Mapping single-cell atlases throughout Metazoa unravels cell type evolution

Alexander J. Tarashansky[1], Jacob M. Musser[2,§], Margarita Khariton[1,§], Pengyang Li[1], Detlev Arendt[2,3], Stephen R. Quake[1,4,5], Bo Wang[1,6*]

[1]Department of Bioengineering, Stanford University, Stanford, CA, USA.

[2]European Molecular Biology Laboratory, Developmental Biology Unit, Heidelberg, Germany.

[3]Centre for Organismal Studies, University of Heidelberg, Heidelberg, Germany.

[4]Department of Applied Physics, Stanford University, Stanford, CA, USA.

[5]Chan Zuckerberg Biohub, San Francisco, CA, USA.

[6]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA, USA.

[§]These authors contributed equally to this work.

*Correspondence: wangbo@stanford.edu.

Abstract

Comparing single-cell transcriptomic atlases from diverse organisms can provide evolutionary definition of cell types, elucidate the origins of cellular diversity, and transfer cell type knowledge between species. Yet, comparison among distant relatives, especially beyond a single phylum, is hindered by complex gene histories, lineage-specific inventions, and cell type evolutionary diversifications. Here, we develop a method to enable mapping cell atlases throughout Metazoa spanning sponge to mouse. Within phyla, we identify homologous cell types, even between distant species, with some even emerging from distinct germ layers. Across phyla, we find ancient cell type families that form densely interconnected groups, including contractile and stem cells, indicating they likely arose early in animal evolution through hierarchical diversifications. These homologous cell types often substitute paralog expressions at surprising prevalence. Our findings advance the understanding of cell type diversity across the tree of life and the evolution of associated gene expression programs.

Introduction

There is much ongoing success in producing single-cell transcriptomic atlases to investigate the cell type diversity within individual organisms[1]. With the growing diversity of such cell atlases spanning the far-reaching corners of the tree of life[2–7], a new frontier is emerging: the use of cross-species comparisons to unravel the origins of cellular diversity by characterizing the conservation and diversification of cell types and to uncover species-specific cellular innovations[8,9]. Furthermore, such comparisons can accelerate cell annotation and cell type discovery at large by transferring knowledge from well-studied model organisms to under-characterized animals.

However, recent comparative single-cell analyses are limited to closely related species within the same phylum[10–13]. Interphyletic comparisons are thought to be prohibitively challenging, as complex gene evolutionary history causes distantly related organisms to share fewer one-to-one gene orthologs[14], which are often relied upon for comparative studies[4,8]. Here, we present the Self-Assembling Manifold mapping (SAMap) algorithm to enable mapping single-cell transcriptomes between phylogenetically remote species. SAMap introduces a new concept by relaxing the tight constraints imposed by sequence orthology, instead using expression similarity between mapped cells to refine the relative contributions of homologous genes which in turn improve the cell type mapping. Our approach is conceptually different from existing cell atlas analyses as it accommodates evolutionary changes that usually blur cell type interrelationships, such as gene duplications and functional substitutions among paralogs.

In this study, we compared seven whole-body cell atlases from species spanning animal phylogeny, which have divergent transcriptomes and complex molecular homologies (**Figure 1A-B** and **Supplementary Table 1**). Since no prior comparison at this scale exists, we began with well-characterized cell types in developing frog and fish embryos. We found broad concordance between transcriptomic signatures and ontogenetic relationships, which validated our mapping results, yet also detected striking examples of homologous cell types emerging from different germ layers. We next extended the comparison to animals from the same phylum but with highly divergent body plans, using a planarian flatworm and a parasitic blood fluke, and found one-to-one homologies even between cell-subtypes. Comparing all seven species from sponge to mouse, we identified interconnected cell type families broadly shared across animals, including contractile and stem cells, along with their respective gene expression programs. Lastly, we noticed that homologous cell types substitute gene paralogs for each other in evolution with surprising prevalence. Overall, our study represents an important step towards understanding the evolutionary origins of specialized cell types and their associated gene expression programs in animals.

Results

The SAMap algorithm

SAMap overcomes several challenges inherent to mapping single-cell transcriptomes between distantly related species (**Figure 1B**). First, gene evolutionary history often contains duplications or other rearrangements, causing many-to-many homologies that also have convoluted functional relationships. Further, frequent gene losses and the acquisition of new genes results in many gene expression signatures being species-specific, limiting the amount of information that is comparable across species. Lastly, evolution of expression programs gradually diminishes their similarity, making cross-species cell type correspondence difficult to resolve.

We begin with a gene-gene bipartite graph, in which cross-species edges are drawn between homologous genes, weighted by their protein sequence similarity (**Figure 1C**). This allows us to merge the two cell atlases in a joint principal component (PC) space, in which we can compute the cross-species expression correlations between homologous genes. Specifically, we first project each dataset into both its own and its partner's respective PC spaces, constructed by the SAM algorithm, which we previously developed to robustly and sensitively identify cell types[15]. For cross-species projections, we translate each species' features into those of its partner, with the expression for individual genes imputed as the weighted average of their homologs specified in the gene-gene bipartite graph. This allows SAMap to include one-to-many gene homologs, rather than just one-to-one orthologs, to enable the mapping of cells between phylogenetically remote species. The mutual projections provide the coordinates of cells from both species in each PC space, which are concatenated to form the joint space that encodes information from all features, including the species-specific genes that are often necessary in discriminating closely related cell types and subtypes.

The combined coordinates are used to calculate each cell's k-nearest cross-species neighbors, through which we quantify the correlation between homologous genes. These correlations are then used to reweight the edges in the homology graph, relaxing SAMap's initial dependence on sequence similarity to account for the convoluted functional relationships between homologous genes[4,14]. We use the reweighted homology graph to reassign each cell's cross-species neighbors and recalculate the gene-gene correlations in order to again update the gene homology graph.

In the final step, we integrate information from each cell's local neighborhood to establish more robust mutual connectivity between cells across species, overcoming the inherent variability present among single-cell transcriptomes (**Figure 1D**). Two cells are defined as mutual nearest cross-species neighbors when their respective neighborhoods have mutual connectivity. Using this new set of cross-species neighbors, we recalculate gene expression correlations and reweight the homology graph, which is used as input to produce the final atlas alignment.

Homologous cell types emerging from distinct germ layers in frog and fish

We first applied SAMap to the *Xenopus* and zebrafish atlases, which both encompass embryogenesis until early organogenesis[4,5]. Previous analysis had linked cell types between these two organisms by matching ontogeny, thereby providing a reference for comparison. Commonly used batch correction methods (e.g., BBKNN[16] and Scanorama[17], the two top performers of cell atlas stitching algorithms according to a recent benchmarking study[18]) failed to recapitulate these results and found minimal alignment between datasets (**Supplementary Figure 1A-C**). Moreover, matching cell types based on marker genes yielded ambiguous relationships among cell types

(**Supplementary Figure 1D**). In contrast, SAMap produced a combined manifold with a high degree of cross-species alignment while maintaining high resolution for distinguishing cell types in each species (**Figure 2A** and **Supplementary Figure 1B**).

SAMap revealed broad agreement between transcriptomic similarity and developmental ontogeny, linking 26 out of 27 expected pairs based on previous annotations (**Figure 2B** and **Supplementary Table 2**)[4]. The only exception is the embryonic kidney (pronephric duct/mesenchyme), potentially indicating that their gene expression programs have significantly diverged. In addition, SAMap succeeded in drawing parallels between the development of homologous cell types and accurately matched time points along several cell lineages (**Figure 2C**). While the concordance was consistent across cell types and included both early and late embryonic stages, we noticed that the exact progression of developmental timing can vary across cell types, suggesting that SAMap can quantify heterochrony with cell type resolution.

We measured the mapping strength between cell types by calculating an alignment score (edges in **Figure 2B** and color map in **Figure 2C**), defined as the average number of mutual nearest cross-species neighbors of each cell relative to the maximum possible number of neighbors. This metric allows us to address two potential sources of error. To rule out the possibility that we may rely on a small set of mis-assigned homologs to link cell types, we randomly resampled the genes and observed that 83% of cell type pairs appeared robustly in replicate trials and those that did not had low alignment scores (<0.2) (**Supplementary Table 2**). To test the influence of incomplete atlases, we systematically removed cell types from one of the datasets. We noticed that cell types with missing partners did not link strongly to other cell types, but those that did only linked to cells of a similar sort with much lower alignment scores compared to the original pairs (e.g., the hindbrain linked to the forebrain when the former was removed from one of the datasets). These results show that while SAMap performs optimally on datasets with high coverage of cell types, the alignment score can be used as a reliable measure of mapping quality.

Beyond the broad concordance between gene expression and ontogeny, which serves as validation for our mapping results, we also made two surprising observations. First, SAMap linked secretory cell types that differ in their developmental origin and even arise from different germ layers (black edges in **Figure 2B**). They are linked through a large set of genes including conserved transcription factors (e.g., *foxa1*[19], *grhl*[20]) and proteins involved in vesicular protein trafficking (**Supplementary Figure 2A**). This observation corroborates the notion that cell types may be transcriptionally and evolutionarily related despite having different developmental origins[21].

Second, we noticed that some gene paralogs exhibit much greater similarity in expression across species than their corresponding orthologs. As a result, weighting homologous gene pairs based on overlapping expressions significantly improved the alignment scores of many cell types (**Figure 2D-E**). To explore this phenomenon further, we used Eggnog to map proteins to orthology groups at different ancestral nodes in the tree of life[22]. We found that SAMap selected 4,271 vertebrate orthologs and 5,191 paralogs for manifold alignment. Among these, 462 genes have higher expression correlations (correlation difference > 0.3) with their paralogs than orthologs (**Figure 2F** and **Supplementary Figure 2B**). We term these events as "paralog substitutions". In 182 of these cases, the ortholog is either absent or lowly-expressed with no cell-type specificity, suggesting that the functions of these orthologs may have prevailed in the paralogs (**Supplementary Table 3**). SAMap linked an additional 648 homologous pairs with no annotated orthology or paralogy but high expression correlations (>0.5 Pearson correlation), which may represent unannotated orthologs/paralogs or isofunctional but distantly related homologous proteins[23]. These results illustrate the potential of SAMap in leveraging single-cell gene expression data to complement protein sequences in inferring gene homology.

Homologous cell types between two flatworm species with divergent body plans

To test if we can identify homologous cell types in animals of different body plans, we mapped the cell atlases of two flatworms, the planarian *Schmidtea mediterranea*[6], and the trematode *Schistosoma mansoni*, which we collected recently[24]. They represent two distant lineages within the same phylum but have remarkably distinct body plans and autecology[25]. While planarians live in freshwater and are known for their ability to regenerate[26], schistosomes live as parasites in humans. The degree to which cell types are conserved between them is unresolved, given the vast phenetic differences caused by the transition from free-living to parasitic habits[25].

SAMap revealed broad cell type homology between schistosomes and planarians. The schistosome had cells mapped to the planarian stem cells, called neoblasts, as well as most of the differentiated tissues: neural, muscle, intestine, epidermis, parenchymal, protonephridia, and *cathepsin*$^+$ cells, the latter of which consists of cryptic cell types that, until now, have only been found in planarians[6] (**Figure 3A**). These mappings are evidenced by both known cell type specific marker genes and numerous homologous transcriptional regulators (**Figure 3B** and **Supplementary Figure 3A-C**). SAMap did not find any schistosome cell types mapped to planarian pharynx populations, consistent with previous anatomic characterizations suggesting the schistosome lacks a pharynx[27].

We next determined if cell type homologies exist at the subtype level. For this, we compared the neoblasts, as planarian neoblasts are known to comprise populations of pluripotent cells and tissue-specific progenitors[6,28]. By mapping the schistosome neoblasts to a planarian neoblast atlas[28], we found that the schistosome has a population of neoblasts (ε-cells[29]) that cluster with the planarian's pluripotent neoblasts, both expressing a common set of TFs (e.g., *soxp2, unc4*, *pax6a*, *gcm1*) (**Figure 3C-D**). We note that ε-cells are closely associated with juvenile development and lost in adult schistosomes[29], indicating pluripotent stem cells may be a transient population restricted to their early developmental stages. This is consistent with the fact that, while schistosomes can heal wounds, they have very limited regenerative ability[30]. SAMap also linked other schistosome neoblast populations with planarian progenitors. In particular, SAMap linked two populations of schistosome neoblasts (denoted as μ[15] and μ') to planarian muscle progenitors, all of which express *myoD*, a canonical master regulator of myogenesis[31]. Because μ-cells do not yet express differentiated muscle markers such as *troponin* whereas μ'-cells do (**Supplementary Figure 3D**), they may represent early and late muscle progenitors, respectively.

Cell type families spanning the animal tree of life

To compare cell types across broader taxonomic scales, we extended our analysis to include freshwater sponge (*Spongilla lacustris*)[2], *Hydra* (*Hydra vulgaris*)[3], and mouse (*Mus musculus*) embryogenesis[32] atlases. SAMap connected large fractions of cells for all 21 pairwise comparisons. Consistent with the evolutionary distances, more cells mapped with higher alignment scores between vertebrates compared to mappings across broader taxa (**Figure 4A**). In total, SAMap linked 910 cross-species pairs of cell types, which are defined according to previous annotations, though the same analysis could be performed using *de novo* cluster assignments should annotations not be available.

We observed interconnected groups of cell types with dense, many-to-many connections between them. This finding corroborates recent suggestions that ancestral cell types may have diversified into families of cell types over long evolutionary distances[9,21]. We used 'transitivity' to quantify the connectivity, in which two mapping partners also link to at least one common cell type in a third dataset, forming a triangle. The transitivity of a cell type pair (edge) or a cell type (node) is defined as the fraction of triads to which they belong that are in triangles (**Figure 4B**). The majority (81%) of cell type pairs have non-zero transitivity independent of alignment score (**Supplementary Figure 4**).

We next used the vertebrate cell types as references to identify highly transitive groups of cell types across phyla. This led us to neurons and contractile/muscle cells (**Figure 4C**), which have high interconnectedness compared to the overall graph connectivity (bootstrap p-value < 5x10^{-3}) (**Figure 4D**). Consistent with the nerve net hypothesis suggesting a unified origin of neural cell types[33], the neural family includes vertebrate brain tissues, both bilaterian and cnidarian neurons, and *Spongilla* choanocytes and apopylar cells, both of which are not considered as neurons but have been shown to express postsynaptic-like scaffolding machinery[2]. The contractile family includes myocytes in bilaterian animals, *Hydra* myoepithelial cells that are known to have contractile myofibrils[34], and sponge pinacocytes and myopeptidocytes, both of which have been implicated recently to play roles in contractility[2]. In contrast to the groups encompassing all seven species, we also found a fully interconnected subgraph that contains invertebrate pluripotent stem cells, including planarian and schistosome neoblasts, *Hydra* interstitial cells, and sponge archeocytes. Our findings suggest these cell types may emerge early in animal evolution.

It is difficult to define potential “false positives” and “false negatives” in mapping cell types, as there is little prior knowledge and mapped pairs typically have significant overlap in gene expression. However, cell type pairs gain additional support via transitivity because they share common relationships to other cell types through independent mappings. We note that ~16% mapped cell type pairs have zero edge transitivity but non-zero node transitivity, meaning that at least one of their cell types connects to only a single member of a cell type family (**Figure 4E**). Such edges may be potential “false positives” as they should connect to other members of the same group. Conversely, cell types that are incorrectly disconnected due to large evolutionary distances could be transitively linked if they belong to the same larger cell type family. Therefore, transitivity provides a quantifiable measure of mapping quality, especially over long evolutionary distances.

Transcriptomic signatures of cell type families

The high interconnectedness between cell types across broad taxonomic scales suggests that they should share ancestral transcriptional programs[11,21]. SAMap identified broad transcriptomic similarity between bilaterian and non-bilaterian contractile cells that extends beyond the core contractile apparatus. It links a total of 11,641 gene pairs, connecting 5,884 unique genes, which are enriched in at least one contractile cell type pair (see **Methods** for the definition of gene enrichment in cell type pairs). Performing functional enrichment analysis on these genes, we found cytoskeleton and signal transduction functions to be enriched (p-value < 10^{-3}) based on the KOG functional classifications[35] assigned by Eggnog (**Figure 5A**). These genes include orthology groups spanning diverse functional roles in contractile cells, including actin regulation, cell adhesion and stability, and signaling (**Figure 5B** and **Supplementary Table 4**), indicating that contractile cells were likely multifunctional near the beginning of animal evolution.

We also identified several transcriptional regulators shared among contractile cells (**Figure 5B**). Previously known core regulators involved in myocyte specification[36] were enriched only in bilaterian (e.g., *myod*, and *tcf4/E12*) or vertebrate contractile cells (e.g., *mef2*). In contrast, we found homologs of Muscle Lim Protein (*Csrp*) and Forkhead Box

Group 1[37] enriched in contractile cells from all seven species. The Fox proteins included FoxC, which is known to regulate cardiac muscle identity in vertebrates[36] and is contractile-specific in all species except schistosome and *Spongilla*. Notably, we also identified FoxG orthologs to be enriched in three of the four invertebrates (**Supplementary Figure 5**), suggesting that FoxG may play an underappreciated role in contractile cell specification outside vertebrates.

For the family of invertebrate stem cells, we identified 3,340 genes that are enriched in at least one cell type pair and observed significant enrichment (p-value < 10^{-3}) of genes involved in translational regulation such as RNA processing, translation, and post-translational modification (**Figure 5C**). These genes form 979 orthology groups, 17% of which are enriched in all cell types of this family (**Supplementary Table 4**). Importantly, other stem cell populations in *Hydra* and planarian lineage-restricted neoblasts have significantly reduced expression of these genes (**Figure 5D**). These results suggest that SAMap identified a large, deeply conserved gene module specifically associated with multipotency.

Discussion

Cell types evolve as their gene expression programs change either as integrated units or split to give rise to separate derived programs that cause new cell types to emerge. while this notion of coupled cellular and molecular evolution has gained significant traction in the past years, it has remained mostly untested over long evolutionary distances. Here, we present the first study mapping single-cell atlases between evolutionarily distant species, which were thought to be prohibitively difficult. SAMap aligns cell atlases in two mutually reinforcing directions, mapping both the genes and the cells, with each feeding back into the other. This method allows us to identify one-to-one cell type concordance between animals in the same phylum, whereas between phyla, we observe interconnected cell types forming distinct families. These finding support the recently proposed hierarchical cell type diversification in evolution[9], where a cell type family is composed of evolutionary related cell types sharing a common gene expression program that originated from a common ancestral cell type. One-to-one cell type homologies persist only if no further cell type diversification has occurred since the species split. In parallel, our results reveal surprising prevalence of paralogs exhibiting greater expression similarity than orthologs across species, resolving previously documented discrepancies between the conservation of protein sequence and function[4].

Besides studying cell type evolution, SAMap can also catalyze the annotation of new cell atlases, which often represents a substantial bottleneck requiring extensive manual curation and prior knowledge that non-model organisms often lack. Its ability to use the existing atlases to inform the annotation of cell types in related species will keep improving as more datasets become available to better sample the diversity of cell types in the tree of life. Moreover, our approach allows leveraging existing and forthcoming single-cell gene expression data to deconvolute homology relationships, predict gene functions, and guide future mechanistic molecular studies.

Acknowledgments

We thank D. Wagner and C. Juliano for sharing the data and essential discussions. We also thank S. Granick, L. Luo, and J. Kebschull for their critical reading of the manuscript. AJT is a Bio-X Stanford Interdisciplinary Graduate Fellow. This work is supported by a Beckman Young Investigator Award to BW.

Author Contributions

A.J.T. and B.W. designed the research, A.J.T. developed and benchmarked the algorithm, A.J.T., J.M.M., and M.K. performed the analysis, P.L. performed experiments on schistosomes, D.A. and S.R.Q. provided conceptual advices, A.J.T., J.M.M., M.K., and B.W. wrote the paper with input from all authors, and B.W. supervised the project.

Declaration of Interests

The authors declare no competing financial interests.

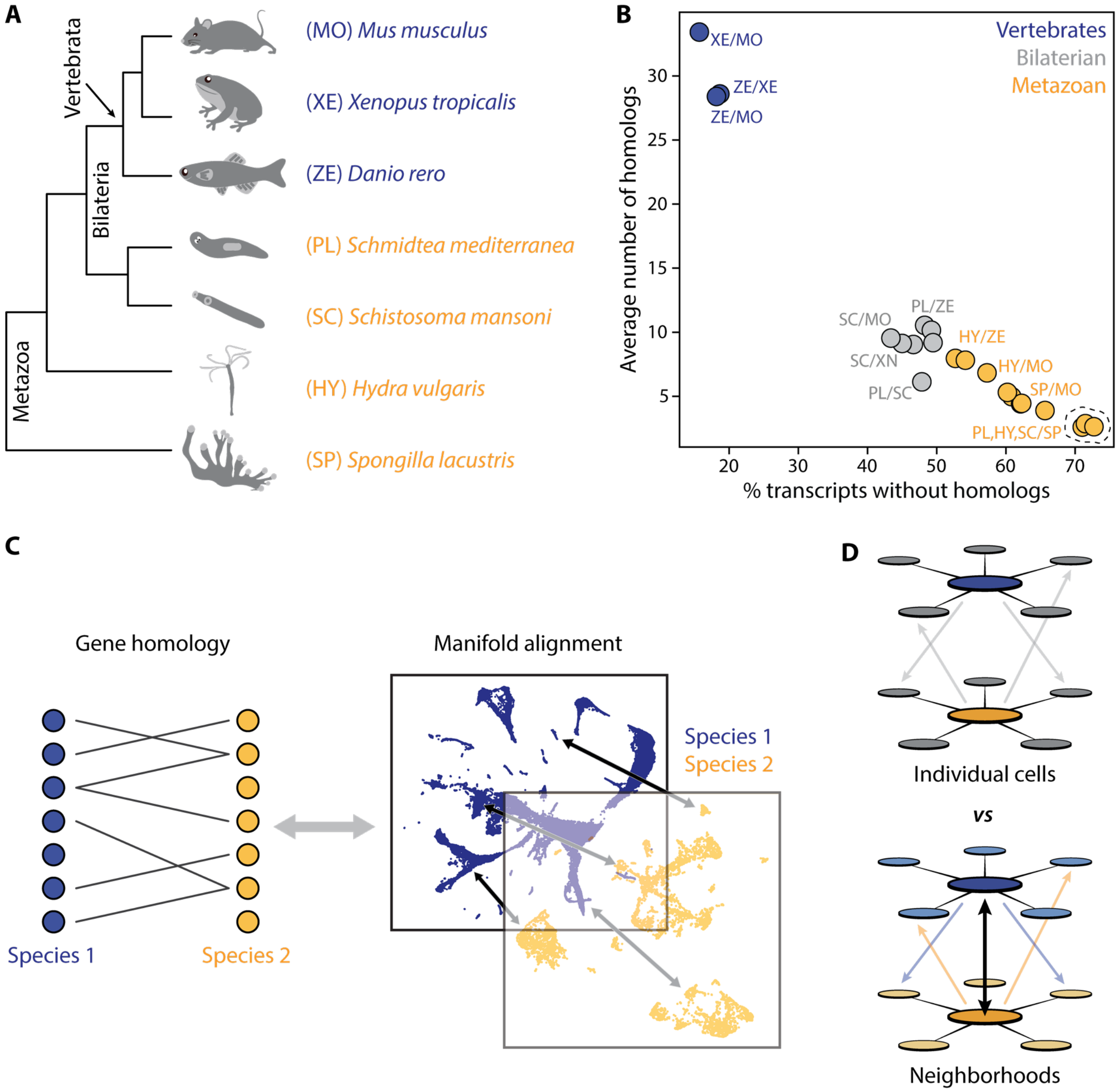


Figure 1: SAMap addresses challenges in mapping cell atlases of distantly related species. (A) Schematic showing the phylogenetic relationships among 7 species analyzed. (B) Two challenges in mapping single-cell transcriptomes. Gene duplications cause large numbers of homologs per gene, determined by reciprocal BLAST (cut-off: e-value < 10^{-6}), and frequent gene losses and the acquisition of new genes results in large fractions of transcriptomes lacking homology, which limits the amount of information comparable across species. (C) SAMap workflow. Homologous gene pairs initially

weighted by protein sequence similarity are used to align the manifolds, low dimensional representations of the cell atlases. Gene-gene correlations calculated from the aligned manifolds are used to update the edge weights in the bipartite graph, which are then used to improve manifold alignment. (D) Mutual nearest neighborhoods improve the detection of cross-species mutual nearest neighbors by connecting cells that target one other's within-species neighborhoods.

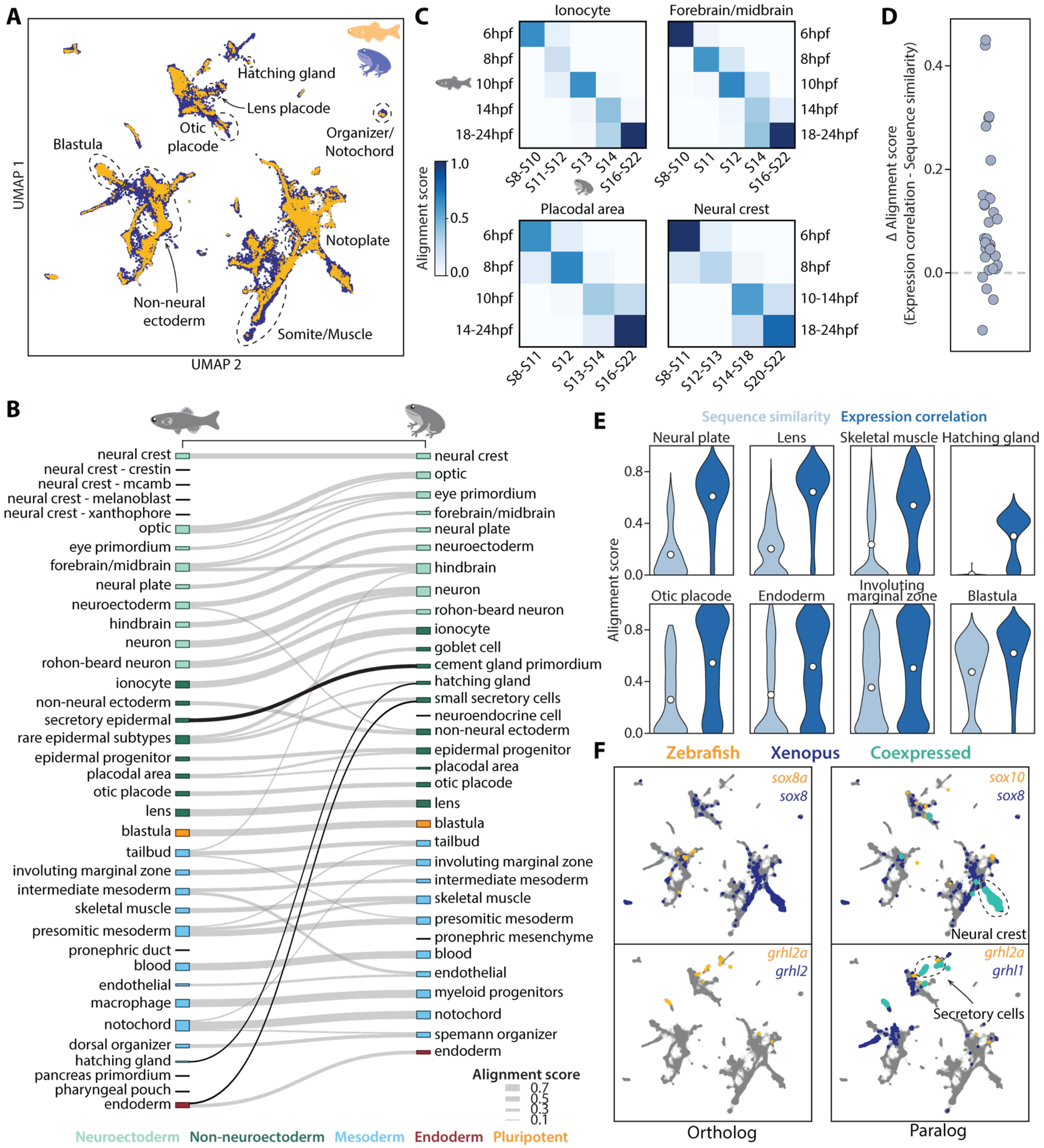


Figure 2: SAMap successfully maps *D. rerio* and *X. tropicalis* atlases and reveals prevalent paralog substitutions. (A) UMAP projection of the combined zebrafish (yellow) and *Xenopus* (blue) manifolds, with cell types highlighted in (E) circled. (B) Sankey plot summarizing the cell type mappings. Edges with alignment score < 0.1 are omitted. Edges that connect developmentally distinct secretory cell types are highlighted

in black. (C) Heatmaps of alignment scores between developmental time points for ionocyte, forebrain/midbrain, placodal, and neural crest lineages. Time points are grouped based on one-to-one correspondence between development stages across species. X-axis: zebrafish. Y-axis: *Xenopus*. (D) Improvement in alignment scores after using gene expression to curate gene homology, compared to the initial alignment based on sequence similarity alone. Each dot represents a cell type pair supported by ontogeny annotations. (E) Distribution of single-cell alignment scores in cell type pairs with the largest improvement. Circles denote the mean. (F) Expression of orthologous (left) and paralogous (right) gene pairs overlaid on the combined UMAP projection. Expressing cells are color-coded by species, with those that are connected across species colored cyan. Cells with no expression are shown in gray. More examples are provided in **Supplementary Figure 2B**.

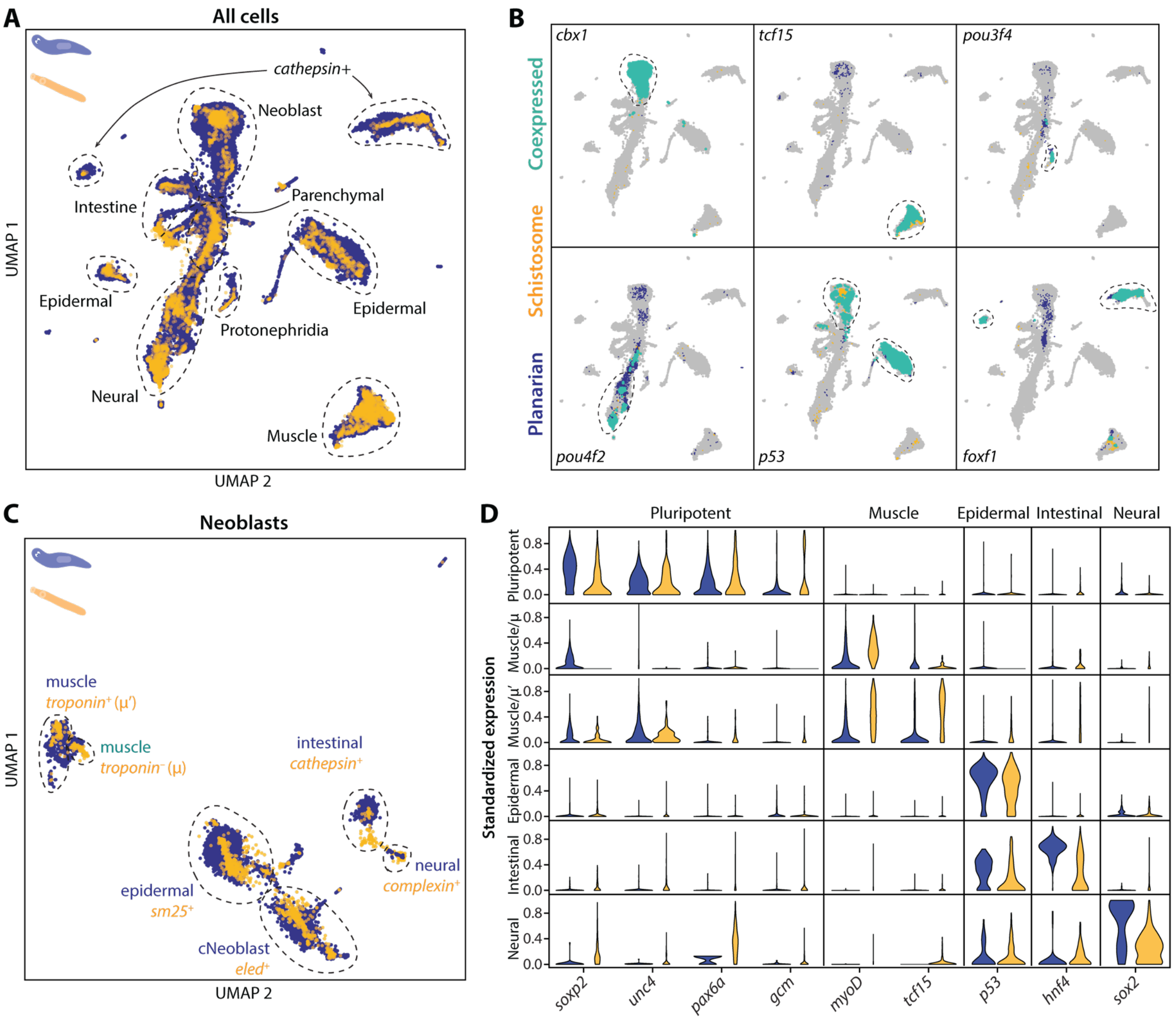


Figure 3: SAMap transfers cell type information from a well-annotated organism (planarian *S. mediterranea*) to its less-studied cousin (schistosome *S. mansoni*) and identifies parallel stem cell compartments. (A) UMAP projection of the combined manifolds. Tissue type annotations are adopted from the *S. mediterranea* atlas[6]. The schistosome atlas was collected from juvenile worms, which we found to contain neoblasts with an abundance comparable to that of planarian neoblasts[24]. (B) Overlapping expressions of selected tissue-specific TFs with expressing cell types circled. (C) UMAP projection of the aligned manifolds showing planarian and schistosome neoblasts, with homologous subpopulations circled. Planarian neoblast data is from[28],

and cNeoblasts correspond to the Nb2 population, which are pluripotent cells that can rescue neoblast-depleted planarians in transplantation experiments. (D) Distributions of conserved TF expressions in each neoblast subpopulation. Expression values are *k*-nearest-neighbor averaged and standardized, with negative values set to zero. Blue: planarian; yellow: schistosome.

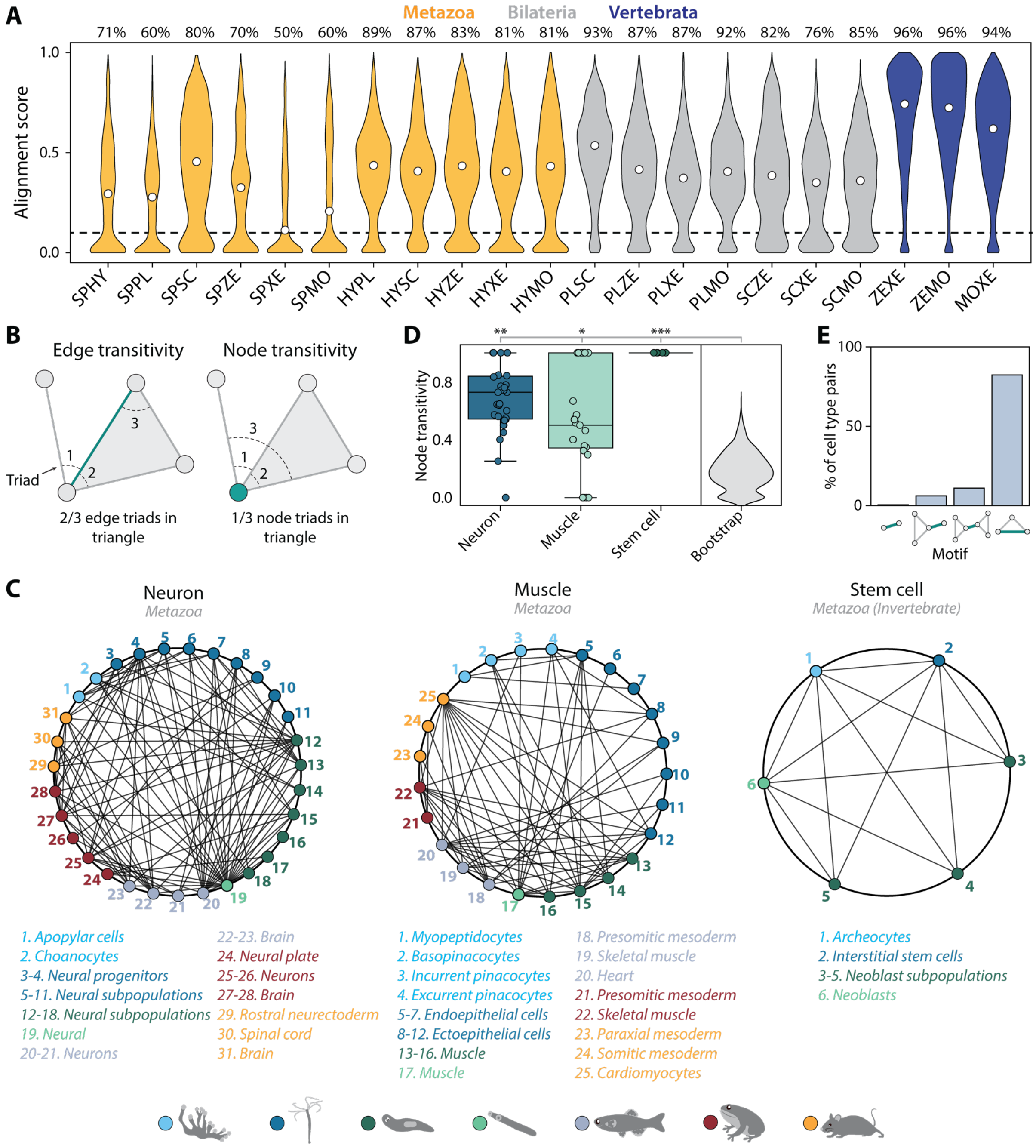


Figure 4: Mapping evolutionarily distant species identifies densely connected cell type groups. (A) Distribution of single-cell alignment scores for all 21 pairwise mappings between the 7 species. Circles denote the median. The percentage of cells with greater than 0.1 alignment score (dotted line) are reported at the top. Species acronyms are the same as in **Figure 1A**. (B) Schematic illustrating edge (left) and node (right) transitivities,

defined as the fraction of triads (set of three connected nodes) in closed triangles. (C) Network graphs showing highly connected cell type groups. Each node represents a cell type, color-coded by species (detailed annotations are provided in **Supplementary Table 5**). Mapped cell types are connected with an edge. (D) Boxplot showing the median and interquartile ranges of node transitivities for highly connected cell type groups. The average node transitivity per group is compared to a bootstrapped null transitivity distribution, generated by repeatedly sampling subsets of nodes in the cell type graph and calculating their transitivities. *p $< 5\times10^{-3}$, ** p $< 5\times10^{-5}$, ***p $< 5\times10^{-7}$. (E) The percentage of cell type pairs that are topologically equivalent to the green edge in each illustrated motif.

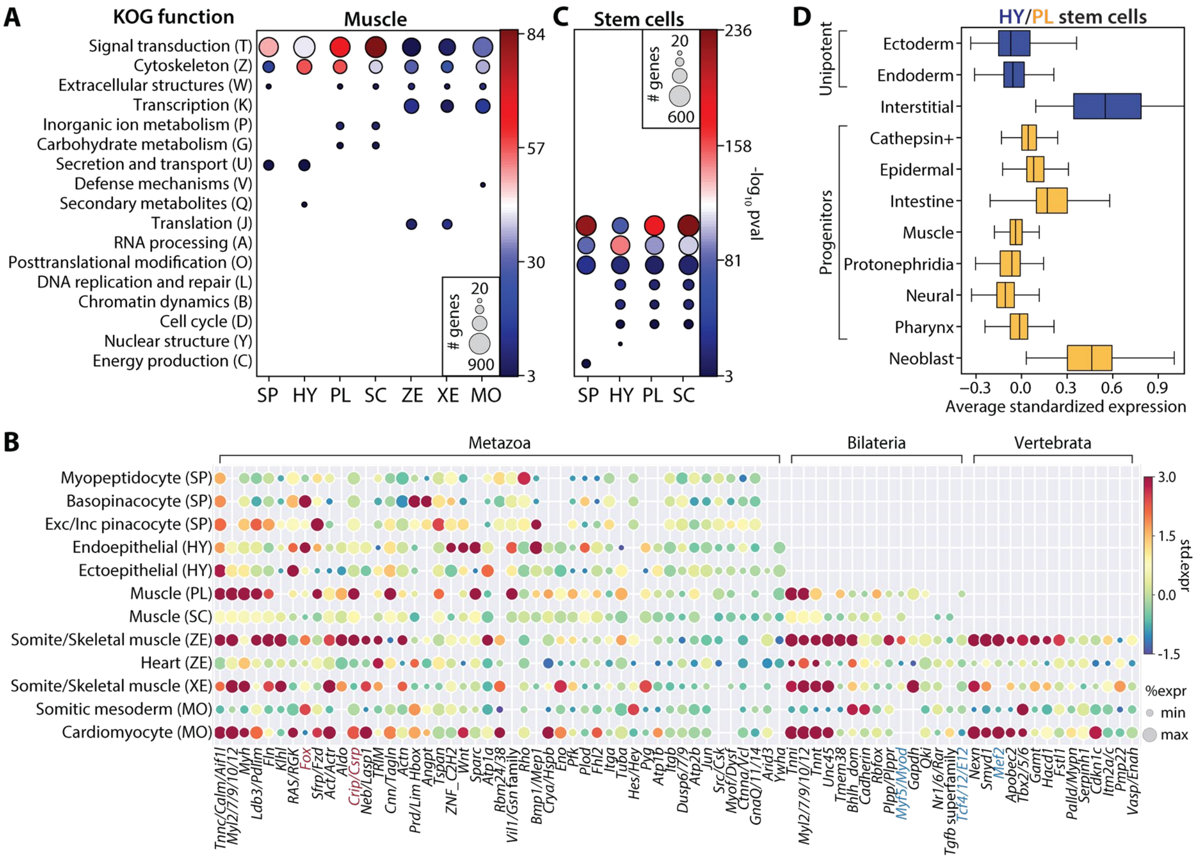
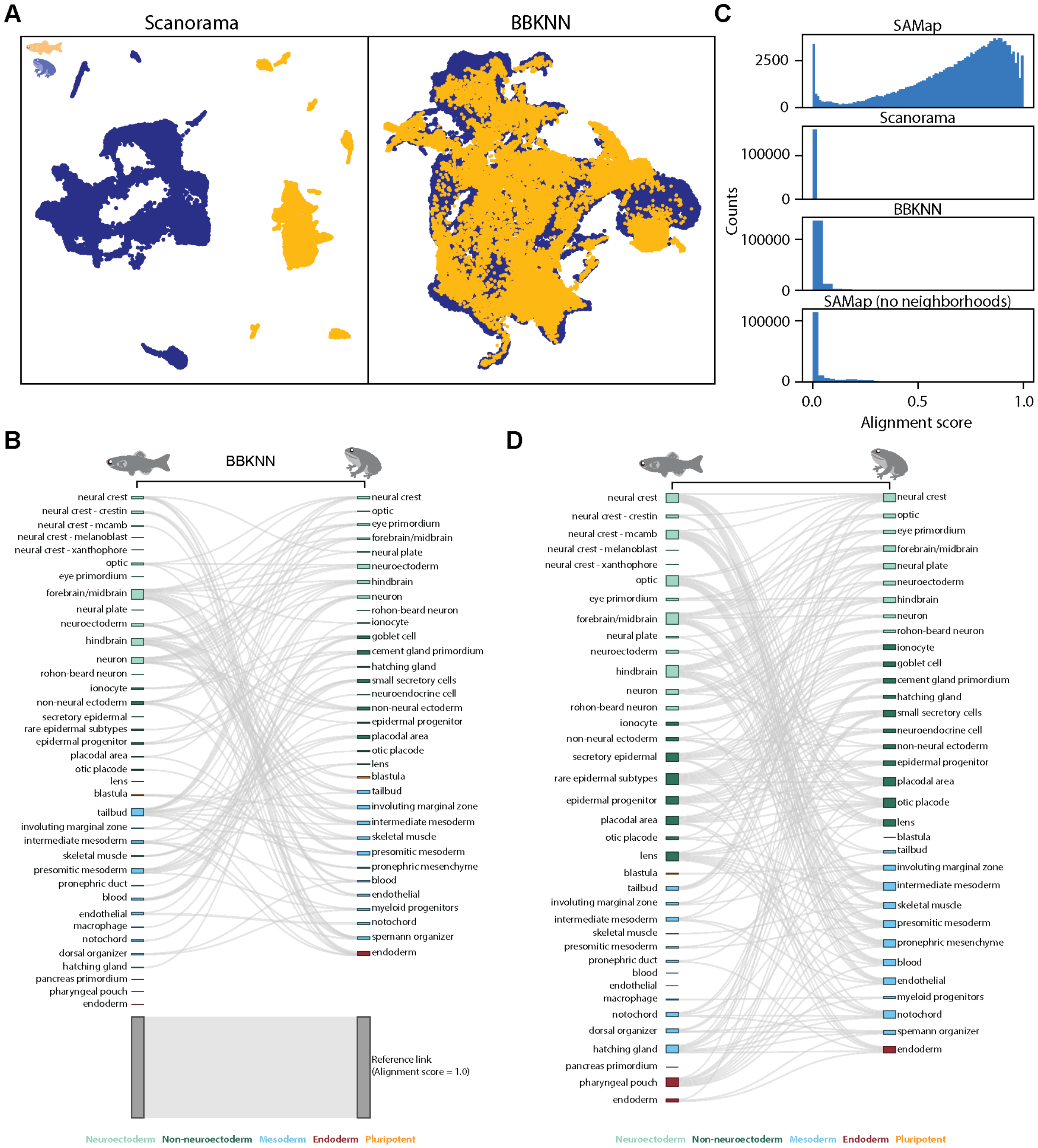


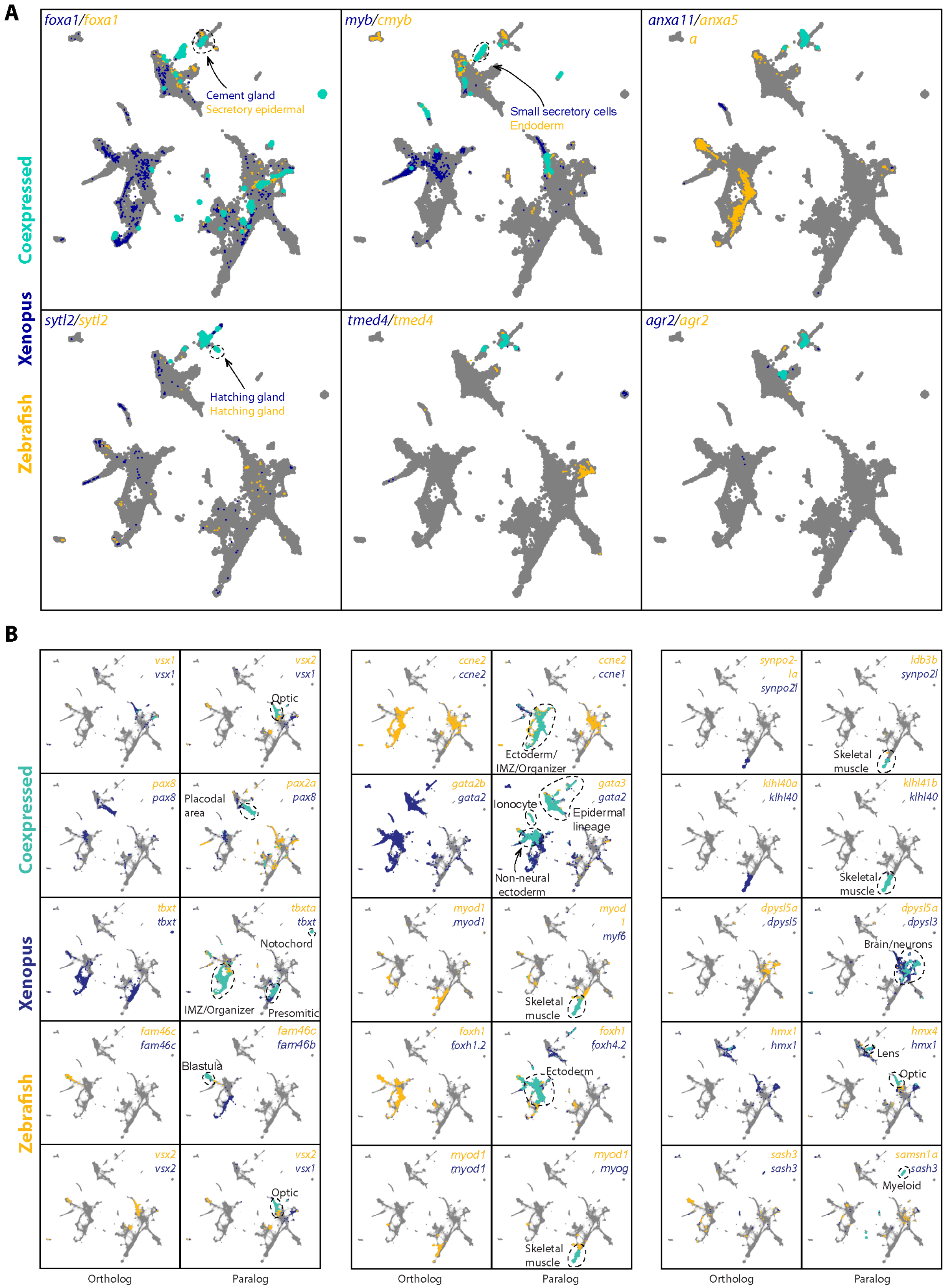
Figure 5: SAMap identifies muscle and stem cell transcriptional signatures conserved across species. (A) Enrichment of KOG functional annotations calculated for genes shared in contractile cell types. For each species, genes enriched in individual contractile cell types are combined. (B) Expression and enrichment of conserved muscle genes in contractile cell types. Color: mean standardized expression. Symbol size: the fraction of cells each gene is expressed in per cell type. Homologs are grouped based on overlapping eukaryotic Eggnog orthology groups. If multiple genes from a species are contained within an orthology group, the gene with highest standardized expression is shown. Genes in blue: core transcriptional program of bilaterian muscles; red: transcription factors conserved throughout Metazoa. (C) Enrichment of KOG functional annotations for genes shared by stem cell types. (D) Boxplot showing the median and

interquartile ranges of the mean standardized expressions of genes in hydra and planarian stem cells/progenitors that are conserved across all invertebrate species in this study. Planarian progenitors: *piwi*+ cells that cluster with differentiated tissues in Fincher et al.[6]. Neoblasts: cluster 0 in Fincher et al.[6] that does not express any tissue-specific markers.

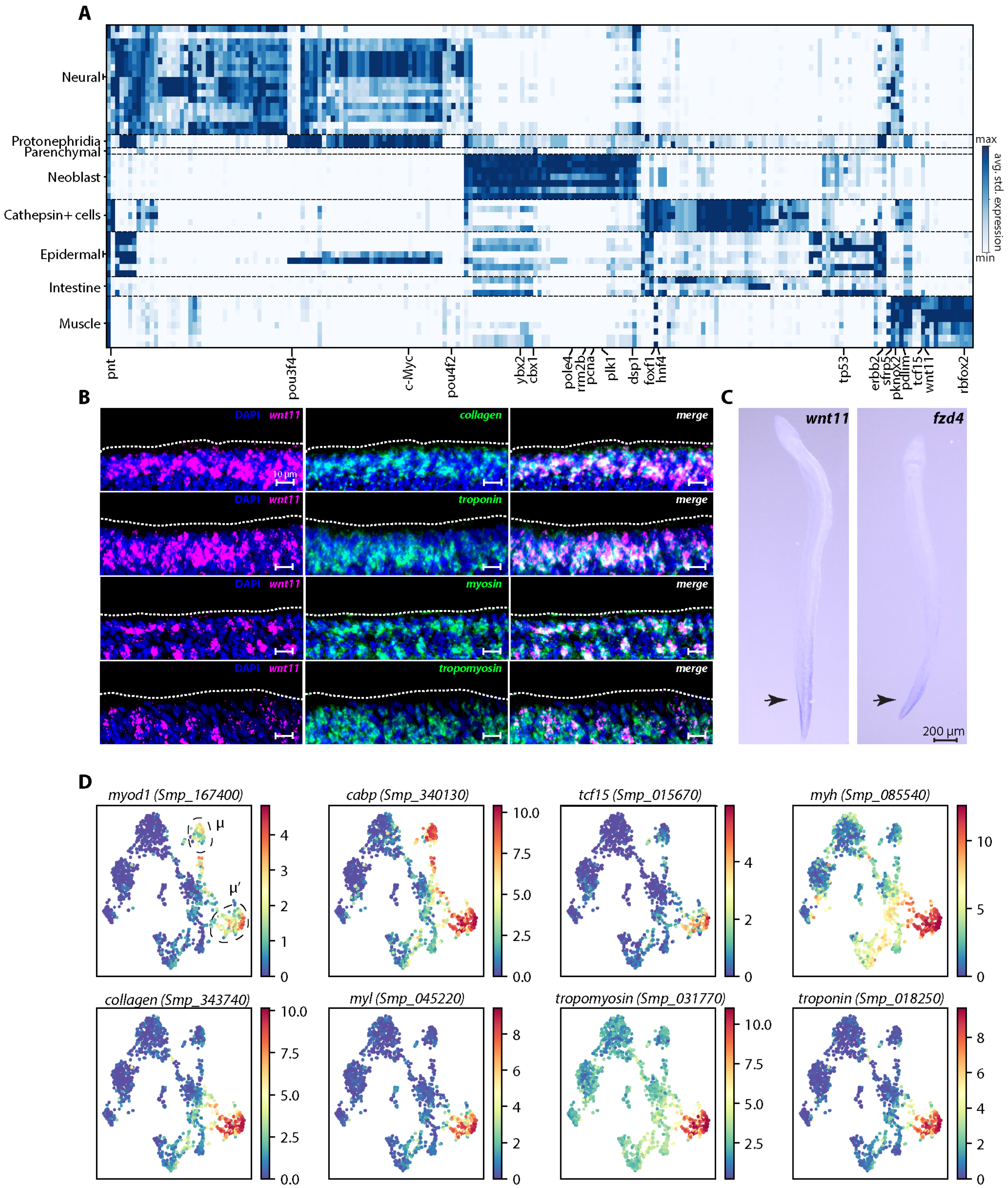
Supplementary Information



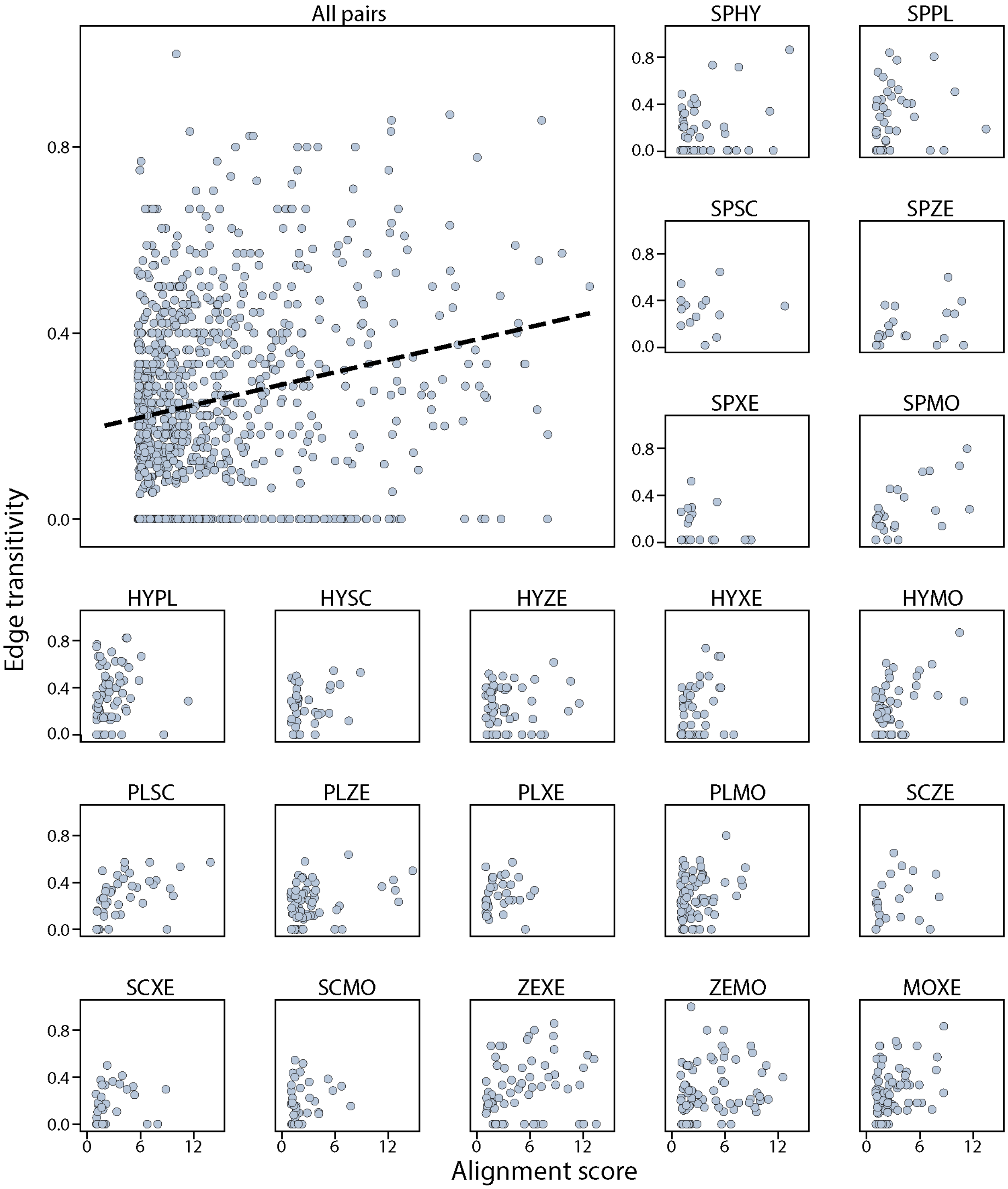
Supplementary Figure 1: Existing methods failed to map *D. rerio* and *X. tropicalis* atlases. (A) UMAP projections of the integration results from Scanorama and BBKNN. (B) Sankey plot of the mapping results output by BBKNN. Edge thickness: alignment score. The alignment is weak so a reference edge with unit alignment score is provided for comparison at the bottom. (C) Distribution of alignment scores between individual cells calculated by SAMap, Scanorama, BBKNN, and SAMap modified to use only mutual nearest neighbors as opposed to neighborhoods. Importantly, since both BBKNN and SAMap use nearest-neighbor graphs as the primary representation of the datasets, the observed contrast in their performance suggests that extending the notion of mutual connectivity between cells to neighborhoods of cells is crucial for SAMap to successfully align manifolds. Indeed, modifying SAMap to only rely on mutual connectivity between individual cells results in dramatically reduced alignment (bottom). (D) Sankey plot of cell type mapping through matching marker genes. Cell types sharing at least 5 homologous marker genes are connected with an edge.


A
Coexpressed
Xenopus
Zebrafish
foxa1/foxa1
Cement gland
Secretory epidermal
myb/cmyb
Small secretory cells
Endoderm
anxa11/anxa5
a
sytl2/sytl2
Hatching gland
Hatching gland
tmed4/tmed4
agr2/agr2
B
Coexpressed
Xenopus
Zebrafish
vsx1
vsx1
vsx2
vsx1
Optic
pax8
pax8
pax2a
pax8
Placodal
area
tbxt
tbxt
tbxta
tbxt
Notochord
IMZ/Organizer
Presomitic
fam46c
fam46c
fam46c
fam46b
Blastula
vsx2
vsx2
vsx2
vsx1
Optic
ccne2
ccne2
ccne2
ccne1
Ectoderm/
IMZ/Organizer
gata2b
gata2
gata3
gata2
Ionocyte
Epidermal
lineage
Non-neural
ectoderm
myod1
myod1
myod
1
myf6
Skeletal
muscle
foxh1
foxh1.2
foxh1
foxh4.2
Ectoderm
myod1
myod1
myod1
myog
Skeletal
muscle
synpo2-
la
synpo2l
ldb3b
synpo2l
Skeletal
muscle
klhl40a
klhl40
klhl41b
klhl40
Skeletal
muscle
dpysl5a
dpysl5
dpysl5a
dpysl3
Brain/neurons
hmx1
hmx1
hmx4
hmx1
Lens
Optic
sash3
sash3
samsn1a
sash3
Myeloid
Ortholog
Paralog
Ortholog
Paralog
Ortholog
Paralog

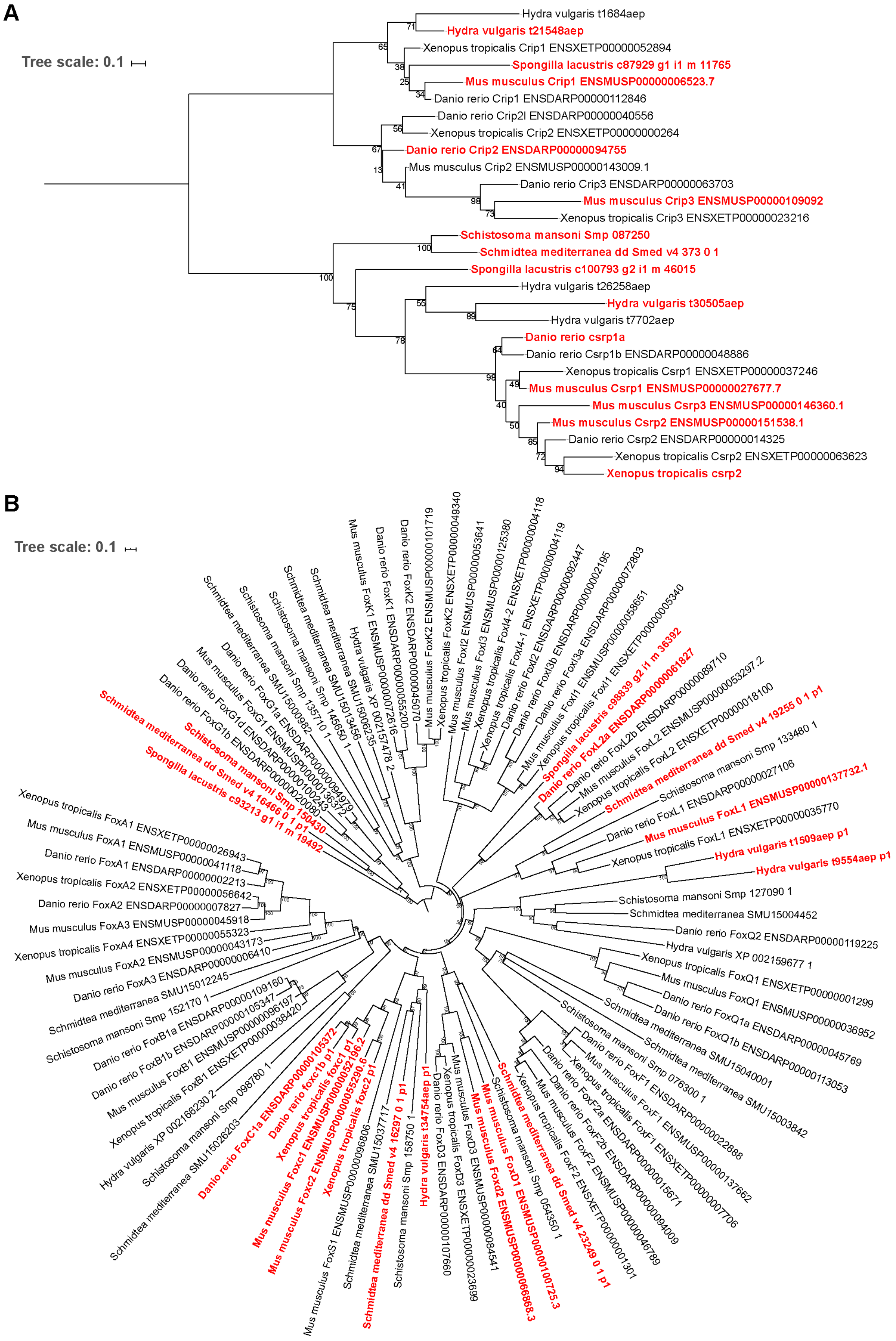
Supplementary Figure 2: Expression of selected genes enriched in secretory cell types and representative examples of paralog substitution events in *D. rerio* and *X. tropicalis* atlases. (A) Expressions of orthologous gene pairs linked by SAMap are overlaid on the combined UMAP projection. Expressing cells are color-coded by species, with those that are connected across species colored cyan. Cells with no expression are shown in gray. The mapped secretory cell types are highlighted with circles. (B) Expressions of orthologous and paralogous gene pairs are overlaid on the combined UMAP projection. Expressing cells are color-coded by species, with those that are connected across species colored in cyan. Cells with no expression are shown in gray.



Supplementary Figure 3: **SAMap-linked gene pairs that are enriched in cell type pairs between *S. mediterranea* and *S. mansoni*.** (A) Rows: linked cell types. Schistosome cell types correspond to leiden clusters. Columns: genes linked by SAMap with overlapping eukaryotic Eggnog orthology groups. We calculate the average standardized expression of each gene in an orthology group for its corresponding cell type in a particular pair and report the highest expression. A selected set of orthology groups corresponding to transcriptional regulators are labeled. (B) Fluorescence *in situ* hybridization shows the co-expression of *wnt* (Smp_156540) and a panel of muscle markers (*collagen*, *troponin*, *myosin* and *tropomyosin*) in *S. mansoni* juveniles. The body wall muscles are expected to be located close to the parasite surface (dashed outline). The images are maximum intensity projections constructed from ~10 confocal slices with optimal axial spacing recommended by the Zen software collected on a Zeiss LSM 800 confocal microscope using a 40× (N.A. = 1.1, working distance = 0.62 mm) water-immersion objective (LD C-Apochromat Corr M27). (C) Whole mount *in situ* hybridization images showing that the expression of *wnt* and *frizzled* (Smp_174350) are concentrated in the parasite tail (arrows) with decreasing gradients extending anteriorly. In planarian muscles, Wnt genes provide the positional cues for setting up the body plan during regeneration[31]. The presence of an anterior-posterior expression gradient of *wnt* and *frizzled* in muscles of schistosome juveniles as well suggests that they may have similar functional roles in patterning during development. (D) UMAP projections of schistosome neoblasts with gene expressions overlaid. μ and μ' cells are circled. Colormap: expression in units of $log_2(D+1)$. For visualization, expression was smoothed via nearest-neighbor averaging using SAM. Note that *myod1* and *cabp* are expressed in both presumptive muscle progenitor populations, whereas all other markers are enriched in μ' cells. All genes displayed are also expressed in fully differentiated muscle tissues.



Supplementary Figure 4: **Alignment scores are independent of edge transitivity in the cell type connectivity graph constructed from all 7 species.** Top left: alignment scores and edge transitivity for all cell type pairs in the connectivity graph. Dotted line: the linear best fit, with the Pearson correlation coefficient reported at the top. Alignment scores and edge transitivity for individual species pairs are shown in the remaining subplots.


A
Tree scale: 0.1
Hydra vulgaris t1684aep
71
Hydra vulgaris t21548aep
65
Xenopus tropicalis Crip1 ENSXETP00000052894
38
Spongilla lacustris c87929 g1 i1 m 11765
25
Mus musculus Crip1 ENSMUSP00000006523.7
34
Danio rerio Crip1 ENSDARP00000112846
Danio rerio Crip2l ENSDARP00000040556
58
Xenopus tropicalis Crip2 ENSXETP00000000264
67
Danio rerio Crip2 ENSDARP00000094755
13
Mus musculus Crip2 ENSMUSP00000143009.1
41
Danio rerio Crip3 ENSDARP00000063703
98
Mus musculus Crip3 ENSMUSP00000109092
73
Xenopus tropicalis Crip3 ENSXETP00000023216
Schistosoma mansoni Smp 087250
100
Schmidtea mediterranea dd Smed v4 373 0 1
100
Spongilla lacustris c100793 g2 i1 m 46015
75
Hydra vulgaris t26258aep
55
Hydra vulgaris t30505aep
89
Hydra vulgaris t7702aep
78
Danio rerio csrp1a
64
Danio rerio Csrp1b ENSDARP00000048886
Xenopus tropicalis Csrp1 ENSXETP00000037246
98
49
Mus musculus Csrp1 ENSMUSP00000027677.7
40
Mus musculus Csrp3 ENSMUSP00000146360.1
50
Mus musculus Csrp2 ENSMUSP00000151538.1
85
Danio rerio Csrp2 ENSDARP00000014325
72
Xenopus tropicalis Csrp2 ENSXETP00000063623
94
Xenopus tropicalis csrp2
B
Tree scale: 0.1
Mus musculus FoxK1 ENSMUSP00000072616
Danio rerio FoxK1 ENSDARP00000055200
Danio rerio FoxK2 ENSDARP00000045070
Mus musculus FoxK2 ENSMUSP00000101719
Xenopus tropicalis FoxK2 ENSXETP00000049340
Mus musculus FoxI2 ENSMUSP00000053641
Mus musculus FoxI3 ENSMUSP00000125380
Xenopus tropicalis FoxI4-2 ENSXETP00000004118
Xenopus tropicalis FoxI4-1 ENSXETP00000004119
Danio rerio FoxI2 ENSDARP00000092447
Danio rerio FoxI3b ENSDARP00000002195
Danio rerio FoxI3a ENSDARP00000072803
Mus musculus FoxI1 ENSMUSP00000058651
Xenopus tropicalis FoxI1 ENSXETP00000005340
Spongilla lacustris c98839 g2 i1 m 36392
Danio rerio FoxL2a ENSDARP00000061827
Danio rerio FoxL2b ENSDARP00000089710
Mus musculus FoxL2 ENSMUSP00000053297.2
Xenopus tropicalis FoxL2 ENSXETP00000018100
Schmidtea mediterranea dd Smed v4 19255 0 1 p1
Schistosoma mansoni Smp 133480 1
Danio rerio FoxL1 ENSDARP00000027106
Mus musculus FoxL1 ENSMUSP00000137732.1
Xenopus tropicalis FoxL1 ENSXETP00000035770
Hydra vulgaris t1509aep p1
Hydra vulgaris t9554aep p1
Schistosoma mansoni Smp 127090 1
Schmidtea mediterranea SMU15004452
Danio rerio FoxQ2 ENSDARP00000119225
Hydra vulgaris XP 002159677 1
Xenopus tropicalis FoxQ1 ENSXETP00000001299
Mus musculus FoxQ1 ENSMUSP00000036952
Danio rerio FoxQ1a ENSDARP00000045769
Danio rerio FoxQ1b ENSDARP00000113053
Schmidtea mediterranea SMU15003842
Schmidtea mediterranea SMU15040001
Schistosoma mansoni Smp 076300 1
Danio rerio FoxF1 ENSDARP00000022888
Mus musculus FoxF1 ENSMUSP00000137662
Xenopus tropicalis FoxF1 ENSXETP00000007706
Danio rerio FoxF2a ENSDARP00000015671
Danio rerio FoxF2b ENSDARP00000094009
Mus musculus FoxF2 ENSMUSP00000046789
Xenopus tropicalis FoxF2 ENSXETP00000001301
Schmidtea mediterranea dd Smed v4 23249 0 1 p1
Schistosoma mansoni Smp 054350 1
Mus musculus FoxD1 ENSMUSP00000100725.3
Mus musculus Foxd2 ENSMUSP00000066868.3
Mus musculus FoxD3 ENSMUSP00000084541
Xenopus tropicalis FoxD3 ENSXETP00000023699
Danio rerio FoxD3 ENSDARP00000107660
Hydra vulgaris t34754aep p1
Schistosoma mansoni Smp 158750 1
Schmidtea mediterranea dd Smed v4 16297 0 1 p1
Schmidtea mediterranea SMU15037717
Mus musculus FoxS1 ENSMUSP00000096806
Xenopus tropicalis foxc2 p1
Mus musculus Foxc2 ENSMUSP00000055290.6
Mus musculus Foxc1 ENSMUSP00000052186.2
Xenopus tropicalis foxc1 p1
Danio rerio foxc1b p1
Danio rerio FoxC1a ENSDARP00000105372
Schmidtea mediterranea SMU15026203
Schistosoma mansoni Smp 098780 1
Hydra vulgaris XP 002166230 2
Xenopus tropicalis FoxB1 ENSXETP00000038420
Mus musculus FoxB1 ENSMUSP00000096197
Danio rerio FoxB1b ENSDARP00000105347
Danio rerio FoxB1a ENSDARP00000109160
Schistosoma mansoni Smp 152170 1
Schmidtea mediterranea SMU15012245
Danio rerio FoxA3 ENSDARP00000006410
Mus musculus FoxA2 ENSMUSP00000043173
Xenopus tropicalis FoxA4 ENSXETP00000055323
Mus musculus FoxA3 ENSMUSP00000045918
Danio rerio FoxA2 ENSDARP00000007827
Xenopus tropicalis FoxA2 ENSXETP00000056642
Danio rerio FoxA1 ENSDARP00000002213
Mus musculus FoxA1 ENSMUSP00000041118
Xenopus tropicalis FoxA1 ENSXETP00000026943
Spongilla lacustris c93213 g1 i1 m 19492
Schmidtea mediterranea dd Smed v4 16466 0 1 p1
Schistosoma mansoni Smp 150430
Danio rerio FoxG1b ENSDARP00000020080
Danio rerio FoxG1d ENSDARP00000100243
Mus musculus FoxG1 ENSMUSP00000136372
Danio rerio FoxG1a ENSDARP00000094979
Schmidtea mediterranea SMU15000982
Schistosoma mansoni Smp 135710 1
Schistosoma mansoni Smp 145650 1
Schmidtea mediterranea SMU15013456
Schmidtea mediterranea SMU15006235
Hydra vulgaris XP 002157478 2

Supplementary Figure 5: Phylogenetic reconstruction of animal contractile cell transcriptional regulators. Trees depict *Csrp/Crip* (A) and Fox group I (B) gene families. Genes labelled red are enriched in at least one contractile gene pair identified via SAMap. Support values indicate bootstrap support from 1,000 nonparametric (*Csrp*) or ultrafast (*Fox*) bootstrap replicates. Besides these two transcriptional regulators, contractile cells in all seven species were found to be also enriched for transcription factors from the C2H2 Zinc Finger, Lim Homeobox, and Paired Homeobox families, although in different cell types we found enrichment of a number of distinct orthologs. Whether this reflects an ancestral role for these transcription factor families in regulating contractility or their independent evolution will require additional taxonomic sampling and broader coverage of muscle cell diversity to resolve.

Supplementary table captions

Supplementary Table 1: Cell atlas metadata and cell annotations. Metadata includes the number of cells, number of transcripts in the transcriptome, median number of transcripts detected per cell, the reference transcriptome used in this study, database through which the transcriptomes are provided, technology used for constructing the cell atlases, atlas data accessions, processing notes, and references. Leiden clusters and cell type annotations are reported for cells in each atlas. The Zebrafish and *Xenopus* tables include both the original cell type annotations and those used in this study. *D. rerio*, *X. tropicalis*, and mouse annotations include developmental stages.

Supplementary Table 2: Bootstrapping results and cell type annotations for the zebrafish-*Xenopus* mapping. For each pair of mapped cell types output by SAMap, the mean, variance, and dispersion of the alignment scores from replicate bootstrap trials are reported. The original alignment score and the fraction of trials in which each edge is observed with greater than 0.1 alignment score are also provided. 83% of cell type pairs appeared in at least 85% of trials. A list of cell type annotations in the original study[4,5] and corresponding annotations used in this study is provided for both *D. rerio* and *X. tropicalis* atlases.

Supplementary Table 3: Identified paralog substitution events in the zebrafish-*Xenopus* mapping. Each row contains a pair of vertebrate-orthologous genes and a corresponding pair of eukaryotic paralogs with higher correlation in expression compared to the orthologs, the expression correlations for ortholog and paralog pairs, the difference between their correlations, and if the substituted ortholog is either absent or lowly-expressed with no cell-type specificity. Highlighted rows are shown in **Figure 2F** and **Supplementary Figure 2B**.

Supplementary Table 4: Genes enriched in contractile cell types and invertebrate stem cells highlighted in Figure 4C. The IDs of the genes enriched in the contractile and invertebrate stem cell types are provided along with the IDs of the Eggnog orthology groups to which they belong. In cases where multiple genes from a species belonging to the same orthology group are enriched, the most differentially expressed gene is shown. The descriptions in the stem cell table are orthology annotations associated with the *Spongilla* genes provided in the original study[2].

Supplementary Table 5: Cell types in the cell type families shown in Figure 4C. For the schistosome cell types, we annotated two neural clusters, both of which express the neural marker *complexin*[24]. One of the clusters expresses the antigen *SmKK7*, so we label the clusters “Neural” and “Neural_KK7”, respectively. The “Muscle” population contains non-neoblast cells expressing *troponin*. The “Tegument_prog” and “Tegument” populations consist of cells expressing tegument progenitor and differentiated marker genes, respectively, as reported in[38].

References

1. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).
2. Musser, J. M. *et al.* Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. http://biorxiv.org/lookup/doi/10.1101/758276 (2019).
3. Siebert, S. *et al.* Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science* **365**, eaav9314 (2019).
4. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
5. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
6. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, eaaq1736 (2018).
7. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).
8. Shafer, M. E. R. Cross-species analysis of single-cell transcriptomic data. *Front. Cell Dev. Biol.* **7**, 175 (2019).
9. Arendt, D., Bertucci, P. Y., Achim, K. & Musser, J. M. Evolution of neuronal types and families. *Curr. Opin. Neurobiol.* **56**, 144–152 (2019).
10. Baron, M. *et al.* A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346-360.e4 (2016).
11. 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).

12. Geirsdottir, L. *et al.* Cross-species single-cell analysis reveals divergence of the primate microglia program. *Cell* **179**, 1609-1622.e16 (2019).
13. Sebé-Pedrós, A. *et al.* Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.* **2**, 1176–1188 (2018).
14. Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput. Biol.* **7**, e1002073 (2011).
15. Tarashansky, A. J., Xue, Y., Li, P., Quake, S. R. & Wang, B. Self-assembling manifolds in single-cell RNA sequencing data. *eLife* **8**, e48994 (2019).
16. Polański, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2019).
17. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
18. Luecken, M. *et al.* Benchmarking atlas-level data integration in single-cell genomics. http://biorxiv.org/lookup/doi/10.1101/2020.05.22.111161 (2020).
19. Dubaissi, E. *et al.* A secretory cell type develops alongside multiciliated cells, ionocytes and goblet cells, and provides a protective, anti-infective function in the frog embryonic mucociliary epidermis. *Development* **141**, 1514–1525 (2014).
20. Miles, L. B. *et al.* Mis-expression of grainyhead-like transcription factors in zebrafish leads to defects in enveloping layer (EVL) integrity, cellular morphogenesis and axial extension. *Sci. Rep.* **7**, 17607 (2017).
21. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).

22. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

23. Gabaldón, T. & Koonin, E. V. Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* **14**, 360–366 (2013).

24. Li, P. *et al.* Single-cell analysis of *Schistosoma mansoni* reveals a conserved genetic program controlling germline stem cell fate. http://biorxiv.org/lookup/doi/10.1101/2020.07.06.190033 (2020).

25. Laumer, C. E., Hejnol, A. & Giribet, G. Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation. *eLife* **4**, e05503 (2015).

26. Reddien, P. W. The cellular and molecular basis for planarian regeneration. *Cell* **175**, 327–345 (2018).

27. Mohammed, A. S. The secretory glands of the cercariae of S. *Haematobium* and *S. Mansoni* from Egypt. *Ann. Trop. Med. Parasitol.* **26**, 7–22 (1932).

28. Zeng, A. *et al.* Prospectively isolated *tetraspanin*+ neoblasts are adult pluripotent stem cells underlying planaria regeneration. *Cell* **173**, 1593–1608.e20 (2018).

29. Wang, B. *et al.* Stem cell heterogeneity drives the parasitic life cycle of Schistosoma mansoni. *eLife* **7**, e35449 (2018).

30. Wendt, G. R. & Collins, J. J. Schistosomiasis as a disease of stem cells. *Curr. Opin. Genet. Dev.* **40**, 95–102 (2016).

31. Scimone, M. L., Cote, L. E. & Reddien, P. W. Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature* **551**, 623–628 (2017).

32. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

33. Tosches, M. A. & Arendt, D. The bilaterian forebrain: an evolutionary chimaera. *Curr. Opin. Neurobiol.* **23**, 1080–1089 (2013).

34. Buzgariu, W., Al Haddad, S., Tomczyk, S., Wenger, Y. & Galliot, B. Multi-functionality and plasticity characterize epithelial cells in *Hydra*. *Tissue Barriers* **3**, e1068908 (2015).

35. Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).

36. Brunet, T. *et al.* The evolutionary origin of bilaterian smooth and striated myocytes. *eLife* **5**, e19607 (2016).

37. Larroux, C. *et al.* Genesis and expansion of Metazoan transcription factor gene classes. *Mol. Biol. Evol.* **25**, 980–996 (2008).

38. Wendt, G. R. *et al.* Flatworm-specific transcriptional regulators promote the specification of tegumental progenitors in *Schistosoma mansoni*. *eLife* **7**, e33221 (2018).