

## **Supplementary material for: Estimates of molecular convergence reveal multiple genes with adaptive variation across teleost fish**

Agneesh Barua<sup>1,2\*</sup>, Malvika Srivastava<sup>1</sup>, Brice Beinstainer<sup>3</sup>, Vincent Laudet<sup>4,5</sup>, Marc Robinson-Rechavi<sup>1,2</sup>

<sup>1</sup>Department of Ecology and Evolution, University of Lausanne

<sup>2</sup>Swiss Institute of Bioinformatics

<sup>3</sup>Helmholtz Pioneer Campus, Helmholtz Zentrum München, Neuherberg, Germany

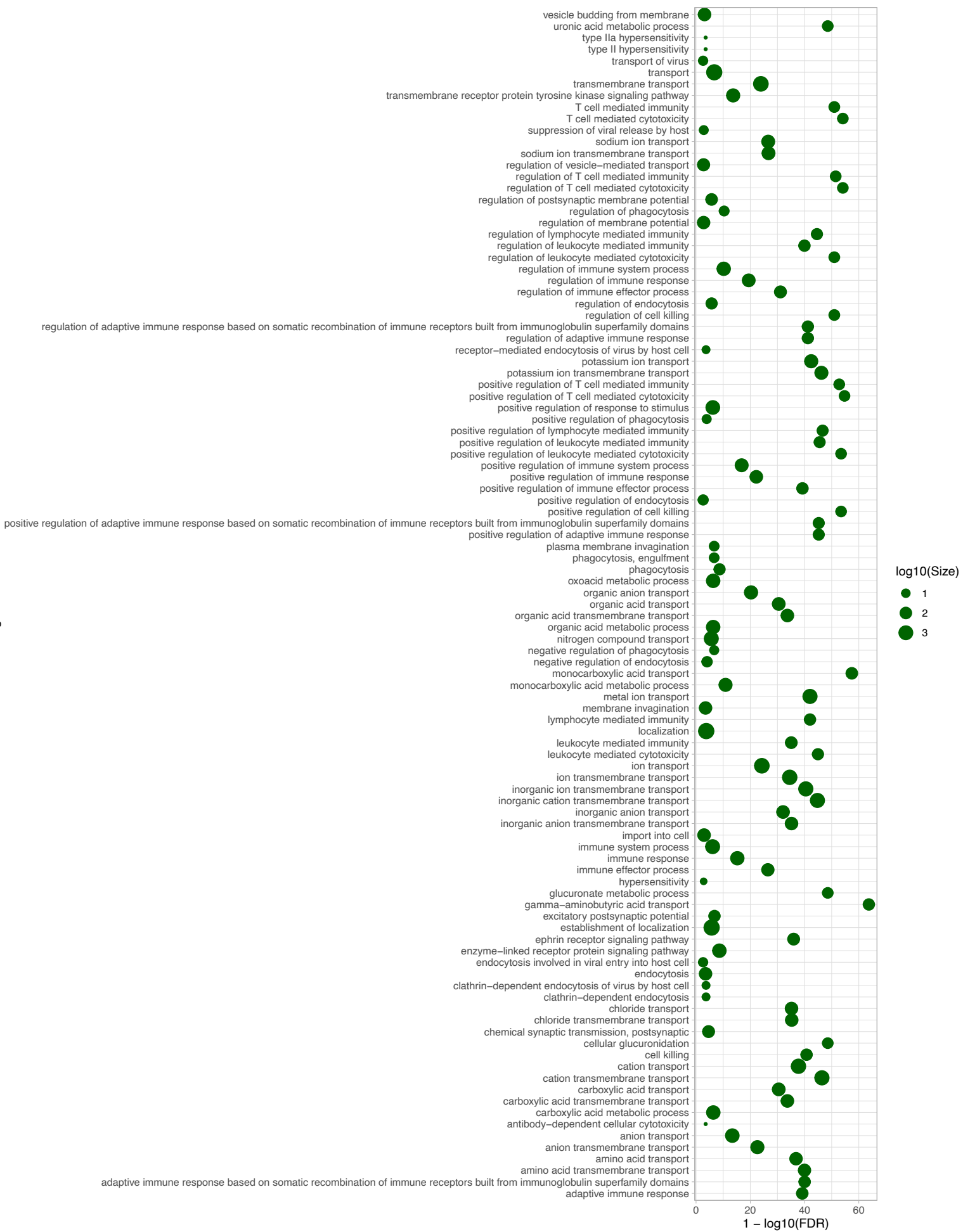
<sup>4</sup>Marine Eco-Evo-Devo Unit, Okinawa Institute of Science and Technology Graduate University, Japan

<sup>5</sup>Marine Research Station, Institute of Cellular and Organismic Biology (ICOB), Academia Sinica, 23-10, Dah-Uen Rd, Jiau Shi, I-Lan 262, Taiwan

An electronic report containing output of R code used in this study can be found at:

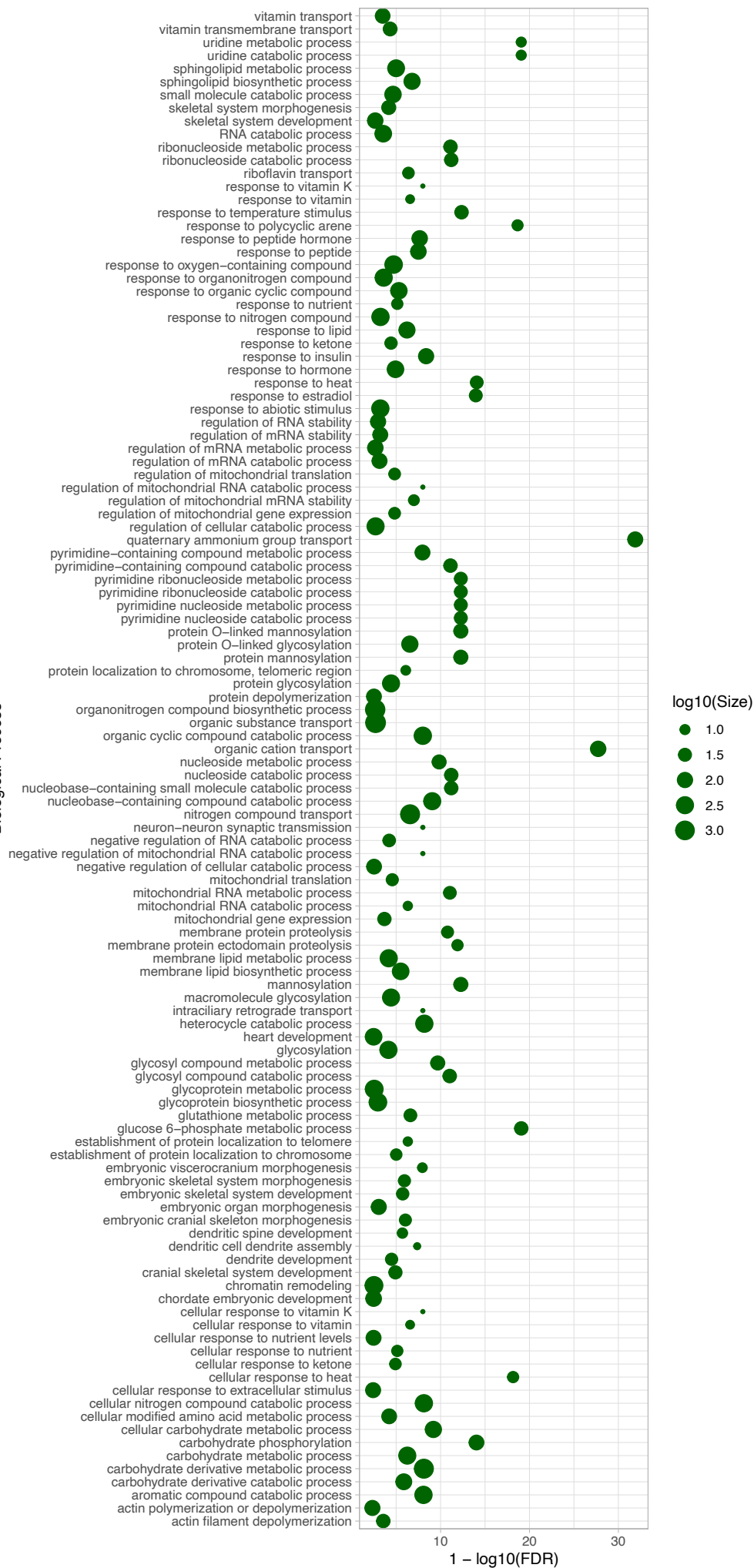
[https://agneeshbarua.github.io/Teleost\\_convergence/](https://agneeshbarua.github.io/Teleost_convergence/)

All data comprising sequencers, output files, figures, and code are available in the Zenodo data repository: <https://doi.org/10.5281/zenodo.17631061>



**Fig S1: Gene Ontology (GO) term enrichment of excluded orthogroups.** To keep computational times reasonable we restricted our analysis to orthogroups containing a maximum of 1500 genes. We annotated and functionally characterised the excluded orthogroups and found that they were mostly associated with immune processes, metabolisms, and cell signalling. Although processes related to cell signalling and immunity have substantial functional significance in organisms, our sampled gene sets encompassed a more diverse array of processes (Fig S6).

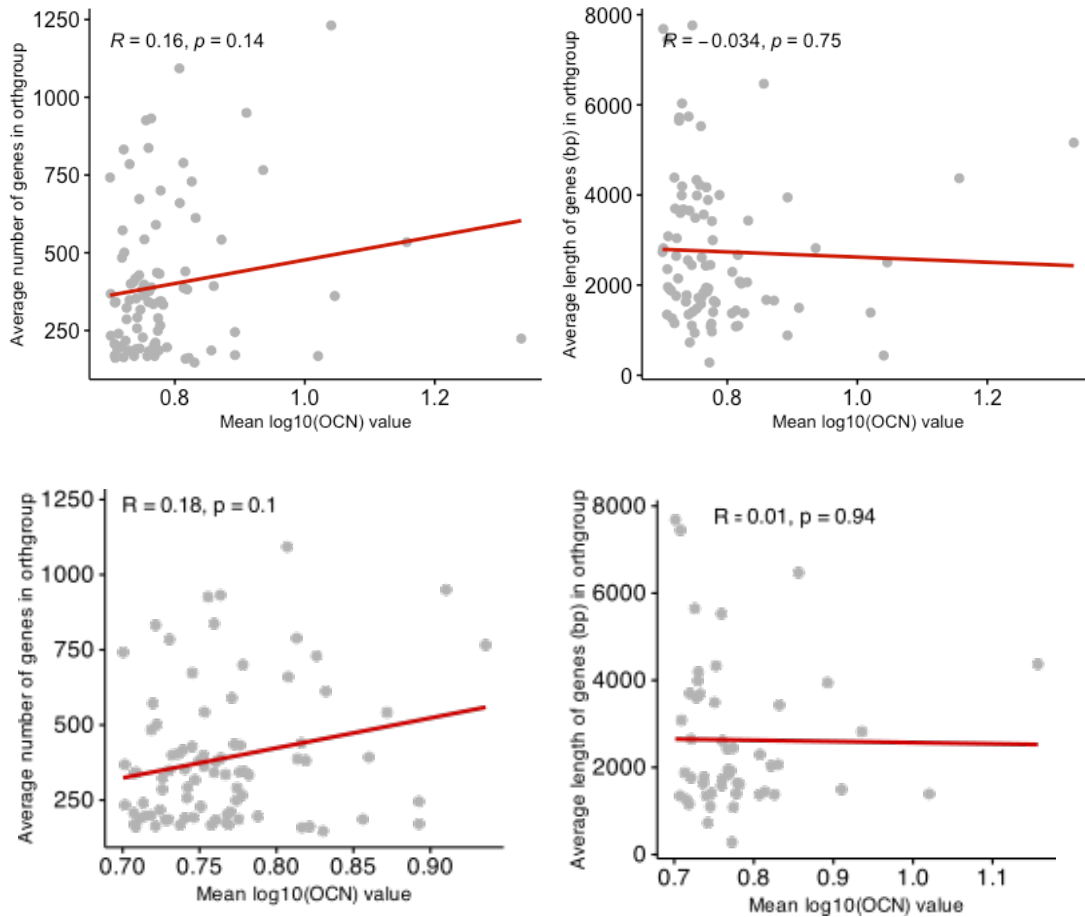
Biological Process



**Fig S2: GO term enrichment of the convergent substitutions.** GO term enrichment of the convergent genes showed that they were involved in processes related to biomolecule metabolism, response to hormones or stimuli such as heat, and processes of embryonic development and tissue morphogenesis. This underscores the potential multifunctional nature of these convergent genes.

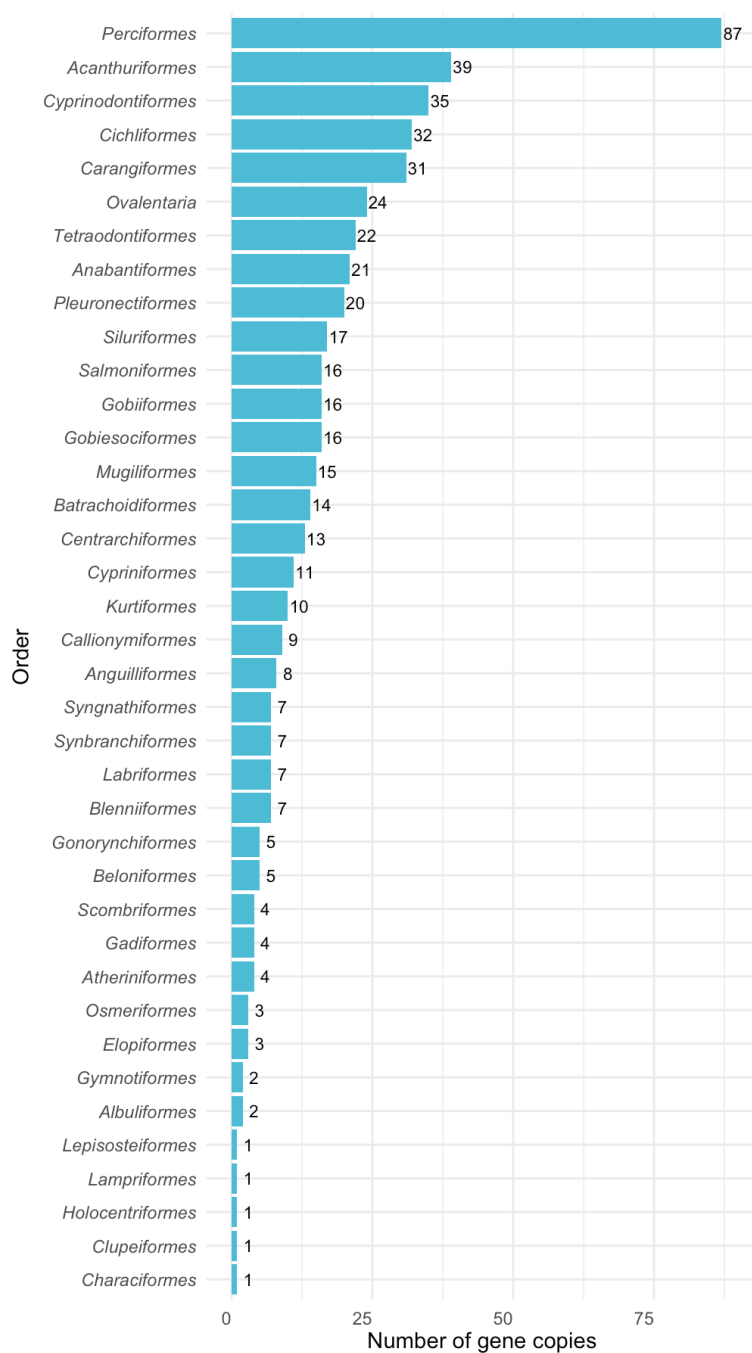
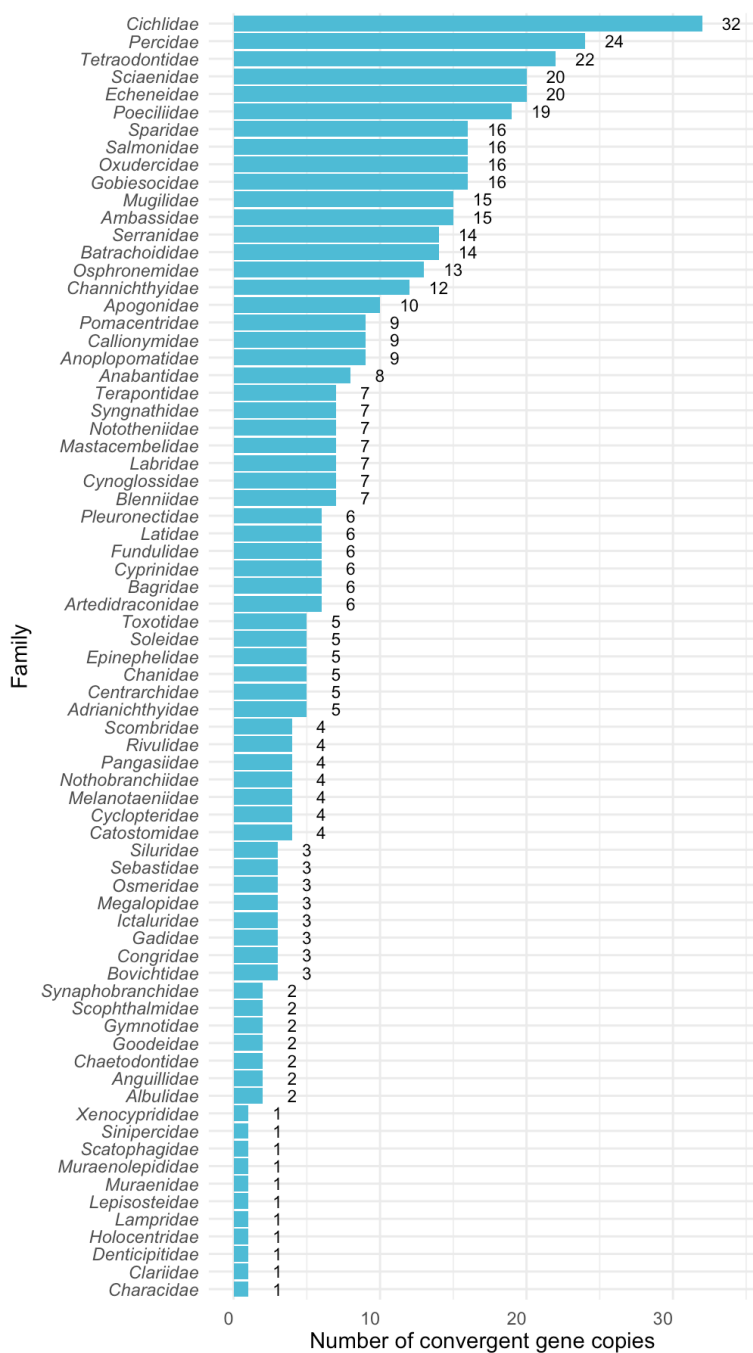


**Fig S3: Check for spurious convergence:** There can be certain instances of spurious convergence events even when stringent convergence metrics are used. This can often occur due to misalignment of sequences, which might be due to splice variants represented inconsistently among different species. Fukushima and Pollock also encountered this in their original publication when analysing human and mouse sequences (Supplementary text 12 in (1)). This suggests that spurious convergence events are not restricted to any specific dataset but is an artefact that has to be corrected. (A) The characteristic of these spurious convergence events is their unnatural localisation on the protein structures that can be detected in the output of the CSUBST site function. A workaround would be to select convergence events that are not located at proximity to one another. We perform this ad-hoc filtering by selecting orthologs where the individual substitutions have a high OCN value and are well separated on the protein structure. (B) This helps us identify reliable patterns of convergence as shown. In the above plots black and grey vertical bars represent non-synonymous and synonymous substitutions respectively. The black horizontal bars in panel A represent gaps in mapping the alignment to the protein structure. The posterior probability of *any2spe* represents the site-wise posterior probabilities of a substitution from a different ancestral amino acid to a specific extant amino acid, i.e. convergent substitutions, while *any2diff* represents a substitution from any ancestral amino acid to a different extant amino acid. These demarkations were used in Fig 2 of the main text.

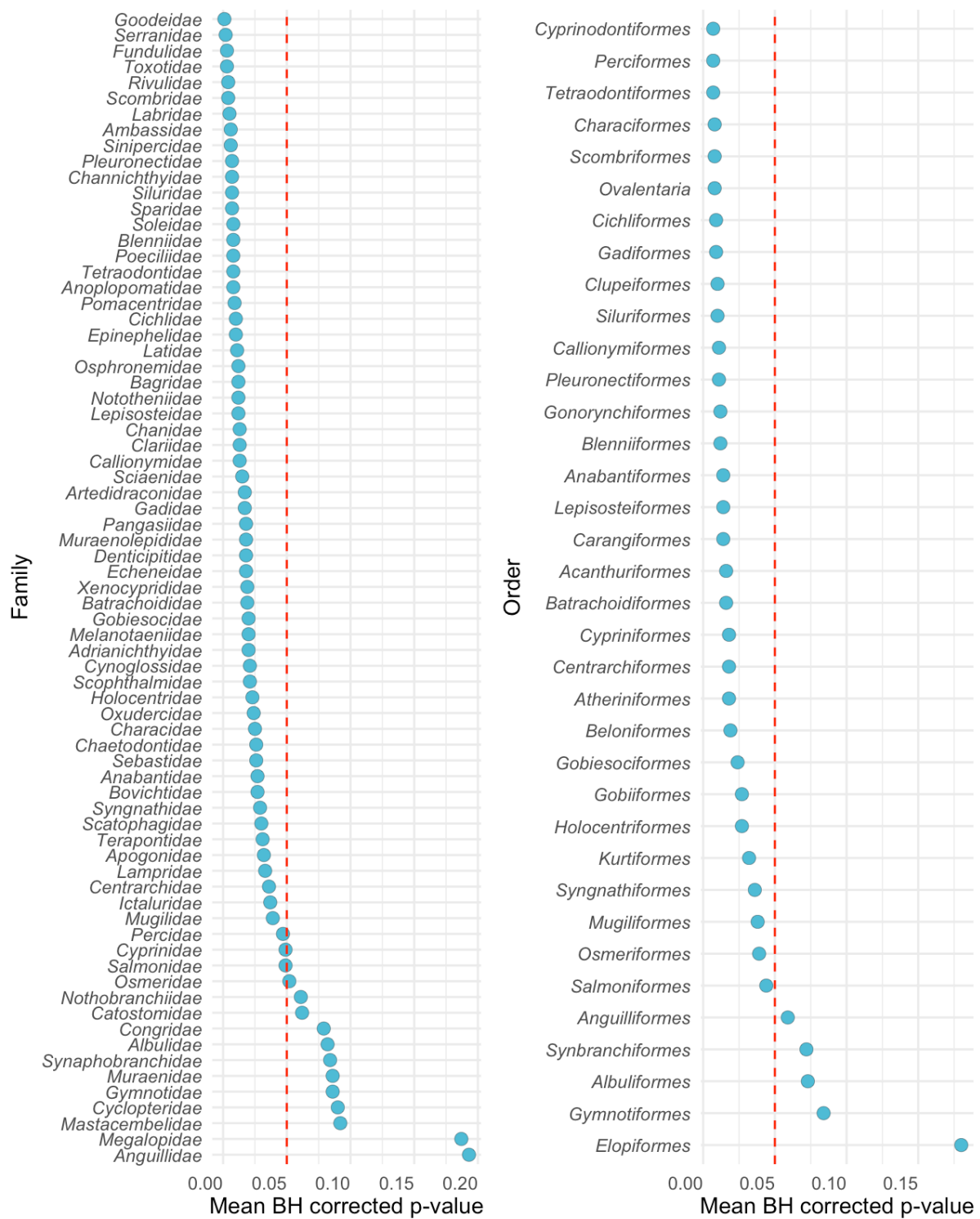


**Fig S4: Observed rate of convergence (OCN) versus orthogroups' characteristics.** We checked for the presence of any potential bias in between OCN values and the average length and number of genes in each orthogroup (top panel). Using a Pearson's correlation test we found no significant relationship between gene number or gene length and the OCN metric, suggesting that the OCN value strictly depends on sequence variation, and that such bias is not a concern in our data. We removed outliers with high OCN values and observed the same lack of correlation (bottom panel).



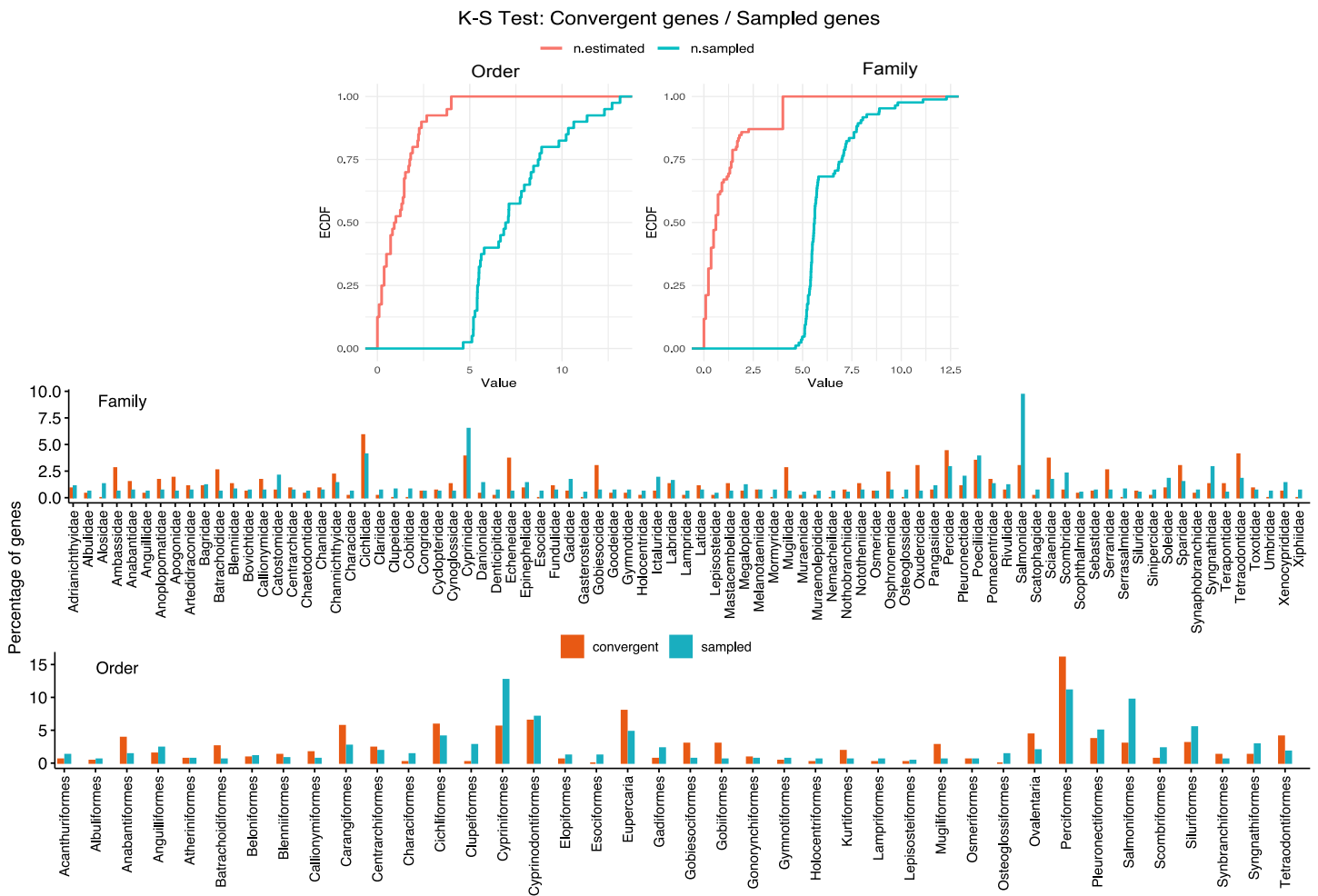


Continued on next page

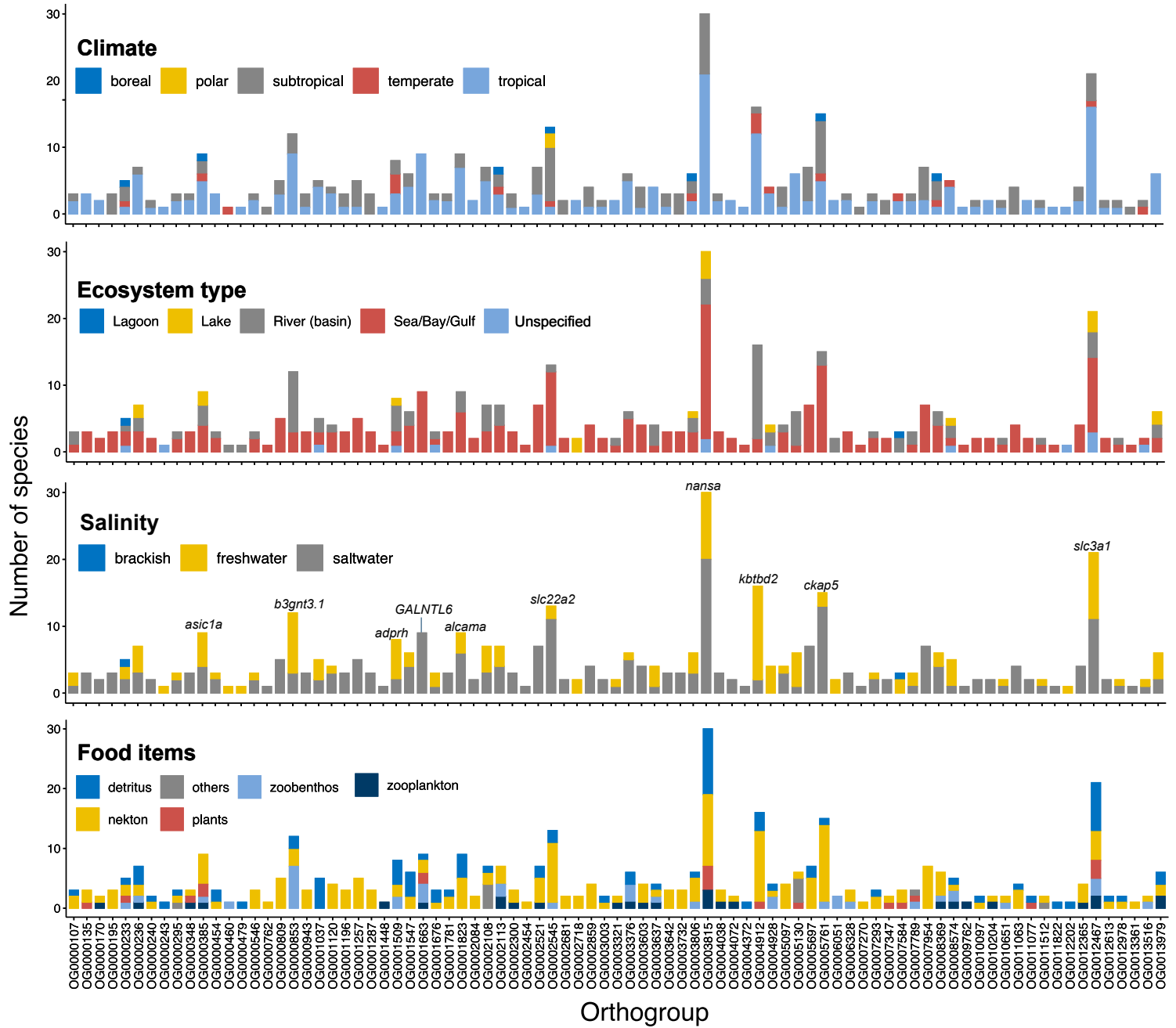


**Fig S5: Copy number distribution of the convergent orthogroups.** Top panel shows the number of gene copies present in the convergent orthogroups across teleost family and order. Bottom panel shows the mean Benjamini-Hochberg corrected P values for a one sided KS test to determine difference in gene copy number distributions between convergent orthogroups and background sets. Beyond the global trend of high copy number in the convergent orthogroups, we also observed similar trends within individual clades. At the order level, 86% (31 out of 36) of clades showed evidence of higher copy numbers among convergent genes, and at the family level, 83% (61 out of

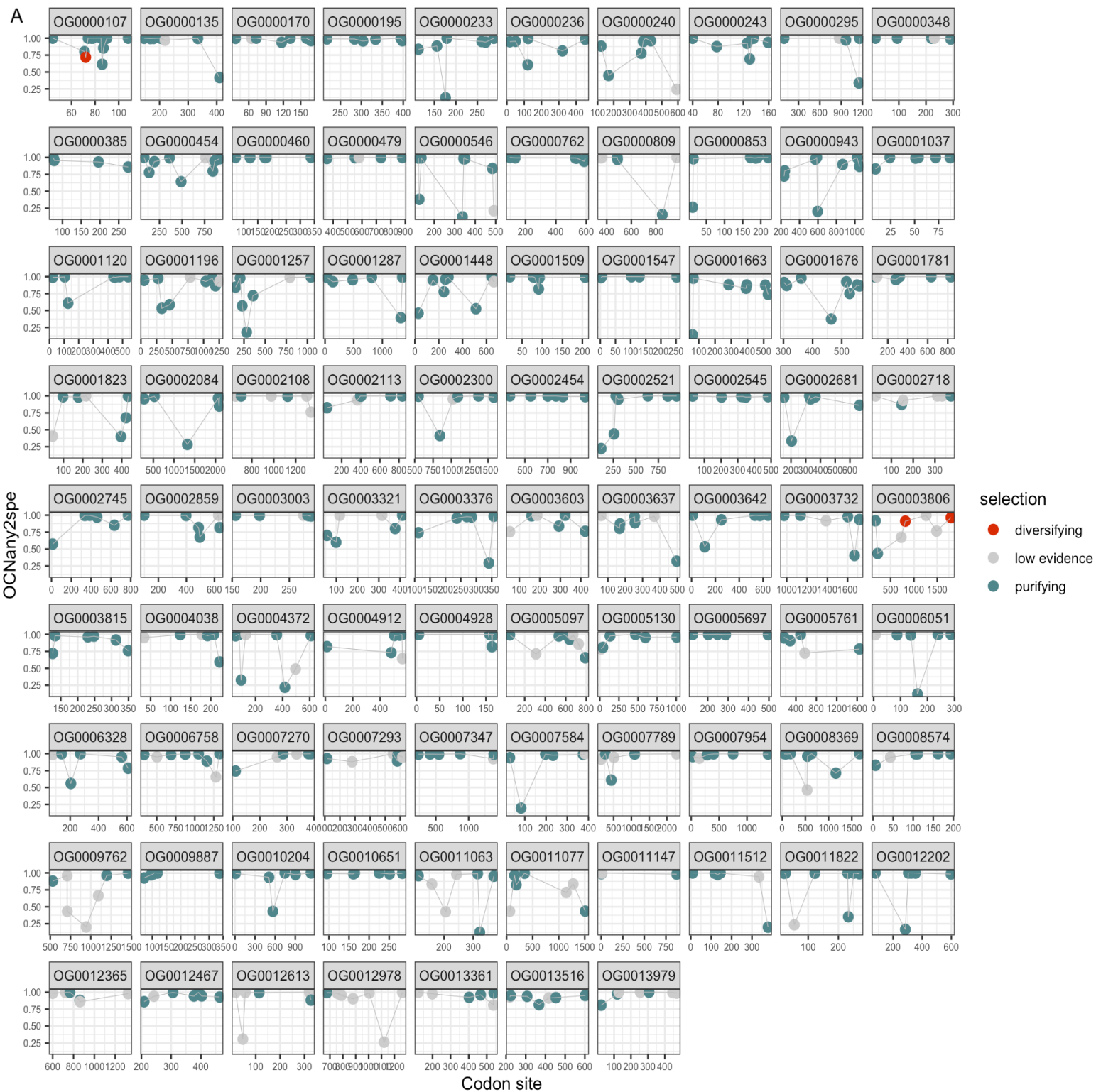
73) showed the same pattern. Thus, while the majority of clades had higher copy numbers in convergent orthogroups, this trend was not universal across all teleost lineages



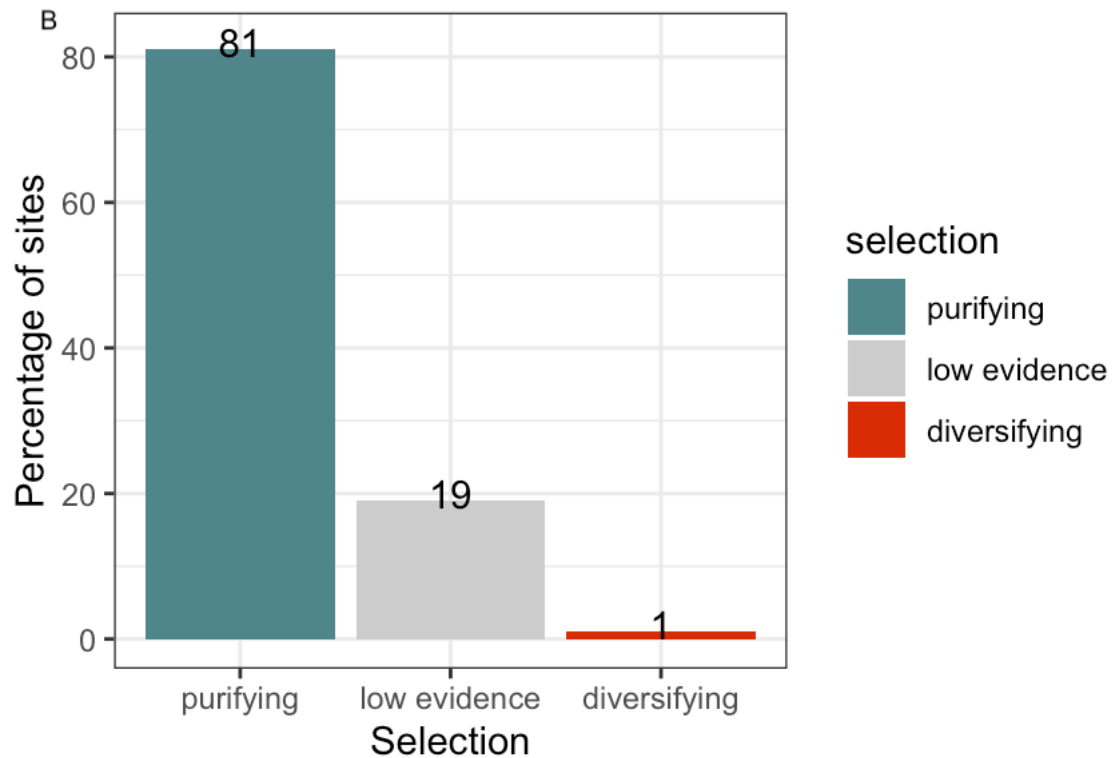
**Fig S6: Comparing the distribution of convergent orthogroups versus sampled genes from each order.** We use the Kolmogorov-Smirnov test (KS test) to compare distributions of total number of genes sampled versus number of convergent genes estimated (top panel). The rationale is to check whether the relative distributions of the convergent genes are different from that of the number of genes that were sampled. This will tell us whether the deviation between the number of gene sampled in each order/family and the number of genes found convergent is meaningful. The KS test is done with bootstrapping which is suitable for frequency distributions of discrete variables (number of genes in each order). See <https://rdrr.io/cran/kldtools/man/ksboot.html>. Observe that the shape of the empirical cumulative probability distributions (ECDF) are different for the convergent genes (red) and total sampled genes (blue). The differences in frequencies is further visualised in the ar plots.



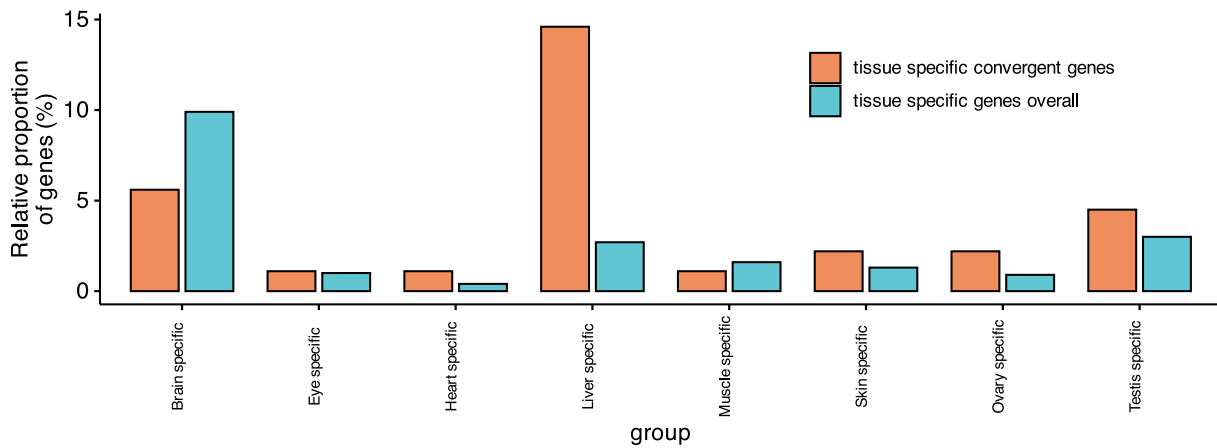
**Fig S7: Ecological characteristics of species with convergent genes.** The bar plots show ecological characteristics of species that harbour convergent substitutions. The data were obtained from the FishBase database. The proportions are based on species with available data. The gene names or orthogroups with a high number of species are labelled. We performed tests of independence to infer whether there is any relationship between convergence in one orthogroup and the ecological characters. After correcting for multiple testing, only 4 orthogroups show a significant relationship. However, these relationships are with the undefined food variable 'other'. As a result, we cannot make any biological relevant inference from this relationship.



Continued on next page

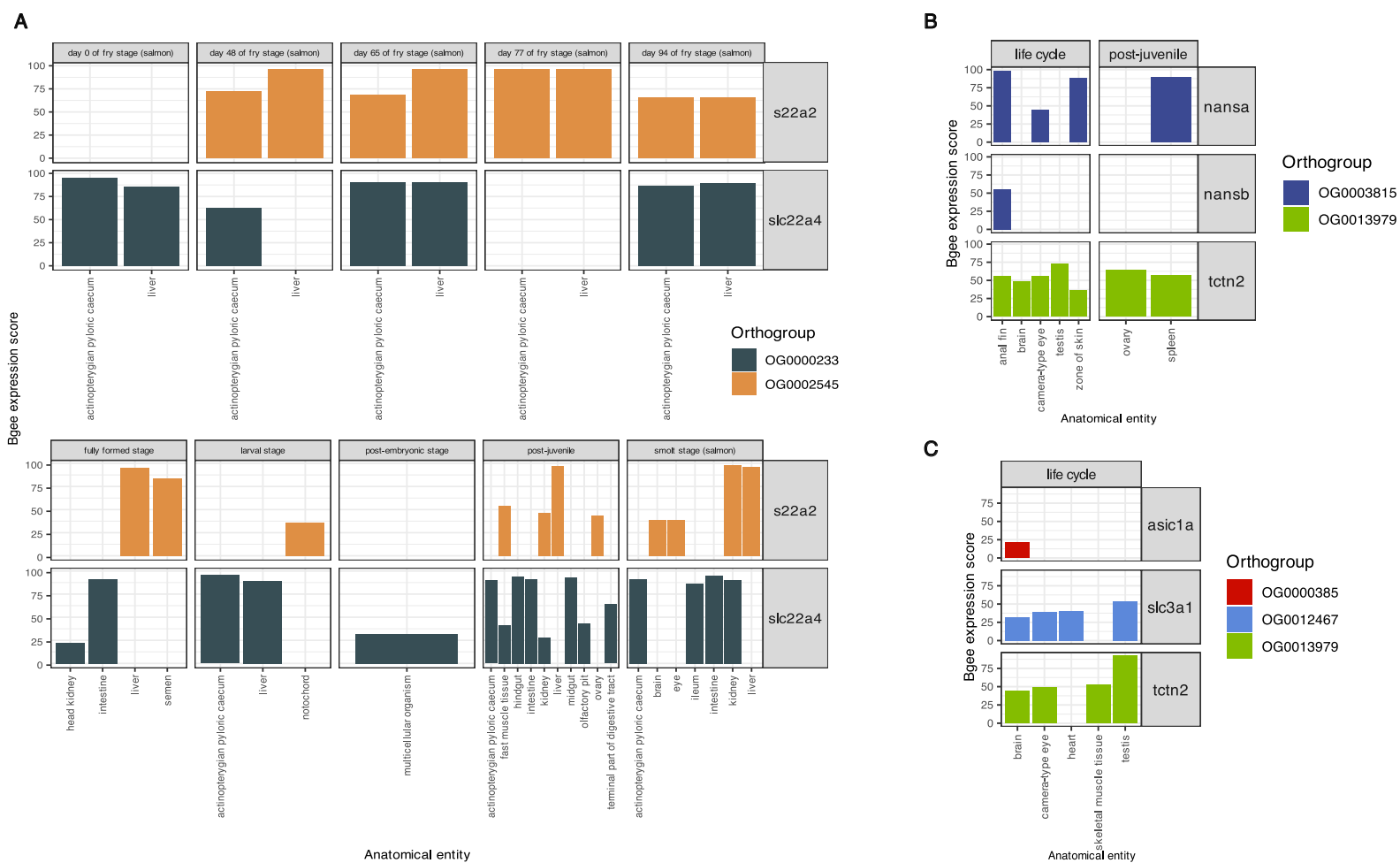


**Fig S8 Fast Unconstrained Bayesian Approximation (FUBAR) of selective regime.** Using FUBAR we estimated selective regime at each convergent site across the teleosts. (A) The  $OCN_{any-to-specific}$  convergent metric value and the selective regime experienced by that site. This figure shows all non-zero  $OCN_{any-to-specific}$  values. In our analyses we only considered sites with an  $OCN_{any-to-specific}$  above 0.5. (B) The proportion of convergent sites and their underlying selective regimes.

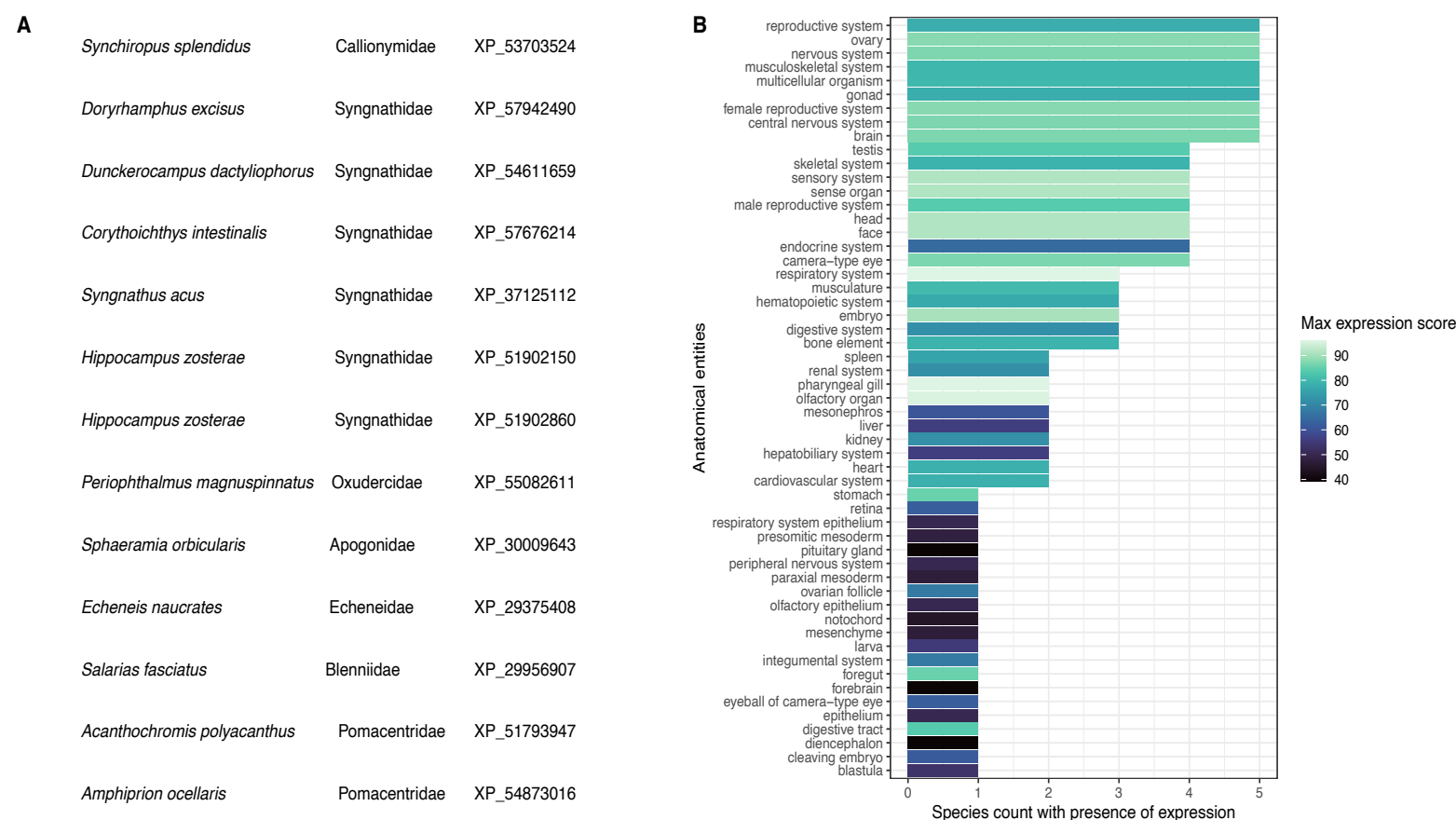


**Fig S9:** The proportion of tissue specific genes in convergent genes and genes overall for different tissue types: We collected gene expression data for the brain, eye, heart, liver, muscle tissue, ovary, and testis across eleven species (*Astyanax mexicanus*, *Astatotilapia calliptera*, *Danio rerio*, *Esox lucius*, *Gasterosteus aculeatus*, *Gadus morhua*, *Neolamprologus brichardi*, *Nothobranchius furzeri*, *Oryzias latipes*, *Salmo salar*, and *Scophthalmus maximus*). We classified a gene as tissue-specific if it had a  $\tau > 0.8$  and its expression in the target tissue was greater than the sum of its expression in other tissues. Since we are comparing tissue specificity across different species, we identified genes that are tissue-specific in the same tissue across all the species sampled. These criteria ensured we captured robust signals for tissue specificity. Around one-third of these genes had signals of convergent evolution with the largest being in the liver. We observed that the proportion of convergent genes that are tissue specific is no different from the total proportion of tissue specific genes in our dataset. However, using Fisher's exact test we found a significant difference in the convergent/non-convergent ratio between tissue-specific and non-tissue-specific genes only for the liver and not for other tissues; in other words, our convergent gene set had a higher proportion of liver-specific genes than expected by chance.

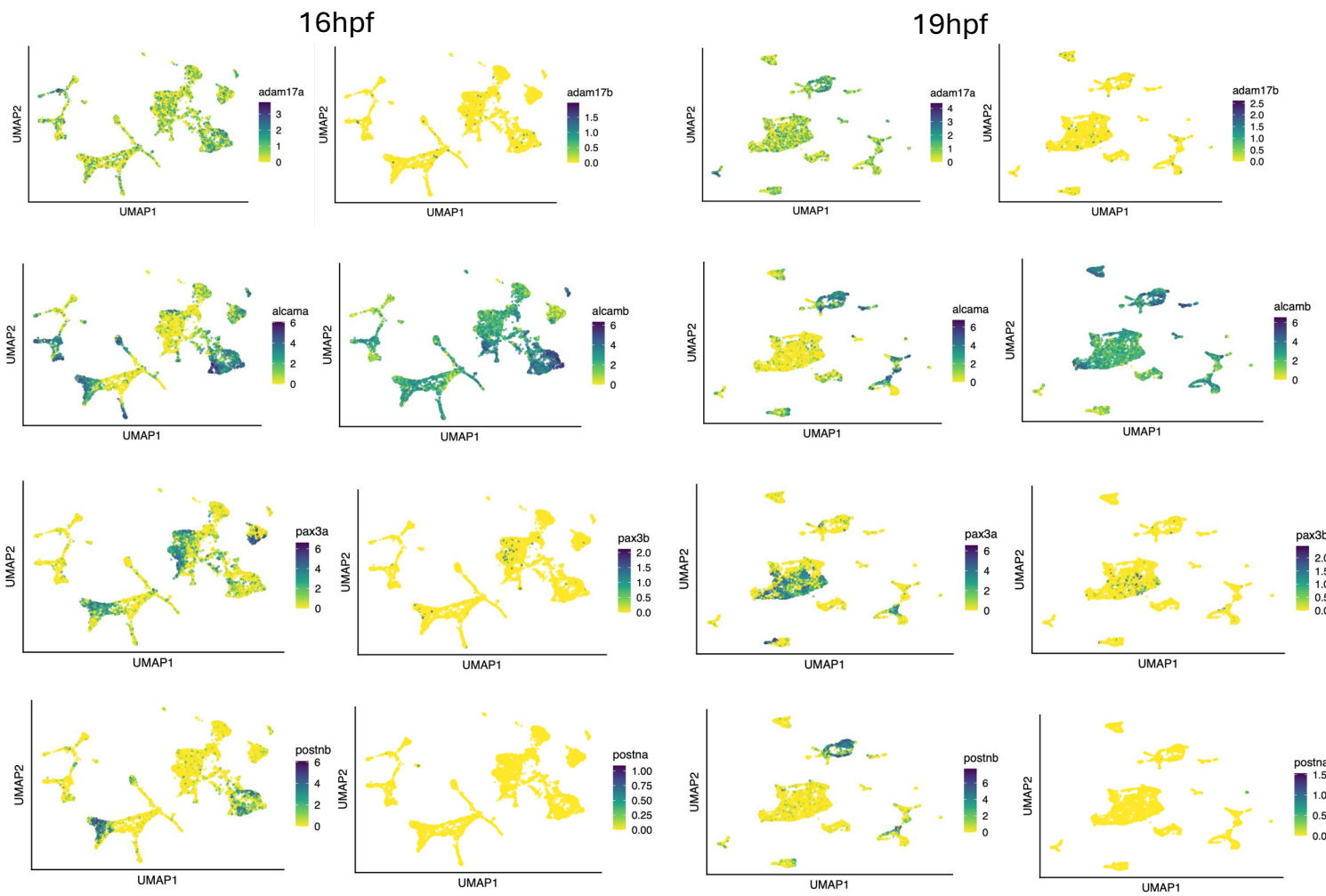




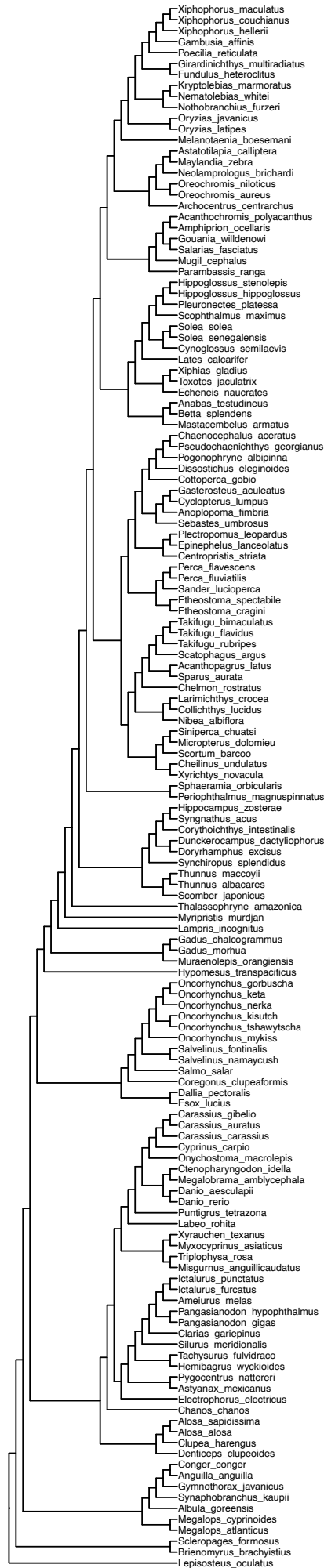
**Fig S10: Expression of convergent orthogroups in species harbouring convergent substitutions.** Using Bgee expression we show that the convergent genes have a wide spatial and temporal breadth, being expressed across multiple tissues and different developmental stages. (A) Solute-carrier protein *slc22a2*, shows high expression in the liver and pyloric caecum across multiple larval stages of *Salmo salar*, with a subsequent expression in the kidney in the post-juvenile stages. (B) N-acetylneuraminic acid synthase (*NANS*) is highly expressed in the anal fin, skin, and eye of *Astatotilapia calliptera*, as well as in the spleen during the early juvenile stages. We also observed a difference in expression between gene copies, with *nansa* showing high expression across the anal fin, eye, and skin, whereas *nansb* was expressed only in the anal fin. This divergence in expression patterns between paralogs may reflect a partitioning of functional roles across gene copies. (C) Tectonic family member 2, *tctn2* was expressed in multiple tissues in *Neolamprologus brichardi*, with the highest expression in testis. The gene had comparable levels of expression in brain and eye across both *Astatotilapia calliptera* and *Neolamprologus brichardi*, suggesting conserved function.



**Fig S11: Case of convergent substitution in seven branches.** Panel (A) shows the species with harbouring the convergent substitution in the polypeptide N-acetylgalactosaminyltransferase 6 (*GALNTL6*) gene along with their family classification and protein ID respectively. (B) Using data from Bgee we observed high expression in the reproductive and skeletal systems across teleost species. The species with available expression data in this dataset are *Anguilla anguilla*, *Astyanax mexicanus*, *Danio rerio*, *Esox lucius*, and *Salmo salar*.



**Fig S12: Expression divergence between paralogs.** Single-cell RNA-seq data of 16 hours post fertilization and 19 hours post fertilization embryos showing difference in gene expression between paralogs. Similar trend was observed for all other time-points.



**Fig S13: Phylogenetic tree of species.** Phylogenetic tree of 143 teleost fish species used in the study.