Supplementary material for: Estimates of molecular convergence reveal pleiotropic genes underlying adaptive variation across teleost fish

Agneesh Barua[1,2*], Brice Beinsteiner[3], Vincent Laudet[4,5], Marc Robinson-Rechavi[1,2]

[1]Department of Ecology and Evolution, University of Lausanne

[2]Swiss Institute of Bioinformatics

[3]Helmholtz Pioneer Campus, Helmholtz Zentrum München, Neuherberg, Germany

[4]Marine Eco-Evo-Devo Unit, Okinawa Institute of Science and Technology Graduate University, Japan

[5]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/

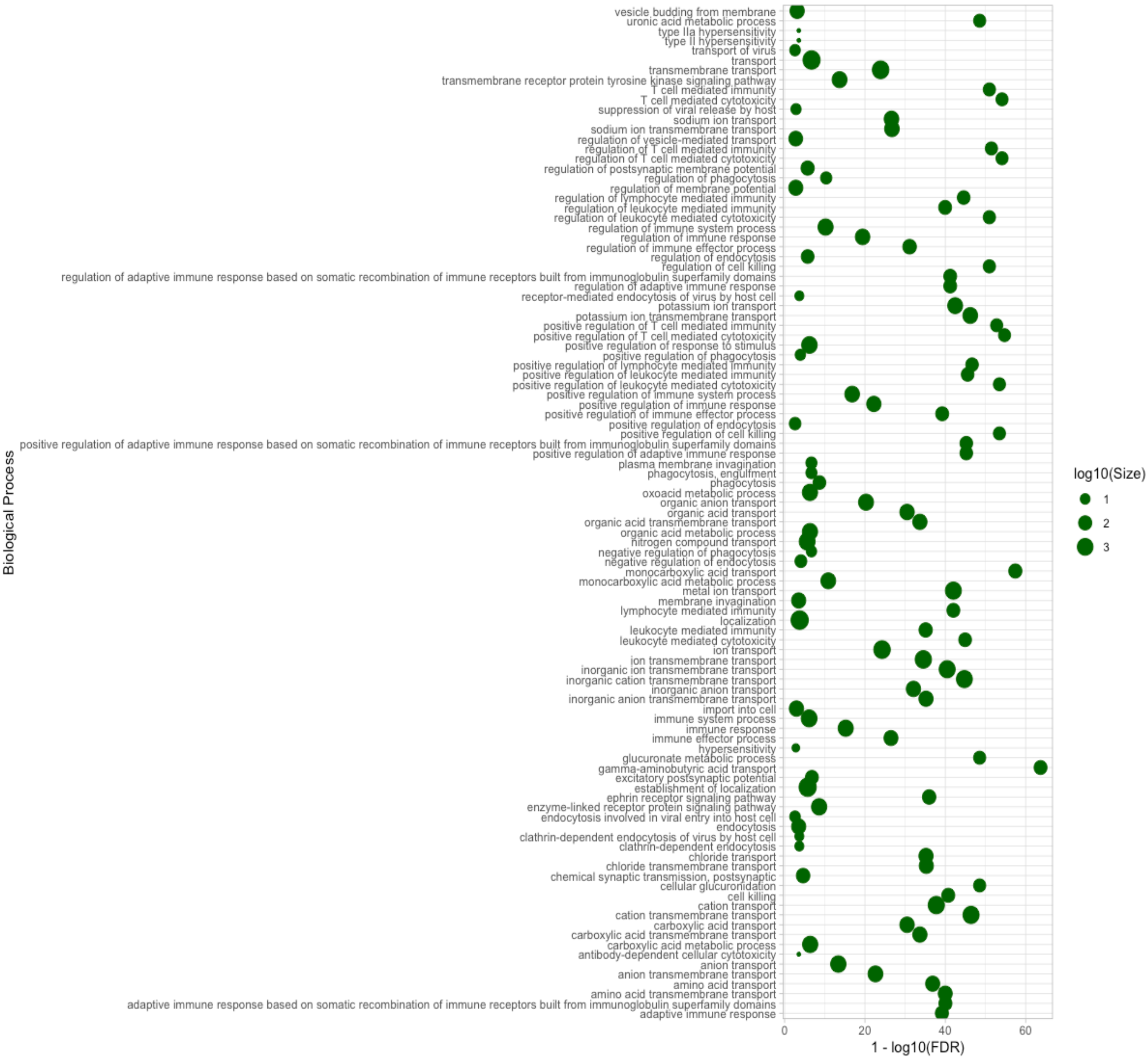

Fig S1: Gene Ontology (GO) term enrichment of excluded orthogroups. To keep computational times reasonable we restricted our analysis to orthogroups containing a maximum of 1500 genes. We annotated and functionally characterised the excluded orthogroups and found that they were mostly associated with 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 S3).

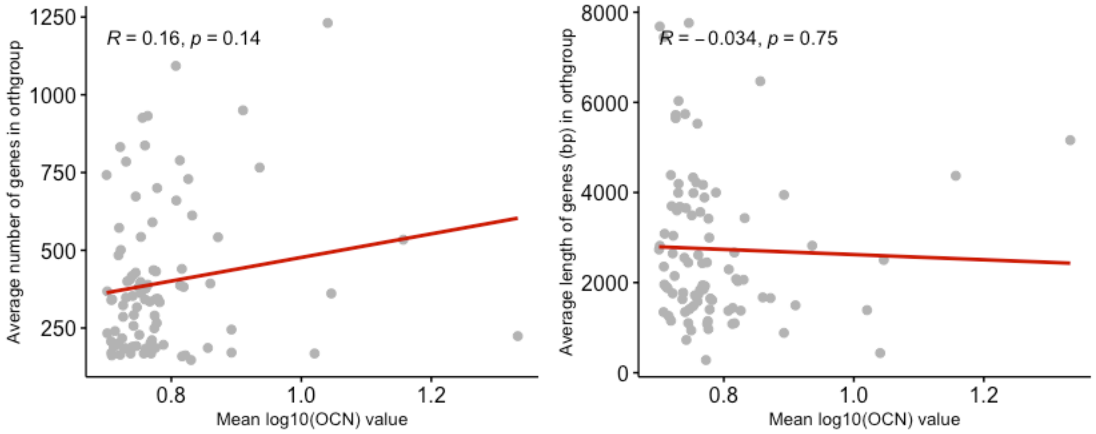

Fig S2: Observed rate of convergence (OCN) versus orthogroups characterisitcs. We checked for the presense of any potential bias in between OCN values and the average length and number of genes in each orthogroup. Using a Pearson's correlation test we found no significant relationship between gene number 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.

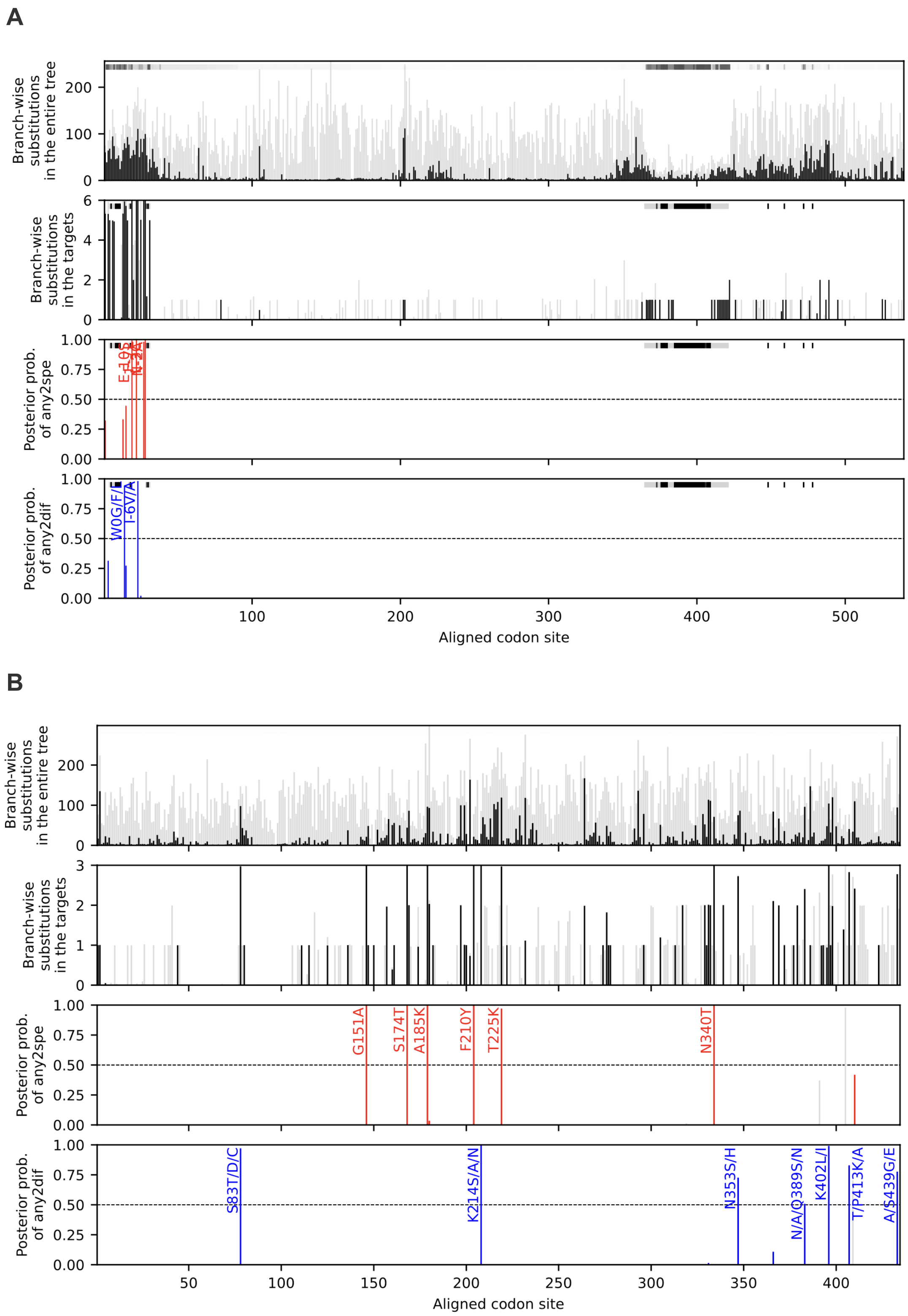
A
Branch-wise substitutions in the entire tree
0
100
200
Branch-wise substitutions in the targets
0
2
4
6
Posterior prob. of any2spe
0.00
0.25
0.50
0.75
1.00
E-10S
Posterior prob. of any2dif
W0G/F/E
T-6V/A
100
200
300
400
500
Aligned codon site
B
Branch-wise substitutions in the entire tree
0
100
200
Branch-wise substitutions in the targets
0
1
2
3
Posterior prob. of any2spe
0.00
0.25
0.50
0.75
1.00
G151A
S174T
A185K
F210Y
T225K
N340T
Posterior prob. of any2dif
S83T/D/C
K214S/A/N
N353S/H
N/A/Q389S/N
K402L/I
T/P413K/A
A/S439G/E
50
100
150
200
250
300
350
400
Aligned codon site

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 a 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 repsectively. 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 differnt acestral 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 extact amino acid. These demarkations were used in Fig 2 of the main text.

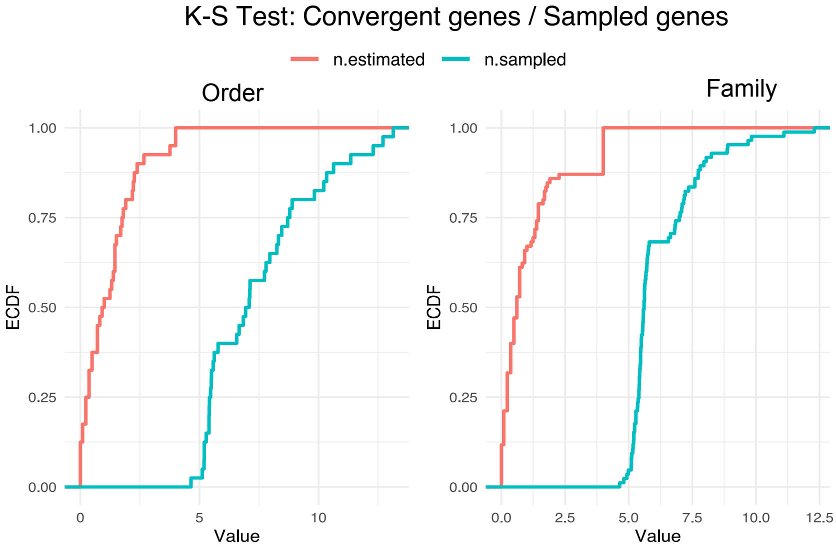

Fig S4: Comparing the distribution of convergent genes versus sampled genes from each Order. We use the Kolmogorov-Smirnov test (KS test) to compare distributions of total number of genes sampled versus number of convergent genes estimated. The rationale is to check whether the relative distributions 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 difference in frequencies is further visualised in Fig 1 of the main manuscript.

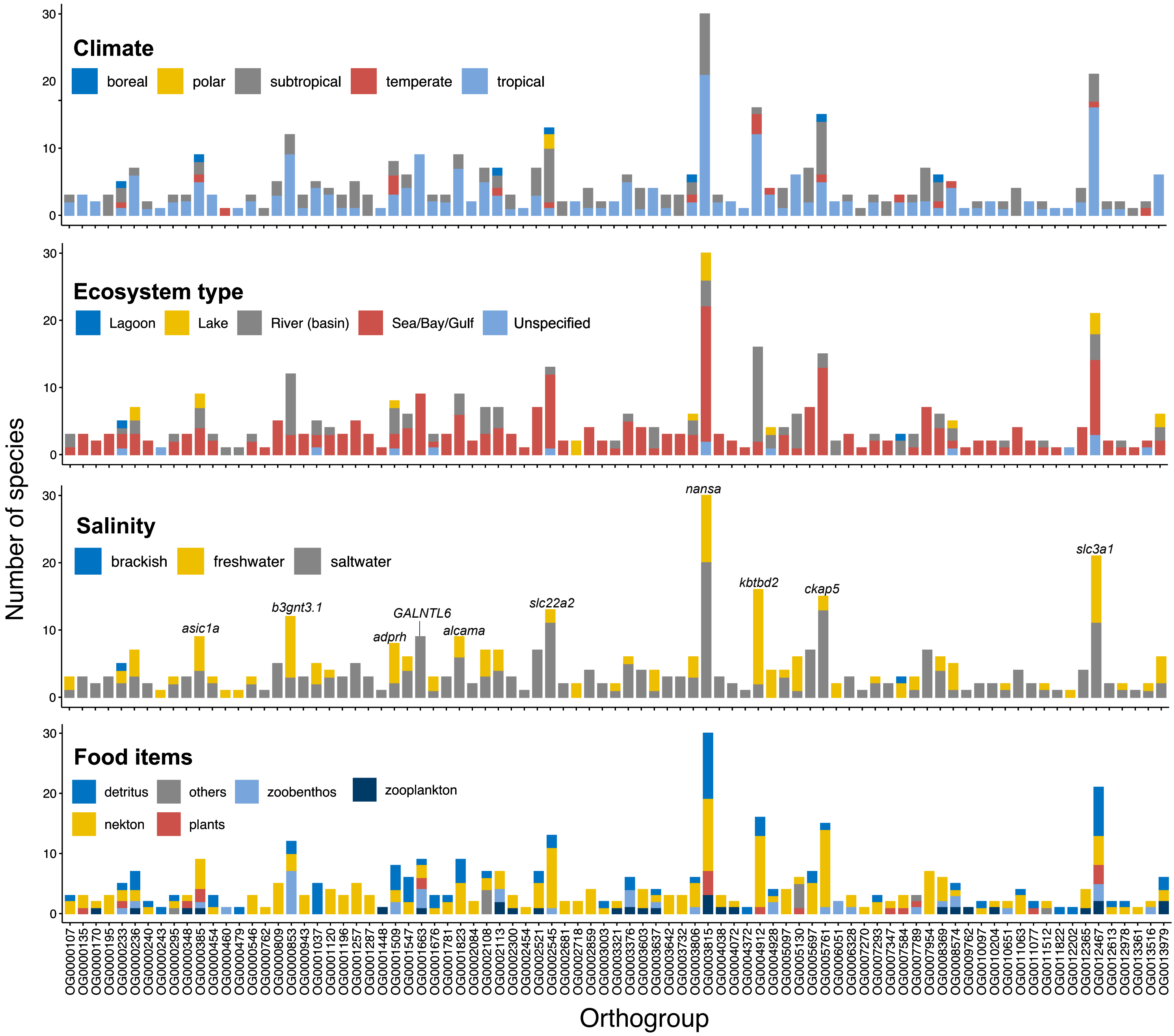

Fig S5: Ecological characteristics of species with convergent genes. The bar plots show ecological characteristics of species that harbour convergent substitutions. The data were obtained from the FishBase database. The proportions are based on species with available data. The gene names or orthogroups 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. See project report in the electronic supplementary material.

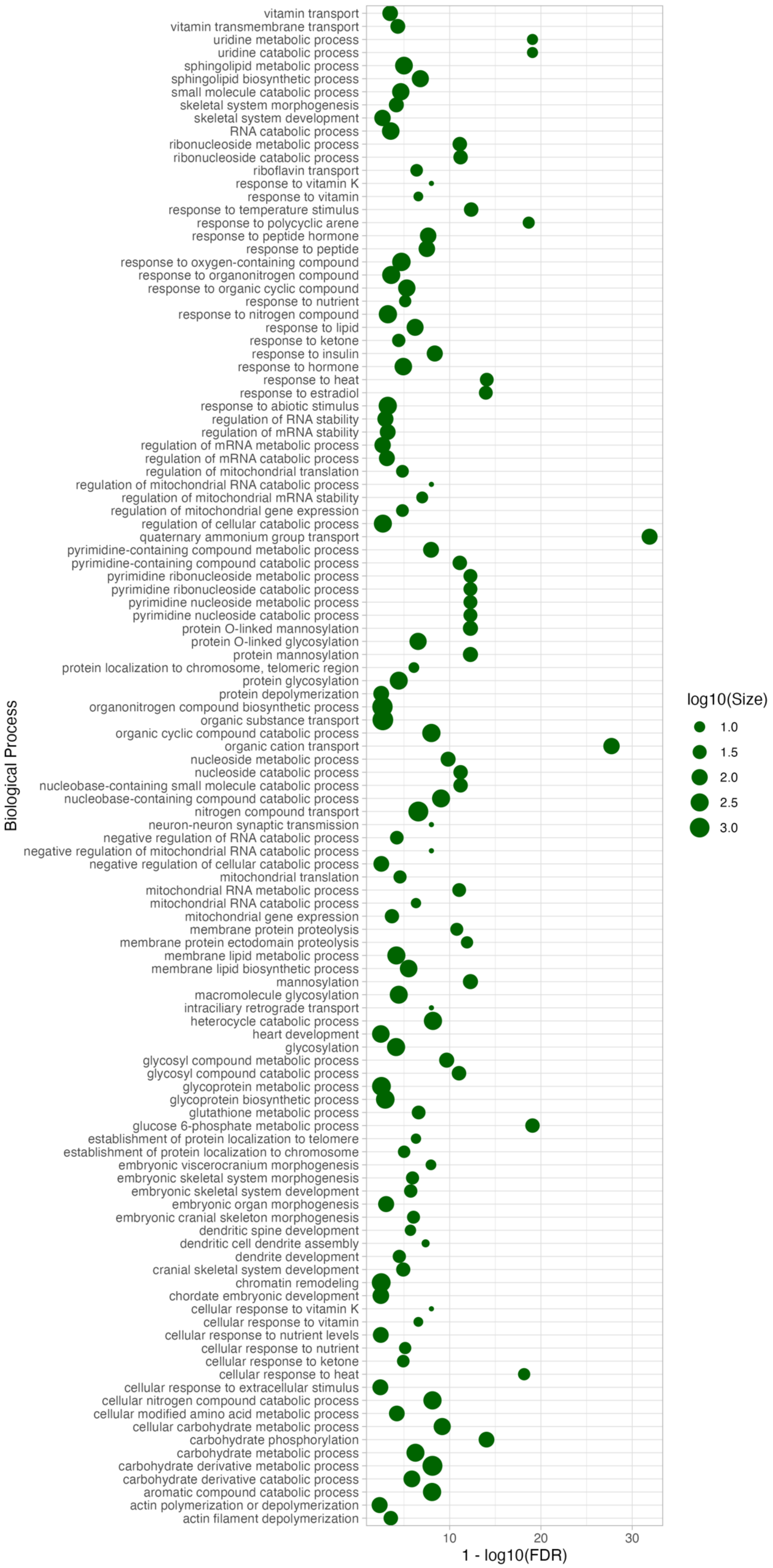
vitamin transport
vitamin transmembrane transport
uridine metabolic process
uridine catabolic process
sphingolipid metabolic process
sphingolipid biosynthetic process
small molecule catabolic process
skeletal system morphogenesis
skeletal system development
RNA catabolic process
ribonucleoside metabolic process
ribonucleoside catabolic process
riboflavin transport
response to vitamin K
response to vitamin
response to temperature stimulus
response to polycyclic arene
response to peptide hormone
response to peptide
response to oxygen-containing compound
response to organonitrogen compound
response to organic cyclic compound
response to nutrient
response to nitrogen compound
response to lipid
response to ketone
response to insulin
response to hormone
response to heat
response to estradiol
response to abiotic stimulus
regulation of RNA stability
regulation of mRNA stability
regulation of mRNA metabolic process
regulation of mRNA catabolic process
regulation of mitochondrial translation
regulation of mitochondrial RNA catabolic process
regulation of mitochondrial mRNA stability
regulation of mitochondrial gene expression
regulation of cellular catabolic process
quaternary ammonium group transport
pyrimidine-containing compound metabolic process
pyrimidine-containing compound catabolic process
pyrimidine ribonucleoside metabolic process
pyrimidine ribonucleoside catabolic process
pyrimidine nucleoside metabolic process
pyrimidine nucleoside catabolic process
protein O-linked mannosylation
protein O-linked glycosylation
protein mannosylation
protein localization to chromosome, telomeric region
protein glycosylation
protein depolymerization
organonitrogen compound biosynthetic process
organic substance transport
organic cyclic compound catabolic process
organic cation transport
nucleoside metabolic process
nucleoside catabolic process
nucleobase-containing small molecule catabolic process
nucleobase-containing compound catabolic process
nitrogen compound transport
neuron-neuron synaptic transmission
negative regulation of RNA catabolic process
negative regulation of mitochondrial RNA catabolic process
negative regulation of cellular catabolic process
mitochondrial translation
mitochondrial RNA metabolic process
mitochondrial RNA catabolic process
mitochondrial gene expression
membrane protein proteolysis
membrane protein ectodomain proteolysis
membrane lipid metabolic process
membrane lipid biosynthetic process
mannosylation
macromolecule glycosylation
intraciliary retrograde transport
heterocycle catabolic process
heart development
glycosylation
glycosyl compound metabolic process
glycosyl compound catabolic process
glycoprotein metabolic process
glycoprotein biosynthetic process
glutathione metabolic process
glucose 6-phosphate metabolic process
establishment of protein localization to telomere
establishment of protein localization to chromosome
embryonic viscerocranium morphogenesis
embryonic skeletal system morphogenesis
embryonic skeletal system development
embryonic organ morphogenesis
embryonic cranial skeleton morphogenesis
dendritic spine development
dendritic cell dendrite assembly
dendrite development
cranial skeletal system development
chromatin remodeling
chordate embryonic development
cellular response to vitamin K
cellular response to vitamin
cellular response to nutrient levels
cellular response to nutrient
cellular response to ketone
cellular response to heat
cellular response to extracellular stimulus
cellular nitrogen compound catabolic process
cellular modified amino acid metabolic process
cellular carbohydrate metabolic process
carbohydrate phosphorylation
carbohydrate metabolic process
carbohydrate derivative metabolic process
carbohydrate derivative catabolic process
aromatic compound catabolic process
actin polymerization or depolymerization
actin filament depolymerization
Biological Process
10
20
30
1 - log10(FDR)
log10(Size)
1.0
1.5
2.0
2.5
3.0

Fig S6: Gene with convergent substitutions are involved in multiple biological processes: GO term enrichment of the convergent genes showed that they were inolved 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.

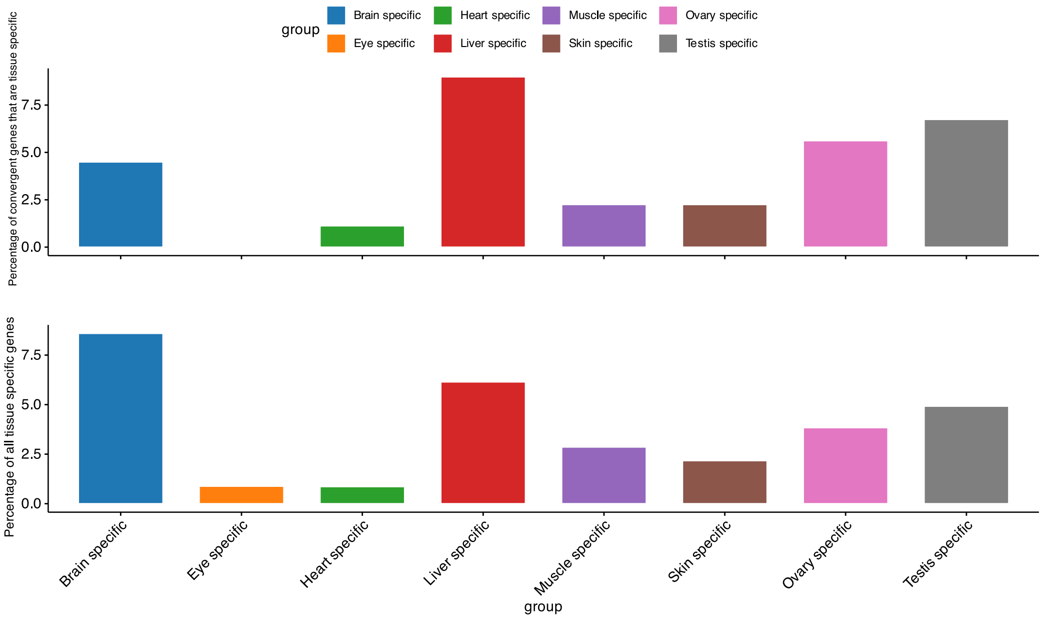

Fig S7: Only one-third of the convergent genes were tissue specific: We collected gene expression data for the brain, eye, heart, liver, muscle tissue, ovary, and testis across eleven species (*Lepisosteus oculatus*, *Danio rerio*, *Astyanax mexicanus*, *Esox lucius*, *Salmo salar*, *Gadus morhua*, *Oryzias latipes*, *Astatotilapia calliptera*, *Neolamprologus brichardi*, *Scophthalmus maximus*, and *Nothobranchius furzeri*). 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 stringent 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. This lack of tissue specificity coupled with the multifunctional nature of these convergent genes suggests that they are likely pleiotropic.

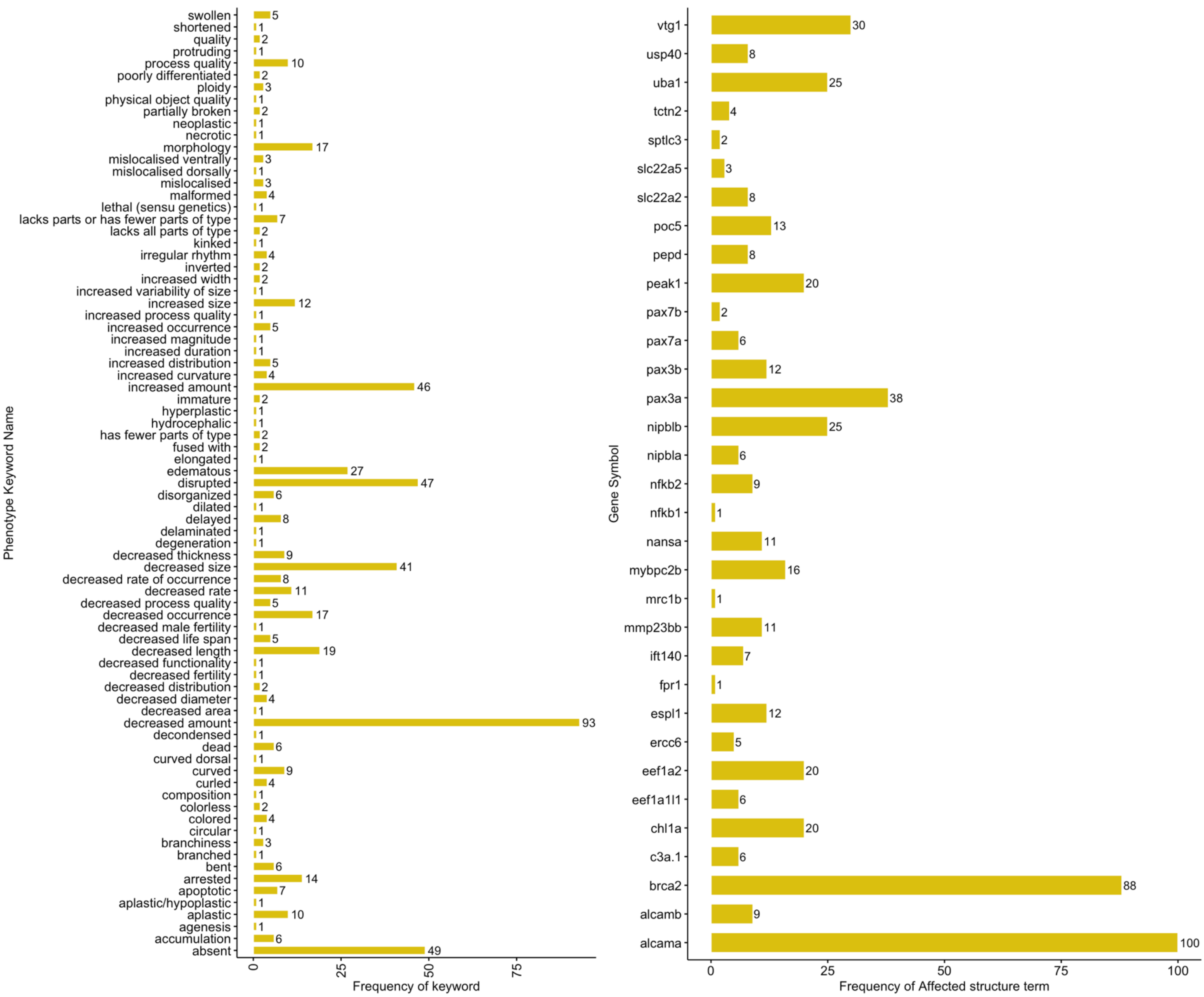

Fig S8: Functional data from the ZFIN database for convergent genes: The plot shows the distribution of phenotype keyword terms and the frequency of affected phenotype terms from ZFIN. Panel on the left shows the effect of genetic perturbation on the convergent genes. Panel on the right shows the number of unique affected structures that show a phenotype after genetic perturbation. The data shows that mutations in the convergent genes can affect several phenotypes and structures, with a majority of the effects showing a decrease or reduction in affected structures or processes, while a few shows an increase. Only convergent genes that have functional data in ZFIN are shown.

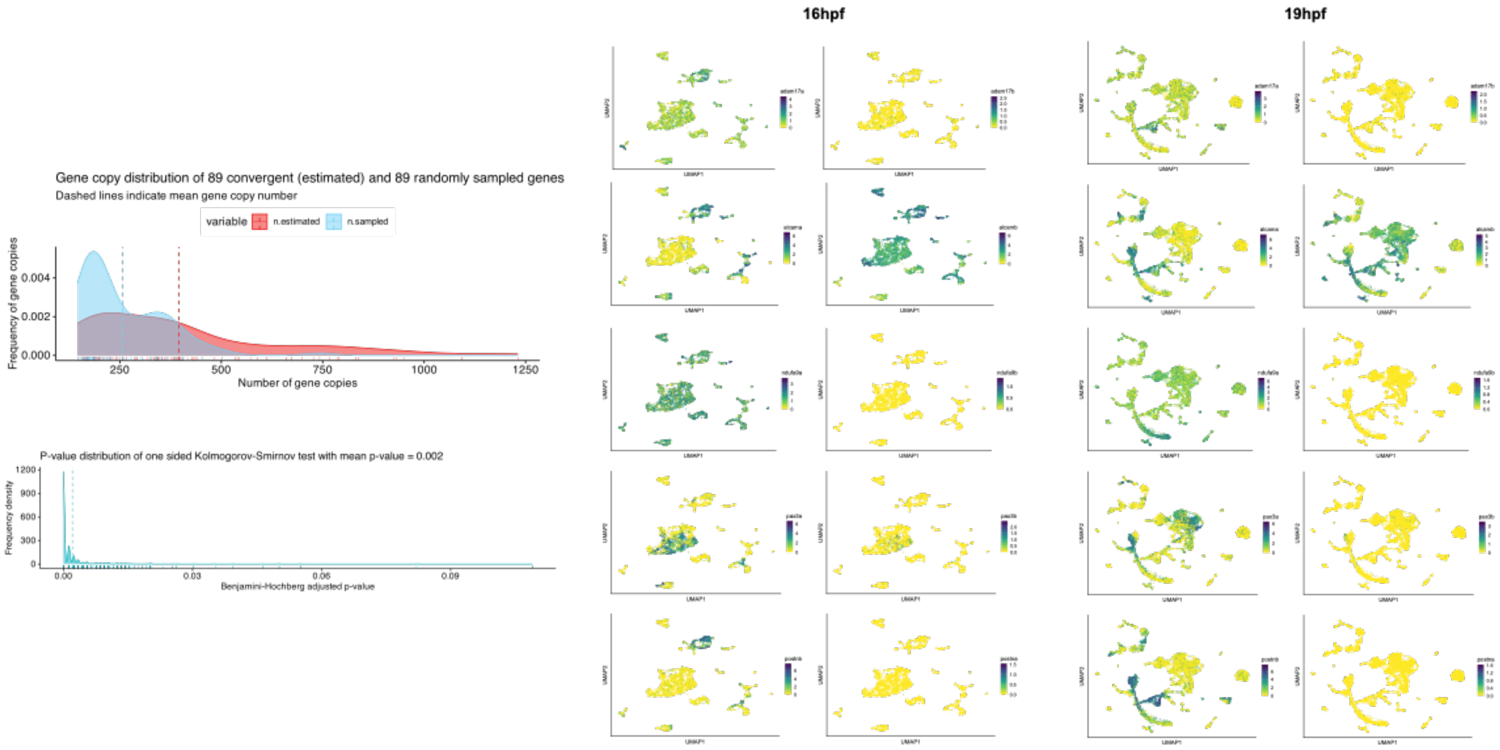

Fig S9: Distribution of gene copies in convergent orthogroups and expression divergence between paralogs. Here we test whether the distribution of gene copy numbers in the set convergent orthogroups is generally higher than the other orthogroups in our data set. To do we make a data frame with the copy number distribution of convergent orthogroups and 89 randomly selected non-convergent orthogroups and perform the KS test. I repeat this test 1000 times with different sets of randomly selected orthogroups. Following this I adjust all the *p-values* for multiple correction and plot the *p-value* distribution. As we can observe an overwhelming number of the *p-values* are closer to zero. 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.

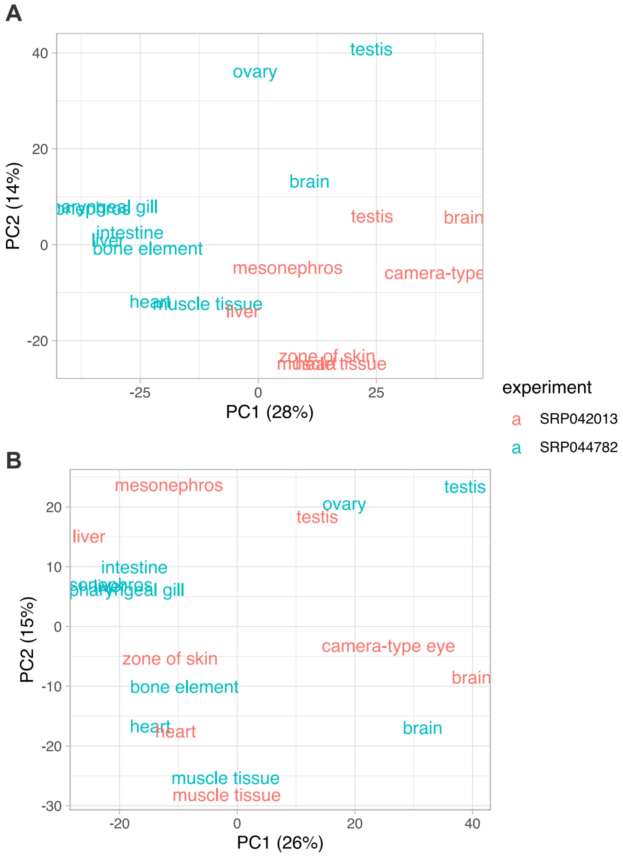

Fig S10: Correcting for batch effect in gene expression: The Bgee data sometimes includes expression levels of tissues from different experiments. (A) We checked for potential batch effects in tissue expression between different experiments and found evidence only in *Lepisosteus oculatus*. (B) We corrected this using the ComBat_seq function from the sva R package followed by quantile normalisation.