How did vertebrate visual opsins diversify?
- putting the last pieces of the puzzle together

David Lagman[1*], Christina A. Bergqvist[1], Shigehiro Kuraku[2,3]

1) Department of Medical Cell Biology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
2) Molecular Life History Laboratory, National Institute of Genetics, Mishima, Japan
3) Department of Genetics, Sokendai (Graduate University for Advanced Studies), Mishima, Japan
*) Correspondence: david.lagman@uu.se

Abstract

A crucial event in the evolution of compex vertebrate vision was the expansion of the visual opsin family in the ancestral vertebrate. However, it has been controversial whether the expansion was achieved through whole genome duplications or in tandem gene duplications. We here performed a vertebrate-wide genome comparison, including jawless vertebrates, which strongly favors the latter scenario. Our analyses uncovered frequent paralogy underlying homologous phenotypes caused by massive gene losses and distinct paralog utilization between jawless and jawed vertebrates.

Main

Vertebrates have diverse visual capacities which primarily stem from differences in the absorption spectra of light detecting proteins expressed in the retina of the eye, the visual opsins. These are expressed in distinct types of photoreceptor cells with different but related molecular architecture, which together with the overall retinal organization contributes to the visual capacity [1].

Vertebrate visual opsins are classified into five groups: long wavelength sensitive opsins (*LWS*), two short wavelength sensitive opsin types (*SWS1* and *SWS2*), medium wavelength sensitive opsins (*RH2*) and the dim-light rhodopsin (*RH1*) [2]. These types are thought to have evolved through either of two different scenarios: 1) a single tandem duplication followed by two rounds of whole genome duplication (WGD) in the vertebrate ancestor [3–5] (Fig. 1a); or 2) four tandem gene duplications before these two WGD events (Fig. 1b) [6]. Dim-light (scotopic) vision might have been in both cases provided by pinopsin before these duplications [7]. Since more gene losses would be required by scenario 2, this scenario could be considered less parsimonious under maximum parsimony, especially in the light of the strong synteny evidence for an expansion in the two whole genome duplications [3]. Resolving this issue is important since the answer has implications for at what point in vertebrate evolution complex features of the vertebrate eye such as color vision, or a more recently suggested ancestral function of different feature channels for visually guided behaviors [8], could evolve. Analyses of visual opsin gene locations and chromosome relationships in jawless vertebrates could provide the important missing piece of evidence. However, this was not possible at the time of these studies due to the lack of chromosome-scale genome assemblies.

Here we have analyzed the genome assemblies of two southern hemisphere lampreys and four publicly available northern hemisphere lampreys [9–12], all at the chromosome-scale. This allowed us to perform exhaustive visual opsin searches in, and chromosomal comparisons between the different species. Our phylogeny inference suggests the best representation of the

jawless vertebrate visual opsin repertoires by pouched lamprey with frequent loss and pseudogenization in other lamprey species (Fig 1c and Extended data Fig. 1).

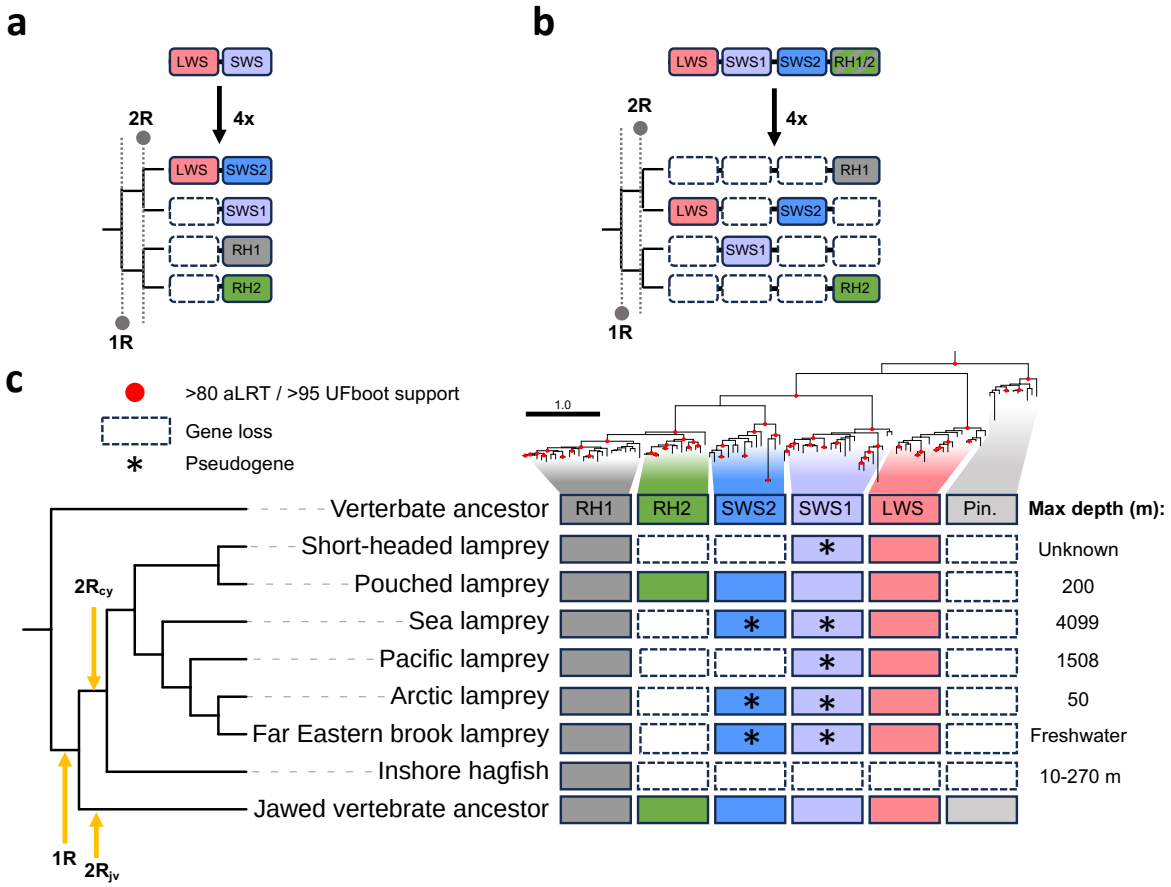


Figure 1: Evolutionary scenario for visual opsins in vertebrates and gene repertoires of jawless vertebrates. a) Evolutionary scenario based on relationship between chromosomal regions in gnathostomes [3–5]. b) More recent evolutionary scenario based on sequence phylogeny of the visual opsins [6,13]. c) Our phylogenetic analysis of the visual opsins (full tree in Extended data fig. 1) combined with observed repertoires of each jawless vertebrate species investigated. The exon structure of the short-headed lamprey RH1 gene, rearranged in the genome assembly, was curated based on transcriptome data to retrieve the full open reading frame. 1R, the first WGD that was shared between jawed and jawless vertebrates; 2R, the previously hypothesized shared second event; $2R_{jv}$,the jawed vertebrate specific WGD event; $2R_{cy}$, the jawless vertebrate-specific triplication event. The phylogenetic relationship among jawless vertebrates is based on [14]. Max observed ecological depth references are based on existing literature[15,16] and FishBase.org.

When investigating the chromosomal locations of the visual opsins we observed that jawed and jawless vertebrates have different opsin gene synteny patterns: *RH2* and *SWS2* are close to each other in the pouched lamprey (~6 Kbp apart) and *RH1* and *LWS* are close (<17 Kbp) in all lamprey species (except in the Pacific lamprey in which *LWS* remains split between different scaffolds). In jawed vertebrates it is *LWS* and *SWS2* that are located close to each other (~10 Kbp and ~25 Kbp in the spotted gar and in the green anole genomes respectively). These observations of gene positioning provide invaluable evidence for a scenario with multiple ancient tandem duplications (Fig 1b), compared to the previously suggested scenario with a single tandem duplication (Fig. 1a).

Next, we generated a 'super-alignment' of concatenated genes flanking the visual opsins in all six lamprey species and performed phylogenetic analysis (as described in [17]). The resultant tree showed reliable grouping into six related chromosomal regions (Fig. 2a). This is consistent with jawless vertebrates being ancient hexaploids due to a lamprey specific genome triplication [11,17,18]. This analysis allowed us to regard the *RH1* and *RH2* split as part of the first WGD as suggested from phylogenetic evidence [6]. The analysis reveals that chromosomes that

have lost visual opsins also have fewer members of the flanking gene families (Fig. 2a). Finally, we could observe that the *SWS2* genes of jawed and jawless vertebrates were separated in the first WGD and thus are paralogs (Fig. 2b).

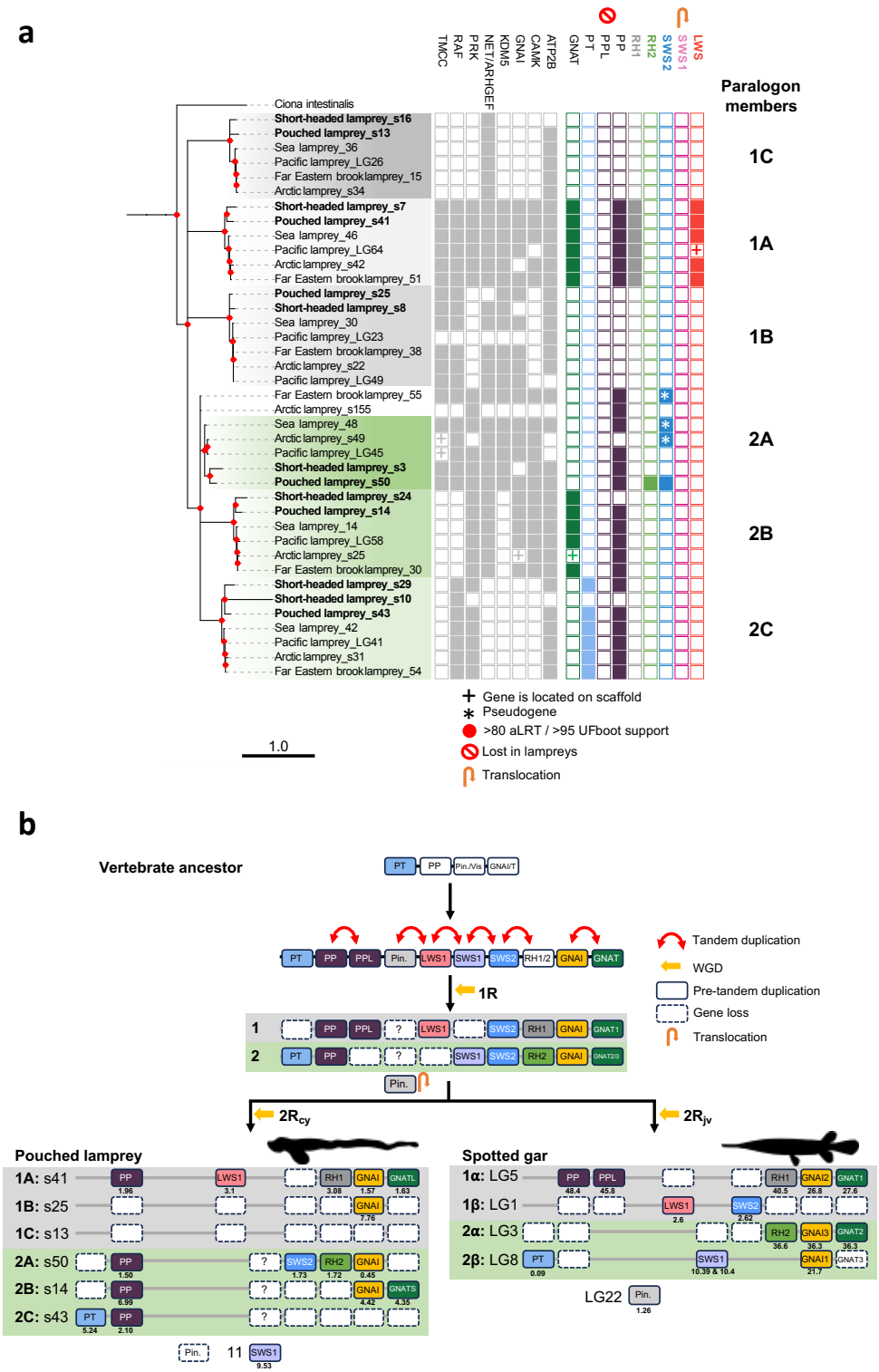


Figure 2: Proposed evolutionary scenario of the visual opsins and their chromosomes in vertebrates. a) Phylogenetic tree made using a 'super-alignment' of eight flanking gene families as well as parapinopsin (PP). VA opsin, another non-visual opsion, was excluded since it is located in a separate genomic region. Abbreviated non-visual opsins are: PT – parietopsin; PPL – parapinopsin-Like; Pin. – pinopsin. Red circles indicate nodes with high support (>80 aLRT / >95 UFBoot support). Filled and open boxes indicate the presence and absence of ortholog, respectively. Grey and green backgrounds represent origins of *RH1* or *RH2* carrying post-1R chromosomes, respectively. The tree was rooted with a concatenated acorn worm sequence. The acorn worm sequence and a concatenated Florida lancet sequence is not shown for simplicity. b) Proposed

evolutionary history of the visual opsins, their non-visual opsin relatives, and *GNAT* and *GNAI* genes in vertebrates. The identifiers 1A-C and 2A-C are given to paralogon members as in (a) and chromosome in pouched lamprey and spotted gar. Numbers under boxes represents gene positions in Mbp in each genome assembly. The two numbers under spotted gar *SWS1* represent the two tandem duplicates of this gene in spotted gar. Paralogon members in jawed vertebrates were named as in [13]. Silhouettes of the animals were downloaded from phylopic.org.

The evolutionary scenario presented in Fig. 2b reconciles all existing phylogenetic and genomic observations. It confirms several ancient tandem gene duplications and indicate massive visual opsin gene losses due to the separate WGD events in the two vertebrate lineages. Scenario 1 assumed only two ancestral visual opsin genes in the last common ancestor of vertebrates before the first WGD. This scenario would be preferred under maximum parsimony, by requiring fewer events (gene losses) than scenario 2. However, based on the genomic positioning of visual opsin genes in lampreys presented here, scenario 2 can now be favored over scenario 1 (Fig. 2). No additional visual opsin gene duplicates survived the independent WGD events in either jawed or jawless vertebrates. Thus, the results we present here further stress the important role tandem duplications has played in visual opsin evolution [19,20]. Our analysis of this genomic region differs from the analysis of [17] in that two paralogon members (named here 1B and 2C) have switched location in our tree compared to theirs, this could be due to differences in family or species representation. However, both theirs and our study agree that the paralogon members housing the visual opsins most likely were separated in the first WGD.

Our analysis also questions the previously proposed relationship of *GNAT* genes, which are syntenic to visual opsins, among jawed and jawless vertebrates [21,22]. GNAT proteins convey the signal from the visual opsin to phosphodiesterase 6 in the signaling cascade of vertebrate photoreceptor cells. Here we show that rod- and cone-like cells of jawless vertebrates use different *GNAT* paralogs compared to rods and cones of jawed vertebrates (Fig. 2b and Extended data Fig. 2). This suggests even more complex scenarios through paralogy in molecular machineries behind seemingly homologous photoreceptor phenotypes.

In conclusion, our observations uncover a complex evolutionary history of vertebrate visual opsins with tandem duplications, massive gene losses and that *RH1* and *RH2* genes as well as *SWS2* genes are paralogs from the first WGD event. These conclusions are only made possible since we in this study have included the pouched lamprey, with its unparalleled visual opsin repertoire among jawless vertebrates, in analyses of gene positions and chromosomal relationships.

Data availability

The genome assemblies of the pouched- and short-headed lampreys are deposited in NCBI under BioProject PRJNA1030065. All sequence alignment files and the phylogenetic inference output files has been deposited at FigShare with DOI: 10.6084/m9.figshare.c.6988044.

Code availability

No new code was generated for this study.

References

1. Baden, T. & Osorio, D. *Annu. Rev. Vis. Sci* **5**, 177–200 (2019).
2. Okano, T., Kojima, D., Fukada, Y., Shichida, Y. & Yoshizawa, T. *Proc Natl Acad Sci U S A* **89**, 5932–6 (1992).
3. Lagman, D. et al. *BMC Evol Biol* **13**, 238 (2013).
4. Nordström, K., Larsson, T.A. & Larhammar, D. *Genomics* **83**, 852–872 (2004).
5. Larhammar, D., Nordström, K. & Larsson, T.A. *Philos Trans R Soc Lond B Biol Sci* **364**, 2867–2880 (2009).

6. Lamb, T.D. & Hunt, D.M. *Dev Biol* **431**, 77–92 (2017).
7. Sato, K. et al. *Commun Biol* **1**, (2018).
8. Baden, T. *Nat Ecol Evol* (2024).doi:10.1038/s41559-023-02291-7
9. Zhu, T. et al. *Mol Ecol Resour* **21**, 448–463 (2021).
10. Timoshevskaya, N. et al. *Cell Rep* **42**, (2023).
11. Nakatani, Y. et al. *Nat Commun* **12**, 1–14 (2021).
12. Hess, J.E. et al. *Mol Ecol* **29**, 3841–3856 (2020).
13. Lamb, T.D. *J Physiol* **0**, (2022).
14. Brownstein, C.D. & Near, T.J. *Current Biology* **33**, 397-404.e3 (2023).
15. Miller, A.K. et al. *Rev Fish Biol Fish* **31**, 201–232 (2021).
16. Potter, I.C., Prince, P.A. & Croxal, J.P. *Em. Biol. Fish* **4**, 65–69 (1979).
17. Marlétaz, F. et al. *Nature* (2024).doi:10.1038/s41586-024-07070-3
18. Yu, D. et al. *Nat Ecol Evol* (2024).doi:10.1038/s41559-023-02299-z
19. Musilova, Z., Salzburger, W. & Cortesi, F. *Annu. Rev. Cell Dev. Biol. 2021* **37**, 441–468 (2021).
20. Hagen, J.F.D., Roberts, N.S. & Johnston, R.J. *Dev Biol* **493**, 40–66 (2023).
21. Lagman, D., Sundström, G., Ocampo Daza, D., Abalo, X.M. & Larhammar, D. *Genomics* **100**, 203–211 (2012).
22. Lamb, T.D. *Prog Retin Eye Res* **76**, 100823 (2019).
23. Birney, E., Clamp, M. & Durbin, R. *Genome Res* **14**, 988–995 (2004).
24. Sievers, F. et al. *Mol Syst Biol* **7**, (2011).
25. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. & Minh, B.Q. *Mol Biol Evol* **32**, 268–274 (2015).
26. Emms, D.M. & Kelly, S. *Genome Biol* **20**, (2019).
27. Letunic, I. & Bork, P. *Nucleic Acids Res* **49**, W293–W296 (2021).

Acknowledgements

We thank John Donald at Deakin University for early access to the unpublished genome assemblies, David Forsberg for commenting on the manuscript draft, Trevor Lamb and Dan Larhammar for insightful discussions.

Author contributions

D.L. and S.K. conceived the study. D.L., C.B. and S.K. analyzed genomes. D.L. and C.B. performed phylogenetic analyses. D.L. and S.K. wrote the original draft and edited the manuscript. All authors approved the final version.

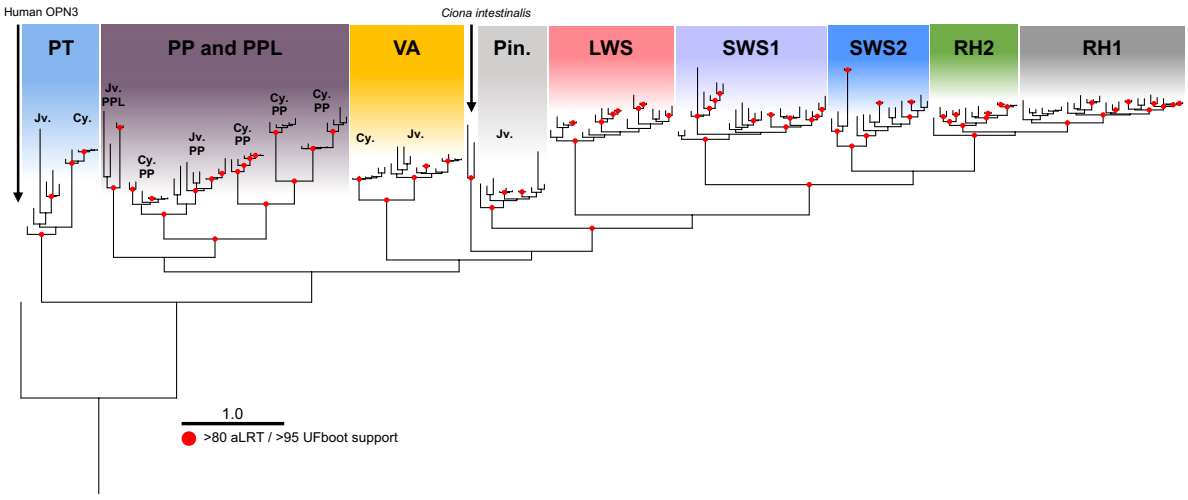
Competing interests

The authors declare no competing interests.

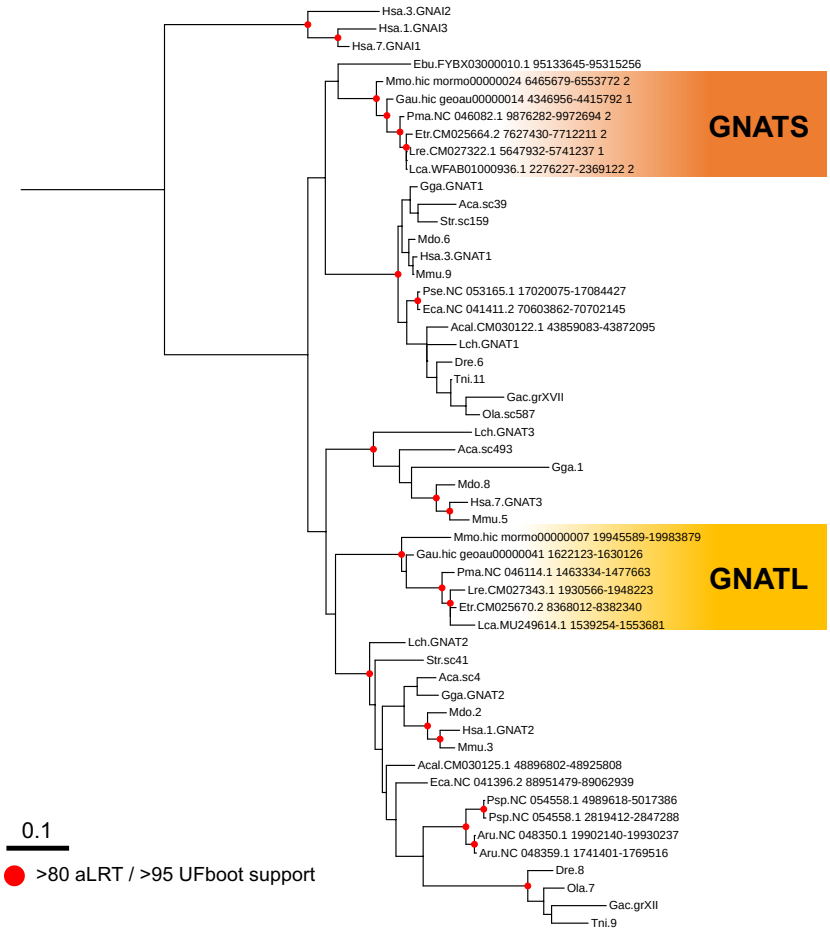
Additional information

Supplementary table 1: Accession codes for genome assemblies queried in this study.
Supplementary table 2: List of identified neighboring gene families and comments.

Extended data Figs.



Extended data Fig. 1: Full phylogenetic tree of the investigated opsins in this study. Phylogenetic analysis of the visual opsins and their close non-visual relatives. Interestingly, this analysis reveals a previously undescribed lamprey specific expansion of parapinopsin with four PP genes. Well supported clades are labelled with the name of the closest jawed vertebrate opsin subtype. Cy. represent jawless vertebrate sequences of non-visual opsins and Jv. represent jawed vertebrate sequences of non-visual opsins. Red circles indicate nodes with high support (>80 aLRT / >95 UFBoot support). The tree was rooted with the human OPN3 sequence.



Extended data Fig. 2: Phylogenetic analysis of GNAT amino acid sequences. Well supported clades are labelled with the name of the closest sea lamprey subtype. Red circles indicate nodes with high support (>80 aLRT / >95 UFBoot support). Leaf names are the species name in tree letter format followed by gene name or scaffold/chromosome accession number and position of the gene.

Methods

Identification of opsins and *GNAT/GNAI* genes and phylogenetic analyses

Visual opsin sequences from the pouched lamprey, non-visual opsins from zebrafish and chicken, and *GNAT1* and *GNAI1* sequences from human were used as TBLASTN queries against the genome assemblies of the pouched lamprey (*Geotria australis*), short-headed lamprey (*Mordacia mordax*), sea lamprey (*Petromyzon marinus*), arctic lamprey (*Lethenteron camtschaticum*), Far Eastern brook lamprey (*Lethenteron reissneri*), pacific lamprey (*Entosphenus tridentatus*), inshore hagfish (*Eptatretus burgeri*), West African lungfish (*Protopterus annectens*), Australian lungfish (*Neoceratodus forsteri*), coelacanth (*Latimeria chalumnae*), green anole lizard (*Anolis carolinensis*), grey short-tailed opossum (*Monodelphis domestica*), human (*Homo sapiens*), reedfish (*Erpetoichthys calabaricus*), Grey bichir (*Polypterus senegalus*), sterlett (*Acipenser ruthenus*) and paddlefish (*Polyodon spathula*) (see Supplementary table 1 for genome assembly accession codes). Lungfish genomes were only searched for visual opsins and *GNAT/GNAI* genes were only searched for in jawless vertebrate genomes.

For each query sequence the top five hits were retained and high scoring pairs (HSPs) that were located less or equal to 50 Kbp apart was merged into bed files. FASTA sequences for each hit was extracted and amino acid sequences were predicted using GeneWise [23] with standard settings and the BLAST query sequences as reference for prediction. The longest predicted peptides per gene was extracted and opsins were added to the visual opsin alignment from Lagman *et al.* (2013) with teleost species excluded and *GNAT/GNAI* genes were added to the alignments from Lagman *et al.* (2012). Sequences with erroneous stretches were edited either by introducing gaps or manual prediction. The resulting alignments were aligned using ClustalO [24] with standard settings. Phylogenetic trees were constructed using IQ-TREE (v. 1.6.12) [25] using the following settings: -t BIONJ -nt 8 -quiet -keep-ident -bb 10000 -alrt 10000 -m TEST -bnni.

Gene family classification and selection

The longest peptide sequences of each gene in the genomes of acorn worm (*Saccoglossus kowalevskii*), Florida lancet (*Branchiostoma floridae*), vase tunicate (*Ciona intestinalis*), pouched lamprey, short-headed lamprey, sea lamprey, Australian ghost shark (*Callorhincus milii*), human (*Homo sapiens*), chicken (*Gallus gallus*), spotted gar (*Lepisosteus oculatus*), zebrafish (*Danio rerio*) and Japanese medaka (*Oryzias latipes*) were used to create gene families using OrthoFinder [26] with the following settings: -M msa. Families with one member in acorn worm, Florida lancet or vase tunicate and more than two members in either sea lamprey, pouched lamprey, short-headed lamprey, human, spotted gar, or chicken (in total 1709 families) were categorized as families with putative origin in early vertebrate whole genome duplications.

To investigate the duplication history of the chromosomal regions harboring the visual opsins in lampreys we selected gene families with members on at least two of the three chromosomes that carry members of the visual opsin gene family in the sea lamprey genome. This resulted in the identification of a total of 15 gene families, four of these families were excluded due to complex topologies (Supplementary table 2). Out of the remaining, seven families with members on at least three chromosomes were used in further analyses.

BLAST searches, gene prediction and phylogenetic analyses of neighboring gene families

Amino acid sequences from each of the selected gene families were used as TBLASTN queries against the genome assemblies of the sea lamprey, arctic lamprey, far eastern brook lamprey,

247 pacific lamprey, pouched lamprey and short-headed lamprey. For each query sequence the top
248 five hits were retained and high scoring pairs (HSPs) that were located less or equal to 50 Kbp
249 apart was merged into bed files. FASTA sequences for each hit was extracted and amino acid
250 sequences were predicted using GeneWise [23] with standard settings and the BLAST query
251 sequences as reference for prediction. The longest predicted peptides per FASTA sequence and
252 family was extracted and combined with the vase tunicate, Florida lancet and acorn worm
253 sequences from the family alignment generated by OrthoFinder. The resulting alignments were
254 aligned using ClustalO [24] with standard settings. Phylogenetic trees for inspection of families
255 were constructed using IQ-TREE (v. 1.6.12) [25] using the following settings: -t BIONJ -nt 8 -
256 quiet -keep-ident -bb 10000 -alrt 10000 -m TEST -bnni. The resulting alignments and trees
257 were inspected for inconsistencies and short sequences were extended manually when possible.
258
259 **Analysis of paralogon evolution**
260 Alignments of neighboring gene families with one member each on at least three
261 chromosomes/scaffolds, *GNAI* and PP were inspected for sequences located on short
262 scaffolds/contigs. These were removed and each family was realigned using ClustalO. All
263 separate alignments (nine families in total) were combined into a concatenated alignment
264 (missing genes was coded as missing in the final alignment) which was subjected to
265 phylogenetic analysis using IQ-TREE (v. 1.6.12) [25] with the following settings: -t BIONJ -nt
266 8 -quiet -keep-ident -bb 10000 -alrt 10000 -m TEST -bnni. The resulting phylogenetic tree was
267 visualized using iTOL by combining with data on presence or absence of opsins, *GNAT* or the
268 neighboring families on each chromosome (https://itol.embl.de) [27].