Fleming, J. F., Feuda, R., Roberts, N. W., & Pisani, D. (2020). A novel approach to investigate the effect of tree reconstruction artifacts in single gene analysis clarifies opsin evolution in non-bilaterian metazoans. *Genome Biology and Evolution*, [evaa015]. https://doi.org/10.1093/gbe/evaa015

Article

A novel approach to investigate the effect of tree reconstruction artifacts in single gene analysis clarifies opsin evolution in non-bilaterian metazoans

James F Fleming[1, 2], Roberto Feuda[1], Nicholas W. Roberts[3], Davide Pisani[1*]

[1]: School of Earth Sciences, University of Bristol, Woodland Road, Bristol, UK

[2]: Institute of Advanced Biosciences, Keio University, Kakuganji 246-2, Tsuruoka, Yamagata, Japan

[3]: School of Biological Sciences, University of Bristol, Tyndall Avenue, Bristol, UK

*:Corresponding Author: Davide Pisani, davide.pisani@bristol.ac.uk

Abstract:

Our ability to correctly reconstruct the topology of a phylogenetic tree is strongly affected by both systematic errors and the amount of phylogenetic signal in the data. Current approaches to tackle tree reconstruction artifacts, such as the use of parameter-rich models, do not translate readily to single-gene alignments. This, coupled with the limited amount of phylogenetic information contained in single-gene alignments, makes single-gene phylogenies particularly difficult to reconstruct. Opsin phylogeny illustrates this problem clearly. Opsins are G-protein coupled receptors utilised in photoreceptive processes across Metazoa and their protein sequences are roughly 300 amino acids long. Because of their relevance to the understanding of the evolution of photoreception, a number of independent single-gene phylogenetic analyses have been performed to understand opsin evolution, but such studies inferred incongruent trees that hampered progress of our understanding of the evolutionary origins of animal vision. Here, we present a novel approach to investigate and potentially circumvent errors in single-gene phylogenies. First, we demonstrate the efficacy of our approach using two well-understood cases of long branch attraction in single gene datasets and simulations. After that, we apply our approach to a large collection of well-characterised opsins, and clarify early opsin evolution and the relationships of the three main opsin subfamilies.

Introduction:

Resolving gene phylogenies is difficult for two reasons. Firstly, single gene alignments are relatively short and might be poor in phylogenetic signal. Secondly, it is more difficult to counter tree reconstruction artifacts in single-gene alignments, as most approaches used to address these problems have been developed for long super-alignments where such artifacts are exacerbated (Jeffroy et al. 2006; Lartillot & Philippe 2004) - e.g. the application of the CAT-based models of Lartillot & Philippe (2004) and methods that remove sites and that will further decrease the length of single-gene alignment (e.g. Brinkmann & Philippe 1999; Pisani 2004; Sperling et al. 2009). On the other hand, in the post-genomic era, we generally have access to an abundance of sequences from a multitude of species for many gene families. Sequence-rich alignments can thus be subsampled to exclude "problematic sequences" that could potentially lead to tree reconstruction artifacts. Following from the conclusions of Felsenstein 1978, Anderson & Swofford (2004) and Xi et al (2015), problematic sequences lack sufficient phylogenetic signal under the current model to reliably determine their topological position in the dataset. Thereby, the identification and removal of "problematic" sequences, i.e. sequences that are likely to have a distorting effect on the final gene-tree topology, could constitute an alternative, viable strategy to improve the accuracy of single gene phylogenies. However, assessing how to objectively identify such sequences is far from trivial: even in situations where a high-quality species tree is available, it is often unclear whether a gene tree that is incongruent with the species tree is the result of duplication and independent loss within the clades or an artefact of tree reconstruction.

The opsins are G-coupled protein receptors fundamental to light sensitive processes across Metazoa. Opsins are present in almost every animal phylum; including Cnidaria and Ctenophora (Feuda et al. 2012, 2014; Ramirez et al. 2016; Schnitzler et al. 2012). Furthermore, opsin-like sequences have been found in Placozoa, known as placopsins (Feuda et al. 2012, 2014), though opsins and opsin-like sequences are currently unknown in sponges (Feuda et al. 2012; Plachetzki et al. 2007). A general agreement exists that most opsins can be ascribed to one of three "canonical" (i.e. widely recognised – see Ramirez et al 2016) groups: the rhabdomeric opsins, the ciliary opsins , and the group 4 opsins (the peropsins/RGRs, Go-opsins and the neuropsins). In addition to these groups, Ramirez et al. (2016) defined three more opsin subfamilies: the bathyopsins, the xenopsins and the chaopsins, though the functions of members of these new families are unclear, and their phylogenetic status requires further testing.

Despite their functional diversity, most opsins function in a similar fashion. They bind to a chromophore, an aldehyde derivative of vitamin A, in its *cis* photoisomerised state. Together, this combination is known as a visual pigment (Terakita 2005). When the chromophore

absorbs a photon of light, it changes from its *cis* state to its *trans* state, which alters the conformation of the opsin-chromophore binding site and activates the G-protein the opsin is coupled to, thus starting a signalling cascade (Terakita 2005). Rhabdomeric visual pigments are bistable: the chromophore stays attached to the opsin when in *cis* or *trans* state (Tsukamoto 2014). The chromophore is able to reversibly switch between the *cis* and *trans* conformational states by absorbing light of different wavelengths (Tsukamoto & Terakita 2010). However, in canonical ciliary visual pigments, the chromophore becomes detached from the opsin when it changes conformation from *cis* to *trans*. This is known as opsin bleaching. RGRs instead bind to all-*trans*-retinal and convert the chromophore back to all-*cis*-retinal; that can then be reattached to bleached ciliary opsins (Terakita 2005). Visual ciliary opsins are therefore dependent on the presence of RGRs to function, although it is still unclear whether some non-visual ciliary opsins undergo bleaching (Terakita et al. 2012).

The relationships among the canonical opsin families are still unclear (see figure 1). Three main scenarios have been proposed for their evolution. The first scenario, proposed by Feuda et al. (2012, 2014), Hering & Mayer (2014) and Schnitzler et al. (2012), suggests a sister group relationship between the group 4 and the ciliary opsins to the exclusion of the rhabdomeric opsins. The second, proposed by Porter et al. (2011) found a sister group relationship between the group 4 and rhabdomeric opsins to the exclusion of the ciliary opsins (Figure 1). Finally, Ramirez et al. (2016) found a sister group relationship between the rhabdomeric and ciliary opsins to the exclusion of the group 4 opsins. All non-canonical opsin families nest within the orthogroup defined by the three canonical opsin families, hence non-canonical visual opsins are irrelevant to understanding early opsin evolution. Indeed, the deepest duplication in the history of the animal visual opsins is defined by the split between the rhabdomeric opsins and either the ciliary or the group 4 opsins (Figure 1). Interestingly, this lack of consensus implies that the identity of the opsins that were part of the ancestral metazoan photoreceptive system remains unclear: it is not yet known whether the ancestral opsins bore more similarity to ciliary, group 4 or rhabdomeric opsins.

Here we present a new approach – the canary sequence approach - which can be applied to single gene phylogenies and aims to identify and remove potentially problematic sequences: sequences that lack sufficient phylogenetic signal under the current model to reliably determine their topological position in the dataset (Felsenstein 1978, Anderson & Swofford 2004, Xi et al 2015). The name of the method derives from the practise of using canaries to detect methane in mine shafts. This approach uses sequences that change position between multiple rounds of tree searches, but do not affect the relationships inferred for all the other sequences in the dataset to identify potentially problematic sequences.

We first demonstrate that our approach is able to identify potentially problematic sequences in two classic case studies – recovering a tree displaying Ecdysozoa using the data of Aguinaldo et al. (1997), and a tree assessing the monophyly of Platyhelminthes (and the relationships of the Lophotrochozoa) using the data of Carranza et al. (1997). Furthermore, we test the method using simulated data sets. Finally, focusing on the cnidarian and ctenophoran opsins and on the three canonical opsin families (ciliary, rhabdomeric and group 4), we used the canary sequence method to investigate the deepest history of duplications in the opsin gene family. Our results corroborate those of Feuda et al. (2012, 2014) and Hering & Mayer (2014), and suggest that the deepest duplication in the history of the bilaterian opsins separates the rhabdomeric opsins from a group composed by the ciliary and the group 4 opsins. In addition, we confirm the existence of cnidarian rhabdomeric opsins, which emerge as the sister of the bilaterian rhabdomeric opsins (canonical and non-canonical). While we could confirm the existence of cnidarian and ctenophoran opsins sharing a common ancestor with the ciliary opsins, we could not confirm the existence of cnidarian and ctenophoran opsins related to the group 4 opsins, or the existence of opsins predating the duplication separating the rhabdomeric opsins from the group 4 plus ciliary opsins.

The canary sequence approach to identify problematic sequences: The canary sequence approach aims to identify and reduce the number of problematic sequences in an alignment, and thereby reduce topological reconstruction artefacts. The logic underlying the canary sequence approach is based on the identification of sequences that are prone to moving within a phylogeny due to poor clustering signals (Brinkmann & Philippe 1999; Dabert et al. 2010): the canary sequences. We then ascertain whether other sequences in the dataset affect the phylogenetic relationships of the canary sequences, to identify potentially problematic sequences, sequences that can attract other sequences in the dataset. Potentially problematic sequences can be excluded from the analyses in order to infer what we define as the "minimal tree" for a protein family. The steps of the canary method are presented in Figure 2 and are summarized below:

1) *Data set creation*: The first step requires the identification of the "full dataset" (the considered dataset) and of two additional sub-datasets. The first sub-dataset is composed of the "sequences of interest", which includes all the sequences that are under examination (these are a set of sequences that we intend to add to a pre-existing gene family dataset). The second set is referred to as the "base dataset", which includes all sequences in the full dataset except the sequences of interest. Trees are constructed from both the base dataset and the full dataset – these are referred to as the "base tree" and "full tree" respectively. The base tree and full tree serve to measure the effect of the sequences of interest on the topology of the gene tree, and

allow for an existing gene phylogeny to act as a basis for the application of the canary method, but (see 18S Results), the method does not necessarily assume that either the base tree or full tree are reliable.

2) *Measuring the effect of the sequences of interest on the base dataset*: In the second step, a series of datasets are generated by separately combining the base dataset with each individual sequence of interest. These datasets and the trees from these datasets are referred to as "checking datasets" and "checking trees". The position of each sequence in each checking tree is noted.
3) *Identification of sequences for further examination*: For each sequence of interest, if the checking tree and the base tree are isomorphic (after the removal of the sequence of interest), the sequence of interest is marked as a "sequence for further examination". If the checking tree and base tree are not isomorphic (after the removal of the sequence of interest), then the sequences are moved to the "non-canary sequences of interest dataset".
4) *Identification of canary and stable sequences:* As inaccurate phylogenies might emerge because of compositional heterogeneity (Roure et al. 2012), a posterior predictive test then ascertains whether each of the "sequences for further examination" are compositionally homogeneous or heterogeneous. If the "sequence for further examination" is found in different positions in the checking tree and the full tree, it is defined as a "canary sequence". However, if the checking tree and base tree are isomorphic (after the removal of the sequence of interest), compositionally homogeneous but found in the same position in the checking tree and full tree, it is marked as a "stable sequence". If the sequence is found to be compositionally heterogeneous, it is moved to the "non-canary sequences of interest dataset".

   After each sequence of interest has been classified it is possible that canary sequences might not be present in a dataset. If that is the case, move to step 8. Otherwise (if canary sequences can be identified), the sequences previously identified as stable sequences are also added to the "non-canary sequences of interest dataset", and the analysis moves to step 5.

   Steps 3 and 4 identify sequences that are unstable within their checking tree and have the expected amino acid composition. Such sequences do not have enough information to precisely cluster within their checking tree, but also do not convey enough clustering information to alter the relationships in the base tree (the compared trees are isomorphic once the canary sequence is removed). Because these sequences do not have sufficient information to cluster firmly in their checking tree, they are more likely to be affected by the presence of "problematic sequences" when compared to other sequences in the dataset. We thus reason that they can be used as

indicators to highlight potentially problematic sequences, which are expected to have the tendency to attract canary sequences.

5) *Definition of the "canary dataset", "canary tree" and of the "non-canary sequences of interest dataset":* All canary sequences identified in Step 4 are added to the base dataset to generate the "canary dataset". A tree is inferred from the canary dataset, which is referred to as the "canary tree". Sequences of interest that do not meet the criteria necessary to become canary sequences are moved to the "non-canary sequences of interest dataset".
6) *Measuring the effect of the non-canary sequences on the canary dataset*: In step 5 the "canary dataset" and the "non-canary sequences of interest dataset" were generated. For each sequence in the "non-canary sequences of interest" dataset, a new alignment is generated where a single non-canary sequence of interest is added to the "canary dataset". These datasets and the trees they generate are referred to as the "canary checking datasets" and "canary checking trees" respectively. For each non-canary sequence of interest, if the "canary checking tree" and the "canary tree" of step 5 are isomorphic (after the removal of the non-canary sequence of interest), the non-canary sequence of interest is defined as "non-problematic". All other non-canary sequences of interest are defined as "potentially problematic".
7) *Generation of the "Minimal dataset" and completion of the canary pipeline:* All "non-problematic sequences" are added to the "canary dataset" to generate the "minimal dataset". The tree generated from the minimal dataset is the final point of the canary sequence approach and is called the "minimal tree". This is to stress that this tree only represents the backbone of the evolutionary history of the family of interest.
8) *No canary sequences Identified*: Previously identified "stable sequences" from step 3, are deemed to be potentially non-problematic. Stable sequences only are added to the base dataset to then generate the minimal tree.

Materials & Methods

To test the reliability of the canary approach we performed analyses using two datasets Aguinaldo et al. (1997) and Carranza et al. (1997) that address problems that were considered hard at the time these datasets were published, but that are now well understood. In addition to that, we used simulated datasets to further understand the behaviour of the canary approach. Finally, we applied the canary approach to understand early opsin evolution. For both case studies (Aguinaldo et al. 1997; Carranza et al. 1997) and all simulation analyses, alignments were performed in MUSCLE (Edgar 2004) and analysed under the JC69 model in

PhyML (Guindon et al. 2010) . JC69 was used to generate results comparable to those of the original studies.

Case Study 1: We used the Aguinaldo and collaborators 18s rRNA dataset to test the performance of the canary method. The original 18s rRNA analysis of Aguinaldo et al. （1997) recovered a monophyletic Ecdysozoa through increased sampling of the Nematoda. We selected this dataset as it represents a key study solving what is now accepted as a notable long branch attraction artefact. Here, we tested whether the canary method was able to recover the monophyly of Ecdysozoa by removing problematic sequences. In this experiment all nematode sequences were designated as "sequences of interest", as these sequences were the focus of the Aguinaldo et al. （1997) study. Following the canary sequence approach (see figure 2), after the construction of the "base dataset" and the "full dataset" and their respective trees, three "checking datasets" were generated, each consisting of 46 18S rRNA sequences – the "base dataset" plus one nematode sequence of interest. Every compositionally homogenous nematode sequence (i.e. sequence of interest) that resolved in a different phylogenetic position in the "checking tree" and the "full tree", while otherwise the "checking tree" was isomorphic with "the base tree", was selected as a canary sequence.

Once canary sequences were identified, they were added to the "base dataset" to form the "canary dataset", which contained 47 sequences, and the sequence that was not determined to be a canary sequence was moved to the non-canary sequences dataset. As there was only one non-canary sequence of interest one "canary checking dataset" was constructed, consisting of 48 sequences. The "canary checking tree" was compared with the "canary tree" (see point 5 above) to evaluate whether the sequence of interest was "potentially problematic" or not, and whether it was to be excluded from the "minimal dataset" and the "minimal tree" that we built to conclude the canary approach (see point 6 above).

Case Study 2: The original 18s rRNA analysis of Carranza et al. (1997) was unable to recover a monophyletic Platyhelminthes (inclusive of Catenulida) despite increased sampling of the Platyhelminthes. Here, we used the Carranza et al. (1997) dataset to attempt to establish whether a monophyletic Platyhelminthes could instead be recovered through application of the canary sequence approach. This dataset was chosen because both the "full tree" and "base tree" (point one above) do not conform to modern understandings of platyhelminth relationships. Accordingly, this test allowed us to evaluate the extent to which the canary approach is robust to the use of an inaccurate "base tree" to identify canary and non-canary sequences.

We started by considering all 15 platyhelminth sequences in the dataset as "sequences

of interest", as these sequences were the focus of the Carranza et al. (1997) original study. We thus defined the "base dataset" as the complete dataset of Carranza et al. (1997), the "full dataset", minus the platyhelminth sequences. We then generated 15 checking datasets, each consisting of 16 species – the base dataset plus one sequence of interest (as in point two above). We followed the rules in points two to four above to partition the sequence of interest in "canary sequences" and "non-canary sequences of interest".

Once canary sequences were identified, they were added to the base dataset to generate the canary dataset (point 3 above), which contained 16 sequences. The non-canary sequences of interest identified were 14, and we thus then generated 14 "canary checking datasets" consisting of 17 sequences each – the canary dataset plus one non-canary sequence of interest. The 14 "canary checking trees" were compared to the "canary tree" to identify and remove the "potentially problematic sequences" to generate the "minimal dataset" and conclude the canary approach (see point 6 above).

Simulation datasets: Fifty simulation datasets were constructed in PAML evolver (Yang 2007), using the Aguinaldo et al. (1997) dataset and the Rev model. Each dataset therefore included 49 sequences 1968 nucleotides long – where 1956 was the length of the shortest sequence in the Aguinaldo et al. (1997) dataset. However, we increased the length of the long-branched sequences by 250% to further exacerbate long branch attraction artifacts and increase the number of datasets where a standard phylogenetic analysis would be expected to recover an incorrect tree. This made the two long branches (*Strongyloides* and *Caenorhabditis*) ~10 times longer than the next longest branches in the simulation. For each simulated dataset we recovered trees using the JC69 model, to increase chances of recovering an incorrect topology, which we identified as any incorrect arrangement of nematode species (i.e. all cases where nematodes were not monophyletic or not members of Ecdysozoa). Simulated datasets that did not recover an incorrect topology, where no canary could be identified or where all sequences emerged as canary sequences were not further considered as we only wanted to evaluate the number of successes in cases in which the full, standard, canary pipeline could be applied (points one to six above). A success in the application of the canary approach was defined as the recovery of a monophyletic Nematoda within the non-arthropod Ecdysozoa.

Opsin dataset: We assembled a dataset of 98 well-characterised bilaterian opsins - downloaded from the NCBI website. This dataset was assembled to avoid biasing the taxonomic composition of our dataset in favour of groups that are overrepresented in sequence databases, such as the Vertebrata in the ciliary opsins, and the Arthropoda in the rhabdomeric opsins (see Heath et al., 2008 for more details). Our dataset included sequences

sampled from all bilaterian C, R and Group 4. We did not include bilaterian sequences from recently proposed opsin families: xenopsins, chaopsins and bathyopsins (Ramirez et al. 2016) as these families invariably share common ancestors with another canonical bilaterian opsin family (Ramirez et al. 2016), and therefore the order of the most basal duplication in the opsin family is fully defined by the order in which the C, R and Group 4 opsins emerged. To this core group of sequences, we added opsins from non-bilaterian lineages sampled from three recent studies: Feuda et al., (2012, 2014); Ramirez et al., (2016); Schnitzler et al., (2012) , for a total of 115 sequences – note that these sequences might include non-bilaterian representatives of the non-canonical opsin families. When sequences that were identical between the datasets were removed, the number of sequences retained dropped to 78; of these sequences, 5 belong to the Ctenophora, and 73 to the Cnidaria. Opsin sequences from Hering and Mayer (2014) were not directly considered, as all the sequences in this study were included in at least one of the other three considered datasets. The 78 ctenophoran and cnidarian sequences constitute our "sequences of interest" (see Figure 2, Figure 3), while the 20 bilaterian opsin sequences considered constitute our "base dataset", whilst the "full dataset" is comprised of 98 sequences: the "base dataset" plus the "sequences of interest" (as in point one above). Opsin sequence alignments were generated using MUSCLE (Edgar 2004) and phylogenetic analyses were performed under the GTR+G (see Feuda et al. 2012, 2014 and Vocking et al., 2017 for the rationale) model in Phylobayes 3 (Lartillot et al. 2009), Comparing the maximum discrepancies observed over the bipartitions and the effective sample size in bpcomp and tracecomp (which are included in the Phylobayes distribution) was used to assess convergence. For all analyses two independent chains were run, and a burnin of 50% of the sample size was used for all analyses, sampling every fiftieth tree following the burnin period. All alignments and Newick tree files of the canary sequence methodology are available at: https://bitbucket.org/flemingj/canarysequencemethodology.

Results and Discussion

The canary approach correctly identifies Ecdysozoa monophyly using the Aguinaldo et al. (1997) dataset: In figure S2 we show that the canary sequence approach can be applied to Aguinaldo et al. (1997) data set to recover a monophyletic Ecdysozoa. The Aguinaldo et al., (1997) dataset is composed of 18s sequences – some of the Nematode representatives in this dataset are long branched and attracted to the root of the tree (Holton & Pisani 2010) under certain analytical conditions. This is a well understood problem that produces a known and replicable phylogenetic artefact when analysed using poorly fitting substitution models. We followed the protocol in Figure 2 and points one to six above to identify "canary" and "non-canary sequences of interest" and to ultimately remove all "potentially problematic" sequences in this dataset. Two sequences emerged as canary

sequences: the 18s rRNA sequences for *Caenorhabditis* and *Trichuris*. One sequence emerged as "potentially problematic": the *Strongyloides* 18S rRNA sequence. The "minimal tree" that excludes the *Strongyloides* 18S rRNA sequence recovered monophyletic Ecdysozoa (see Figure S2).

The canary method correctly resolves Platyhelminthes using the Carranza et al (1997) dataset: To more firmly assess the capabilities of the canary approach, a second dataset was analysed – Carranza et al. (1997). Carranza et al. (1997) undertook a study of eighteen 18S rRNA "flatworm" sequences (3 Acoela and 15 Plathelminthes). They found a monophyletic Platyhelminthes separated from a monophyletic Acoelomorpha. Acoelomorpha emerged as the sister to the other Bilateria (but not in all their analyses). However, they failed to recover a monophyletic Lophotrochozoa, inclusive of the catenulid flatworms. However, current molecular consensus indicates that Platyhelminthes are a monophyletic member of the Lophotrochozoa (Halanych 2004), with the position of the Acoelamorpha still being disputed (e.g. Philippe et al., 2019). We focused on the "flatworms" (Platyhelminthes plus Catenulida, minus Acoelomorpha, considering the current controversy over their current phylogenetic placement), which we considered to be our "sequences of interest". We followed the protocol in Figure 2 (and points 1 to 6 above) to identify "canary sequences" and "non-canary sequences of interest" from our flatworm sequences. Only the 18S rRNA of *Discocelis tigrina* was found to be a canary sequence, and of the non-canary sequences of interest, only the 18S rRNA of *Planocera* emerged as "not problematic". A "minimal dataset" (see Figure 2) was derived including these two flatworm sequences only (*Planocera* and *Discocelis tigrina)*. The minimal tree recovered monophyletic Platyhelminthes, and Lophotrochozoa, in accordance with current molecular consensus (see Figure S3).

Simulation Datasets: We then applied the approach described in Figure 2 (and points one to six above) to 50 simulated datasets (see supplementary information for datasets). We found that the canary approach has a 66% success rate against our relevant datasets. While a 66% success rate is not overwhelmingly high, it should be noted that (1) we are aware of no other approaches that are available to identify problematic sequences in single-gene analyses, and that (2), in the 34% of the cases where the method did not improve the analytical result, failure of the canary approach was caused by its inability to identify and thus exclude problematic sequences. Accordingly, the canary approach seems conservative and, based on our current set of results, when it fails, it is not because it identifies false positives (i.e. it does not seem to erroneously identify non problematic sequences as "potentially problematic"). Accordingly, even in the worst case scenario, the application of the canary method does not seem to lead to results that are worse than those that would have been obtained if the method was not applied.

Caenorhabditis elegans 18S was not rejected in the original dataset, however, the sequences simulated to represent this taxon were rejected in 57.5% of the simulations, identifying the long branch (and potentially suggesting that the long branch in the particular simulation dataset was problematic in a way that the 'real' *Caenorhabditis elegans* 18S is not). Similarly, sequences simulated to represent *Strongyloides stercorali* (which was rejected in the original dataset) were rejected in 63.6% of the simulations. 78.8% of the successful simulations reject at least one of these two simulated sequences, with the remaining sequences being able to resolve a correctly positioned monophyletic Nematoda. As the canary sequence approach scales with the capabilities of the models used to resolve the "checking tree" and "canary checking tree", better results could be expected in simulation using more sophisticated models that were not used here to maintain comparability with the original results of Aguinaldo et al. (1997).

Identifying problematic non-bilaterian opsin: We sampled 115 cnidarian and ctenophoran sequences from Schnitzler et al., (2012) (19 sequences), Feuda et al., (2014) (31 sequences) and Ramirez et al. (2016); (65 sequences). Of these sequences, 37 were found to be identical (the same sequence but possessing different names between the datasets) leaving a total of 78 non-bilaterian opsins (sequences of interest) and 85 bilaterian opsins (base dataset). The canary approach found 37 of the 78 non-bilaterian opsin sequences to be problematic (see Table S1, supplementary information for further details). Of the 37 discarded, 10 were present in Feuda et al. (2014), 32 in Ramirez et al (2016) and 10 inSchnitzler et al. (2012). The starlet sea anemone *Nematostella vectensis* provided the highest number of sequences of interest, but also the highest number of problematic sequences, whereas the anthomedusan *Cladonema radiatum* and the box jellyfish *Tripedalia cystophora* provided the largest proportion of non-problematic sequences. Only two of eight opsins were problematic for *Cladonema radiatum*, whilst five of eighteen opsins were problematic in the case of the box jellyfish (see Table S1).

The "Minimal" Opsin tree: Once "potentially problematic" cnidarian and ctenophoran sequences were excluded from the analyses, the "minimal opsin tree" showed that the remaining non-bilaterian opsins were related to two groups: the rhabdomeric opsins and the ciliary opsins (Figure 4, Figure 5). More precisely, non-bilaterian sequences that in Ramirez et al. (2016) emerged as xenopsins (sharing a common ancestor with the group 4 opsins – see Figure 1) and as "canonical cnidarian visual opsins" (sharing a common ancestor with the ciliary and rhabdomeric opsins – Figure 1) were all recovered as sharing a common ancestor with the bilaterian ciliary opsins. In Feuda et al. (2014) these sequences resolve as members of either the group 4 opsins or the ciliary opsins. In Schnitzler et al. (2012), these same sequences either emerge as group 4 opsins or as the sister of both the group 4 and ciliary

opsins. It is notable that our "Minimal opsin tree" has elements in common with the trees of Feuda et al. (2014), Ramirez et al. (2016), and Schnitzler et al. (2012), whilst also differing from all of these trees, suggesting some sort of consensus solution instead. Cnidarian sequences that are resolved as rhabdomeric in our minimal opsin tree also emerged as rhabdomeric in Feuda et al. (2012, 2014), whilst Schnitzler et al. (2012) these sequences emerged as the sister group of all the other opsins. In Ramirez et al. (2016) these same non-bilaterian opsins emerged as members of the newly proposed chaopsins group together with four echinoderm opsins, i.e. they were suggested to have a common ancestor with the group 4 opsins instead.

Cnidarian and ctenophoran group 4 opsins are not recovered in our minimal opsin tree. Accordingly, our results suggest either an independent loss of the group 4 opsins in the non-bilaterians or that all non-bilaterian group 4 opsin sequences are problematic according to the canary approach. The latter hypothesis is supported by Schnitzler et al (2012) and Feuda et al. (2014), both of whom recovered cnidarian and ctenophoran sequences within the group 4 opsins that were excluded as problematic by the canary sequence approach. However, the suggestion of a real loss of the group 4 opsins within non-bilaterians is supported by Ramirez et al. (2016), in which sequences recovered as group 4 opsins by the previously cited studies were instead recovered as members of the non-canonical opsin families. In any case, it is clear that the presence of group 4 opsins in non-Bilateria deserves further investigation.

Two particularly important non-bilaterian opsins are mnemiopsis3 and acropsin3. The first was found at the root of the opsin tree in Schnitzler et al. (2012), in presumed agreement with the Ctenophora-sister hypothesis. However, Feuda et al. (2014) suggested that the placement was a phylogenetic artefact and that this sequence was more likely linked to the group 4 opsins. Here, we found mnemiopsis3 to be problematic, and thus likely to be involved in the generation of tree reconstruction artefacts. This conclusion is in accordance with Feuda et al. (2014). However, as this sequence was removed by the canary sequence method we could not confirm this sequence as a Group 4 Opsin.

Acropsin3 was found by Mason et al. (2012) to link to a G-protein of the Gq type (as expected from rhabdomeric opsins), and there is thus biochemical evidence suggesting that this protein might be a rhabdomeric opsin. Indeed, Feuda et al. (2014) found acropsin3 to be a rhabdomeric opsin nesting with two more sequences from *Nematostella* that Feuda et al. (2012) and Suga et al. (2008) previously resolved as cnidarian rhabdomeric opsins. However, Ramirez et al. (2016) found these sequences to be the sister of both the ciliary and rhabdomeric opsins, raising doubts about whether cnidarian rhabdomeric opsins exist.

Acropsin3 emerged as a canary sequence in our study. This suggests that its position might be affected by the inclusion of problematic sequences in the dataset. The application of the canary approach suggested that the putative *Nematostella* rhabdomeric opsins of Feuda et

al. (2012, 2014) are problematic and could have had a negative impact also on the placement of acropsin3 in Feuda et al. (2014). However, also in the minimal opsin tree, which excludes all potentially problematic sequences, acropsin3 emerged (together with two more non problematic *Nematostella* sequences) as a rhabdomeric opsin, strengthening the evidence for the existence of this opsin type in Cnidaria (Feuda et al. 2012, 2014; Suga et al. 2008) and further suggesting that cnidarians might possess rhabdomeric opsins (Figure 5).

Conclusions: We develop a method that can identify potentially problematic sequences in single gene datasets. We validated the test using case studies and simulation and then applied it to the problem of understanding opsin evolution. While we investigated the removal of potentially problematic sequences from the dataset, it is clear that such sequences could be retained, and we do not necessarily advocate their exclusion from an analysis. If one was to retain all the sequences from a dataset, the result of the canary pipeline would still be useful, as knowledge of which sequences in the dataset are "potentially problematic", and which are "canary sequences" (i.e. unstable) would still be useful when interpreting the results of a phylogenetic analysis. Our minimal opsin tree confirms that the three main canonical opsin lineages emerged before the separation of Cnidaria, Ctenophora and Bilateria (Figure 5). Ctenophora possesses sequences that share a common ancestor with the bilaterian ciliary opsins, and the position of the ciliary opsins in the minimal opsin tree suggests that the shared ancestor of Ctenophora, Cnidaria and Bilateria possessed three opsins. These opsins emerged from two duplications in the stem lineage subtending the crown defined by these taxa. Whether that lineage is the stem metazoan lineage or the stem eumetazoan lineage will depend on whether Porifera represent the sister group of all the other animals (Feuda et al. 2017; King & Rokas 2017; Pett et al. 2019; Pisani et al. 2015; Zhao et al. 2019) . Irrespective of that, according to our minimal opsin tree, the first duplication in opsin history separated the rhabdomeric opsin from the common ancestor of the ciliary and of the group 4 opsins. The second separated the ciliary opsins from the group 4 opsins (Feuda et al. 2012, 2014; Hering & Mayer 2014). Accordingly, we argue that the absence of rhabdomeric opsins in Ctenophora and of group 4 opsins in Cnidaria and Ctenophora can be attributed to either a secondary loss or a failure to unambiguously detect genes belonging to this opsin family. We suggest the latter possibility to be more likely.

Acknowledgements

This work is supported by a NERC-GW4 PhD studentship (to JF) and a NERC grant (NE/P013678/1). The authors would like to thank Todd Oakley for useful comments about our manuscript.

Figure Captions

Figure 1. Competing hypothesis on the phylogenetic affinities of canonical opsin families. (A) Porter et al. 2011; **(B)** Feuda et al. 2012, Feuda et al. 2014 and Hering and Meyer 2014 **(C)** Ramirez et al. 2016.

Figure 2. A flowchart illustrating the canary sequence methodology. The red stage represents the first stage of the methodology, in which sequences are classified as members of the base dataset or sequences of interest. The orange stage represents the intermediate stage of the methodology, in which sequences are assessed using checking datasets to determine whether or not they are canary sequences. The green stage represents the final stage of the methodology, in which non-canary sequences are assessed using canary checking datasets in order to produce the minimal tree. The stages are numbered with respect to the stages described in the methods section of this paper. A more detailed description of the methodology is available as Figure S1.

Figure 3. The canary sequence methodology applied to the opsin dataset. The number of sequences at each stage of the canary sequence approach, when applied to the non-bilaterian opsin sequences. Each stage is colour-coded to correspond to the stages depicted in figure 2.

Figure 4. The Minimal opsin tree recovered under GTR+G. Support values (Bayesian PPs) are reported only for key nodes.

Figure 5. Synopsis of opsin evolution (A) Phylogenetic distribution of canonical opsin in Eumetazoans. **(B**) Duplication pattern of opsin genes in Eumetazoa. Dashed lines indicates lineage-specific, losses.

Aguinaldo AM et al. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 387:489–493. doi: 10.1038/387489a0.

Anderson, F.E. and Swofford, D.L., 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. Molecular phylogenetics and evolution, 33(2), pp.440-451.

Brinkmann H, Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol. 16:817–25.

Carranza S, Baguñà J, Riutort M. 1997. Are the Platyhelminthes a monophyletic primitive group? An assessment using 18S rDNA sequences. Mol. Biol. Evol. 14:485–497. doi: 10.1093/oxfordjournals.molbev.a025785.

Dabert M, Witalinski W, Kazmierski A, Olszanowski Z, Dabert J. 2010. Molecular phylogeny of acariform mites (Acari, Arachnida): strong conflict between phylogenetic signal and long-branch attraction artifacts. Mol. Phylogenet. Evol. 56:222–241. doi: 10.1016/j.ympev.2009.12.020.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–7. doi: 10.1093/nar/gkh340.

Felsenstein J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. Systematic Zoology. 27:401–410.

Feuda R et al. 2017. Improved Modeling of Compositional Heterogeneity Supports Sponges as Sister to All Other Animals. Current Biology. 27:3864-3870.e4. doi: 10.1016/j.cub.2017.11.008.

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

Feuda R, Rota-Stabelli O, Oakley TH, Pisani D. 2014. The comb jelly opsins and the origins of animal phototransduction. Genome biology and evolution. 6:1964–1971.

Guindon S et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–21. doi: 10.1093/sysbio/syq010.

Halanych KM. 2004. The New View of Animal Phylogeny. Annual Review of Ecology, Evolution, and Systematics. 35:229–256. doi: 10.1146/annurev.ecolsys.35.112202.130124.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. doi: 10.3724/SP.J.1002.2008.08016.

Hering L, Mayer G. 2014. Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in Panarthropoda. Genome biology and evolution. 6:2380–2391.

Holton TA, Pisani D. 2010. Deep Genomic-Scale Analyses of the Metazoa Reject Coelomata: Evidence from Single- and Multigene Families Analyzed Under a Supertree and Supermatrix Paradigm. Genome Biol Evol. 2:310–324. doi: Doi 10.1093/Gbe/Evq016.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–31. doi: 10.1016/j.tig.2006.02.003.

King N, Rokas A. 2017. Embracing Uncertainty in Reconstructing Early Animal Evolution. Current Biology. 27:R1081–R1088. doi: 10.1016/j.cub.2017.08.054.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics. 25:2286–8. doi: btp368 [pii] 10.1093/bioinformatics/btp368.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095–109. doi: 10.1093/molbev/msh112.

Mason B et al. 2012. Evidence for multiple phototransduction pathways in a reef-building coral. PLoS One. 7:e50371. doi: 10.1371/journal.pone.0050371.

Pett W et al. 2019. The Role of Homology and Orthology in the Phylogenomic Analysis of Metazoan Gene Content. Mol Biol Evol. 36:643–649. doi: 10.1093/molbev/msz013.

Philippe H et al. 2019. Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. Current Biology. 29:1818-1826.e6. doi: 10.1016/j.cub.2019.04.009.

Pisani D et al. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proceedings of the National Academy of Sciences. 112:15402–15407.

Pisani D. 2004. Identifying and Removing Fast-Evolving Sites Using Compatibility Analysis: An Example from the Arthropoda. Systematic Biology. 53:978–989.

Plachetzki DC, Degnan BM, Oakley TH. 2007. The Origins of Novel Protein Interactions during Animal Opsin Evolution. PLoS One. 2. doi: ARTN e1054 DOI 10.1371/journal.pone.0001054.

Porter ML et al. 2011. Shedding new light on opsin evolution. Proc Biol Sci. doi: rspb.2011.1819 [pii] 10.1098/rspb.2011.1819.

Ramirez MD et al. 2016. The Last Common Ancestor of Most Bilaterian Animals Possessed at Least Nine Opsins. Genome Biol Evol. 8:3640–3652. doi: 10.1093/gbe/evw248.

Roure B, Baurain D, Philippe H. 2012. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol Biol Evol. doi: 10.1093/molbev/mss208.

Schnitzler CE et al. 2012. Genomic organization, evolution, and expression of photoprotein and opsin genes in Mnemiopsis leidyi: a new view of ctenophore photocytes. BMC Biol. 10:107. doi: 10.1186/1741-7007-10-107.

Sperling EA, Peterson KJ, Pisani D. 2009. Phylogenetic-Signal Dissection of Nuclear Housekeeping Genes Supports the Paraphyly of Sponges and the Monophyly of Eumetazoa. Mol Biol Evol. 26:2261–2274. doi: Doi 10.1093/Molbev/Msp148.

Suga H, Schmid V, Gehring WJ. 2008. Evolution and functional diversity of jellyfish opsins. Curr Biol. 18:51–55. doi: Doi 10.1016/J.Cub.2007.11.059.

Terakita A. 2005. The opsins. Genome Biol. 6:213. doi: gb-2005-6-3-213 [pii] 10.1186/gb-2005-6-3-213.

Terakita A, Kawano-Yamashita E, Koyanagi M. 2012. Evolution and diversity of opsins. Wiley Interdisciplinary Reviews: Membrane Transport and Signaling. 1:104–111. doi: 10.1002/wmts.6.

Tsukamoto H. 2014. Diversity and Functional Properties of Bistable Photopigments. In: Evolution of Visual and Non-visual Pigments. Hunt, DM, Hankins, MW, Collin, SP, & Marshall, NJ, editors. Springer Series in Vision Research Springer US: Boston, MA pp. 219–239. doi: 10.1007/978-1-4614-4355-1_7.

Tsukamoto H, Terakita A. 2010. Diversity and functional properties of bistable pigments. Photochem. Photobiol. Sci. 9:1435–1443. doi: 10.1039/c0pp00168f.

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

Xi, Z., Liu, L. and Davis, C.C., 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Molecular Phylogenetics and Evolution, 92, pp.63-71.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591. doi: 10.1093/molbev/msm088.

Zhao Y et al. 2019. Cambrian Sessile, Suspension Feeding Stem-Group Ctenophores and Evolution of the Comb Jelly Body Plan. Current Biology. 29:1112-1125.e2. doi: 10.1016/j.cub.2019.02.036.

Figure 1

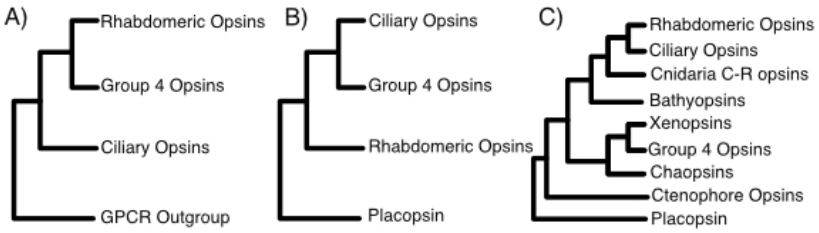

Figure 1. Competing hypothesis on the phylogenetic affinities of canonical opsin families. (A) Porter et al. 2011; (B) Feuda et al. 2012, Feuda et al. 2014 and Hering and Meyer 2014 (C) Ramirez et al. 2016.

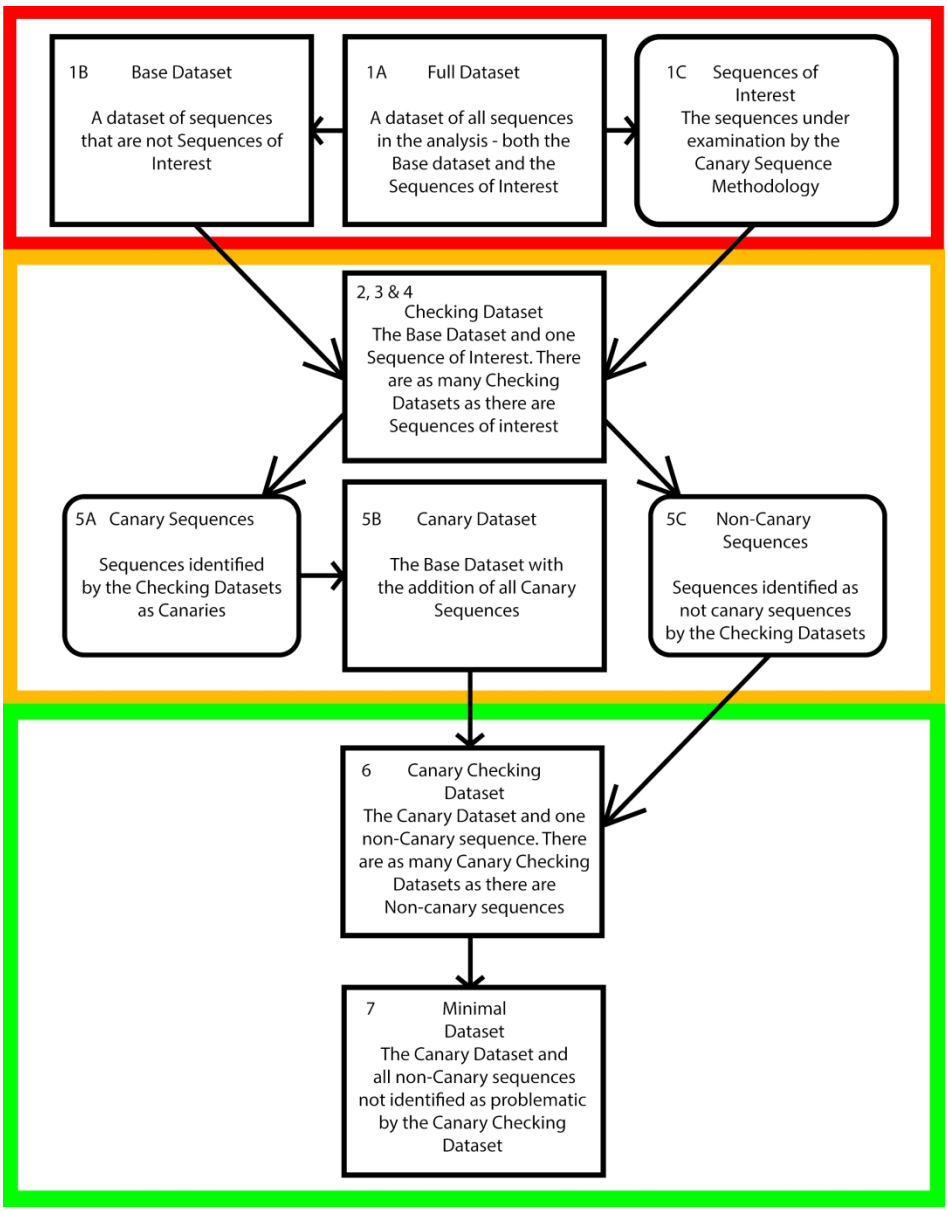

Figure 2. A flowchart illustrating the canary sequence methodology. The red stage represents the first stage of the methodology, in which sequences are classified as members of the base dataset or sequences of interest. The orange stage represents the intermediate stage of the methodology, in which sequences are assessed using checking datasets to determine whether or not they are canary sequences. The green stage represents the final stage of the methodology, in which non-canary sequences are assessed using canary checking datasets in order to produce the minimal tree. The stages are numbered with respect to the stages described in the methods section of this paper. A more detailed description of the methodology is available as Figure S1.

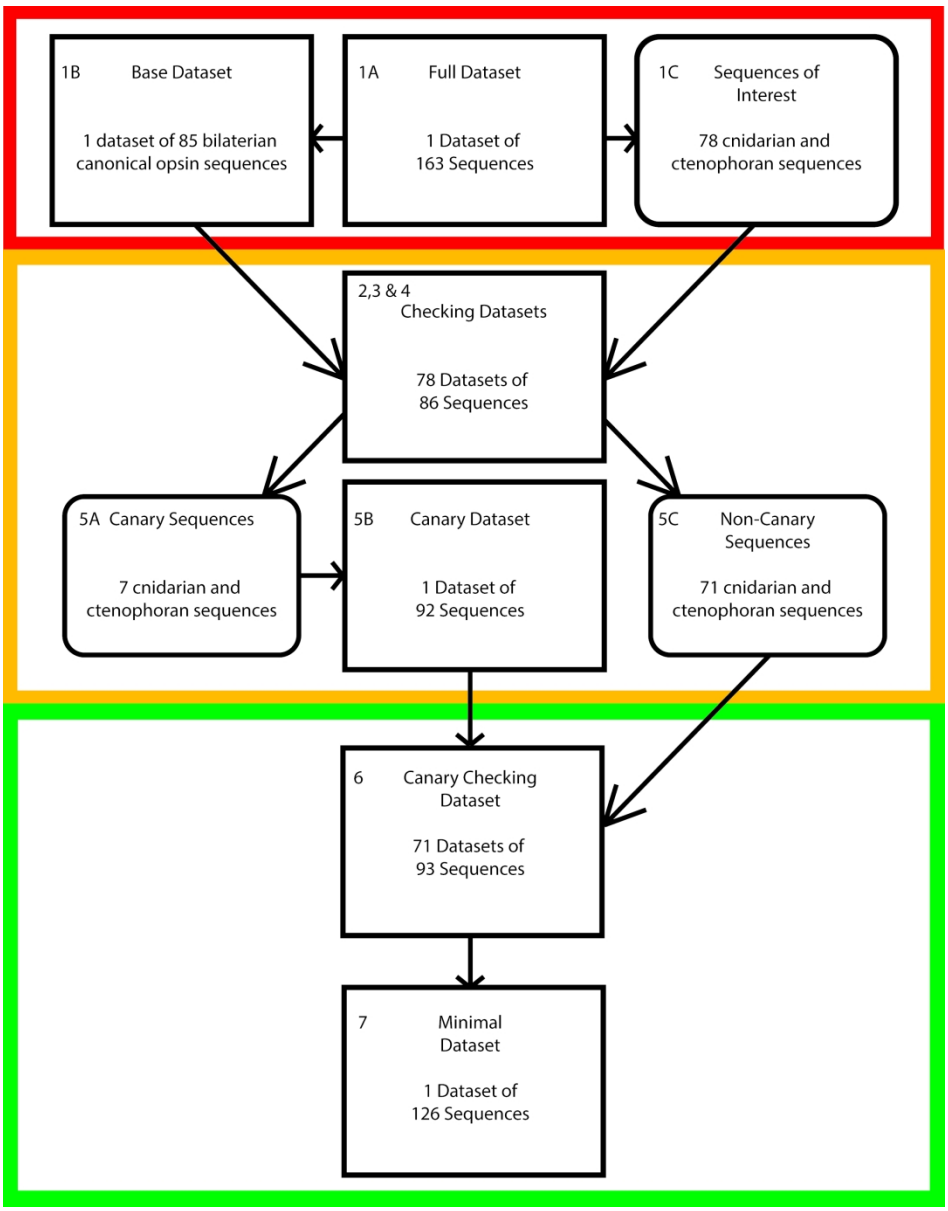

Figure 3. The canary sequence methodology applied to the opsin dataset. The number of sequences at each stage of the canary sequence approach, when applied to the non-bilaterian opsin sequences. Each stage is colour-coded to correspond to the stages depicted in figure 2.

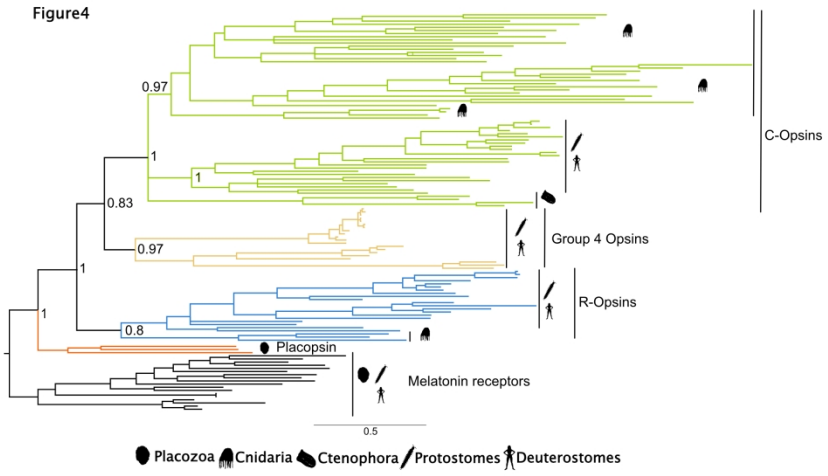

Figure 4. The Minimal opsin tree recovered under GTR+G. Support values (Bayesian PPs) are reported only for key nodes.

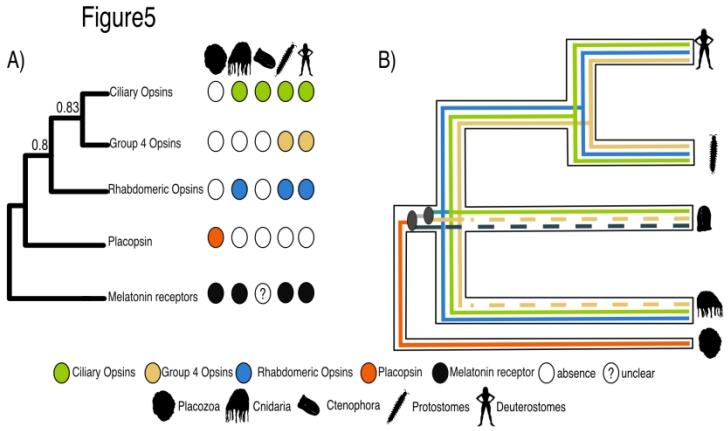

Figure 5. Synopsis of opsin evolution (A) Phylogenetic distribution of canonical opsin in Eumetazoans. (B) Duplication pattern of opsin genes in Eumetazoa. Dashed lines indicates lineage-specific, losses.