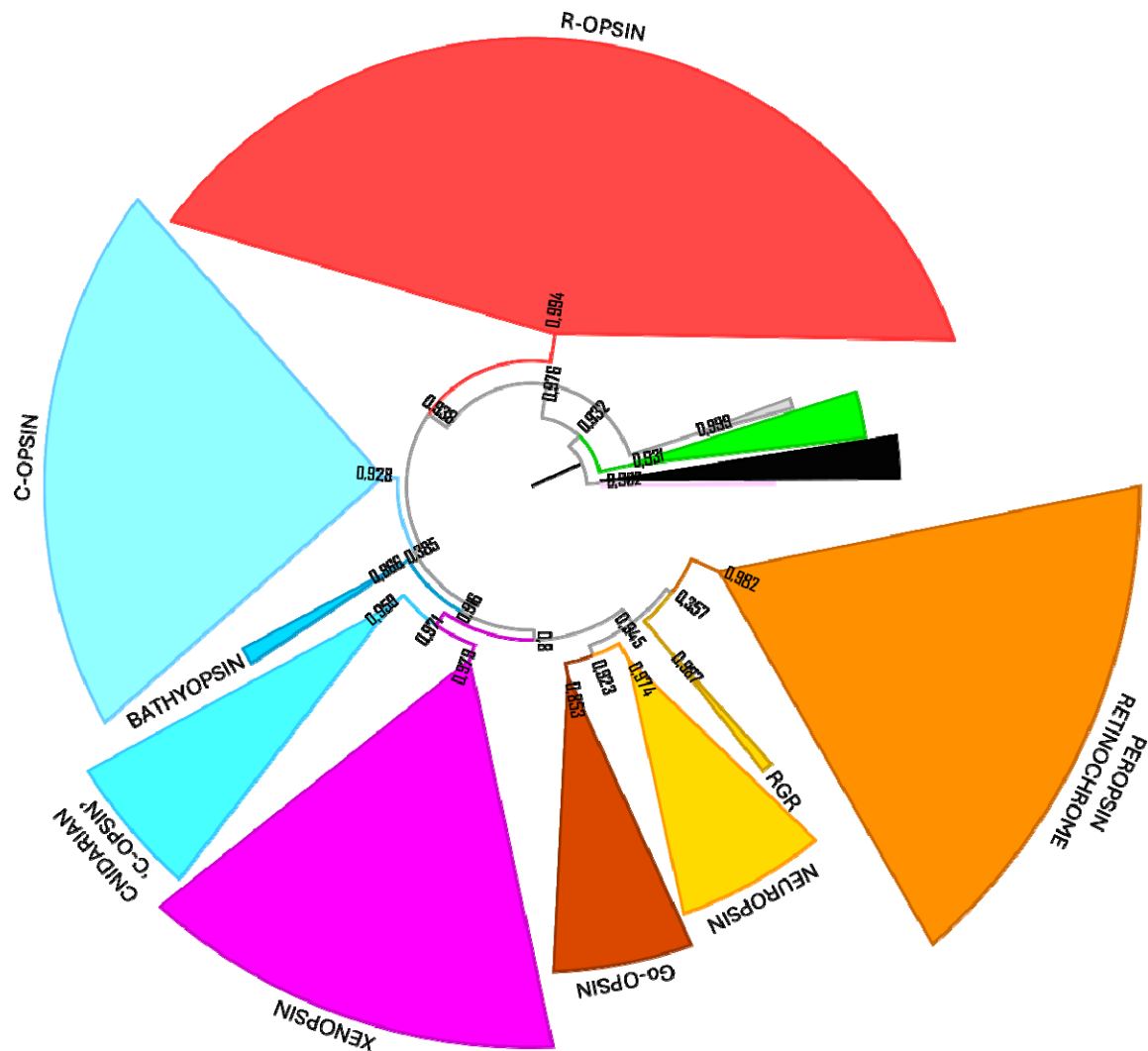


Figure 2. Maximum Likelihood phylogenetic tree (WAG+CAT) showing the relationships between the main opsin subfamilies: r-opsin (red), Group-4 opsin (orange, yellow, straw green, brown), xenopsin (magenta), c-opsin (cyan). Divergent cnidarian opsins are represented in grey, while the outgroups placopsin and pseudopsin are in pink and green, respectively. In this tree 95% of arthropod sequences have been removed. Bootstrap (B) values are indicated at the nodes.

Discussion

Our results contribute to the study of metazoan opsin diversity and evolution and indicate that, despite mining 158 million unbiased eukaryotic gene sequences from plankton, we did not discover any new opsin gene subfamilies. This suggests that current knowledge of metazoan opsin diversity is approaching completeness, at least within the marine biome. However, the study highlights that this dataset, while extensive, may underrepresent certain opsin subfamilies.

Opsin subfamilies in our dataset appear to be unequally distributed. For instance, excluding the highly expressed visual opsins (c- and r-opsins), subfamilies such as peropsins and xenopsins are notably underrepresented. This could be linked to the extensive gene duplications observed in opsins involved in visual tasks, such as those that mediate colour vision. For instance, 77 out of 106 vertebrate c-opsins identified in this study belong to green-sensitive opsins of the rhodopsin-like 2 (Rh2) group, which have undergone substantial duplications in fishes (Hagen et al., 2023).

Moreover, our study supports the monophyly of c- and Group-4 opsins, consistent with previous research (Fleming et al., 2020; Feuda et al., 2015; Feuda et al., 2012), though certain evolutionary relationships, particularly for cnidarian and urochordate opsins, remain unresolved. This phylogenetic instability emphasizes the need for continued investigation into highly divergent opsin lineages and their evolutionary trajectories.

The functional roles of non-visual opsins, such as peropsins and retinochromes, also remain underexplored, despite their potential ecological importance. While the classification of major opsin subfamilies appears robust and well-supported, continued investigation into the functional roles and evolutionary histories of a large opsin diversity will be crucial to clarify their function as moonlighting protein (Feuda et al., 2022).

Methods

Data filtering

We initially mined opsin genes from a database containing 158 million gene sequences derived from meta-genomic and meta-transcriptomic samples collected during the TARA Oceans and TARA Arctic expeditions. Translated sequences from the TARA dataset, both genomic (SMAGs) and transcriptomic (MATOU-v1.5), were blasted (BLASTp) using Diamond against a precompiled opsin dataset representative on all the entire known opsin subfamilies as a query composed of 50 sequences. As a result, we obtained 14,176 putative opsin sequences were obtained (SMAGs: 5,266; MATOU: 8,910).

To filter out all non-opsin sequences that were collected by the initial BLAST search and to check if potential new opsin gene families were present, we proceeded through the following steps: 1) The sequences were merged with a larger opsin dataset consisting of 35,745 sequences obtained from UniProt and NCBI 2) Sequences shorter than 150 amino acids were removed from the dataset; 3) Only sequences with more than 4 and less than 9 transmembrane domains (TM), as predicted by SCAMPI (Peters et al., 2015), were

retained for subsequent analysis. After these steps, a total of 43,228 sequences remained, of which 7,836 were from the TARA dataset.

To classify putative opsin in TARA dataset and given the computational limitation of performing a phylogenetic analysis using 43,228 sequences, we opted to split the dataset into subgroups based on protein similarity. The rationale behind this approach is that by grouping sequences with similar proteins, we should be able to identify all TARA sequences that exhibit some level of similarity to known opsins. To achieve this goal, we conducted an all-versus-all BLASTp in Diamond (e-value threshold: $1e^{-10}$). Sequences were then clustered based on their similarity using the Markov Cluster Algorithm (MCL) (van Dongen and Abreu-Goodger, 2019) with an inflation value set to 1.4. Each cluster was annotated using EggNOG v5.0 (Huerta-Cepas et al., 2019), and only the clusters that contained at least an annotated opsin were retained. This filtering process resulted in 369 clusters out of 735, of which 127 contained at least one opsin from TARA and 16 clusters were exclusively from TARA. This resulted in 2,317 putative opsins (1,801 from MATOU and 516 from SMAG). To confirm the identity of these 2,317 sequences, we BLASTP against the NCBI database. A total of 223 sequences were discarded, leaving a final set of 2,093 opsin sequences for subsequent analysis. The entire pipeline is outlined in Fig. S1.

Phylogenetic analysis

For the phylogenetic analysis, we generated a combined dataset that included all 2,093 TARA opsin sequences and 567 previously annotated opsin sequences (Dataset 1; File S1). The 567 sequences included placopsins, pseudopsins, and melatonin receptors to serve as outgroup (Feuda et al, 2012; De Vivo et al, 2023; McElroy et al. 2024). The datasets were aligned using MAFFT-DASH v7.3 (server <https://mafft.cbrc.jp/alignment/software/>, option mafft --averagelinkage --reorder --anysymbol --allowshift --unalignlevel 0.8 --dash --dashserver https://sysimm.ifrec.osaka-u.ac.jp/dash/REST1.0/ --maxiterate 0 --globalpair input) and trimmed using TrimAL with gap threshold of 0.1 (Capella-Gutiérrez et al., 2009; Rozewicki et al., 2019). Finally, phylogenetic analyses have been performed using FastTree (model WAG+CAT) (Price et al., 2010). Finally, taxonomic data were assigned by looking the tree topology and confirmed using reverse BLASTp on NCBI database.

To address the potential impact of the large number of similar sequences from closely related organisms—especially arthropods—on the phylogenetic results (e.g., node support values), we conducted a second phylogenetic analysis. In this analysis, we

removed 95% of the TARA sequences assigned to arthropods from Dataset 1, resulting in Dataset 2 (File S2), and recomputed the phylogeny using the same methods as described above.

254

255 **Figure S1**

256 Bioinformatic pipeline was adopted to efficiently mine, cluster, and annotate opsins from
257 TARA metaT (MATOU) and metaG (SMAGs) to be used for the phylogenetic analyses.

258

259

260 **Figure S2**

261 Phylogenetic analysis computed using all the 2093 sequences from TARA and an already
262 annotated dataset of 567 reference sequences (i.e. Dataset 1).

263

264 **Figure S3**

265 Phylogenetic analysis computed using sequences from TARA where 95% of arthropod
266 sequences were randomly subsampled (only 5% retained) and an already annotated
267 dataset of 567 reference sequences (i.e. Dataset 2).

268

269 **File S1**

270 Alignment containing all the 2093 sequences from TARA and an already annotated
271 dataset of 567 reference sequences (Dataset 1).

272

273 **File S2**

274 Alignment containing sequences from TARA where arthropod sequences were
275 subsamples (only 5% retained) and an already annotated dataset of 567 reference
276 sequences (Dataset 2).

277

278

279 **Data availability**

280 Data are available on GitHub: https://github.com/Xodroont/Tara_Opsins

281 MATOU-v1.5 and SMAGs data are available at <http://www.genoscope.cns.fr/tara/>

282

283 **Acknowledgements**

284 We are grateful to Elijah Lowe, Luigi Caputi, Daniele Iudicone, Maria Ina Arnone for
285 discussions on initial versions of the project. G.D.V. has been supported by a PhD
286 fellowship funded by the Stazione Zoologica Anton Dohrn (Open University – Stazione
287 Zoologica Anton Dohrn PhD Program). R.F. is supported by a University Research

Fellowship (UF160226 and URF/R/221011) a Research Grant from the Royal Society (RGFR1\181012).

Our survey was made possible by the sampling and sequencing efforts of the *Tara Oceans* Project. We are indebted to all who contributed to these efforts and other open-source bioinformatics tools for their commitment to transparency and openness. *Tara Oceans* (which includes the *Tara Oceans* and *Tara Oceans* Polar Circle expeditions) would not exist without the leadership of the *Tara Oceans* Foundation and the continuous support of 23 research institutes (<https://oceans.taraexpeditions.org/>). We also acknowledge the commitment of the CNRS and Genoscope/CEA. Some computations were performed thanks to the TGCC computing facility in France.

References

Aberer AJ, Krompass D, Stamatakis A. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst Biol*. 2013 Jan 1;62(1):162-6. doi: 10.1093/sysbio/sys078.

Bonadè M, Ogura A, Corre E, Bassaglia Y, Bonnaud-Ponticelli L. 2020. Diversity of light sensing molecules and their expression during the embryogenesis of the cuttlefish (*Sepia officinalis*). *Front Physiol*. 11:521989.

Bork P, Bowler C, de Vargas C, Gorsky G, Karsenti E, Wincker P. *Tara Oceans*. *Tara Oceans* studies plankton at planetary scale. Introduction. *Science*. 2015 May 22;348(6237):873. doi: 10.1126/science.aac5605.

Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F, Tirichine L, Labadie K, Kirilovsky A, Bertrand A, Engelen S, Madoui MA, Méheust R, Poulain J, Romac S, Richter DJ, Yoshikawa G, Dimier C, Kandels-Lewis S, Picheral M, Searson S; *Tara Oceans* Coordinators, Jaillon O, Aury JM, Karsenti E, Sullivan MB, Sunagawa S, Bork P, Not F, Hingamp P, Raes J, Guidi L, Ogata H, de Vargas C, Iudicone D, Bowler C, Wincker P. A global ocean atlas of eukaryotic genes. *Nat Commun*. 2018 Jan 25;9(1):373. doi: 10.1038/s41467-017-02342-1.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 2009 Aug 1;25(15):1972-3. doi: 10.1093/bioinformatics/btp348. Epub 2009 Jun 8. PMID: 19505945; PMCID: PMC2712344.

D'Aniello S, Delroisse J, Valero-Gracia A, Lowe EK, Byrne M, Cannon JT, Halanynch KM, Elphick MR, Mallefet J, Kaul-Strehlow S, Lowe CJ, Flammang P, Ullrich-Lüter E, Wanninger A, Arnone MI. Opsin evolution in the Ambulacraria. *Mar Genomics*. 2015 Dec;24 Pt 2:177-83. doi: 10.1016/j.margen.2015.10.001.

Davies WL, Hankins MW, Foster RG. Vertebrate ancient opsin and melanopsin: divergent irradiance detectors. *Photochem Photobiol Sci*. 2010 Nov;9(11):1444-57.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S; Tara Oceans Coordinators, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*. 2015 May 22;348(6237):1261605. doi: 10.1126/science.1261605.

De Vivo G, Crocetta F, Ferretti M, Feuda R, D'Aniello S. Duplication and Losses of Opsin Genes in Lophotrochozoan Evolution. *Mol Biol Evol*. 2023 Apr 4;40(4): msad066.

Feuda R, Hamilton SC, McInerney JO, Pisani D. Metazoan opsin evolution reveals a simple route to animal vision. *Proc Natl Acad Sci U S A*. 2012 Nov 13;109(46):18868-72. doi: 10.1073/pnas.1204609109.

Feuda R, Rota-Stabelli O, Oakley TH, Pisani D. The comb jelly opsins and the origins of animal phototransduction. *Genome Biol Evol*. 2014 Jul 24;6(8):1964-71. doi: 10.1093/gbe/evu154.

355 Feuda R, Menon AK, Göpfert MC. Rethinking Opsins. *Mol Biol Evol.* 2022 Mar
356 2;39(3):msac033. doi: 10.1093/molbev/msac033.

357

358 Fleming JF, Feuda R, Roberts NW, Pisani D. A Novel Approach to Investigate the Effect of
359 Tree Reconstruction Artifacts in Single-Gene Analysis Clarifies Opsin Evolution in
360 Nonbilaterian Metazoans. *Genome Biol Evol.* 2020 Feb 1;12(2):3906-3916. doi:
361 10.1093/gbe/evaa015.

362

363 Hagen JFD, Roberts NS, Johnston RJ Jr. The evolutionary history and spectral tuning of
364 vertebrate visual opsins. *Dev Biol.* 2023 Jan;493:40-66.

365

366 Hagen JFD, Roberts NS, Johnston RJ Jr. The evolutionary history and spectral tuning of
367 vertebrate visual opsins. *Dev Biol.* 2023 Jan; 493:40-66. doi: 10.1016/j.ydbio.2022.10.014.
368 Epub 2022 Nov 9. PMID: 36370769; PMCID: PMC9729497.

369

370 Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H,
371 Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. eggNOG 5.0: a
372 hierarchical, functionally and phylogenetically annotated orthology resource based on
373 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D309-D314. doi:
374 10.1093/nar/gky1085.

375

376 Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D,
377 Benzoni F, Claverie JM, Follows M, Gorsky G, Hingamp P, Iudicone D, Jaillon O, Kandels-
378 Lewis S, Krzic U, Not F, Ogata H, Pesant S, Reynaud EG, Sardet C, Sieracki ME, Speich
379 S, Velayoudon D, Weissenbach J, Wincker P; Tara Oceans Consortium. A holistic
380 approach to marine eco-systems biology. *PLoS Biol.* 2011 Oct;9(10):e1001177. doi:
381 10.1371/journal.pbio.1001177.

382

383 Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the
384 size of large protein databases. *Bioinformatics.* 2001 Mar;17(3):282-3. doi:
385 10.1093/bioinformatics/17.3.282.

386

McCulloch KJ, Babonis LS, Liu A, Daly CM, Martindale MQ, Koenig KM. *Nematostella vectensis* exemplifies the exceptional expansion and diversity of opsins in the eyeless Hexacorallia. *Evodevo*. 2023 Sep 21;14(1):14.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* (2020) 37: 1530-1534. <https://doi.org/10.1093/molbev/msaa015>.

Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Troublé R, Dimier C, Searson S; Tara Oceans Consortium Coordinators. Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data*. 2015 May 26;2:150023. doi: 10.1038/sdata.2015.23.

Peters C, Tsirigos KD, Shu N, Elofsson A. Improved topology prediction using the terminal hydrophobic helices rule. *Bioinformatics*. 2016 Apr 15;32(8):1158-62. doi: 10.1093/bioinformatics/btv709.

Porter ML, Blasic JR, Bok MJ, Cameron EG, Pringle T, Cronin TW, Robinson PR. Shedding new light on opsin evolution. *Proc Biol Sci*. 2012 Jan 7;279(1726):3-14. doi: 10.1098/rspb.2011.1819.

Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. 2010 Mar 10;5(3):e9490. doi: 10.1371/journal.pone.0009490.

Ramirez MD, Pairett AN, Pankey MS, Serb JM, Speiser DI, Swafford AJ, Oakley TH. The Last Common Ancestor of Most Bilaterian Animals Possessed at Least Nine Opsins. *Genome Biol Evol*. 2016 Dec 1;8(12):3640-3652. doi: 10.1093/gbe/evw248.

Rawlinson KA, Lapraz F, Ballister ER, Terasaki M, Rodgers J, McDowell RJ, Girstmair J, Criswell KE, Boldogkoi M, Simpson F, et al. 2019. Extraocular, rod-like photoreceptors in a flatworm express xenopsin photopigment. *eLife* 8:e45465.

Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res.* 2019 Jul 2;47(W1):W5-W10. doi: 10.1093/nar/gkz342.

Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh HJ, Cuenca M, Field CM, Coelho LP, Cruaud C, Engelen S, Gregory AC, Labadie K, Marec C, Pelletier E, Royo-Llonch M, Roux S, Sánchez P, Uehara H, Zayed AA, Zeller G, Carmichael M, Dimier C, Ferland J, Kandels S, Picheral M, Pisarev S, Poulain J; Tara Oceans Coordinators; Acinas SG, Babin M, Bork P, Bowler C, de Vargas C, Guidi L, Hingamp P, Iudicone D, Karp-Boss L, Karsenti E, Ogata H, Pesant S, Speich S, Sullivan MB, Wincker P, Sunagawa S. Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell.* 2019; 179(5):1068-1083.e21. doi: 10.1016/j.cell.2019.10.014.

Sunagawa S, Acinas SG, Bork P, Bowler C; Tara Oceans Coordinators; Eveillard D, Gorsky G, Guidi L, Iudicone D, Karsenti E, Lombard F, Ogata H, Pesant S, Sullivan MB, Wincker P, de Vargas C. Tara Oceans: towards global ocean ecosystems biology. *Nat Rev Microbiol.* 2020 Aug;18(8):428-445. doi: 10.1038/s41579-020-0364-5.

Terakita A. The opsins. *Genome Biol.* 2005;6(3):213. doi: 10.1186/gb-2005-6-3-213.

van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. *Methods Mol Biol.* 2012; 804:281-95. doi: 10.1007/978-1-61779-361-5_15.

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

Vöcking O, Leclère L, Hausen H. 2021. The rhodopsin-retinochrome system for retinal re-isomerization predates the origin of cephalopod eyes. *BMC Ecol Evol.* 21:1–14.

Wollesen T, Rodriguez Monje SV, Oel AP, Arendt D. Characterization of eyes, photoreceptors, and opsins in developmental stages of the arrow worm *Spadella cephaloptera* (Chaetognatha). *J Exp Zool B Mol Dev Evol.* 2023 Jul;340(5):342-353. doi: 10.1002/jez.b.23193.

454 Yoshida MA, Ogura A, Ikeo K, Shigeno S, Moritaki T, Winters GC, Kohn AB, Moroz LL.
 455 2015. Molecular evidence for convergence and parallelism in evolution of complex brains
 456 of cephalopod molluscs: insights from visual systems. *Integr. Comp. Biol.* 55: 1070–1083.