Correspondence:

Supplemental data: Molecular evidence for dim-light vision in the last common ancestor of the vertebrates

Davide Pisani[1,2*], Samantha, M. Mohun[2], Simon, R. Harris[3], James, O. McInerney[1] and Mark Wilkinson[2]

Experimental Procedures

A data set of 47 amino acid sequences, representing all the known vertebrate visual opsins for a wide range of taxa, was downloaded from the NCBI website (see below for accession numbers). These sequences were aligned using ClustalW [S1], default alignment options, but correcting for multiple substitutions when building the guide tree. A second alignment was generated using Muscle [S2] default options. The two alignments, although extremely similar, were not identical, differing slightly in the N- and C- terminal regions. Both the ClustalW and the Muscle alignment were subjected to two GBlocks analyses [S3] to remove poorly aligned sites. The first GBlocks analysis used the default options, but allowing gapped sites. The second GBlocks analysis was more stringent with (1) the minimum length of a block, to be retained increased from 10 to 20 positions, and (2) the maximum number of contiguous non-conserved positions allowed halved from eight to four. The alignments can be downloaded at: http://bioinf.nuim.ie/davide/index.html/.

Phylogenetic analyses were performed on all the four generated datasets, and also excluding all gapped sites. The results of these analyses were always very similar in terms of both recovered trees and nodal support. Accordingly, only the results of the ClustalW generated data set, as cleaned under the less stringent (first) GBlocks analysis are presented (see main text). This alignment was 346 amino acid positions long with very few gaps.

The χ^2 test, as implemented in Tree-Puzzle 5.2 [S4] was used to test the alignment for sequence heterogneity in amino acid composition which can result in spurious phylogenetic results [S5]. No significant heterogeneity was found.

Maximum Likelihood (ML) analyses were performed using both quartet puzzling (Tree-Puzzle5.2 default puzzling options) and standard ML. Standard ML analyses were performed using the software PHYML [S6], PHYML-SPR [S7] and SPR [S7]. Support for the nodes in the ML tree were estimated using the bootstrap (100 replicates) as implemented in PHYML. The Bayesian analysis was performed using the parallel version of MrBayes 3.1.1 [S8]. For the Bayesian analysis an initial value of two million generations was run, sampling each 1000 generations. As suggested in the MrBayes 3.1.1 manual, two independent runs (and four chains for each run) were simultaneously performed, and convergence was tested comparing the average standard deviation of the split frequencies for the two independent runs. The resulting Bayesian tree was derived as the majority rule consensus of the trees sampled from both runs after convergence was reached (c. 300,000 generations).

All likelihood calculations were performed under the best fitting substitution model (WAG + G + I) selected using the Akaike Information Criterion as implemented in MultiPhyl [S9]. In the Bayesian analysis the values of G and I were estimated during the tree search.

Maximum likelihood distances among the 47 sequences were calculated using Tree-Puzzle5.2 and used to build a Minimum Evolution (ME) Tree using PAUP* [S10] (heuristic search). PAUP* was also used to perform a Maximum Parsimony (MP) analysis and a Weighted MP (WMP) analysis of the data. In both cases a heuristic search was performed (100 random addition sequences), and support for the recovered trees was estimated using the bootstrap (1000 replicates, with a single random addition sequence). For the WMP analysis the GPCR specific amino acid substitution matrix of [S11] (see also [S12]) was used.

Potential Long Branch Attraction (LBA) artefacts were investigated using the method of Pisani [8] using the C software Boildown (written and is distributed by SRH). The protocol proposed in [8] was strictly followed and all the sites with LeQuesne Probabilities (LQP; [S13]) $\geq$ 0.1 were sequentially removed. New phylogenetic analyses were preformed and the results obtained were then compared with those of the analysis of the 346 position data set.

The standard ML tree was compared with the quartet-puzzling tree and a tree displaying the Collin et al. [3] arrangement of the Rh sequences (Figure S1). CODEML, which is part of the PAML package [S14], was used to obtain site-specific likelihood values for each of the three topologies of Figure S1. These values were fed to the software CONSEL [S15] to compare the three alternative topologies using the Approximately Unbiased (AU) test [9].

Supplemental Results and Discussion

Figure 1A and B (main text) summarise the results of the quartet-puzzling, ML (PHYML, PHYML-SPR, SPR), Bayesian, minimum evolution, and equally and differentially weighted parsimony analyses. Complete results of the quartet-puzzling and PHYML analyses are reported in Figure S2. Results of the Bayesian analysis are reported in Figure S3. Because PHYML, PHYML-SPR and SPR all returned the same tree Figure S2 only reports the results of the PHYML analysis. All these analyses agree in providing strong support for the clustering of RhA within the Rh1 group but disagree in the position of Rh2. Very strong support for the clustering of RhA within the Rh1 group was also obtained also using MP (bootstrap support = 95%) and WMP (bootstrap support = 99%) whereas MP provided only 19% and WMP 9% bootstrap support for RhB as an Rh2. We did not perform a bootstrap ME analysis, but the ME tree was consistent with the quartet puzzling ML analysis in supporting RhA as an Rh1 and RhB as an Rh2. No analysis supported a monophyletic RhA plus RhB. Relationships within some putatively orthologous clusters, notably within the LWS cluster are not as expected given the species interrelationships and we suspect this is due to either paralogy or noise within this cluster.

We conjecture the correct classification of RhB is probably as an Rh2, and suggest the position of RhB in the standard ML (non quartet-puzzling), MP, WMP and Bayesian analyses is suggestive of a phylogenetic artefact [S5]. The clustering of RhB as an Rh2 implies a simpler pattern of gene duplication and losses, than its more basal position in the standard ML, Bayesian and parsimony trees. One single gene duplication in the ancestral Rh gene (predating the

agnathans-gnathostomians split) and resulting in the Rh1 and Rh2 genes is needed to explain RhB as an Rh2 (and RhA as an Rh1). Differently, if RhB is assumed to be basal to a ((Rh1,RhA), Rh2) group, then two duplications predating the agnathans-gnathostomians split, followed by two independent gene losses (of RhB within gnathostomians and Rh2 within agnathans) are needed to explain the observed pattern. Whatever the correct position of RhB is, it is clear that accurately placing this sequence proved difficult even using ML-based and Bayesian methods and may have hampered previous analyses.

We must await the isolation and sequencing of further agnathan RhB/Rh2 sequences for this gene to be conclusively and correctly classified. However, this is inconsequential for our primary conclusions that RhA is a member of the Rh1 group. Pisani's method [8] found only 22 potentially fast evolving sites and their sequential removal did not cause RhA to change its position. Also the bootstrap for the RhA plus Rh1 group was unchanged further strengthening our conclusions that RhAs are Rh1, and hence that true (Rh1 mediated) dim-light vision predated the Agnatha Gnathostomata split.

Supplemental References

S1. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673-4680.

S2. Edgar, R.C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

S3. Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540-552.

S4. Schmidt, H.A., K. Strimmer, M. Vingron, and von Haeseler A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 18, 502-504.

S5. Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the

reconstruction of the tree of life. Nat. Rev. Gen. 6, 361-375.

S6. Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696-704.

S7. Hordijk, W and Gascuel, O. (2005). Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics Advance Access published online on October 18.

S8. Ronquist, F. and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 12, 1572-1574.

S9. Keane, T.M., Creevey, C.J., Pentony, M.M., Naughton, T.J., and McInerney, J.O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol. Biol. *6*, 29.

S10. Swofford, D.L. (1998). PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. (Sunderland, Massachusetts: Sinauer Associates).

S11. Rice, K. A. (1994). The origin, evolution and classification of G protein-coupled receptors, PhD thesis. (Cambridge, Massachusetts: Harvard University).

S12. Spaethe, J., and Briscoe, A.D. (2004). Early duplication and functional diversification of the opsin gene family in insects. Mol. Biol. Evol. 21, 1583–1594.

S13. Wilkinson, M. (1997). Characters, congruence and quality: A study of neuroanatomical and traditional data in caecilian phylogeny. Biol. Rev. 72, 423-470.

S14. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13, 555-556.

S15. Shimodaira, H., Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17, 1246-1247.

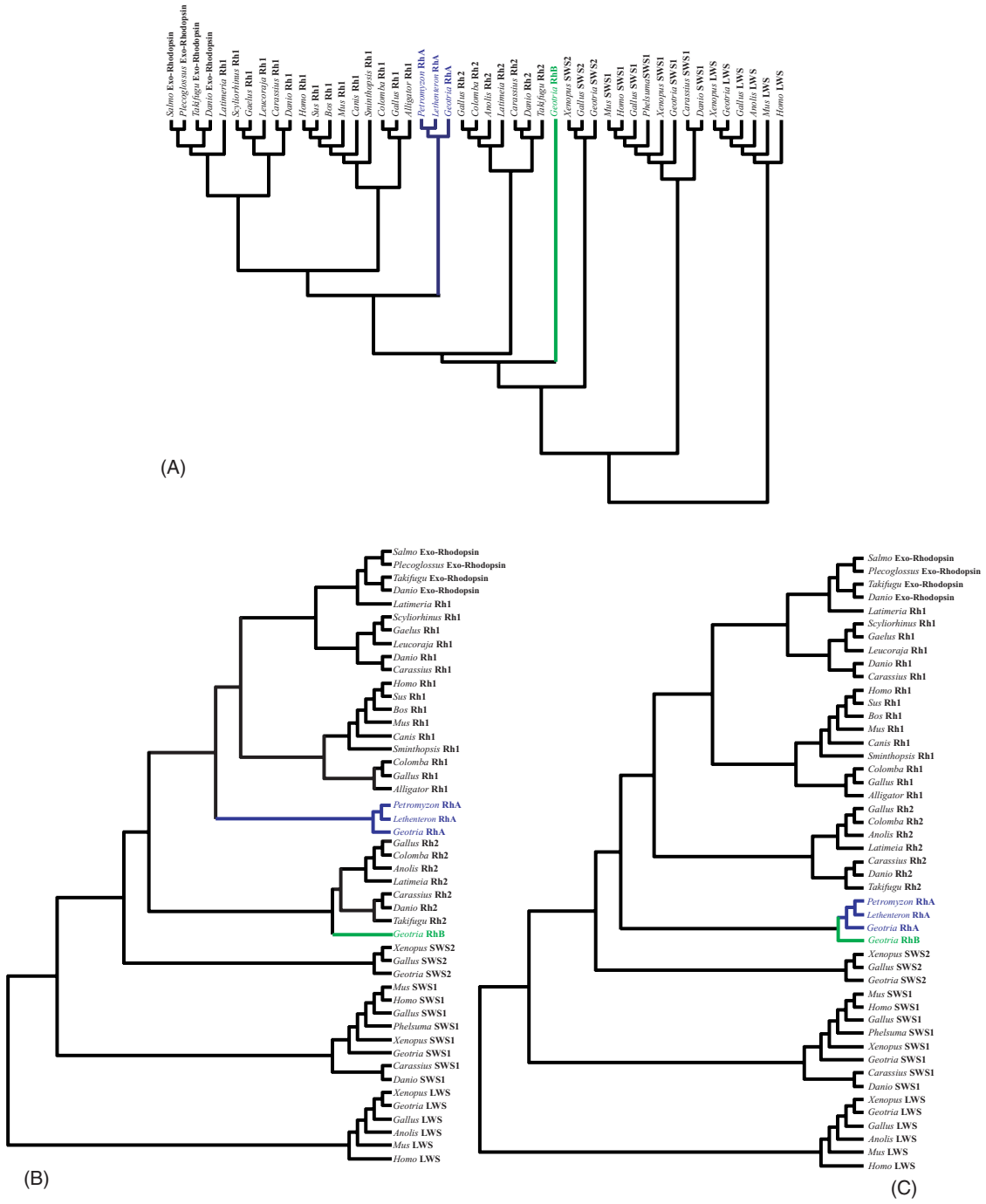

Figure S1. (A) The standard ML (PHYML, SPR-PHYML & SPR) tree (displaying Yokoyama [2] hypothesis). (B) A modification of the standard ML tree illustrating the relationships among the Rh sequences as inferred using quartet-puzzling ML. Note that also this tree display Yokoyama [2] Hypothesis. (C) A modification of the optimal standard ML tree illustrating the relationships among the Rh

sequences as inferred by Collin et al. [3]. These trees has been used to compare the two alternative hypotheses [2 and 3] using the AU test. In Blue: RhA. In Green: RhB. See main text and materials and methods for abbreviations.

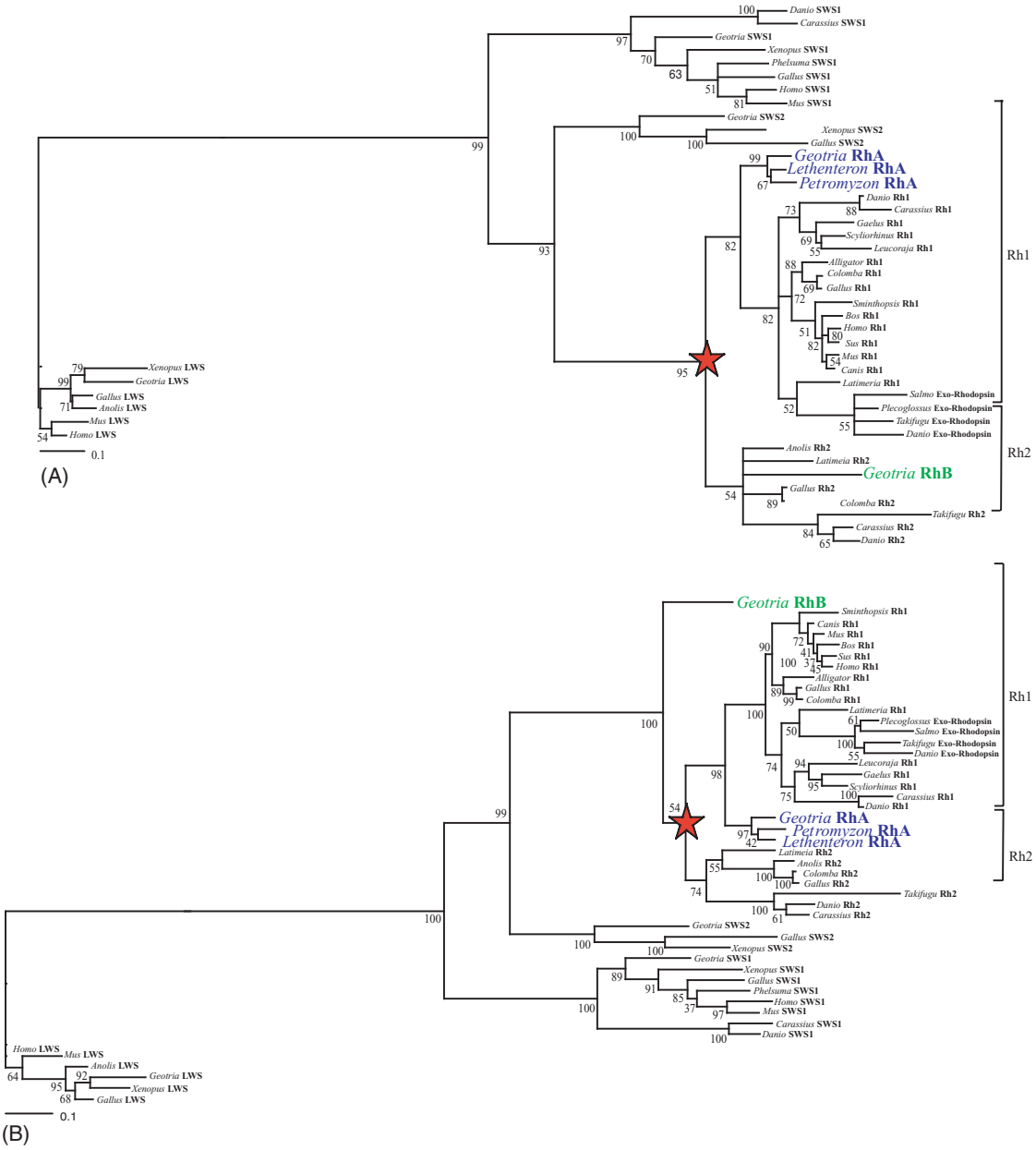

Figure S2. Results of the Maximum Likelihood Analyses. (A) Quartet Puzzling (B) PHYM (see also supplemental results and discussion and main text). Numbers at the nodes represent, respectively, quartet puzzling support values and bootstrap proportions. The star represents the Gene duplication resulting in

the origin of Rh1. In Green: The Jawless Vertebrates RhB sequences, in Blue the Jawless Vertebrates RhA sequences.

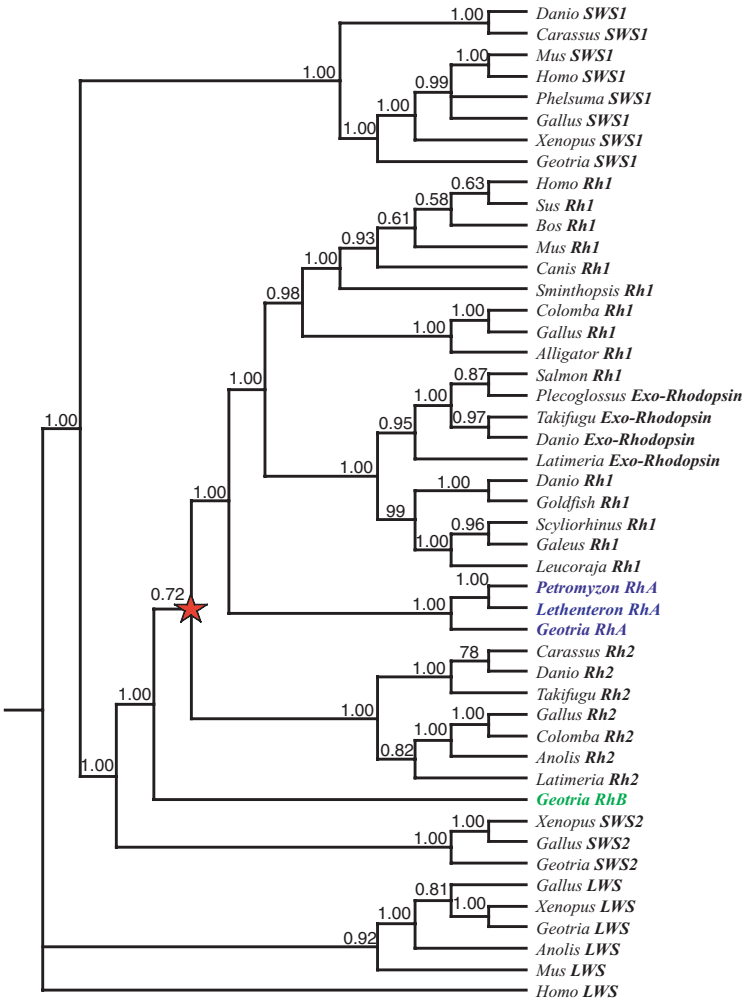

Figure S3. Results of the Bayesian Analysis. The star represents the gene duplication resulting in the origin of Rh1. Number at the nodes represent Posterior Probabilities. In Green the Jawless Vertebrates RhB, in Blue the Jawless Vertebrates RhA.

Supplemental Information (Accession Numbers & Alignments)

SWS1: Danio rerio Q9W6A9; Carassius auratus Q90309; Mus musculus P51491; Homo sapiens P03999; Phelsuma madagascariensis AAD45183; Gallus gallus P28684; Xenopus laevis P51473; Geotria australis AAR14684.

Rh1: Homo sapiens NP_000530; Sus scrofa O18766; Bos Taurus 1F88_B; Mus musculus NP_663358; Canis familiaris P32308; Sminthopsis crassicaudata Q8HY69; Columba livia AAD32241; Gallus gallus P22328; Alligator mississippiensis P52202; Salmo salar AAF44619; Plecoglossus altivelis BAC56700; Takifugu rubripes AAF44622; Danio rerio NP_571287; Latimeria chalumnae AAD30519; Danio rerio BAC21668; Carassius auratus P32309; Scyliorhinus canicula O93459; Galeus melastomus O93441; Leucoraja erinacea P79863; Lethenteron japonicum P22671; Geotria australis AAR14682; Petromyzon marinus Q98980.

Rh2: Carassius auratus P32311; Takifugu rubripes AAF44648; Danio rerio Q9W6A6; Gallus gallus P28683; Columba livia AAD32242; Anolis carolinensis AAB35062; Latimeria chalumnae AAD30520; Geotria australis AAR14683.

SWS2: Xenopus laevis AAO38746; Geotria australis AAR14681; Gallus gallus P28682.

LWS: Gallus gallus P22329; Anolis carolinensis P41592; Xenopus laevis O12948; Geotria australis AAR14680; Mus musculus O35599; Homo sapiens P04001.

[1]Bioinformatics Laboratory, The National University of Ireland, Maynooth, Ireland. [2]Department of Zoology, The Natural History Museum. Cromwell Road, SW7 5BD, London, UK. [3]School of Biology and Psychology, Division of Biology, University of Newcastle-upon-Tyne, NE1 7RU, UK. *E-mail: davide.pisani@nuim.ie