Figure #	Figure title One sentence only	Filename This should be the name the file is saved as when it is uploaded to our system. Please include the file extension. i.e.: *Smith_ED_Fig1.jpg*	Figure Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	Hypothalamic and POA cell types in zebrafish and Mexican tetra	Figure E1 - Hypothalamic and POA cell types in zebrafish and Mexican tetra.eps	(**a**) UMAP of zebrafish cells coloured and labelled by annotated cell type. (**b**) UMAP of Mexican tetra surface- and cave-morphs coloured and labelled by annotated cell type. (**c**) DotPlot of the top 2 maker genes for each zebrafish cluster from (a). (**d**) DotPlot of the top 2 marker genes for each Mexican tetra cluster from (b). Examples of potentially homologous cell types and their top marker genes share a colour (blue, green, red) in (c) and (d). (**e**) UMAP of merged but not batch-corrected zebrafish and Mexican tetra single-cell datasets.
Extended Data Fig. 2	Marker genes for cell types shared between zebrafish and Mexican tetra	Figure E2. Marker genes for cell types shared between zebrafish and Mexican tetra.eps	(**a**) DotPlot of the top 5 marker genes for each integrated cluster. (**b**) Proportion of cells from each cluster by species or species-morph (height of each bar along the x-axis). Width of each bar along the y-axis indicates the proportion of that cluster in the integrated data. Red outlines indicate the Mexican tetra-specific Ciliated cluster, and the integrated Immune clusters which are over-represented in the Mexican tetra dataset. (**c**) Density plot of the number of subclusters versus the fraction of each subcluster that is either from the zebrafish or Mexican tetra dataset. Subclusters with the majority of cells from the zebrafish dataset are shown in purple, and those with the majority of cells from the Mexican tetra dataset in yellow.
Extended Data Fig. 3	Shared subclusters are highly similar due to paralogous gene expression	Figure E3. Shared subclusters are highly similar due to paralogous gene expression.eps	(**a**) Gene orthology confidence from Ensembl for all marker genes, or those marker genes which were paralogs of a marker gene in the other species. (**b**) Gene order score from Ensembl for all marker genes, or those marker genes which were paralogs of a marker gene in the other species. (**c**) The percentage of conserved, species-specific, and species-specific paralogous subcluster marker genes corrected by SCORPiOS synteny-correction. (**d**) The percentage of morph-specific marker genes for each subcluster which were paralogs of either the conserved or opposite species-specific marker gene for surface- and cave-morphs of

			Mexican tetra. (**e**) The odds ratio for the enrichment of paralogs in the species-specific genes for each subcluster for zebrafish and Mexican tetra. (**f**) The row-scaled Δ*SI* for all subclusters between zebrafish and Mexican tetra. Yellow indicates the highest Δ*SI* value between Mexican tetra and zebrafish subclusters. For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5 times the interquartile range, thicker line represents the median.
Extended Data Fig. 4	Paralog shifts are associated with loss of ancestral gene expression patterns	Figure E4. Paralog shifts are associated with loss of ancestral gene expression patterns.eps	(**a-b**) Empirical cumulative distribution function (ECDF) for expression divergence (*dT*) for paralogous gene pairs. (**c-d**) ECDF of the number of cell types that have overlapping expression patterns within ancestral cell types for paralogous genes pairs (redundancy score, orange highlight in b). (**e-f**) ECDF of the number of non-ancestral cell types expressing each individual paralogous gene. Results for c-e are grouped by the age of the duplication inferred from the last common ancestor (LCA) which had both genes - from the oldest (Opisthokonta, yellow), to the most recent common ancestor (Otophysi, red), and to those gene duplicates which are only found in either *Danio rerio* or *Astyanax mexicanus* (dark red). Results from **b**, **d**, and **f** are filtered and grouped by the originating whole genome duplication event (WGD), either vertebrate (2R) or teleost (3R).
Extended Data Fig. 5	Gene regulatory networks identified by GENIE3/SCENIC	Figure E5. Gene regulatory networks identified by GENIE3.eps	(**a**) Comparison of the random forest weights for orthologous transcription factors in the zebrafish (y-axis) and Mexican tetra (x-axis) data for example terminal effector genes. Colours indicate whether those transcription factors are in the top 2% of transcription factors for each gene in either zebrafish (blue) and Mexican tetra (red), both (yellow), or none (black).
Extended Data Fig. 6	Species-specific subcluster identities are not dependent on species-specific genes	Figure E6. Species-specific subcluster identities are not dependent on species-specific genes.eps	(**a**) tSNEs of cells from clusters containing a species-specific neuronal subcluster coloured by the original subcluster identity. (**b**) tSNEs of cells from clusters containing a species-specific neuronal subcluster coloured by subcluster identity derived from subclustering without species-specific genes. (**c**) Sankey diagrams illustrating the relationship between original subcluster identities and identities from subclustering without species-specific genes. Box heights and line widths are proportional to the number of cells in each subcluster and connection, respectively.

			Shaded connections represent cells from species-specific subclusters.
Extended Data Fig. 7	Comparison of subcluster identities between independent and integrated analysis	Figure E7. Comparison of subcluster identities between independent and integrated analysis.eps	(**a**) Sankey diagram of Mexican tetra surface-morph specific subclusters and their relationship to integrated subclusters, and zebrafish subclusters. Box heights and line widths are proportional to the number of cells in each subcluster and connection, respectively. (**d**) Sankey diagram of Mexican tetra cave-morph specific subclusters and their relationship to integrated subclusters, and zebrafish subclusters. Box heights and line widths are proportional to the number of cells in each subcluster and connection, respectively. (**c**) Sankey diagram of the Zebrafish species-specific subclusters (middle) and their relationship to subclusters independently identified in the zebrafish (right) or Mexican tetra datasets (left). Box heights and line widths are proportional to the number of cells in each subcluster and connection, respectively. (**d**) Sankey diagram of the subclusters shared by, ("Shared (147)") or specific to, surface- and/or cave-morphs ("Cave-specific" or "Surface-specific"). The middle column depicts whether each subcluster is found in all cave-morph samples ("All Caves"), different combinations of multiple caves, or only in the datasets from specific cave-lineages ("Pachon" or "Molino"). Box heights and line widths are proportional to the number of cells in each subcluster and connection, respectively.
Extended Data Fig. 8	Comparison of neuropeptides and gene regulatory networks between surface- and cave-morphs	Figure E8. Comparison of neuropeptides and gene regulatory networks between surface- and cave-morphs.eps	(**a**) DotPlot showing expression of *galn* in the cells from the *galn*$^{+}$ cluster (Neuronal_07), and expression of *oxt*, *avp*, and *ENSAMXG00000021172* in the Neuronal_19 cluster. Cells are grouped by species morph and cave-lineage. (**b**) Similarity Index between the transcription factor sets for surface- and cave-morphs of Mexican tetra for neuropeptides, neurotransmitters, synaptic genes, and ion channels. (**c-f**) Random forest weights for orthologous transcription factors in the Mexican tetra surface-morph (y-axis) and Mexican tetra cave-morph (x-axis) data for the neuropeptides *galn*, *hcrt*, *oxt*, and *avp*. Colours indicate whether those transcription factors are in the top 2% of transcription factors for each gene in either surface-morphs (green) and cave-morphs (yellow), both (purple), or none (black). For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5 times the

			interquartile range, thicker line represents the median.
Extended Data Fig. 9	Transcriptional signatures of neuroinflammation resistance in cave-morphs	Figure E9. Transcriptional signatures of neuroinflammation resistance in cave morphs.eps	(**a**) tSNE reduction of immune clusters (Tcells, Bcells, Microglia, Macrophages, Mast cells, Thrombocytes, Neutrophils, and Erythorcytes) from surface- and cave-morph Mexican tetra coloured and labelled by species-morph. (**b**) tSNE reduction of immune cell types from surface- and cave-morph Mexican tetra coloured by cluster. (**c**) Marker genes for surface- and cave-morph versions of each immune cell type. Red outlines indicate differential expression of neuroinflammation associated genes in cave-morph immune cells. Gene expression is quantified by both the percentage of cells which express each gene (dot size) and the average expression in those cells (colour scale). (**d**) tSNE reduction showing expression of *ccr9a* in Mexican tetra immune cells. (**e**) Proportion of cells within each immune subcluster which come from Choy surface-morphs, or Molino, Tinaja, or Pachon cave-morphs.
Extended Data Fig. 10	A permanent stress-response in a cave-morph specific neuronal subcluster	Figure E10. A permanent stress-response in a cave-morph specific neuronal subcluster.eps	(**a**) tSNE reduction of Neuronal_03 cluster from Mexican tetra coloured and labelled by subcluster. (**b**) tSNE reduction of Neuronal_03 cluster from Mexican tetra coloured by species-morph. (**c**) DotPlot of the top 5 marker genes for each subcluster of the Neuronal_03 cell type (x-axis), and their expression across all subclusters (y-axis). Gene expression is quantified by both the percentage of cells which express each gene (dot size) and the average expression in those cells (colour scale). (**d**) Dendrogram of the Neuronal_03 subclusters based on the Variable Features of the Neuronal_03 cluster, and the proportion barplot of cells from each species-morph per subcluster. (**e**) GO analysis of genes differentially expressed between Neuronal_03_1 and Neuronal_03_4. (**f**) tSNE reduction of Neuronal_03 cluster from Mexican tetra coloured by *hspb1* expression. Neuronal_03_4 subcluster is highlighted by a dotted line. (**g**) Sankey diagram of the relationships between the Mexican tetra subclusters (left-hand side), integrated subclusters (middle), and zebrafish subclusters (right-hand side). Box heights and line widths are proportional to the number of cells in each subcluster and connection, respectively.

1

Item	Present?	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_Information_Shafer.pdf	Combined Supplementary Methods/Results, Supplementary References, Supplementary Figures 1-9
Reporting Summary	No		
Peer Review Information	Yes	*Shafer_PFfile*	

2

3

Type	Number If there are multiple files of the same type this should be the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_Supplementary_Video_1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Data		Supplementary_data.zip	Supplementary data for Cavefish single-cell sequencing publication ========== This archive contains the supplementary data for the paper "Gene family evolution underlies cell type diversification in the hypothalamus of teleosts", which includes all of the raw and partially processed data produced by the analyses presented. The archive contains: 1) The raw count data for both the zebrafish (Danio rerio) and Mexican tetra (Astyanax mexicanus) single-cell experiments, as compressed .csv files. 2) The Seurat object meta data for the zebrafish (Danio rerio), Mexican tetra (Astyanax mexicanus), and integrated Seurat objects, containing sample, species, and cell type cluster labels for each cell.

			3) CSVs for all marker gene lists used in the publication. 4) CSVs for all pseudobulk expression data for all cell type labels. 5) The raw data used for calculating the SI for each cluster and subcluster identity in the integrated data 6) Results from SCENIC/GENIE3 analysis, including the Linklists and tfModules outputs from SCENIC. 7) Results of the weir fst analysis between cave and surface populations, for both INDELs and SNPs 8) Ensembl biomart export files for determine paralogy relationships between genes within and across species 9) Results of trinarization of gene expression across all identities, an the uniquely expressed genes per identity. These are provided as R object files (.rds) Supplemental_data/3-marker_gene_lists ========== This folder contains marker gene lists for clusters and subclusters (".sub"), for the zebrafish ("Drerio"), Mexican tetra ("Amexicanus"), or integrated ("Integrated") datasets. Supplemental_data/4-pseudobulk_expression ========== This folder contains psuedobulk expression profiles for clusters ("Clusters") and subclusters ("Subclusters"), for the zebrafish ("Drerio"), combined Mexican tetra, and the surface and cave morphs of Mexican tetra, and integrated datasets.

Gene family evolution underlies cell type diversification in the hypothalamus of teleosts

Maxwell E.R. Shafer[1,2,*,†], Ahilya N. Sawh[1,2,*], Alexander F. Schier[1,2,3,4,5,6,*,†]

[1] Biozentrum, University of Basel, Switzerland
[2] Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts, USA
[3] Allen Discovery Center for Cell Lineage Tracing, Seattle, Washington, USA
[4] Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
[5] Harvard Stem Cell Institute, Harvard University, Cambridge, Massachusetts, USA
[6] Center for Brain Science, Harvard University, Cambridge, Massachusetts, USA

* Current address: Biozentrum, University of Basel, Switzerland

† Corresponding authors: MERS: max.shafer@gmail.com; AFS: alex.schier@unibas.ch

ABSTRACT:

Hundreds of cell types form the vertebrate brain, but it is largely unknown how similar cellular repertoires are between or within species or how cell type diversity evolves. To examine cell type diversity across and within species, we performed single-cell RNA sequencing of ~130,000 hypothalamic cells from zebrafish (*Danio rerio*) and surface- and cave-morphs of Mexican tetra (*Astyanax mexicanus*). We found that over 75% of cell types were shared between zebrafish and Mexican tetra, which diverged from a common ancestor over 150 million years ago. Shared cell types displayed shifts in paralog expression

that were generated by sub-functionalization after genome duplication. Expression of terminal effector genes, such as neuropeptides, was more conserved than the expression of their associated transcriptional regulators. Species-specific cell types were enriched for the expression of species-specific genes, and characterized by the neo-functionalization of expression patterns of members of recently expanded or contracted gene families. Comparisons between surface- and cave-morphs revealed differences in immune repertoires and transcriptional changes in neuropeptidergic cell types associated with genomic differences. The single-cell atlases presented here are a powerful resource to explore hypothalamic cell types, and reveal how gene family evolution and shifts in paralog expression contribute to cellular diversity.

INTRODUCTION:

The homology of neuronal cell types was first revealed by Ramón y Cajal, who observed that morphologically similar neurons were present in the brains of many species[1]. Since then, the comparison of cell types has largely relied on morphological criteria and, more recently, data from select marker genes[2]. These studies have led to the definition of major neuronal classes and subclasses[2,3] but it is still unknown how molecularly similar or different brain cell types are between species. Moreover, it is unclear how cell types diversify during evolution or adaptation to extreme environments. Biological novelty may arise as a result of gene expansion[4], but it is unknown how the evolution of gene families influences the diversification of cell types in the brain.

Single-cell sequencing has recently emerged as a powerful tool to study and map the cell types of individual species, and has allowed the identification of hundreds of transcriptionally unique cell types in vertebrate tissues, including the brain[5]. Recently, cross-species comparisons using single-cell RNA-seq have identified shared and species-specific cell types, as well as mechanisms for neuronal evolution[5,6]. These studies have identified conserved cell types during vertebrate development[7] and mammalian neurogenesis[8,9], as well as primate-specific adaptations[10,11]. Extension of these approaches to more diverse phylogenies is necessary for understanding the molecular and evolutionary basis of cell type conservation and diversification across the tree of life.

A powerful model for comparative studies of biological diversification are the teleosts. This group of nearly 30,000 described ray-finned fish species represents the largest clade within vertebrates and has undergone a taxon-specific whole genome duplication (WGD)[12,13]. It has been hypothesised that the vast diversification in morphology, physiology, and behaviour observed across teleost species was driven by gene family expansions associated with the teleost-specific WGD[4,12,14]. Most duplicated genes lose their functions through deleterious mutations (non-functionalization), but genes that are retained may undergo either sub-functionalization (partitioning of functions or gene expression patterns), or neo-functionalization (gain of novel functions or gene expression patterns). Little is known about the fate of these duplicated genes in teleosts, their roles in the vertebrate brain, or their links to cellular diversification.

In this study we analyze the conservation and diversification of teleost brain cell types using the zebrafish (*Danio rerio*) and the Mexican tetra (*Astyanax mexicanus*) as model systems. Zebrafish is the leading fish model system in developmental and neurobiology, whereas Mexican tetra is a powerful system for comparative studies. Mexican tetra has two morphs, an eyed surface-morph, and an eye-less and pigment-less cave-morph[15–17]. Comparisons between species, and between species-morphs represent two informative evolutionary distances: between distantly related species (150-200 million years, zebrafish and Mexican tetra), and within a species with large phenotypic differences (250-500 thousand years, species-morphs of Mexican tetra), that have been linked to changes in the development and gene expression patterns of the nervous syste[15,18].

To characterise cell type diversity at a high resolution in both zebrafish and Mexican tetra, we focus on the hypothalamus. The hypothalamus is a highly conserved forebrain region that is responsible for the generation and secretion of hormones and neuropeptides involved in diverse behaviours. Within the hypothalamus these functions are partitioned into specific neuropeptidergic cell populations regulating sleep/wake (*hcrt*+ and *galn*+ neurons), food intake (*agrp*, *npy*, *pomc*), aggression and sexual behaviours (*oxt*, *avp*, *npy*), and physiological homeostasis[19–21]. It is thought that hormone-secreting brain centres are ancient, and were present in the last common ancestor of all metazoans[22]. However, the level of homology in the cellular populations of the hypothalamus has not been comprehensively compared between species.

We used single-cell transcriptomics followed by high resolution clustering and cross-species integration to systematically identify the molecular similarities in the cellular repertoire of the teleost hypothalamus. First, we observe high conservation of cell types between species over 150 million years of evolution. Second, our results suggest that shared cell types have undergone shifts in paralog gene expression, and divergence in gene regulatory networks. Third, we link cellular novelty with genetic

novelty and the species-specific expression of paralogous genes. Fourth, we identify transcriptional and genomic differences between surface- and cave-morphs of Mexican tetra that are candidates to be associated with behavioural phenotypes of cave-adaptation.

RESULTS

Iterative clustering identifies shared and divergent cell types in the hypothalamus and preoptic area of D. rerio and *A. mexicanus*

To characterise similarities and differences of brain cell types between and within species, we performed scRNA-seq on ~130,000 cells from the hypothalamus and preoptic area (POA) of D. rerio (zebrafish), and from surface and 3 different cave species-morphs of A. mexicanus (Mexican tetra) (**Figure 1a-c**, **Extended Data Figure 1-2**, **Supplementary Information**, and **Supplementary Figures 1**). To resolve cell populations at high resolution, we performed iterative subclustering resulting in 194 subclusters with distinct gene expression patterns (**Figures 1d,** see **Supplementary Information, Supplementary Figures 2-4** and Methods for details on our clustering approach and the comparison between clusters and subclusters). Subclustering resolved rare cell populations such as the *hcrt*$^{+}$ subcluster which also expressed the neuropeptide *npvf* and the transcription factor *lhx9* (Neuronal_01_10) (Supplementary **Figure 2e**)[19,26]. The majority of subclusters were shared between species (151 out of 194 subclusters composed of > 10% of cells from both species), whereas 43 subclusters were specific to either zebrafish or Mexican tetra (species-specific, > 90% of cells from either species in our analysis) (**Figure 1d**). Analysis of the similarity between subclusters from each of the cave-morphs reflected the known phylogenetic relationship between surface-morphs, and Pachon, Tinaja, and Molino cave-morphs (Figure 1e). Thus, subclustering identified both similar and divergent cell types between zebrafish and Mexican tetra.

In the following sections we analyze this dataset to (1) identify the similarities and differences in gene expression for cell types shared between zebrafish and Mexican tetra (**Figure 2-3**), (2) compare gene regulatory networks across species (**Figure 4**), (3) define gene expression signatures associated with species-specific cell types (**Figure 5**), and (4) examine cell type similarities and differences between the surface- and cave-morphs of Mexican tetra (**Figure 6**). We then discuss our findings in relation to the molecular and evolutionary basis of cell type conservation and diversification.

Transcriptional Similarity Index measures gene expression similarities and differences between cell types

To quantify the similarities and differences in the subclusters shared between species we developed a transcriptional Similarity Index (*SI*) that compares the presence or absence of cell type specific marker genes (Extended Data Figure 2a, Supplementary Information). We calculated the *SI* between the marker gene sets for zebrafish and Mexican tetra for each of the 151 shared subclusters (Figure 2a-b). *SI* was consistently the highest between shared subclusters, and between subclusters from the same or related clusters (**Figure 2a-b**).

Analysis of the patterns of *SI* between subclusters revealed that progenitor subclusters had significantly higher levels (average *SI* of 0.327) than differentiated neuronal cells (average *SI* of 0.224) (**Figure 2c**). The lower divergence for progenitor cells between species could be due to their function in generating many different cell types, with pleiotropic effects expected from changing gene expression patterns in progenitor cell types. Indeed, the *SI* for the larger parental clusters (marker genes shared by multiple subclusters) was significantly higher than the similarity for subclusters (marker genes specific to subclusters) also suggesting pleiotropic effects (**Figure 2d**). For example, the cluster Neuronal_01 had a *SI* of 0.305, whereas its subclusters, Neuronal_01_0 - Neuronal_01_11, had *SI*s between 0.286 and 0.146 (mean of 0.206). To test whether the amount of transcriptional similarity scales with the evolutionary time between species, we also calculated *SI* for the same subclusters between cave and surface morphs of Mexican tetra (Figure 2e). Higher *SI* was observed for surface versus cave morph cells

than between zebrafish and Mexican tetra, indicating that *SI* reflects divergence time. These results highlight that the transcriptomes of progenitors (versus differentiated cells), and of cell clusters (versus subclusters), have changed the least during evolution.

Expression of paralogous and functionally similar genes contributes to transcriptional divergence in shared cell types

To determine what factors contribute to the transcriptional divergence of shared cell types, we compared the identity of marker genes that were specific to one or the other species. Examination of the zebrafish-specific marker genes revealed that many were paralogs of Mexcian tetra-specific marker genes, and vice versa. For example, erythrocytes were marked by *hbaa2* in Mexican tetra, but its paralog *ba1* in zebrafish. Neuronal_00 cells were marked by *zic2b*/*zic5* in Mexican tetra, but their paralogs *zic1*/*zic3* in zebrafish (Extended Data Figure 1). The best markers for GABAergic and Glutamatergic cell types were also paralogs of each other in the two species. Zebrafish cells expressed *slc17a6b*, and *gad2*, whereas Mexican tetra cells expressed the paralogs *slc17a6a* and *gad1b*, for Glutamatergic and GABAergic cells respectively (Supplementary Figure 4). These results suggest that instead of expressing orthologous genes, shared cell types often express paralogous genes.

To determine how frequent such shifts in paralog expression are, we calculated for each subcluster the percentage of zebrafish-specific marker genes that were paralogs of either a shared marker gene or a Mexican tetra-specific marker gene, and vice versa (Figure 2a). Up to 14% of the species-specific marker genes for each subcluster were paralogous to another gene expressed in the same subcluster in either species (Figure 2f). These levels represented a 4-15 fold or higher enrichment for paralogous genes (odds ratio, Fisher's test) (**Figure 2g**), and were not due to mis-identification of orthology/paralogy relationships between genes (Extended Data Figure 3a-c). The vast majority of the observed changes in paralog expression between zebrafish and Mexican tetra were also conserved between surface- and cave-morphs of Mexican tetra (**Figure 2h**). Differences between surface- and cave-morph marker genes were also enriched for paralogous genes (Extended Data Figure 3d-e) Progenitor cells, which had the least amount of transcriptional divergence, also had the highest enrichment for paralogous genes (Figure 2i). Additionally, the enrichment for paralog expression was positively correlated with the *SI* across all subclusters (**Figure 2i**). Non-paralogous differentially expressed genes between species were also enriched for similar gene ontology terms suggesting conservation of function across species (Supplementary Information and Supplementary Figure 6). To account for shifts in paralog expression and to produce a more accurate estimation of the similarity of subclusters between species we calculated a corrected-*SI*, which considers paralogs as functionally equivalent. We also calculated the difference between corrected-*SI* and *SI* (Δ*SI*) (Extended Data Figure 3f). The mean of the corrected-*SI* of subclusters was twice that of the mean of the *SI* for the same subclusters (**Figure 2j**). The Δ*SI* was highest between shared subclusters (Extended Data Figure 3f). These results indicate that gene expression differences between shared cell types are largely due to shifts in the expression of functionally similar paralogs both between species and between species-morphs.

Enrichment of paralogous genes in shared subclusters is due to gene duplication followed by differential retention of expression patterns

Possible explanations for the observed enrichment for paralogous gene expression in shared subclusters include the differential retention or gain of expression patterns in the two species following gene duplication. In this scenario, the expression patterns of newly duplicated genes are initially identical or highly similar, but over time undergo differential sub- or neo-functionalization in gene expression in different species. Consider the hypothetical Gene X, which is duplicated in the ancestor of Species 1 and 2. After time, expression of Gene Xa is retained in the cell types of interest in Species 1, whereas expression of Gene Xb may be retained in Species 2 (**Figure 3a-c**). One prediction of this paralog sub-

functionalization model is that more recently duplicated genes will have more similar expression patterns within each species.

We used a previously described metric to calculate the expression divergence (*dT*) for each pair of paralogous genes within each species[31]. We found that paralog gene pairs that were generated through more recent duplication events showed lower *dT*s (empirical cumulative distribution function) than paralogs from more ancient duplication events (**Figure 3d-e, and Figure S6b**). To gain further insight into the divergence patterns of paralogous genes within species we determined putative ancestral gene expression patterns based on the minimal group of cell types that express either or both paralogs in both species (ancestral cell types) (Figure S12b'). This analysis revealed that divergence of paralog gene pairs was predominantly due to loss of co-expression rather than expansion of expression patterns into new cell types (Extended Data Figure 4 and Supplementary Information).

To test whether genes diverged differentially in zebrafish and Mexican tetra, we compared the *dT* for each pair of paralogous genes which were conserved in both species. The expression patterns of most gene pairs have diverged in both species, but many have diverged in only zebrafish or Mexican tetra (**Figure 3f**). The most recently duplicated gene pairs also had *dT* patterns which were different between species. For example, 14 of 22 gene pairs that arose in the last common ancestor of zebrafish and Mexican tetra (*Otophysi*) have different *dT* levels and expression patterns in zebrafish and Mexican tetra (**Figure 3g**). For example, *etv5a* is expressed across several glutamatergic neuronal clusters in both Mexican tetra and zebrafish (**Figure 3h**). The expression of its paralog *etv5b* in the same subclusters was retained in zebrafish, but lost in Mexican tetra (**Figure 3h**).

For all gene pairs in both species, we determined the correlation of their expression patterns between zebrafish and Mexican tetra (**Figure 3i**). The oldest gene pairs examined had the most correlated expression patterns across species, including those gene pairs which arose prior to the last common ancestors of vertebrates (Bilateria, Chordata, and Vertebrata) (**Figure 3i**). No difference was observed between genes arising from either the 2R or 3R WGDs (**Figure 3i**). This is consistent with the functions and expression patterns of older genes having diverged long ago, and are therefore less likely to change further in more closely related species. These results indicate that the divergence in gene expression patterns for duplicated genes has occurred differently in zebrafish and Mexican tetra.

The transcription factors associated with neuropeptide expression have diverged between species

Our comparison of gene expression signatures in cell types between species revealed maintenance of cellular function by either reducing transcriptional divergence and/or promoting the expression of functionally similar and paralogous genes (**Figures 2-3**). We therefore wondered whether the upstream regulatory mechanisms controlling the expression of functional terminal effector genes, such as neuropeptides, were also conserved between zebrafish and Mexican tetra. To test this hypothesis we identified putative gene regulatory networks using SCENIC/GENIE3 and compared these between species.

The putative gene regulatory networks (TF sets) for the same terminal effector genes were different between zebrafish and Mexican tetra, and many of the top TFs for each gene were not shared between species (Extended Data Figure 5). For example, the neuropeptide *vip* is highly associated with the TFs *gsx1*, *otpa*, *six3a*, *mllt11*, and *xbp1* in zebrafish, but with the TFs *klf17*, *id4*, *nr4a3*, *hes6*, and *ets2* in Mexican tetra (**Figure 4a**). To quantify how similar TF sets were between species, we calculated the *SI* between zebrafish and Mexican tetra for the TF sets of each terminal effector gene (**Figure 4b**). TF sets associated with neuropeptides had significantly higher *SI* than TF sets associated with neurotransmitters, synaptic, or ion channel genes (**Figure 4b**). Thus, more of the same TFs remain associated with specific neuropeptides across species compared to other terminal effector genes.

We then quantified the similarity in the relative contribution of each TF (TF weights) for each terminal effector between species. No statistical difference was observed between the terminal effector

classes, and all gene sets had low or negative correlation between species (mean Pearson correlation between 0 and 0.15) (**Figure 4c**). For example, the TFs *prox1b* and *id4* appear in the TF sets for *vip* in both species, but are more predictive of *vip* expression in Mexican tetra than in zebrafish (**Figure 4a**). These results suggest that even in cases where the same TFs are associated with specific terminal effector genes across species, their relative contributions to putative target gene expression are not maintained.

Transcription factors diverge more than target genes and can be replaced by non-paralogs

Divergence in TF sets between species may be compensated by expression of paralogous TFs, similar to what we observed for subcluster transcriptomes. For example, the expression of the highly conserved neuropeptide *oxt* was correlated with the TFs *sim1b*, *otpa*, and *otpb* in zebrafish, while in Mexican tetra, *oxt* was highly correlated with *sim1a* and *otpb* (Extended Data Figure 5). *sim1b* is specific to the zebrafish genome, indicating that the paralogs *sim1a* and *sim1b* underwent sub-functionalisation in zebrafish, with *sim1a* losing its function and co-expression with *oxt* in zebrafish[34]. In contrast, in the case of the neuropeptide *vip*, there were no highly weighted TFs associated with its expression that were paralogs in the two species (**Figure 4a**). However, in general we did not observe compensation across species through the association of paralogous TFs with terminal effectors (Figure 4d).

The above results suggest that the expression patterns of TFs may be less conserved between species than the expression patterns of their target genes (non-TFs). To test this prediction, we calculated the *SI* across all subclusters using all marker genes, only those marker genes that were TFs, or only those marker genes that were neuropeptides and neurotransmitters (terminal effector genes). SI was significantly higher for transcription factors compared to all marker genes, and was highest for neuropeptides and neurotransmitters (**Figure 4e**). For example, the neuropeptidergic subclusters of the Neuronal_01 parent cluster (including the galn+, hcrt+, and oxt+ subclusters) have lower SI when considering only TFs then when considering NP/NT genes (**Figure 4f**). All together, both the SCENIC and SI results suggest that specific classes of genes in specific cell types may be more conserved between species, and may even be more conserved than the TFs which regulate them.

Species-specific cellular novelty is associated with species-specific genetic novelty

We next sought to define which genes were associated with species-specific subclusters. There were 43 species-specific subclusters in our integrated dataset, each composed of > 90% of cells from one species. Five of these species-specific subclusters were from the Mexican tetra-specific “Ciliated” cell type (**Figure 1** & Extended Data Figure 2b), 19 were hematopoietic subclusters (see below), and 6 were from the Oligodendrocyte, Endothelial, and Lymphatic parental clusters. There were 8 Mexican tetra-specific and 5 zebrafish-specific neuronal subclusters, which expressed a variety of genes indicating that they may represent different temporal or spatial cell states, captured in one species or the other (Figure 5a, Supplementary Information and Supplementary Figure 7).

It has previously been reported that species- or lineage-specific genes may contribute to species- or lineage-specific morphological and cellular innovations[35]. Though they may be shared by other species of fish, for simplicity we refer here to genes that are only found in the genome of zebrafish or Mexican tetra as species-specific genes. In both zebrafish and Mexican tetra, a higher percentage of the genes expressed in species-specific subclusters were species-specific, compared to the genes expressed in shared subclusters (**Figure 5b-c**). Importantly, all of these subclusters were still identified in the absence of species-specific genes, suggesting that they have distinct expression patterns of orthologous genes as well (Extended Data Figure 6). Mexican tetra non-neuronal cells expressed significantly more species-specific genes as compared to neuronal subclusters, with the immune subclusters expressing the highest percentages (**Figure 5d**). In contrast, all zebrafish neuronal subclusters expressed more species-specific genes as compared to non-neuronal subclusters (**Figure 5e**).

Enrichment for species-specific genes was also apparent in the species-specific neuronal subclusters (**Figure 5f**). For example, the zebrafish-specific Neuronal_04_7 subcluster was distinguished by 5 members of the jacalin family of lectins (*jac2*, *jac3*, *jac6*, *jac8*, and *jac9*) (**Figure 5f**). This gene family has undergone an extensive species-specific gene expansion, resulting in 14 known genes in zebrafish, compared to only 2 genes in Mexican tetra[36]. Additionally, the Mexican tetra-specific subcluster Neuronal_12_5 expressed the neuronal calcium sensor (NCS) *HPCAL1, which was generated by a Mexican tetra-specific duplication* (**Figure 5f**). [37]. One subcluster expressed the Mexican tetra-specific guanylate cyclase *ENSAMXG00000017498*, and *vipb*, which is a paralog of the neuropeptide *vip* that arose in a common ancestor of teleosts. *vip* is expressed in several subclusters shared by zebrafish and Mexican tetra, whereas *vipb* is expressed only in the Mexican tetra-specific Neuronal_13_2 (**Figure 5f**). This result suggests that the expression patterns of *vipb* and *ENSAMXG00000017498* have undergone neo-functionalization in Mexican tetra, but not zebrafish.

Three other species-specific subclusters were characterized by genes which were duplicated in a common ancestor of zebrafish and Mexican tetra, but subsequently lost in zebrafish (**Figure 5f**). These include Neuronal_10_5, which expresses the c-type lectin *COLEC12*, and Neuronal_07_0 and Neuronal_07_5 which both express *NPTX1* and *PPP3CA* (**Figure 5f**). Altogether, these results suggest that the main driver of cellular diversification may be species-specific expansion, retention, and neo-functionalization of the expression patterns of gene families.

Comparisons of cell types between species-morphs

Transcriptional differences in shared neuropeptidergic cell types

Zebrafish and Mexican tetra last shared a common ancestor roughly 200 million years ago, yet we found that the degree of cell type conservation between these two teleost species was extensive. We therefore wondered whether the surface- and cave-morphs of Mexican tetra, which shared a common ancestor 250-500 thousand years ago, had any detectable cell type differences. To identify the repertoire of cellular diversity in Mexican tetra, we performed subclustering on the Mexican tetra dataset alone, resulting in 166 subclusters, including 19 subclusters that were species-morph specific, and 147 subclusters shared between species-morphs (Extended Data Figure 7a-c and Supplementary Figure 8; See Supplementary Data for the full list of subclusters and associated marker genes for both zebrafish and Mexican tetra).

Neuronal subclusters shared by cave- and surface-morphs were characterized by low dendrogram distance, high *SI*, and a lack of enrichment for genes associated with divergent genomic windows (**Figure 6a**). The most transcriptionally different subclusters between morphs were the *galn*$^{+}$ and *otpa*$^{+}$/*oxt*$^{+}$ subclusters (**Figure 6a**). Further examination of these subclusters revealed that surface-morph *galn*$^{+}$ cells expressed *galn* at a significantly higher level than cave-morph cells (Extended Data Figure 8a). Similarly to what we observed between species (**Figure 5**), surface- and cave-morph *oxt*$^{+}$ cells were distinguished by the differential expression of gene duplications. Surface-morph *oxt*$^{+}$ cells co-expressed *oxt* and its paralogs *avp* and *ENSAMXG00000021172*, whereas cave-morph *oxt*$^{+}$ cells only expressed *oxt* (Extended Data Figure 8b). Co-expression of *oxt* and *avp* was not observed in zebrafish, with each neuropeptide expressed in its own subcluster. These results highlight transcriptional changes in conserved cell types which may be associated with cave-adaptation.

It was recently reported that the expression of the neuropeptide *hcrt* is upregulated in Pachon cave-morphs, and is associated with increased sleep/wake activity compared to surface-morphs[38,39]. Additionally, our genetic analysis suggested that genes associated with circadian rhythm and neuropeptidergic cell types were under selection in cave morphs (Supplementary Information and Supplementary Figure 9). We wondered if we could use our single-cell data to identify changes in the transcription factors or regulatory network underlying the expression of *hcrt* and other neuropeptides and terminal effector genes between morphs. The majority of terminal effector GRNs were more conserved between species-morphs than between species, including the TFs associated with *galn* (Extended Data

Figure 8c-d). The TFs associated with the *hcrt* were poorly correlated between species-morphs, with the TF *creb3l1* more highly associated with *hcrt* expression in surface-morph cells, compared to cave-morph cells (Extended Data Figure 8e). High association between *hcrt* and *creb3l1* was not observed in the zebrafish data, indicating that this association may be specific to Mexican tetra, and responsible for the increased *hcrt* expression previously observed. This analysis also provided a potential mechanism for the co-expression of *oxt and avp* in Mexican-tetra compared to zebrafish. Three of the top TFs associated with *oxt* (*creb3l1*, *otpb*, and *sim1a*), were also predictive of *avp* expression (Extended Data Figure 8f-g). Differential expression of neuropeptides within conserved cell types may therefore be a common cave adaptation strategy across morphs.

Species-morph specific subclusters are species-specific and express cell-state transcriptomes

Of the 19 species-morph specific subclusters, 4 were neuronal, 3 were from glial populations, and 12 were from the hematopoietic lineage (**Figure 6a-b** and Extended Data Figure 7a-b). The majority (11/19) of these species-morph specific subclusters mapped to integrated identities that were also specific to Mexican tetra. This included 3 of the 4 neuronal subclusters and 7 of the 12 immune subclusters that were species morph-specific (Extended Data Figure 7a-b). This suggests that many of the cell types specific to Mexican tetra are associated with or were co-opted during adaptation to the cave environment. Similarly, expression of Mexican tetra-specific genes was enriched in cell types from the hematopoietic lineage, which represented the majority of the species and species-morph specific subclusters (Figure 5a). Pachon cave morphs have been reported to have a smaller and less active immune system than surface morphs[40]. Though concluding changes in cell type proportions is difficult in single-cell experiments, we consistently observed fewer immune cells across independent cave-morph samples than in surface-morph samples (Extended Data Figure 2b, Figure 6a, Extended Data 9). Furthermore, hematopoietic lineage cells from cave-morphs expressed high levels of *ccr9a* and *sat1b*, and low levels of *fabp11a*, conditions which have been linked to inflammation resistance (Extended Data Figure 9c-d)[41–44]. Altogether, these results suggest that cave-morphs have a reduced immune system that expresses a neuro-inflammation resistance cell state transcriptome.

Three of the four species-morph specific neuronal subclusters mapped to integrated subclusters that were also species-specific: surface-morph specific Neuronal_09-4 mapped to Mexican tetra specific Neuronal_03_13 (*pou4f2* and *etv1* positive), cave-morph specific Neuronal_00_1 and Neuronal_03_6 mapped to Mexican tetra specific Neuronal_07_6 (*ENSAMXG00000025407*+), and Neuronal_12_5 (*HPCAL1*+) respectively (Extended Data Figure 7a-b). The identity of the cave-specific neuronal subcluster Neuronal_03_4 was less clear. Cells from this subcluster mapped to a cell type shared between species, and expressed a set of marker genes that was conserved across species (*rtn4rl2a*, *rtn4rl2b*, *cd9b*, and *penkb*) (Extended Data Figure 10a-d). However, Neuronal_03_4 cells also expressed an additional gene signature, which included the genes *rcan1a* and *prelid3b* (Extended Data Figure 10c). GO analysis of these differentially expressed genes between Neuronal_03_4 and the highly similar shared Neuronal_03_1 revealed enrichment for terms related to stress response, protein folding, and translation, including the heat-shock genes *hspb1*, *hspa4a*, and prolyl isomerase *fkbp4* (Extended Data Figure 10e-f). These results indicate that an ancestral cell type found in both zebrafish and Mexican tetra acquired a stress response transcriptional program in the cave lineage, resulting in a morph-specific cell state (Extended Data Figure 10g).

DISCUSSION:

How evolution generates and shapes cellular diversity is largely unknown. In this study we used single-cell transcriptomics, high resolution clustering, and cross-species integration to compare cell types of the teleost hypothalamus between two divergent teleosts, zebrafish and Mexican tetra. First, we observe extensive conservation of cell-types across roughly 150 million years of evolution between zebrafish and Mexican tetra (>75% of all subclusters were shared), providing a high resolution

quantification of the molecular similarity between cell types across such a large phylogenetic distance. Second, we show that cell types conserved between species are characterised by subfunctionalization of paralogous gene expression patterns and by gene regulatory divergence. Third, we find that species-specific cell types were associated with the evolution of gene families, linking genetic novelty with cellular novelty. Fourth, we identify transcriptomic, cellular and genomic changes associated with cave-adaptation in Mexican tetra.

Shared cell types are characterized by regulatory divergence and shifts in paralog expression

Hundreds of cell types have been cataloged in the brains of vertebrates, including fish, mice and humans, but their conservation between species is unclear[8,11,23,45]. We observed extensive conservation of 75% of cell-types between zebrafish and Mexican tetra, who last shared a common ancestor more than 150 million years ago, before the break-up of Pangea[46,47]. In our analysis, shared cell types were even more similar when taking paralog expression into account. Up to 20% of the transcriptomic divergence of shared cell types between species was from preferential expression of functionally similar paralogous genes. These expression pattern differences, or paralog shifts, suggest that shared cell types often express paralogous genes. Similarity and shifts in paralog expression between species was highest for progenitor cell types, and for clusters compared to subclusters. Changes to cluster and progenitor populations would likely have pleiotropic effects that may have prevented transcriptional divergence. A comparison of single-cell atlases across animal phyla has also demonstrated shifts in paralog expression for homologous cell types[48]. In that study the authors argue that paralog shifts may be due to genetic compensation by paralog substitution. However, our analysis suggests that divergence patterns of paralogous genes were mostly due to loss of redundancy, differed between species, and scaled with evolutionary gene age. This observation suggests that following ancestral gene duplication, expression patterns of paralogous genes are shifted, caused by independent sub-functionalization of gene expression patterns in each species. Further work will be necessary to determine the exact evolutionary and molecular mechanisms that generate paralog shifts in shared and homologous cell types.

We found that the expression patterns of transcription factors and their putative associations with specific classes of terminal effectors were less conserved than the expression patterns of the terminal effectors themselves. This observation contrasts with the high inter-species conservation of 'core' TFs expressed during early lineage determination events[7]. It agrees, however, with the low inter-species conservation in the expression of TFs 're-used' multiple times in different tissues[7]. The TF code associated with specific cell types, such as hypothalamic neurons, may therefore not be highly conserved between species. Alternatively, the differences in GRNs we observe might be caused by convergence in the GRNs of non-orthologous cell types to regulate terminal effector genes, as has been postulated for neurotransmitters in the Drosophila brain[52]. We note, however, that neuropeptidergic cell types and effector gene expression patterns were highly conserved between zebrafish and Mexican Tetra. We therefore favor a process akin to developmental systems drift, where conserved homologous traits between species can have divergent gene regulatory underpinnings caused by neutral drift[53]. We speculate that there might be cellular systems drift, where selection acts to maintain the functional output of cell types, rather than the regulatory mechanisms which generate or maintain them.

Together, our results paint a picture of the evolutionary history of the hypothalamic cell types in two teleost species. Cell types are highly conserved between species, yet divergence in paralog expression and regulatory associations is common. These patterns suggest an interplay between dosage compensation and subfunctionalization of expression patterns after genome duplication, neutral evolution causing shifts in paralog expression and regulatory divergence, and stabilizing selection maintaining cell type functions.

Species specific cellular novelty is associated with species-specific genetic novelty and paralog neo-functionalization

Cross-species comparisons using single-cell sequencing data typically only consider orthologous genes between the species of interest, limiting the identification of species-specific innovations[5]. Here we find that the majority of species-specific cell types between zebrafish and Mexican tetra were enriched for the expression of non-homologous genes between species. This observation extends previous studies that have linked the evolution and diversification of biological traits with genetic novelty[35,54]. For example, expression of human specific genes in radial glia has been linked to cortical evolution and the expansion of the neocortex in primates[55]. Indeed, we found that the expression of jacalin lectins, which are specific to the zebrafish lineage[36], are associated with a zebrafish-specific neuronal cell type. These results illustrate how species-specific genetic novelty underlies species-specific cellular novelty.

Moreover, our results suggest that the generation of new cell types within teleosts may be driven by species-specific neo-functionalization of paralogous genes. Many of the species-specific cell types were associated with expression of genes generated by recent duplication events. Furthermore, the loss of ancestrally duplicated paralogs in zebrafish (*HPCAL1*, *COLEC12*, *NPTX1*, and *PPP3CA*) was also associated with Mexican tetra specific cell types. These genes had expression patterns that differed from their paralogs, suggesting they have gained new functions since their duplication. Previous studies have suggested that new cell types are generated first through the birth of similar or homologous sister cell types[56]. Genetic individuation of sister cell types through the generation of distinct core regulatory complexes would then allow subsequent divergence through acquisition of different pre-existing gene modules[3]. Our results suggest an alternative scenario wherein gene duplication may precede or even drive the partitioning of cellular functions into distinct cell types (sister cell types). For example, amino acid substitutions between *vip* and *vipb* may have endowed different functionality, promoting the generation of the *vipb* subcluster in Mexican tetra. This scenario is reminiscent of the evolution of rod and cone cells following opsin gene duplication[57,58]. These observations suggest paralog neo-functionalization as a basis for cell type diversification.

Similar to the relationships between homologous genes, homologous cell types (shared cell types or sister-cell types) could refer to populations separated by a speciation event (orthologous cell types), or through a cell type duplication event (paralogous cell types)[56]. The shared populations we observed between zebrafish and Mexican tetra may therefore represent orthologous cell types which were present in the last common ancestor of both species. Species-specific cell types derived from cell type duplication events within species may be paralogous to cell types shared between species. Future work will be necessary to unravel the complicated evolutionary history of gene and cell type diversification.

Single-cell transcriptomic signatures associated with cave-adaptation

Mexican tetra Pachon cave-morphs have previously been reported to have a smaller and differentially active immune system[40]. Our results extend these observations to the Tinaja and Molino cave-lineages. In addition, we observe expression of a neuro-inflammation resistance signature in the immune cells of all three cave-morphs. Inflammation and neurodegeneration are intricately connected, and associated with aging in many species, including humans[59]. Negligible senescence has been reported in cave-morphs compared to surface-morphs[60]. It is therefore intriguing to speculate that the lack of immune inflammation in the nervous system may contribute to the lack of age-related senescence in cave-morphs.

The species-morphs of the Mexican tetra have divergent behavioural phenotypes which have previously been linked to the hypothalamus[38,39,61,62]. We observed differences in the expression patterns of several neuropeptides associated with these behaviours. For example, decreased *galn* expression in cave-morphs versus surface-morphs could partially explain the loss of sleep or changes in appetite or aggression in cave-morphs[63–65]. Alterations in oxytocin cells might also be linked to changes in appetite, or the lack of social interactions (schooling) observed in cave-morphs[63,66]. In the future, the single-cell atlases presented here will be a powerful resource to explore the behavioural differences between both species and species-morphs.

METHODS

Husbandry of zebrafish and Mexican tetra

All animal work was performed at the facilities of Harvard University, Faculty of Arts & Sciences (HU/FAS). Mexican tetra husbandry was performed as previously described[67]. This study was approved by the Harvard University/Faculty of Arts & Sciences Standing Committee on the Use of Animals in Research & Teaching under Protocol No. 25–08. The HU/FAS animal care and use program maintains full AAALAC accreditation, is assured with OLAW (A3593-01), and is currently registered with the USDA.

Processing of samples for scRNA-seq

Wild type adult zebrafish, and wild type adult Mexican tetra surface- and cave-morphs were used for scRNA-seq analysis. All zebrafish used were approximately 2-3 months old, and all Mexican tetra were between 1-2 years of age. For all zebrafish samples, tissues from 4-6 individual zebrafish were pooled for downstream dissociation then split into 4 samples for single-cell encapsulation. As their brains are much larger, each sample for Mexican tetra was composed of a single individual fish. A total of 16 samples of zebrafish (8 males, 8 females), and 16 individual Mexican tetra were used (8 male, and 8 female), including 8 Choy surface-morphs, 4 Pachon cave-morphs, 2 Tinaja cave-morphs, and 2 Molino cave-morphs split evenly between males and females.

The same procedure was used to collect and dissociate single-cells from both zebrafish and Mexican tetra. Animals were sacrificed by first placing them on ice, followed by decapitation. Whole brains were removed and immediately placed in 4% low-melt agarose mixed 50:50 with Neurobasal media plus B27 supplement (2% agarose final solution) (Thermofisher). Once solidified, 500 µm sections were obtained from whole brains mounted in agarose using a vibratome (Leica VT1000S). The hypothalamus and pre-optic area were then dissected from vibratome sections and dissociated into single cells using the Papin Dissociation Kit (Worthington) as previously described[23]. Cells were counted using a hemocytometer and resuspended at a final concentration of 1000 cells/µl in Neurobasal media (Thermofisher). Samples were run on the 10X Genomics scRNA-seq platform according to the manufacturer's instructions (Single Cell 3' v2 kit). Libraries were processed according to the manufacturer's instructions (Single Cell 3' v2 kit). Transcriptome libraries were sequenced using Nextera 75 cycle kits at the Bauer Core Facility (Harvard). Protocol for cell dissociation is available at https://github.com/maxshafer/Cavefish_Paper.

We recovered between 2998 and 5490 cells per sample for the zebrafish dataset, and 3029 and 5919 cells per sample for the Mexican tetra dataset. Samples had a minimum of 18,347 reads per cell with averages of 31,656 and 29,536 reads per cell for the zebrafish and Mexican tetra datasets, respectively. Sequencing saturation was between 68%-90%, with means of 78% for zebrafish samples and 83% for Mexican tetra samples. Between 30% and 60% of reads per sample for both datasets were mapped confidently to their respective transcriptome (78% - 88% to the genome).

Bioinformatic processing of raw sequencing data and independent cell type clustering and subclustering analysis

Transcriptome sequencing data were processed using Cell Ranger 2.1.0 according to the 10X guidelines to obtain cell by gene expression matrices for each sample. For zebrafish, reads were mapped to a transcriptome constructed using the GRCz10 genome assembly annotated using the RefSeq genome annotation for GRCz10 (NCBI). For Mexican tetra, reads were mapped to a transcriptome constructed using the AstMex102 genome assembly annotated using the Ensembl genome annotation for AstMex102 (Ensembl). Clustering analysis was performed using Seurat v3.2.0[24]. Due to the lack of a mitochondrial genome for *A. mexicanus*, we opted not to remove cells with high mitochondrial content from the Zebrafish dataset. The following options were used for PCA, knn graph construction, and

clustering for both zebrafish and Mexican tetra. Only cells with between 200 and 2500 expressed genes were used (nFeature_RNA). Variable features were obtained using the *mean.var.plot* (*mvp*) selection method as in Seurat v2.3.4. The identified variable features were used for PCA, and the top 50 PCs were used for clustering, though similar results were obtained with variable PC numbers. A k of 30 (*k.param*), and an error bound of 0.5 (*nn.eps*) were used for constructing the Shared Nearest Neighbor (SNN) graph. Clusters were called using a resolution (*resolution*) of 0.6 using the original Louvain algorithm. Shared marker genes for each cluster were obtained using Seurat's FindConservedMarkers function, and species-specific marker genes were identified by first subsetting the Seurat object by species before running the FindMarkers function for each cell cluster and subcluster. These genes, as well as genes with known expression patterns in neuronal and hypothalamic cell types were used to annotate subclusters. Clusters were identified as GABAergic or Glutaminergic based on which marker genes they expressed most highly (*slc17a6a*/*slc17a6b* or *gad1b/gad2*). In many cases, clusters expressed markers of both, due to having both GABAergic and Glutamatergic cells and were therefore all annotated as Neuronal.

Independent subclustering analysis was performed by first subsetting the zebrafish or Mexican tetra data into individual clusters, then performing all of the steps of Seurat clustering on each cluster independently, including finding highly variable genes and principal component analysis. Parameters used for subclustering were the same as for clusters, except we used the resolution 0.4 (*resolution*) and 15 PCs derived from the variable features of the cells in each cell cluster. Because different sets of variable genes were used for subclustering and construction of the tSNE projection for the full dataset, the positions of subcluster labels are not necessarily representative of the true differences between subclusters. Full analysis scripts for cell type clustering, R objects, and raw sequencing, including all variables used are available on GitHub (https://github.com/maxshafer/Cavefish_Paper). Raw count data is available in the **Supplemental Data**, and raw sequencing data is available on NCBI GEO.

Dataset integration and integrated clustering and subclustering analysis

Datasets were initially integrated using Seurats' MergeObjects function. Given the large biological batch effects between the species, cells first clustered by species, then by cell type. To correct for species-specific batch effects, and identify shared and species-specific cell types we used Seurat v3.0.0 to integrate the zebrafish and Mexican tetra datasets. Seurat uses Canonical Correlation Analysis (CCA) to identify correlated changes in the transcriptomes of cell types between species, and identifies the most similar clusters across species using Mutual Nearest Neighbour (MNN) analysis. This allows identification of cluster specific batch correction vectors, which are used to correct the expression values of a subset of genes between species. These genes and their corrected expression values are then used for dimensionality reduction and clustering analysis. All genes from both datasets were used in the integration process, and orthologous genes were identified by matching gene names. The best results were obtained by first clustering using 100 *dims*, a *k.param* of 20, a *res* of 0.15, and an *nn.eps* of 0, which segregated all non-neuronal cells into appropriate clusters. Following this, we used expression of the neuronal marker gene *gng3* to identify and combine all neuronal cell clusters together. To cluster the neuronal cells, we used 10 *dims*, a *k.param* of 20, and an *nn.eps* of 0 which generated the 14 neuronal populations used in this study. Integrated subclustering analysis was performed by first subsetting the integrated data into individual clusters, then performing all of the steps of Seurat integration and clustering on the zebrafish and Mexican tetra cells from each cluster independently. Integration, including CCA and MNN analysis, was performed independently on each integrated cluster to maximize the gene information used to identify shared and species-specific cellular heterogeneity. For subclustering we used between 5 and 50 *dims* for each cluster depending on the number of cells, and a *res* of 0.25. Following integration, several subclusters were identified as aberrant, and expressed marker genes from both non-neuronal and neuronal subclusters. These populations appeared to be created by the integration and batch correction process, and were derived mainly from Erythrocyte cells from both species. These subclusters were removed from downstream analysis.

Trinarization of gene expression patterns

To identify genes robustly expressed by each population (clusters and subclusters) we calculated trinarization scores for each gene per cluster and subcluster[27]. Trinarization scores represent the probability that each gene is actually expressed by each population, based on the detection frequency of each gene in each population, and the posterior distribution of the underlying population frequency of expression (□). We used a Bayesian beta-binomial model to trinarized the data and calculate the posterior distribution (□) using hyperparameter values of 1.5 for *a*, and 1 for *b* as previously described[27]*.* We called a gene expressed if P(□ > *f*) > (1 - PEP), where the fraction of cells expressing each gene (*f*) is 0.1, and the Posterior Error Probability (PEP) of 0.05. Similar results were obtained using values of 0.2 and 0.35 for *f*. Trinarization scores were calculated for all clusters and subclusters, for the zebrafish, Mexican tetra, surface-morph, and cave-morph datasets (**Supplemental Data**). These scores were also used to determine expression patterns of duplicated genes for the calculation of expression divergence (*dT*).

Marker gene identification and calculation of the transcriptional similarity index (*SI*)

To annotate subclusters and identify genes whose expression was enriched within clusters and subclusters we used Seurat to find marker genes for each population. This was done for all clusters and subclusters in both the zebrafish and Mexican tetra datasets, as well as for all of the integrated clusters and subclusters in the combined dataset. For the integrated clusters and subclusters, we used the uncorrected expression data (*DefultAssay(object)* <- *"RNA"*), which allowed the detection of species-specific gene expression patterns. For each population, we identified marker genes independently for the zebrafish and Mexican tetra cells within that cluster or subcluster. To identify shared marker genes for each population, we used Seurat's FindConservedMarkers function, which uses meta analysis of statistical values for each gene in the marker genes for each species. For all cases we used the following variables for *FindMarkers* and *FindConservedMarkers*; *logfc.threshold* of 0.25 (default), *min.pct* of 0.1 (default), *min.cells.per.ident* of 1000, and "wilcox" for *test.use*. Species-specific marker genes for each population were defined as the set difference between the marker genes for one species and conserved marker genes (Figure 2a). These lists were then used to calculate the Similarity Index (*SI*) for each cluster and subcluster between zebrafish and Mexican tetra. *SI* was calculated with the following equation, where G_T is the shared set of marker genes, and G_A and G_B are the total number of marker genes for species *A* and *B,* including both species-specific and shared marker genes[68].

$$SI = 1 - \sqrt{(1 - \frac{G_T}{G_B}) * (1 - \frac{G_T}{G_A})}$$

The same procedure was used to identify species-morph specific marker genes, and marker genes conserved between species-morphs for each population within the Mexican tetra data for calculation of the *SI* between species-morphs. To calculate *SI* between across all sets of integrated clusters and subclusters, we used the conserved marker genes for each population, and compared their p-values using the same procedure as in *Seurat's* FindConservedMarkers function - using the *minimump* function from metap package - to determine shared marker genes for each pair of cluster or subclusters. We then calculated *SI* as above. All marker gene sets were filtered to contain only those genes which also passed the trinarization threshold for that population, and for each dataset, prior to calculation of *SI.*

Paralog Identification and enrichment analysis across species

Paralogous gene pairs and orthology confidence and gene order scores were identified using the Ensembl BioMart service, and accessed using the biomaRt R package[69]. For each gene that was

specifically expressed in one species, we identified all corresponding paralogous genes and determined if any of these genes were present in the conserved marker genes, or the marker genes specific to the other species (**Figure 2a**). This was done for all clusters and subclusters shared between zebrafish and Mexican tetra. Fisher's exact test was performed to calculate statistical enrichment for paralogous genes for each cluster and subcluster using the fdrtool R package[70]. The remaining species specific genes (those that were not paralogs of a conserved, or opposite species-specific gene) were then subjected to gene ontology analysis. Species-specific marker genes for each subcluster were pooled by cluster, and the RDAVIDWebService R package was used to submit each list for GO analysis by DAVID[71].

Calculation of expression divergence (*dT*), redundancy, and gene expression expansion

For both zebrafish and Mexican tetra, we used Ensembl's Biomart tool to identify paralogous gene pairs in both species,and the last common ancestor which shares each gene duplication (**Supplemental Data**). To calculate the expression divergence for paralogous gene pairs, we calculated the following for each paralogous gene pair in each species. The subclusters which expressed each gene above the trinarization threshold were used as input for the calculation of expression divergence (*dT*) as previously described[31]. Expression divergence was calculated for both zebrafish and Mexican tetra separately by comparing the number of subclusters that express either paralog (n_{tu}) to the number of subclusters that express both paralogs (n_{ti}), with the following equation.

$$dT = \frac{(n_{tu} - n_{ti})}{n_{ui}}$$

Gene pairs where neither gene was expressed in our datasets were not included in this analysis To determine the expression pattern of the ancestral gene prior to duplication in the common ancestor of zebrafish and Mexican tetra (putative ancestral gene expression patterns), we used the intersection of the subclusters that expressed either or both paralogs in both species (**Figure 3c**) Redundancy of expression for paralogous gene pairs was calculated as 1 - *dT* within ancestral cell types, representing how much of the ancestral gene expression pattern was conserved between paralog gene pairs. The number of non-ancestral subclusters which only expressed either or both paralogs in only one species was used to determine the amount of expansion of paralogous gene expression patterns. Importantly, these metrics cannot account for the possibility that ancestral expression in one or more cell types was lost for both paralogs in one species, but retained for at least one paralog in the other. Paralogous genes generated by the vertebrate 2R or teleost 3R whole genome duplication events (Ohnologs) were identified from the OHNOLOG repository using the "Strict 2R" and "Strict 3R" datasets for zebrafish (*D. rerio*) (https://ohnologs.curie.fr)[72].

Identification and analysis of putative gene regulatory networks using SCENIC/GENIE3

We used SCENIC/GENIE3 to identify transcription factors (TFs) that were predictive of the expression of terminal effector genes associated with the functions of the hypothalamus, including neuropeptides, neurotransmitter or synapse associated genes, and ion channel genes33,34. This analysis outputs numerical weights for the association between each TF and each terminal effector gene in the two species, which are used to determine TF sets for each terminal effector gene. We downsampled the zebrafish and Mexican tetra datasets to ensure equal cell numbers across subclusters. This analysis uses single-cell information, but is independent of cell cluster and subcluster identities. Lists of transcription factors, neuropeptides, neurotransmitter related, synaptic, and ion channel genes were identified using ZebrafishMine, and used to identify orthologous genes in Mexican tetra with the same ontology[73]. Gene lists used in this study are available at (https://github.com/maxshafer/Cavefish_Paper). For the list of transcription factors, we used search terms "transcription" and "transcription factor activity", which resulted in a combined list of 3141 unique gene names. Datasets were first downsampled such that

the same number of cells from each subcluster were included to reduce the effects of differential subcluster abundances between species on the GRN analysis. Cutoff values for *minSamples* and *minCountsPerGene* for the *geneFiltering* function used to filter out lowly detected or expressed genes were determined such that *hcrt* was included in all analyses. SCENIC then uses the GENIE3 algorithm to generate random forest weights for each transcription factor and target gene, based on the predictive power of each transcription factor in determining the expression level for each target gene. The lists of transcription factors and their corresponding weights for each target gene were used in downstream analysis. We used the "top50" cutoff from SCENIC to determine transcription factors to calculate the *SI* between species or species-morphs. The same procedure was used for analysis of Mexican tetra surface- and cave-morphs. We used customized versions of some SCENIC functions, including *geneFiltering*, *runGenie3*, and *runSCENIC_1_coexNetwork2modules*, to allow use of our gene lists, and to allow easier implementation on a laptop (specifically the ability to stop and restart the analysis).

To test whether divergence in all TF sets were mitigated by association with paralogous TFs, we calculated paralog enrichment in the species-specific TFs for each terminal effector gene as done previously (Figure 2). The majority of TF sets (260 out of 435, 60%) were composed of roughly the number of paralogs expected by random chance (odds ratio ~ 1), and 26 terminal effector genes had significantly fewer paralogs than expected (Figure 4d). Therefore, divergence in the gene regulatory networks of neuropeptides and other terminal effector classes is not compensated through the expression of paralogous TFs.

Analysis of species-specific cell types and identification of species-specific genes

Clusters which were specific to either a species or a species-morph were identified by calculating the proportion of cells which came from each species or species-morph for each cluster or subcluster. Clusters which were composed of 90% or more cells from one species or species-morph were considered specific to that species. Cells belonging to other species or species-morphs in those clusters were likely incorrectly assigned due to the limited gene information used for integrated clustering and subclustering analysis. These subclusters were not enriched for technical artifacts due to the encapsulation or clustering methods used (Supplementary Figure 7). We examined the genes that differentiated each species-specific subcluster from the other cells in the same parent cluster for clues to their origins or functions (Figure 5). Two of these subclusters were distinguished by the expression of unique TFs (Neuronal_03_13 and Neuronal_04_5). For example, the zebrafish-specific Neuronal_04_5 subcluster expressed meis1b, six6a, and six6b in addition to the parent cluster marker genes cbln1 and adcyap1b. Other species-specific neuronal subclusters were characterized by cell cycle genes (Neuronal_02_3), genes related to axonal guidance or remodelling (Neuronal_04_6), or expressed different neurotransmitters (Neuronal_05_0 and Neuronal_05_1) (Figure 5). These subclusters may therefore reflect different temporal or spatial cell states, captured in one species or the other. The presence, absence, and orthology of the specific duplicated genes discussed in the current report, including *vipb*, *HPCAL1*, and the jacalin lectin genes, was confirmed using the most recent Ensembl release (Ensembl Release 101), which includes newer versions of both the zebrafish (GRCz11) and Mexican tetra (Astyanax_mexicanus-2.0) genome assemblies[34]. For analysis of subclustering in the absence of non-homologous genes, all non-homologous genes were removed from the Variable Features for each cluster prior to subclustering using the same parameters as above.

Construction of Sankey diagrams and other plots

Sankey diagrams were constructed using the *networkD3* R package. We used the Seurat wrappers for *ggplot2* functions to construct tSNE graphs and DotPlots of expression values across clusters or subclusters. Custom R scripts were used to construct the rest of the plots using *ggplot2*, including the gene ontology analysis, and the multi-layered circular plots (**Figure 6a**). Graphical tables

were constructed using the formatabble R package. All scripts used to construct figures are available on GitHub (https://github.com/maxshafer/Cavefish_Paper). Final figures were assembled using Affinity Designer (Serif Europe).

Species and species-morph dendrograms, as well as subcluster dendrograms were constructed using both the *Seurat*, *ggtree*, and *phylogram* R packages[24,77,78]. Pseudo-bulk expression data for each cluster and subcluster were used to calculate the dendrogram dissimilarity values for **Figure 6a**. For calculation of the similarity between species and species-morph, pseudo-bulk expression data was generated for zebrafish, Choy surface, and Pachon, Tinaja, and Molino cave-morph samples by averaging the expression of each gene across all cells within each cluster and subcluster. Species and species-morph dendrograms were then generated for each population, based on the similarity in whole transcriptomes using the *BuildClusterTree* function in Seurat. The ggtree package was used to construct the density dendrogram, where the colour of the edges corresponds to the number of subclusters which support each arrangement of the dendrogram. The distance between surface- and cave-morph versions of each subcluster on the dendrogram was used for plotting.

DATA AVAILABILITY

Processed single-cell RNAseq counts and metadata, marker gene lists, trinarized gene lists, SI results, SCENIC results, results from genetic analysis, and GO lists are available as supplementary data. Raw sequencing results are available at the Sequence Read Archive (SRA) under BioProject ID PRJNA754013.

ACKNOWLEDGEMENTS

We thank Clifford J. Tabin for providing *Astyanax mexicanus* samples, and advice on experimental design. We thank members of the Schier lab for discussion and advice, including Drs. B. Raj, J. Liu, P. Abitua, and A. Nichols, and the Harvard zebrafish and cavefish facilities staff, including Brian Martineau, for technical support. We thank Drs. Gray Camp, Walter Salzburger and Nathalie Jurisch-Yaksi for helpful comments on the manuscript. This work was supported by a postdoctoral fellowship from the Canadian Institutes of Health Research (CIHR) to M.E.R.S., a grant from the Swiss National Science Foundation (SNSF) to M.E.R.S. (SPARK 196313), grants from SNSF (SPARK 195955) and the University of Basel to A.N.S., and an NIH grant (DP1HD094764), an ERC Advanced grant (834788), an Allen Discovery Center grant, and a McKnight Foundation Technological Innovations in Neuroscience Award to A.F.S..

AUTHOR CONTRIBUTIONS STATEMENT

M.E.R.S. and A.F.S. conceived and designed the study. A.N.S. and M.E.R.S conceived and performed Similarity Index analysis. M.E.R.S. performed all other experiments and analysis, including scRNA-seq experiments, and all bioinformatic analysis. M.E.R.S, A.N.S., and A.F.S. wrote the manuscript. All authors read and approved of the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1. Integration of zebrafish and Mexican tetra single-cell data reveals extensive conservation of cell types

(**a**) UMAP reduction of integrated zebrafish and Mexican tetra cells coloured by species. Datasets were integrated with Mutual Nearest Neighbour (MNN) and Canonical Correlation Analysis (CCA) using Seurat. (**b**) UMAP reduction of integrated zebrafish and Mexican tetra cells coloured by annotated cell type. (**c**)

Sankey diagram of relationships between zebrafish, integrated, and Mexican tetra annotated clusters from Figure S3. Heights of squares and thickness of connecting lines are relative to the number of cells per identity or connection, respectively. (**d**) Circular heatmap of the proportion of zebrafish (dark blue), or Mexican tetra (yellow) cells per integrated subcluster. Subclusters are grouped first by cluster, and clusters are arranged by the dendrogram of cluster similarity, shown in the center of the circular heatmap. Red outlines indicate subclusters with > 90% of cells from one species (species-specific). (**e**) Density dendrogram for all shared subclusters across species and species-morphs. The density dendrogram was constructed using dendrograms for the similarity between species and species-morph versions of each subcluster identity shared between zebrafish and Mexican tetra. Darkness of lines indicate the level of support for each branch.

Figure 2. Shared subclusters are highly similar between species and express paralogous genes

(**a**) Diagram illustrating the relationships between gene sets for each subcluster used in the current study for calculation of the Similarity Index (*SI*). G_A and G_B represent the total marker gene sets in the two species examined. G_T are genes that are found in both G_A and G_B. Species-specific marker genes are those which do not overlap with the other species (G_A or G_B minus G_T), though may be paralogs of a marker gene from the other species. (**b**) The row-scaled *SI* for all subclusters between zebrafish and Mexican tetra based on marker genes filtered for genes which pass the trinarization threshold. Yellow indicates the highest SI value among Mexican tetra subclusters for each Zebrafish subcluster. (**c**) The Similarity Index (*SI*) for progenitor and differentiated neuronal subclusters between zebrafish and Mexican tetra based on marker genes filtered for genes which pass the trinarization threshold. Two sample t-test p-value = 0.007661. (**d**) The *SI* for clusters and the mean of the *SI* for subclusters grouped by cluster between zebrafish and Mexican tetra coloured by cluster based on marker genes filtered for genes which pass the trinarization threshold. Paired t-test p-value = 0.003012. (**e**) Comparison of the SI for the same subclusters between species (purple), and between species-morphs (yellow), calculated using marker gene sets. (**f**) The percentage of species-specific marker genes for each subcluster which were paralogs of either the conserved or opposite species-specific marker gene for zebrafish and Mexican tetra. (**g**) The odds ratio for the enrichment of paralogs in the species-specific genes for each subcluster for zebrafish and Mexican tetra. (**h**) The percentage of paralog shifting events shared by both surface- and cave-morphs of Mexican tetra. Paralog shifts are separated by whether they were from the zebrafish or Mexican tetra species-specific marker genes. (**i**) Relationship between the Similarity Index and the mean of the percentage paralogs (for zebrafish and Mexican tetra) for each subcluster. (**j**) Comparison of the *SI* (blue) and corrected-*SI* (yellow) for subclusters between zebrafish and Mexican tetra. For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5 times the interquartile range, thicker line represents the median.

Figure 3. Paralog shifts are due to differential divergence after duplication between species

(**a**) Example gene tree for a gene that is duplicated once in the common ancestor of two extant species (Species 1 and Species 2). (**b**) Model for paralog gene expression pattern divergence after gene duplication in a common ancestor. Over time (**c**), expression of Gene Xa is retained in the cell types of interest (filled rectangles) in Species 1 and Gene Xb is lost (empty rectangles), whereas expression of only Gene Xb is retained in the same cell types in Species 2. Ancestral cell types are defined as any cell type that expresses either paralog in both species (yellow highlight). Gene expression pattern expansion is represented by cell types that express only 1 paralog in 1 species (green highlight). (**d**) Empirical cumulative distribution function (ECDF) for expression divergence (*dT*) for gene pairs grouped by their last common ancestor in zebrafish. From the oldest (Opisthokonta, yellow), to the most recent common ancestor (Otophysi, red), and to those gene duplicates which are only found in Danio rerio (dark red). (**e**) ECDF for expression divergence (*dT*) for gene pairs grouped by their last common ancestor in Mexican tetra. From the oldest (Opisthokonta, yellow), to the most recent common ancestor (Otophysi, red), and to

those gene duplicates which are only found in Astyanax mexicanus (dark red). (**f**) Relationship between the expression divergence (*dT*) for gene pairs in zebrafish and the expression divergence (*dT*) for gene pairs in Mexican tetra for all gene pairs (black dots). (**g**) Relationship between the expression divergence (*dT*) for gene pairs which arose in the last common ancestor of zebrafish and Mexican tetra (Otophysi, pink). (**h**) DotPlot of the expression of the paralog pairs *etv5a* / *etv5b* in zebrafish (top) and Mexican tetra (bottom) across cell clusters. (**i**) Ridge plots of the Pearson correlation of the binarized expression patterns across subclusters for gene pairs shared by zebrafish and Mexican tetra grouped by their last common ancestor or by their originating WGD event.

Figure 4. Divergence of gene regulatory networks underlying neuronal genes

(**a**) Random forest weight for orthologous transcription factors in the zebrafish (y-axis) and Mexican tetra (x-axis) data for the neuropeptide *vip.* Colours indicate whether those transcription factors are in the top 2% of transcription factors for each gene in either zebrafish (blue) and Mexican tetra (red), both (yellow), or none (black). (**b**) Similarity Index between the transcription factor sets for zebrafish and Mexican tetra for neuropeptides, neurotransmitters, synaptic genes, and ion channels. (**c**) Correlation between the random forest weights for transcription factor sets associated with each for neuropeptide, neurotransmitter, synaptic, or ion channel genes between zebrafish and Mexican tetra. (**d**) Odds ratio from Fisher's exact test for the enrichment for paralogous genes in the transcription factors associated with each gene, red dots indicated significant enrichment. (**e**) Similarity Index for all subclusters shared between zebrafish and Mexican tetra using either only neuropeptides and neurotransmitter related genes (purple), only transcription factors (green), or all marker genes (yellow). (**f**) Similarity Index for individual neuropeptidergic GABA_1 subclusters between zebrafish and Mexican tetra using either only neuropeptides and neurotransmitter related genes (purple), only transcription factors (yellow), or all marker genes. For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5 times the interquartile range, thicker line represents the median.

Figure 5. Species-specific subclusters are associated with species-specific genes

(**a**) Sankey diagram of shared and species-specific subclusters, indicating the species (zebrafish or Mexican tetra) and the cellular lineage they belong to (Ciliated, Glial, Hematopoietic, or Neuronal). (**b**) The percentage of expressed genes (counts > 10) in each zebrafish subcluster which are non-homologous genes between species in shared or species specific cell subclusters. (**c**) The percentage of expressed genes (counts > 10) in each Mexican tetra subcluster which are non-homologous genes between species in shared or species specific cell subclusters. (**d**) The percentage of expressed genes (counts > 10) in each zebrafish subcluster which are non-homologous genes between species in neuronal or non-neuronal subclusters. (**e**) The percentage of expressed genes (counts > 10) in each Mexican tetra subcluster which are non-homologous genes between species in neuronal or non-neuronal subclusters. (**c**) DotPlot of the species-specific subcluster marker genes (y-axis) across subclusters (x-axis). Blue boxes highlight expression of specific paralogous genes in different subclusters. Gene expression is quantified by both the percentage of cells which express each gene (dot size) and the average expression in those cells (colour scale). For all boxplots, box bounds represent the first and third quartiles and whiskers 1.5 times the interquartile range, thicker line represents the median.

Figure 6. Divergence in subcluster repertoires and transcriptomes across Pachon, Tinaja, and Molino cave-morphs

(**a**) Graphical summary of subcluster and transcriptional differences between Mexican tetra surface- and cave-morphs, and between Pachon, Tinaja, and Molino cave-morphs. The first layer indicates the cluster identity (from Figure S2b), and the text label indicates the subcluster (from Figure S8). The second layer indicates the proportion of cells in each subcluster that come from a surface-morph (green) or a cave-morph (yellow). Red outlines indicate morph-specific subclusters (> 90% of cells from either surface- or

cave-morphs). Third layer indicates the proportion of cave-morph cells from each subcluster that come from the Pachon (orange), Tinaja (blue), or Molino (green) cave-morph samples. The fourth layer displays the Similarity Index between the surface-morph, cave-morph for shared marker genes for each subcluster. The fifth layer displays the percentage of marker genes for each subcluster that is also associated with a divergent genomic window (F_{ST} genes). Finally, the sixth layer displays the Dendrogram Distance, which is the distance between the surface- and cave-morph versions of each subcluster on a dendrogram based on the subcluster transcriptomes. (**b**) Dendrogram of the relationships between species and species-morphs in this study coloured by the number of cell type changes normalized by the evolutionary time determined from[18].

REFERENCES

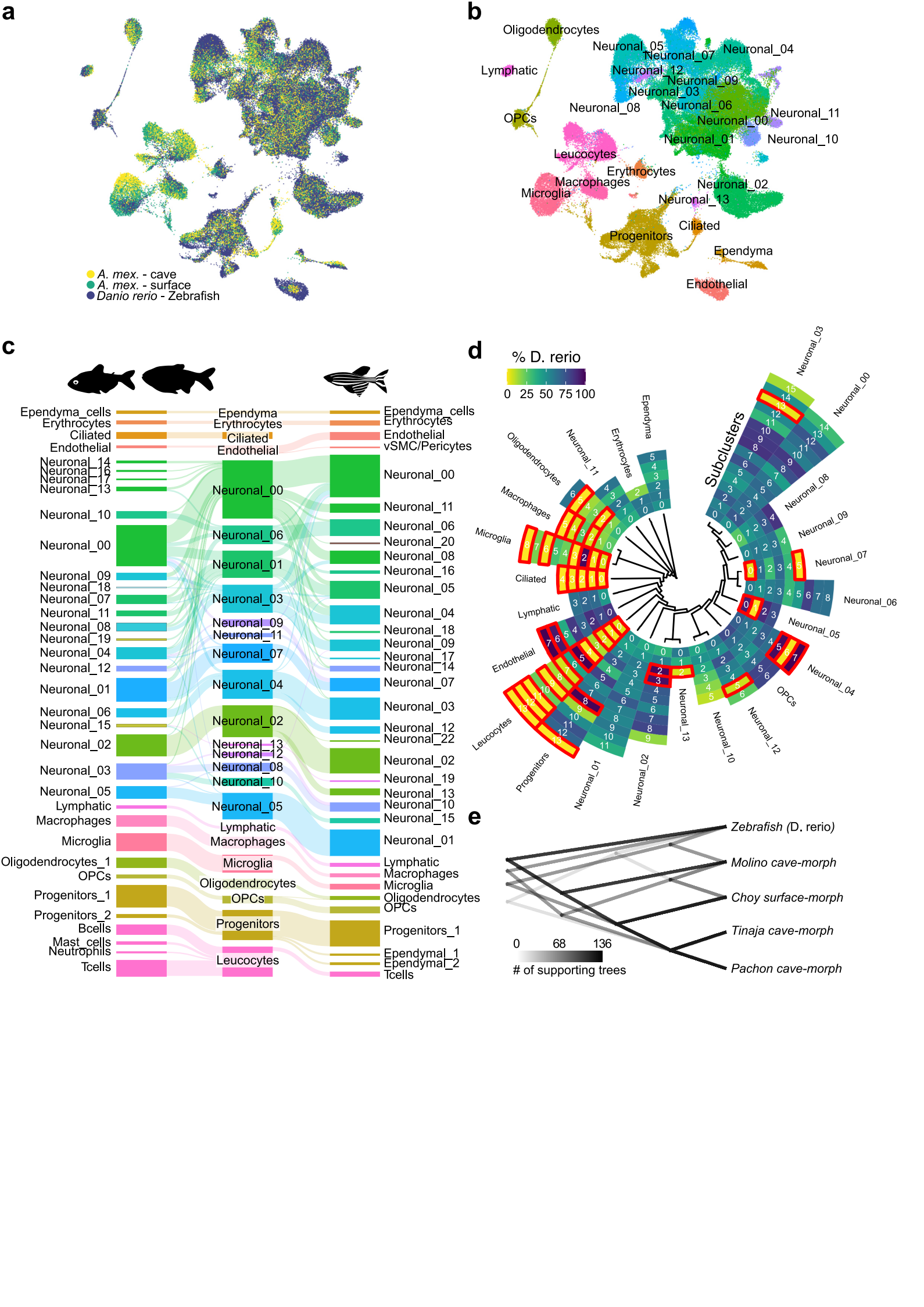
1. S. Ramón y Cajal. Histology of the human nervous system & vertebrates. *Maloine Paris* (1911).
2. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18**, 530–546 (2017).
3. Arendt, D., Bertucci, P. Y., Achim, K. & Musser, J. M. Evolution of neuronal types and families. *Curr. Opin. Neurobiol.* **56**, 144–152 (2019).
4. Ohno, S. *Evolution by Gene Duplication*. (Springer-Verlag, 1970). doi:10.1007/978-3-642-86659-3.
5. Shafer, M. E. R. Cross-Species Analysis of Single-Cell Transcriptomic Data. *Front. Cell Dev. Biol.* **7**, (2019).
6. Tosches, M. A. *et al.* Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881–888 (2018).
7. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
8. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e19 (2016).
9. Pollen, A. A. *et al.* Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell* **176**, 743-756.e17 (2019).
10. Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574**, 418–422 (2019).
11. Krienen, F. M. *et al.* Innovations present in the primate interneuron repertoire. *Nature* 1–8 (2020) doi:10.1038/s41586-020-2781-z.
12. Pasquier, J. *et al.* Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. *BMC Genomics* **17**, 368 (2016).
13. Ravi, V. & Venkatesh, B. The Divergent Genomes of Teleosts. *Annu. Rev. Anim. Biosci.* **6**, 47–68 (2018).
14. Voldoire, E., Brunet, F., Naville, M., Volff, J.-N. & Galiana, D. Expansion by whole genome duplication and evolution of the sox gene family in teleost fish. *PLOS ONE* **12**, e0180936 (2017).
15. Gross, J. B. The complex origin of Astyanax cavefish. *BMC Evol. Biol.* **12**, 105 (2012).
16. Keene, A., Yoshizawa, M. & McGaugh, S. E. *Biology and Evolution of the Mexican Cavefish*. (Academic Press, 2015).
17. Mitchell, R. W., Russell, W. H. & Elliott, W. R. *Mexican Eyeless Characin Fishes, Genus Astyanax: Environment, Distribution, and Evolution*. (Texas Tech Press, 1977).
18. Herman, A. *et al.* The role of gene flow in rapid and repeated evolution of cave-related traits in Mexican tetra, Astyanax mexicanus. *Mol. Ecol.* **27**, 4397–4416 (2018).
19. Prober, D. A., Rihel, J., Onah, A. A., Sung, R.-J. & Schier, A. F. Hypocretin/Orexin Overexpression Induces An Insomnia-Like Phenotype in Zebrafish. *J. Neurosci.* **26**, 13400–13410 (2006).

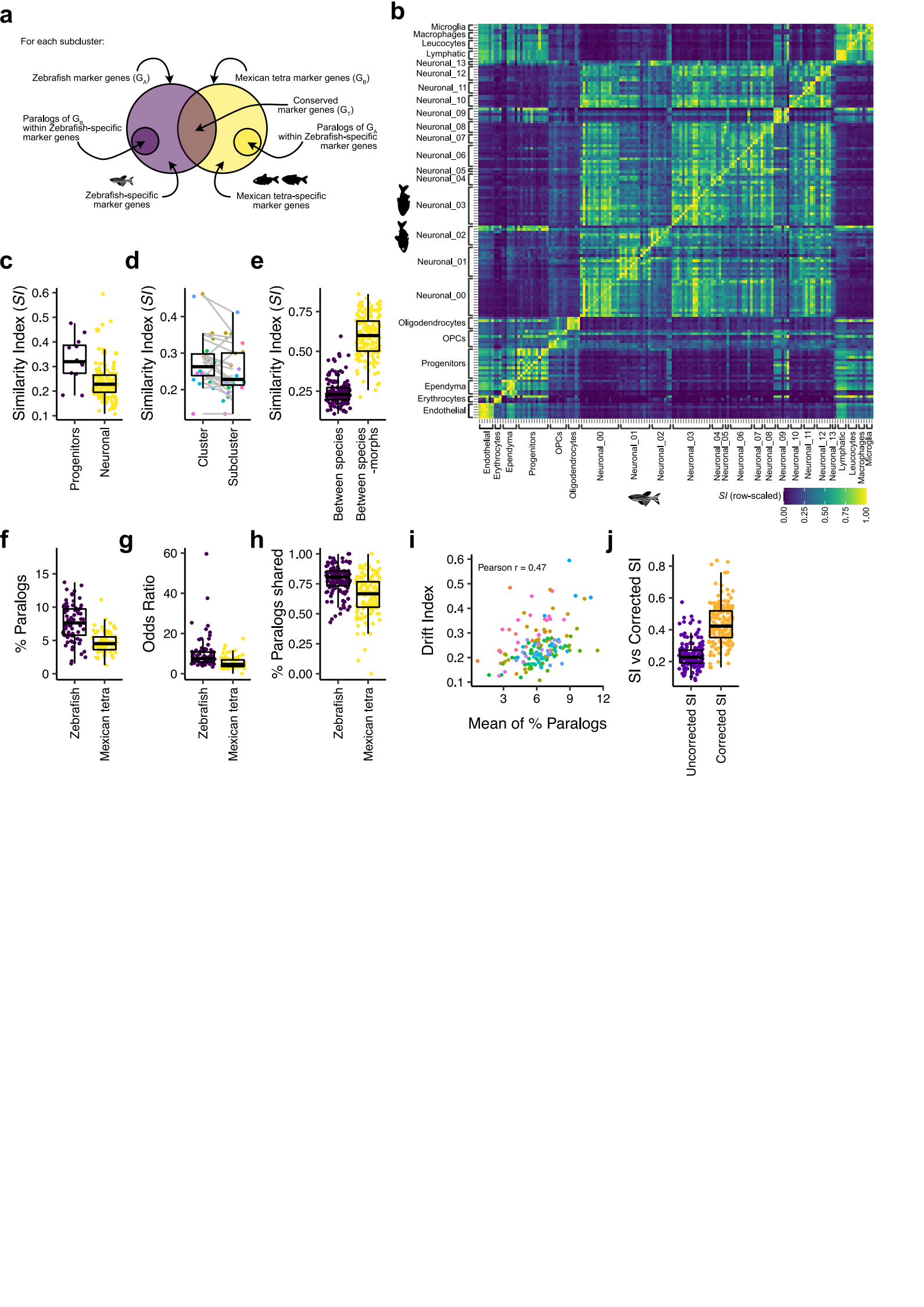
20. Richter, C., Woods, I. G. & Schier, A. F. Neuropeptidergic Control of Sleep and Wakefulness. *Annu. Rev. Neurosci.* **37**, 503–531 (2014).
21. Xie, Y. & Dorsky, R. I. Development of the hypothalamus: conservation, modification and innovation. *Development* **144**, 1588–1599 (2017).
22. Denes, A. S. *et al.* Molecular Architecture of Annelid Nerve Cord Supports Common Origin of Nervous System Centralization in Bilateria. *Cell* **129**, 277–288 (2007).
23. Raj, B. *et al.* Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).
24. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
25. D'Gama, P. P. *et al.* Diversity and Function of Motile Ciliated Cell Types within Ependymal Lineages of the Zebrafish Brain. *bioRxiv* 2021.02.17.431442 (2021) doi:10.1101/2021.02.17.431442.
26. Liu, J. *et al.* Evolutionarily conserved regulation of hypocretin neuron specification by Lhx9. *Development* **142**, 1113–1124 (2015).
27. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).
28. Liang, C., Musser, J. M., Cloutier, A., Prum, R. O. & Wagner, G. P. Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes. *Genome Biol. Evol.* **10**, 538–552 (2018).
29. Gu, X. Understanding tissue expression evolution: from expression phylogeny to phylogenetic network. *Brief. Bioinform.* **17**, 249–254 (2016).
30. Dunn, C. W., Zapata, F., Munro, C., Siebert, S. & Hejnol, A. Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proc. Natl. Acad. Sci.* **115**, E409–E417 (2018).
31. Farré, D. & Albà, M. M. Heterogeneous Patterns of Gene-Expression Diversification in Mammalian Gene Duplicates. *Mol. Biol. Evol.* **27**, 325–335 (2010).
32. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods* **17**, 147–154 (2020).
33. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
34. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
35. Santos, M. E., Bouquin, A. L., Crumière, A. J. J. & Khila, A. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. *Science* **358**, 386–390 (2017).
36. Cao, J. & Lv, Y. Evolutionary analysis of the jacalin-related lectin family genes in 11 fishes. *Fish Shellfish Immunol.* **56**, 543–553 (2016).
37. Burgoyne, R. D. Neuronal calcium sensor proteins: generating diversity in neuronal Ca 2+ signalling. *Nat. Rev. Neurosci.* **8**, 182–193 (2007).
38. Jaggard, J. B. *et al.* Hypocretin underlies the evolution of sleep loss in the Mexican cavefish. *eLife* **7**, e32637 (2018).

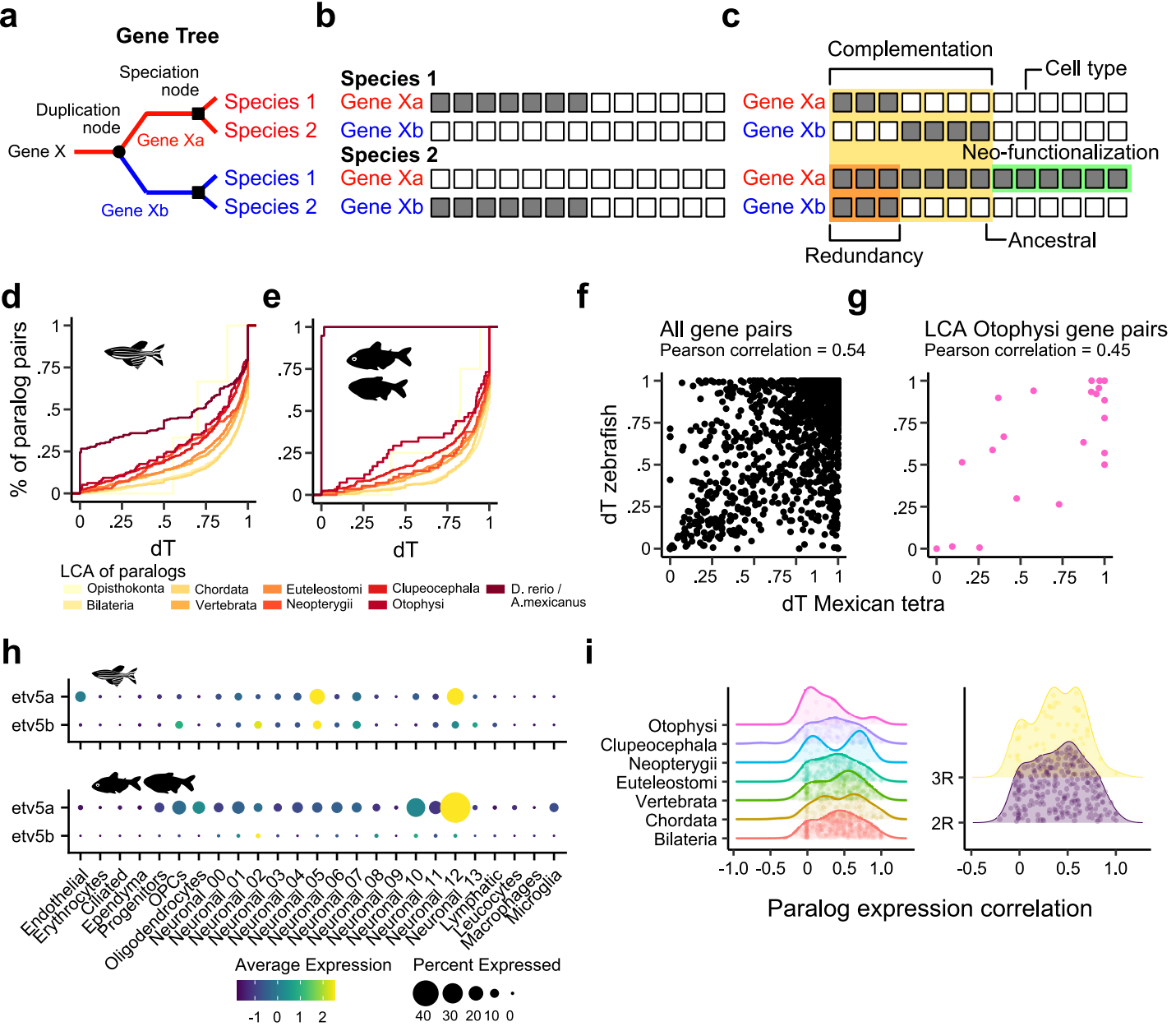
39. Alié, A. *et al.* Developmental evolution of the forebrain in cavefish, from natural variations in neuropeptides to behavior. *eLife* **7**, e32808 (2018).
40. Peuß, R. *et al.* Adaptation to low parasite abundance affects immune investment and immunopathological responses of cavefish. *Nat. Ecol. Evol.* **4**, 1416–1430 (2020).
41. Eisenberg, T. *et al.* Induction of autophagy by spermidine promotes longevity. *Nat. Cell Biol.* **11**, 1305–1314 (2009).
42. Duffy, C. M., Xu, H., Nixon, J. P., Bernlohr, D. A. & Butterick, T. A. Identification of a fatty acid binding protein4-UCP2 axis regulating microglial mediated neuroinflammation. *Mol. Cell. Neurosci.* **80**, 52–57 (2017).
43. Zhang, H. *et al.* Polyamines Control eIF5A Hypusination, TFEB Translation, and Autophagy to Reverse B Cell Senescence. *Mol. Cell* **76**, 110-125.e9 (2019).
44. Rodrigues, F. T. S. *et al.* Major depression model induced by repeated and intermittent lipopolysaccharide administration: Long-lasting behavioral, neuroimmune and neuroprogressive alterations. *J. Psychiatr. Res.* **107**, 57–67 (2018).
45. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015-1030.e16 (2018).
46. Hughes, L. C. *et al.* Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 6249–6254 (2018).
47. Nakatani, M., Miya, M., Mabuchi, K., Saitoh, K. & Nishida, M. Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. *BMC Evol. Biol.* **11**, 177 (2011).
48. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).
49. Kondrashov, F. A. & Koonin, E. V. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* **20**, 287–290 (2004).
50. Thomson, G. J. *et al.* Metabolism-induced oxidative stress and DNA damage selectively trigger genome instability in polyploid fungal cells. *EMBO J.* **38**, e101597 (2019).
51. Gillard, G. B. *et al.* Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. *Genome Biol.* **22**, 103 (2021).
52. Konstantinides, N. *et al.* Phenotypic Convergence: Distinct Transcription Factors Regulate Common Terminal Features. *Cell* **174**, 622-635.e13 (2018).
53. True, J. R. & Haag, E. S. Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.* **3**, 109–119 (2001).
54. Hilgers, L., Hartmann, S., Hofreiter, M. & von Rintelen, T. Novel Genes, Ancient Genes, and Gene Co-Option Contributed to the Genetic Basis of the Radula, a Molluscan Innovation. *Mol. Biol. Evol.* **35**, 1638–1652 (2018).
55. Florio, M. *et al.* Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of

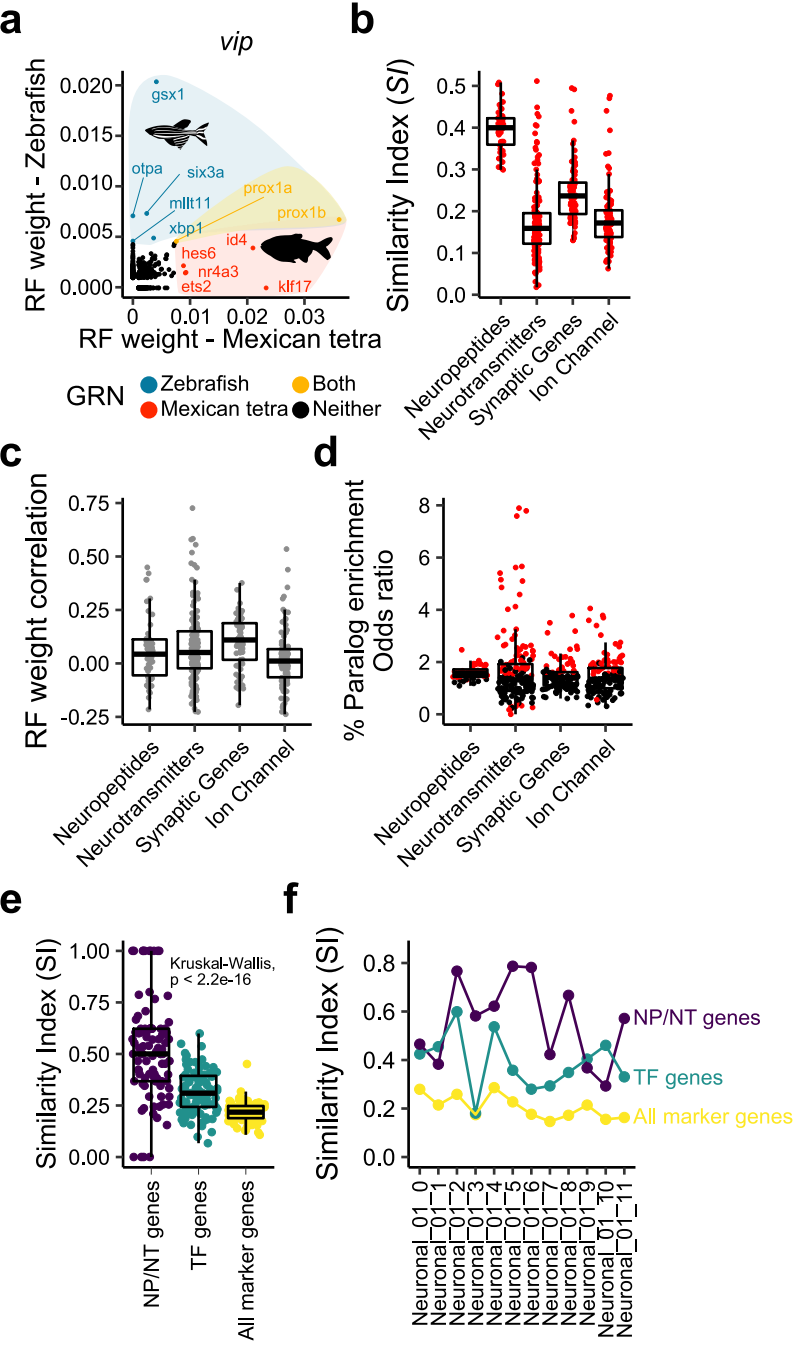
969 fetal neocortex. *eLife* **7**, e32332 (2018).
970 56. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
971 57. Viets, K., Eldred, K. C. & Johnston, R. J. Mechanisms of Photoreceptor Patterning in Vertebrates and Invertebrates.
972 *Trends Genet.* **32**, 638–659 (2016).
973 58. Bowmaker, J. K. Evolution of vertebrate visual pigments. *Vision Res.* **48**, 2022–2041 (2008).
974 59. Chitnis, T. & Weiner, H. L. CNS inflammation and neurodegeneration. *J. Clin. Invest.* **127**, 3577–3587 (2017).
975 60. Riddle, M. R. *et al.* Insulin resistance in cavefish as an adaptation to a nutrient-limited environment. *Nature* **555**, 647–
976 651 (2018).
977 61. Elipot, Y. *et al.* A mutation in the enzyme monoamine oxidase explains part of the Astyanax cavefish behavioural
978 syndrome. *Nat. Commun.* **5**, 3647 (2014).
979 62. Yoshizawa, M. *et al.* Distinct genetic architecture underlies the emergence of sleep loss and prey-seeking behavior in
980 the Mexican cavefish. *BMC Biol.* **13**, 15 (2015).
981 63. Fischer, E. K. & O'Connell, L. A. Modification of feeding circuits in the evolution of social behavior. *J. Exp. Biol.* **220**,
982 92–102 (2017).
983 64. Kroeger, D. *et al.* Galanin neurons in the ventrolateral preoptic area promote sleep and heat loss in mice. *Nat.*
984 *Commun.* **9**, 4129 (2018).
985 65. Yamashita, J. *et al.* Male-predominant galanin mediates androgen-dependent aggressive chases in medaka. *eLife* **9**,.
986 66. Wee, C. L. *et al.* Social isolation modulates appetite and defensive behavior via a common oxytocinergic circuit in larval
987 zebrafish. *bioRxiv* 2020.02.19.956854 (2020) doi:10.1101/2020.02.19.956854.
988 67. Elipot, Y., Legendre, L., Père, S., Sohm, F. & Rétaux, S. Astyanax Transgenesis and Husbandry: How Cavefish Enters
989 the Laboratory. *Zebrafish* **11**, 291–299 (2014).
990 68. Sawh, A. N. *et al.* Lamina-Dependent Stretching and Unconventional Chromosome Compartments in Early C. elegans
991 Embryos. *Mol. Cell* **78**, 96-111.e6 (2020).
992 69. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the
993 R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
994 70. Strimmer, K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*
995 **24**, 1461–1462 (2008).
996 71. Fresno, C. & Fernández, E. A. RDAVIDWebService: a versatile R interface to DAVID. *Bioinformatics* **29**, 2810–2811
997 (2013).
998 72. Singh, P. P. & Isambert, H. OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome
999 duplication in vertebrates. *Nucleic Acids Res.* **48**, D724–D730 (2020).
1000 73. Smith, R. N. *et al.* InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous
1001 biological data. *Bioinformatics* **28**, 3163–3165 (2012).

74. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
75. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
76. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
77. Wilkinson, S. P. & Davy, S. K. phylogram: an R package for phylogenetic analysis with nested lists. *J. Open Source Softw.* **3**, 790 (2018).
78. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* **69**, e96 (2020).

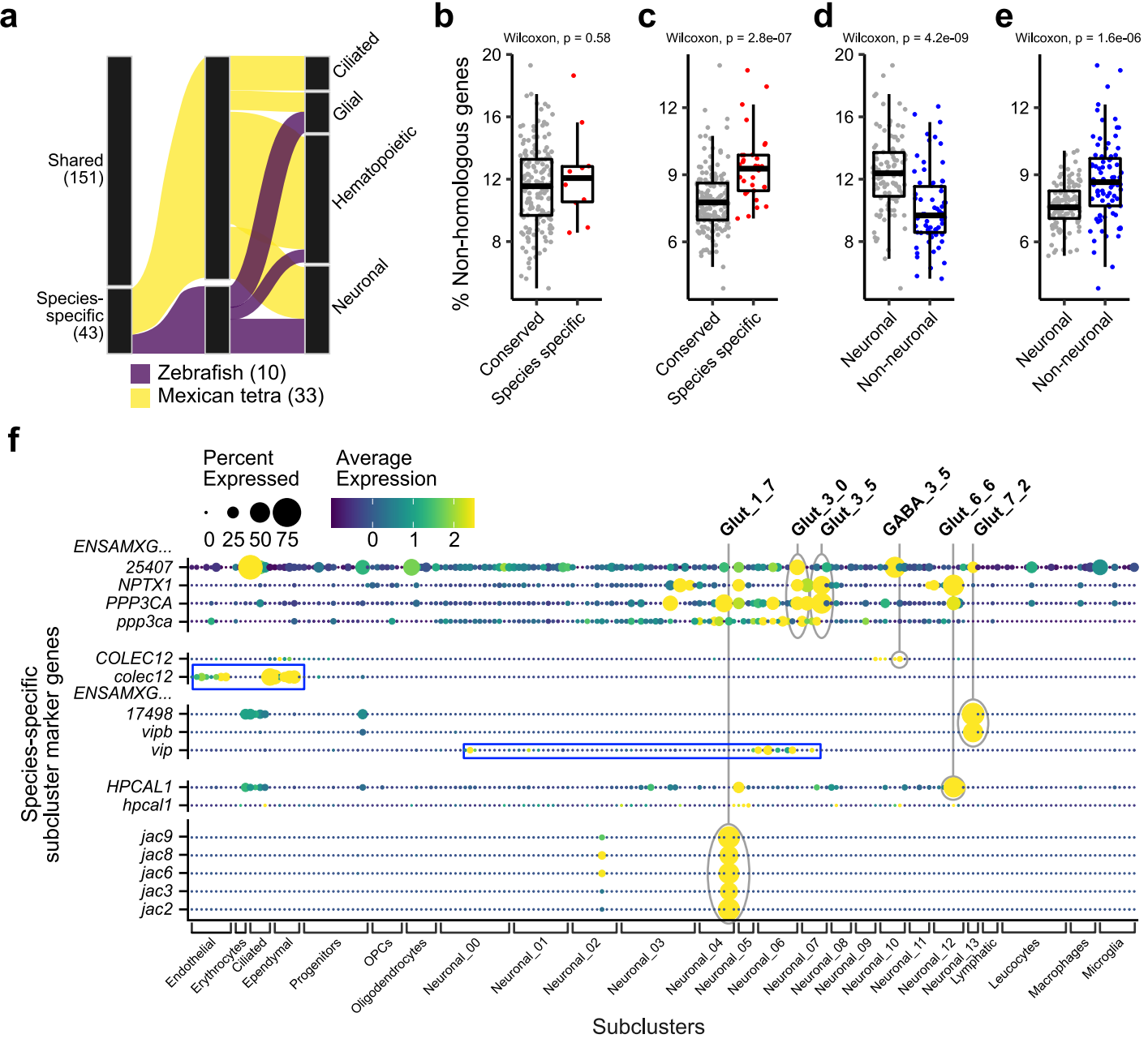

a
A. mex. - cave
A. mex. - surface
Danio rerio - Zebrafish
b
Oligodendrocytes
Lymphatic
OPCs
Neuronal_05
Neuronal_07
Neuronal_04
Neuronal_12
Neuronal_09
Neuronal_03
Neuronal_08
Neuronal_06
Neuronal_11
Neuronal_00
Neuronal_01
Neuronal_10
Leucocytes
Erythrocytes
Macrophages
Microglia
Neuronal_02
Neuronal_13
Ciliated
Progenitors
Ependyma
Endothelial
c
Ependyma_cells
Erythrocytes
Ciliated
Endothelial
Neuronal_14
Neuronal_16
Neuronal_17
Neuronal_13
Neuronal_10
Neuronal_00
Neuronal_09
Neuronal_18
Neuronal_07
Neuronal_11
Neuronal_08
Neuronal_19
Neuronal_04
Neuronal_12
Neuronal_01
Neuronal_06
Neuronal_15
Neuronal_02
Neuronal_03
Neuronal_05
Lymphatic
Macrophages
Microglia
Oligodendrocytes_1
OPCs
Progenitors_1
Progenitors_2
Bcells
Mast_cells
Neutrophils
Tcells
Ependyma
Erythrocytes
Ciliated
Endothelial
Neuronal_00
Neuronal_06
Neuronal_01
Neuronal_03
Neuronal_09
Neuronal_11
Neuronal_07
Neuronal_04
Neuronal_02
Neuronal_13
Neuronal_12
Neuronal_08
Neuronal_10
Neuronal_05
Lymphatic
Macrophages
Microglia
Oligodendrocytes
OPCs
Progenitors
Leucocytes
Ependyma_cells
Erythrocytes
Endothelial
vSMC/Pericytes
Neuronal_00
Neuronal_11
Neuronal_06
Neuronal_20
Neuronal_08
Neuronal_16
Neuronal_05
Neuronal_04
Neuronal_18
Neuronal_09
Neuronal_17
Neuronal_14
Neuronal_07
Neuronal_03
Neuronal_12
Neuronal_22
Neuronal_02
Neuronal_19
Neuronal_13
Neuronal_10
Neuronal_15
Neuronal_01
Lymphatic
Macrophages
Microglia
Oligodendrocytes
OPCs
Progenitors_1
Ependymal_1
Ependymal_2
Tcells
d
% D. rerio
0 25 50 75 100
Subclusters
Ependyma
Erythrocytes
Neuronal_11
Oligodendrocytes
Macrophages
Microglia
Ciliated
Lymphatic
Endothelial
Leucocytes
Progenitors
Neuronal_01
Neuronal_02
Neuronal_13
Neuronal_10
Neuronal_12
OPCs
Neuronal_04
Neuronal_05
Neuronal_06
Neuronal_07
Neuronal_09
Neuronal_08
Neuronal_00
Neuronal_03
e
Zebrafish (D. rerio)
Molino cave-morph
Choy surface-morph
Tinaja cave-morph
Pachon cave-morph
0 68 136
# of supporting trees


a
For each subcluster:
Zebrafish marker genes (GA)
Mexican tetra marker genes (GB)
Conserved marker genes (GT)
Paralogs of GB within Zebrafish-specific marker genes
Paralogs of GA within Zebrafish-specific marker genes
Zebrafish-specific marker genes
Mexican tetra-specific marker genes
b
Microglia
Macrophages
Leucocytes
Lymphatic
Neuronal_13
Neuronal_12
Neuronal_11
Neuronal_10
Neuronal_09
Neuronal_08
Neuronal_07
Neuronal_06
Neuronal_05
Neuronal_04
Neuronal_03
Neuronal_02
Neuronal_01
Neuronal_00
Oligodendrocytes
OPCs
Progenitors
Ependyma
Erythrocytes
Endothelial
SI (row-scaled)
0.00
0.25
0.50
0.75
1.00
c
Similarity Index (SI)
Progenitors
Neuronal
d
Similarity Index (SI)
Cluster
Subcluster
e
Similarity Index (SI)
Between species
Between species -morphs
f
% Paralogs
Zebrafish
Mexican tetra
g
Odds Ratio
Zebrafish
Mexican tetra
h
% Paralogs shared
Zebrafish
Mexican tetra
i
Drift Index
Pearson r = 0.47
Mean of % Paralogs
j
SI vs Corrected SI
Uncorrected SI
Corrected SI

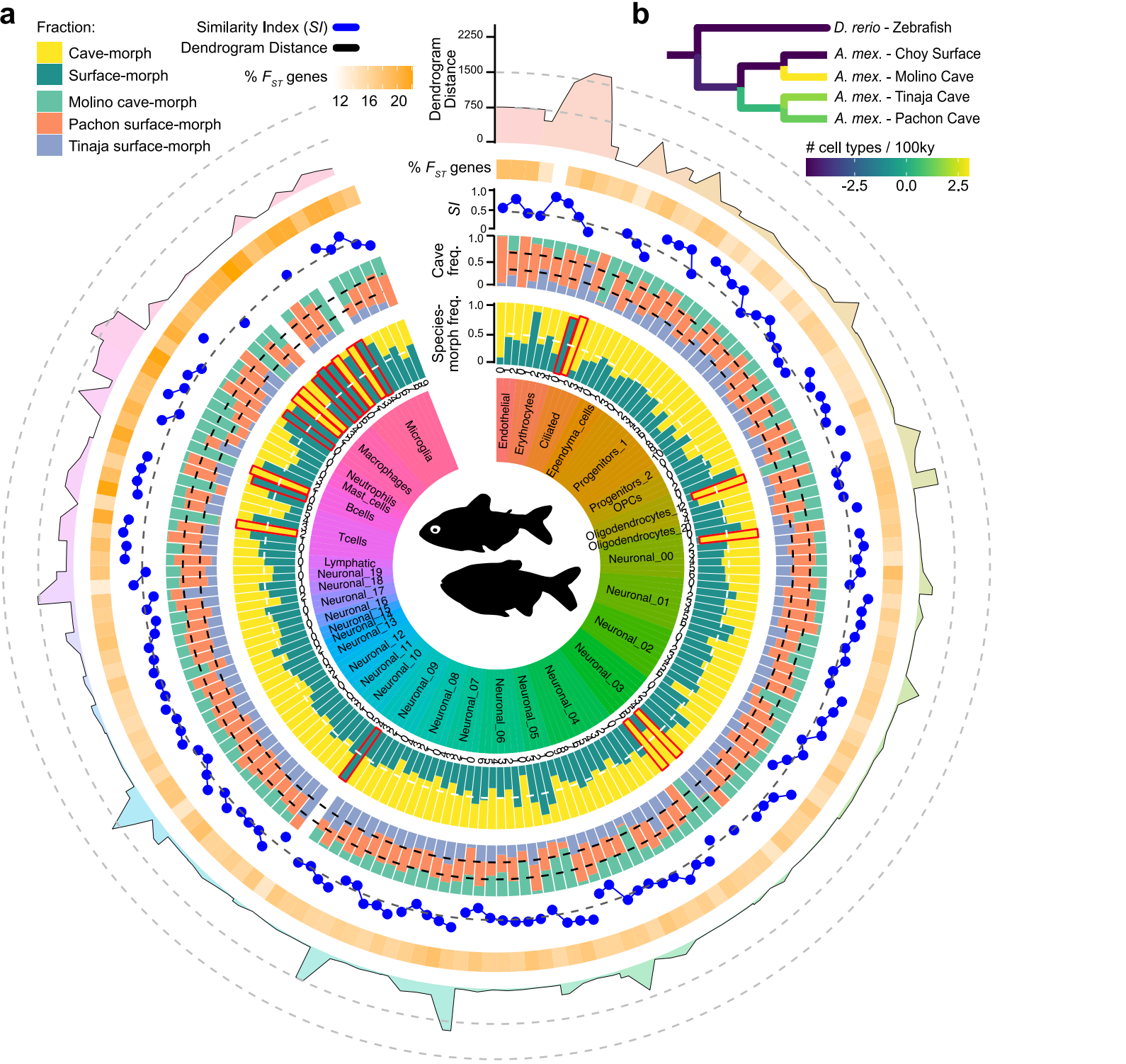

a
Gene Tree
Speciation node
Duplication node
Gene X
Gene Xa
Gene Xb
Species 1
Species 2
Species 1
Species 2
b
Species 1
Gene Xa
Gene Xb
Species 2
Gene Xa
Gene Xb
c
Complementation
Cell type
Gene Xa
Gene Xb
Neo-functionalization
Gene Xa
Gene Xb
Redundancy
Ancestral
d
% of paralog pairs
dT
e
dT
LCA of paralogs
Opisthokonta
Bilateria
Chordata
Vertebrata
Euteleostomi
Neopterygii
Clupeocephala
Otophysi
D. rerio / A.mexicanus
f
All gene pairs
Pearson correlation = 0.54
dT zebrafish
dT Mexican tetra
g
LCA Otophysi gene pairs
Pearson correlation = 0.45
h
etv5a
etv5b
etv5a
etv5b
Endothelial
Erythrocytes
Ciliated
Ependyma
Progenitors
Oligodendrocytes
OPCs
Neuronal 00
Neuronal 01
Neuronal 02
Neuronal 03
Neuronal 04
Neuronal 05
Neuronal 06
Neuronal 07
Neuronal 08
Neuronal 09
Neuronal 10
Neuronal 11
Neuronal 12
Neuronal 13
Lymphatic
Leucocytes
Macrophages
Microglia
Average Expression
Percent Expressed
i
Otophysi
Clupeocephala
Neopterygii
Euteleostomi
Vertebrata
Chordata
Bilateria
3R
2R
Paralog expression correlation


a
vip
RF weight - Zebrafish
RF weight - Mexican tetra
gsx1
otpa
six3a
mllt11
xbp1
prox1a
prox1b
id4
hes6
nr4a3
ets2
klf17
GRN
Zebrafish
Mexican tetra
Both
Neither
b
Similarity Index (SI)
Neuropeptides
Neurotransmitters
Synaptic Genes
Ion Channel
c
RF weight correlation
d
% Paralog enrichment
Odds ratio


e
Similarity Index (SI)
Kruskal-Wallis,
p < 2.2e-16
NP/NT genes
TF genes
All marker genes
f
Similarity Index (SI)
NP/NT genes
TF genes
All marker genes
Neuronal_01_0
Neuronal_01_1
Neuronal_01_2
Neuronal_01_3
Neuronal_01_4
Neuronal_01_5
Neuronal_01_6
Neuronal_01_7
Neuronal_01_8
Neuronal_01_9
Neuronal_01_10
Neuronal_01_11


a
Shared (151)
Species-specific (43)
Ciliated
Glial
Hematopoietic
Neuronal
Zebrafish (10)
Mexican tetra (33)
b
Wilcoxon, p = 0.58
% Non-homologous genes
Conserved
Species specific
c
Wilcoxon, p = 2.8e-07
Conserved
Species specific
d
Wilcoxon, p = 4.2e-09
Neuronal
Non-neuronal
e
Wilcoxon, p = 1.6e-06
Neuronal
Non-neuronal

f
Percent Expressed
0 25 50 75
Average Expression
0 1 2
Glut_1_7
Glut_3_0
Glut_3_5
GABA_3_5
Glut_6_6
Glut_7_2
ENSAMXG...
25407
NPTX1
PPP3CA
ppp3ca
COLEC12
colec12
ENSAMXG...
17498
vipb
vip
HPCAL1
hpcal1
jac9
jac8
jac6
jac3
jac2
Species-specific subcluster marker genes
Endothelial
Erythrocytes
Ciliated
Ependymal
Progenitors
OPCs
Oligodendrocytes
Neuronal_00
Neuronal_01
Neuronal_02
Neuronal_03
Neuronal_04
Neuronal_05
Neuronal_06
Neuronal_07
Neuronal_08
Neuronal_09
Neuronal_10
Neuronal_11
Neuronal_12
Neuronal_13
Lymphatic
Leucocytes
Macrophages
Microglia
Subclusters


a
Fraction:
Cave-morph
Surface-morph
Molino cave-morph
Pachon surface-morph
Tinaja surface-morph
Similarity Index (SI)
Dendrogram Distance
% F_{ST} genes
12 16 20
Dendrogram Distance
2250
1500
750
0
% F_{ST} genes
SI
1.0
0.5
0
Cave freq.
Species-morph freq.
Endothelial
Erythrocytes
Ciliated
Ependyma_cells
Progenitors_1
Progenitors_2
OPCs
Oligodendrocytes_1
Oligodendrocytes_2
Neuronal_00
Neuronal_01
Neuronal_02
Neuronal_03
Neuronal_04
Neuronal_05
Neuronal_06
Neuronal_07
Neuronal_08
Neuronal_09
Neuronal_10
Neuronal_11
Neuronal_12
Neuronal_13
Neuronal_14
Neuronal_15
Neuronal_16
Neuronal_17
Neuronal_18
Neuronal_19
Lymphatic
Bcells
Tcells
Mast_cells
Neutrophils
Macrophages
Microglia
b
D. rerio - Zebrafish
A. mex. - Choy Surface
A. mex. - Molino Cave
A. mex. - Tinaja Cave
A. mex. - Pachon Cave
cell types / 100ky
-2.5 0.0 2.5